Corpora, phraseology and academic discourse

Ute Römer

ELC 2011, Belo Horizonte, Brazil – 11 Nov 2011

uroemer@gsu.edu
www.gsu.edu/alesl

Presentation outline

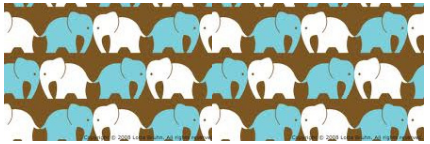
1. Introduction
2. Corpus analysis and phraseology
3. Phraseology and academic corpora (2 case studies)
   - 3.1 The phraseological profile of book reviews
   - 3.2 Textual distribution of phraseological items
4. Concluding thoughts

Ute Römer (uroemer@gsu.edu)

Georgia State University

1. Introduction

- A major finding of corpus-linguistic research:
  Language is highly patterned.



- To a high degree, language is made up of fixed or semi-fixed units (clusters, phrases, chunks, lexical bundles, n-grams, collocational frameworks, formulaic sequences, multi-word units...)
- Meaning-carrying unit in language is not the word in isolation but a larger unit (phrase)
  "the normal carrier of meaning is the phrase" (Sinclair 2005)
  "Language as phraseology" (Hunston 2002: 137)
- We need phraseological items to locate meaning

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

1. Introduction

- A second point of departure:
  English as an academic language
- Nowadays: a large and growing number of academic texts produced by non-native speakers of English
  - Research world is becoming more Anglicized
  - Large numbers of "non-Anglophones" (Swales 2004: 46) produce academic English
  - Important for novices to be familiar with the phraseology/patterning of academic writing
- Explore part of the phraseology of academic English

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

1. Introduction

This talk will...

- ... emphasize the importance of phraseology in language research
- ... focus on corpus methodology in research on phraseology
- ... present tools and techniques for phraseological analyses of language
- ... work with corpora of academic discourse and highlight aspects of different text types
- ... consider the implications of two case studies for language pedagogy and linguistic description

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

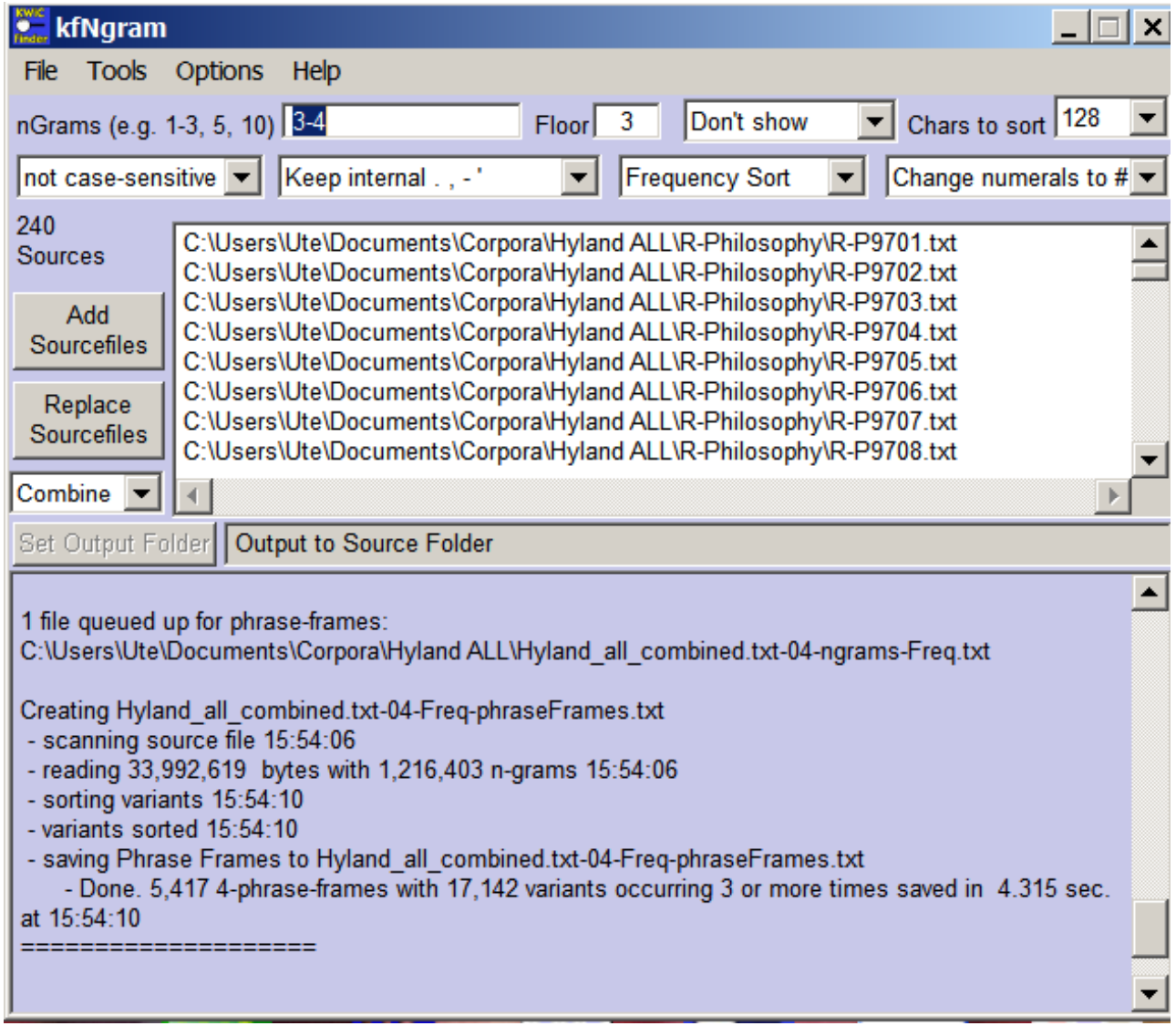
2. Corpus analysis and phraseology

- How can corpus tools that help with phraseological analyses?
- Several concordance packages automatically extract repeated word sequences from corpora:
  - kfNgram
  - Collocate
  - ConcGram
  - AntConc
  - WordSmith Tools
- Look at examples of the types of lists these tools produce...

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

2. Corpus analysis and phraseology



kfNgram

Ute Römer (uroemer@gsu.edu)

Georgia State University

2. Corpus analysis and phraseology

kfNgram

Hyland_all_combined.txt-04-ngrams-Freq.txt - Notepad

File Edit Format View Help

in the case of	210
on the other hand	186
at the same time	138
on the basis of	113
as a function of	94
in the united states	94
in terms of the	91
the extent to which	89
as shown in fig	80
with respect to the	80
as a result of	79
at the end of	79
is shown in fig	77
in the presence of	74
the fact that the	74
in the context of	73
the end of the	71
as well as the	68
figure # shows the	67
the results of the	67
the ship of theseus	66

Hyland_all_combined.txt-04-Freq-phraseFrames.txt - Notepad

File Edit Format View Help

the * of the	3234	391
the end of the	71	
the results of the	67	
the magnitude of the	60	
the nature of the	59	
the size of the	55	
the case of the	48	
the effect of the	48	
the use of the	40	
the basis of the	39	
the center of the	37	
the beginning of the	35	
the role of the	32	
the thickness of the	31	
the influence of the	30	
the shape of the	30	
the effects of the	29	
the performance of the	27	
the analysis of the	27	
the top of the	25	
the importance of the	24	
the slope of the	23	
the sum of the	23	
the course of the	23	
the surface of the	23	

Ute Römer (uroemer@gsu.edu)

Georgia State University

2. Corpus analysis and phraseology

Collocate - [N-Grams]

File Extract Full Extract Display Sort Options Window Info

Freq	Mutual Inf.	Collocation
90	17.938947	at the end of the
82	25.229688	at the university of michigan
54	28.217883	journal of personality and social
53	29.181265	of personality and social psychology
53	21.867350	due to the fact that
51	26.225277	it is important to note
47	16.975991	as a result of the
42	40.450503	new york oxford university press
39	39.984979	body color and wing venation
39	37.686354	chicago university of chicago press
39	28.186826	it should be noted that
39	24.580373	is important to note that
39	18.040300	at the beginning of the
38	15.105693	on the part of the
36	12.886534	in the case of the
35	36.132151	number of children ever born
35	20.173421	in the united states and
33	50.264726	bruhm-schlegel et al junge frauen
33	48.764653	et al junge frauen heute
33	47.900500	babasaheb ambedkar writings and sp

829 files in current corpus | 2,628,665 words, 87,799 types

Collocate

Ute Römer (uroemer@gsu.edu)

Georgia State University

2. Corpus analysis and phraseology

Collocate - [important]

File Extract Full Extract Display Sort Options Window Info

Freq	Mutual Inf.	Collocation
214	10.704406	it is important
188	8.238845	is important to
104	9.700638	the most important
76	9.707751	is an important
55	12.554289	important to note
31	7.147681	is important for
28	10.724368	important role in
27	13.167948	an important role
25	6.348914	is important that
24	11.516441	more important than
23	9.489448	is important because
23	6.208611	be important to
22	14.160563	an important factor
22	8.703900	is also important
21	10.996584	important to consider
20	9.198859	important part of
20	8.782544	be an important
20	6.967026	this is important
18	12.305160	an important part
18	10.122443	is very important

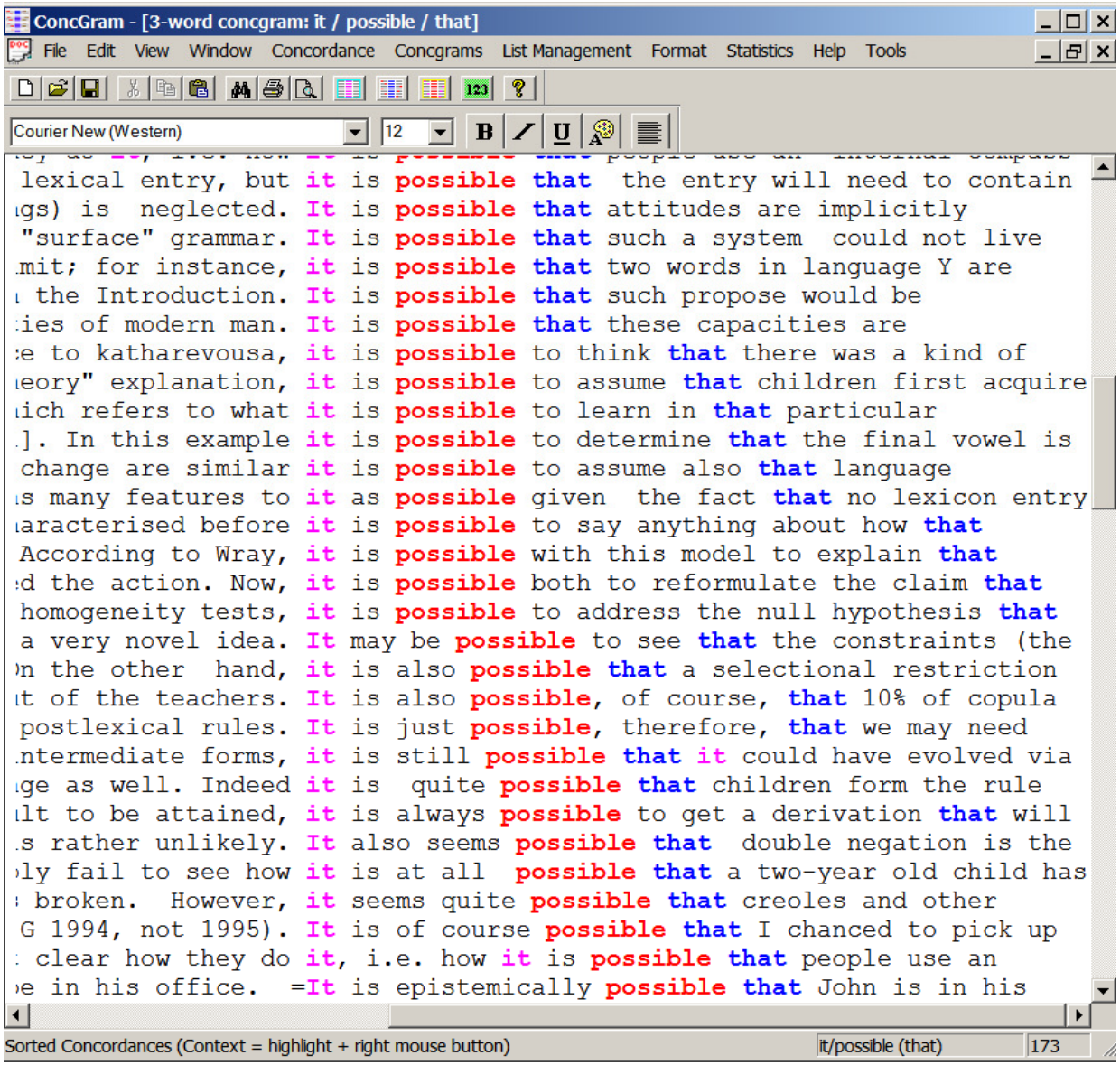
829 files in current corpus | 2,628,665 words, 87,799 types

Collocate

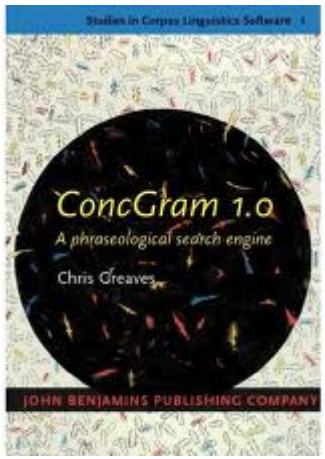
Ute Römer (uroemer@gsu.edu)

Georgia State University

2. Corpus analysis and phraseology



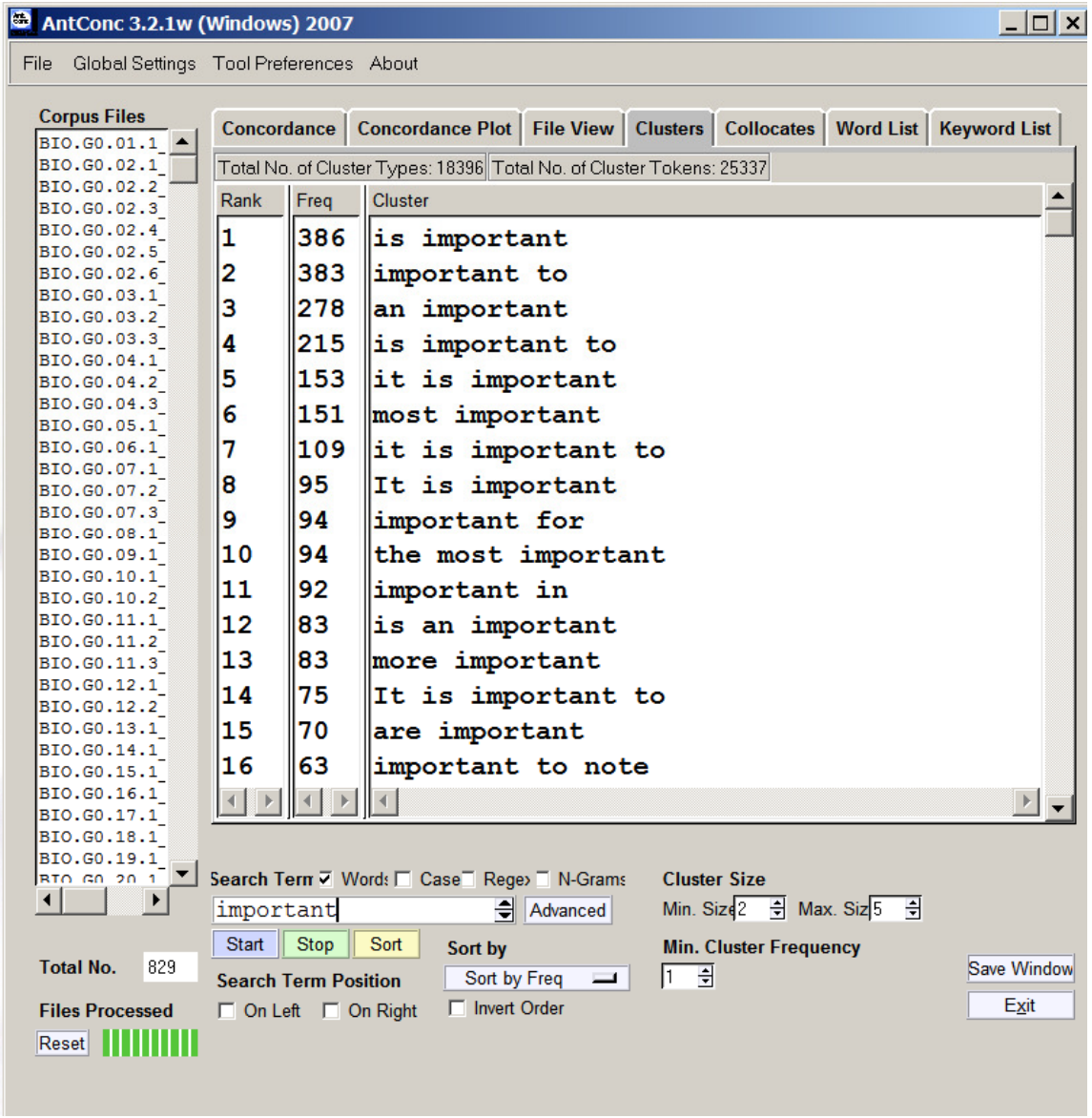
ConcGram



Ute Römer (uroemer@gsu.edu)

Georgia State University

2. Corpus analysis and phraseology

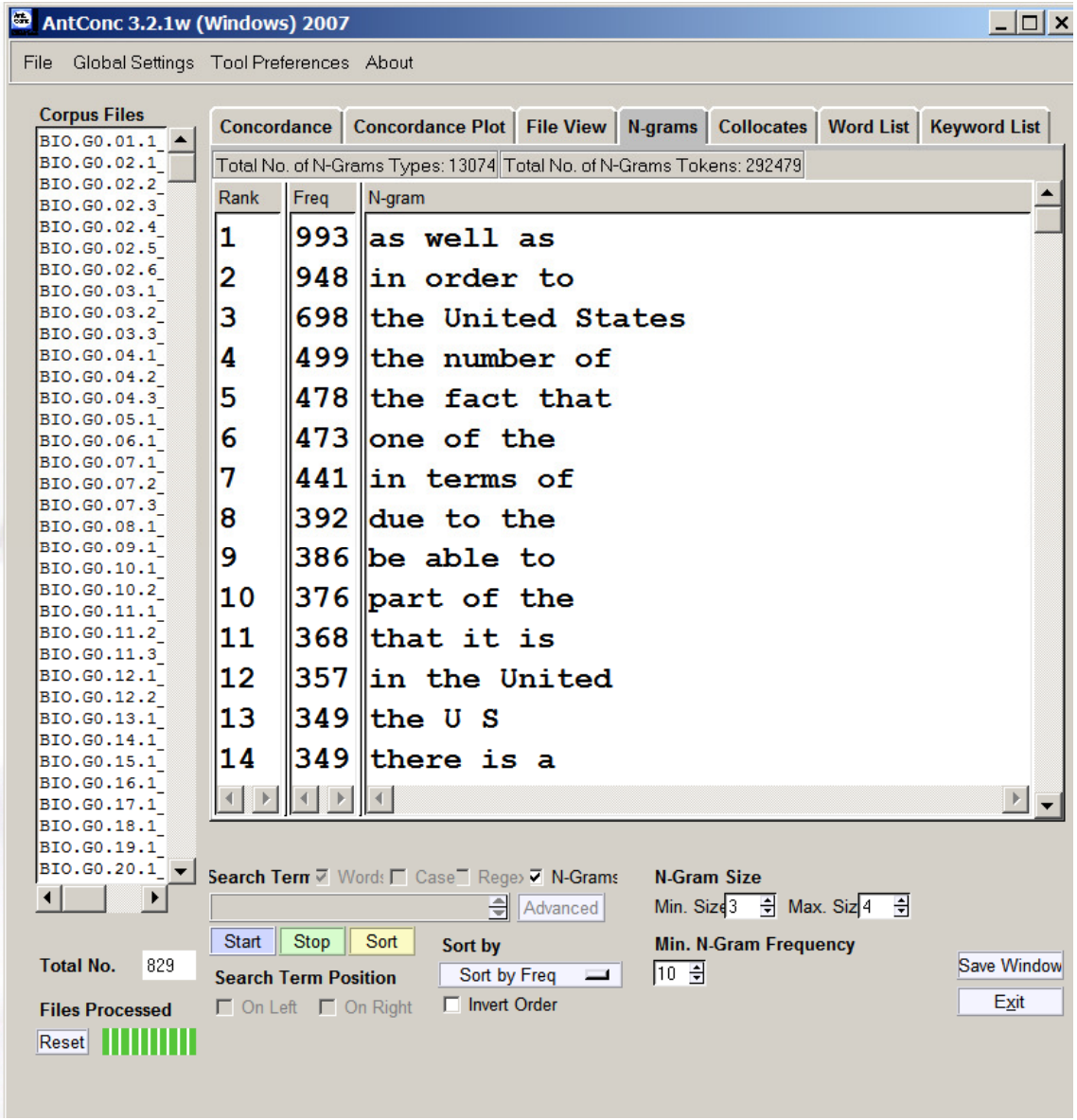


AntConc

Ute Römer (uroemer@gsu.edu)

Georgia State University

2. Corpus analysis and phraseology

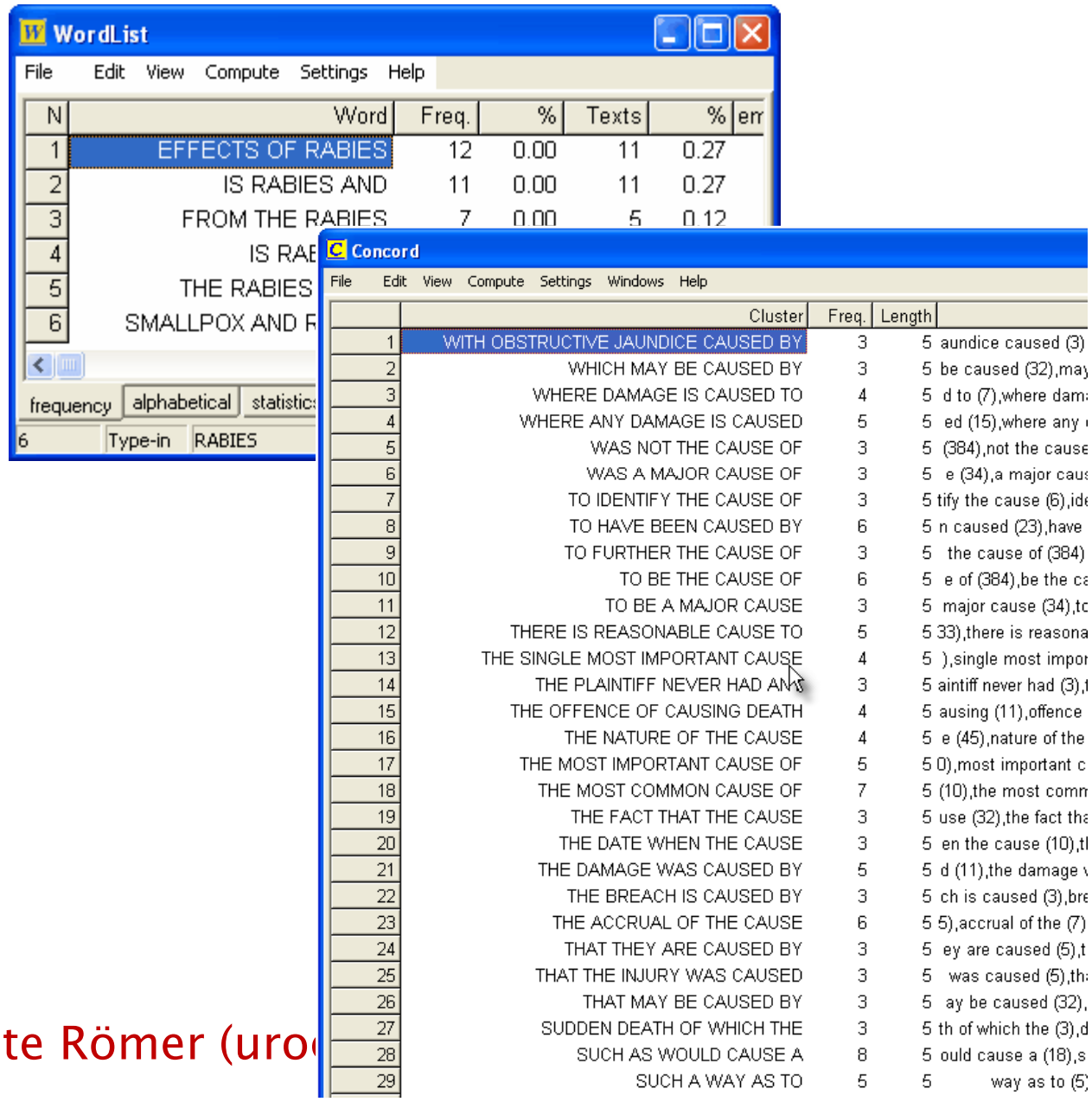


AntConc

Ute Römer (uroemer@gsu.edu)

Georgia State University

2. Corpus analysis and phraseology

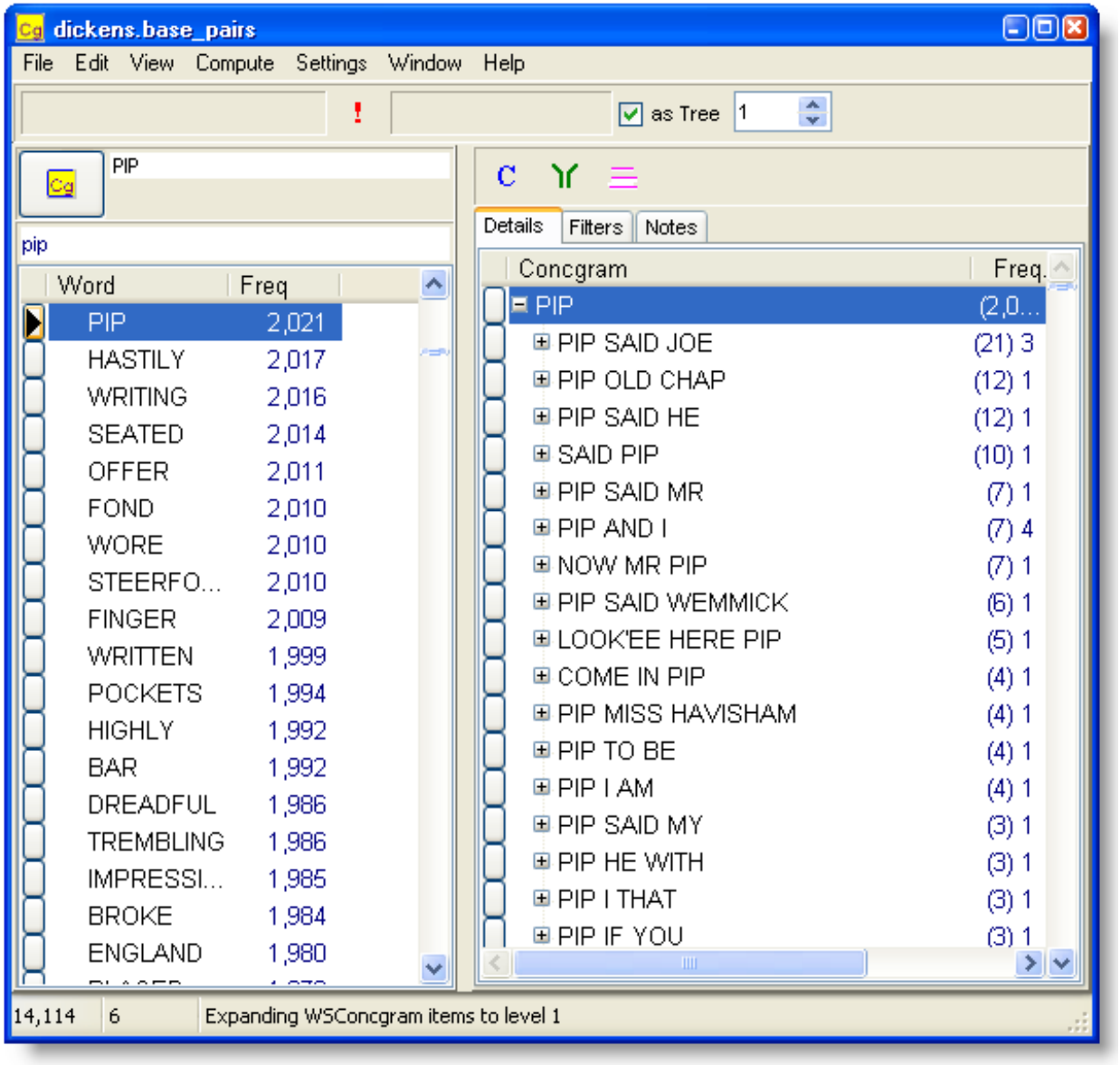


WordSmith Tools Cluster function

Ute Römer (uro

Georgia State University

2. Corpus analysis and phraseology



WordSmith Tools Concgram function

Ute Römer (uroemer@gsu.edu)

Georgia State University

3. Phraseology and academic corpora

- How can corpora of academic discourse be used in phraseology research?
- Present findings from two case studies on:

  ...the phraseological profile of book reviews in linguistics

  ...the distribution of phraseological items across student papers

Establishing the phraseological profile of a text type

The construction of meaning in academic book reviews

Ute Römer
University of Michigan

Starting from the observation that meaning does not primarily reside in individual words but in the phrase, this paper focuses on the examination of recurring phrases in language. It introduces a new analytical model that leads corpus researchers to a profile of the central phraseological items in a selected text or text collection. In this paper, the model is applied to a 3.5-million word corpus

Investigating the interaction between phraseological items and textual position

Matthew Brook O'Donnell & Ute Römer

1. Introduction

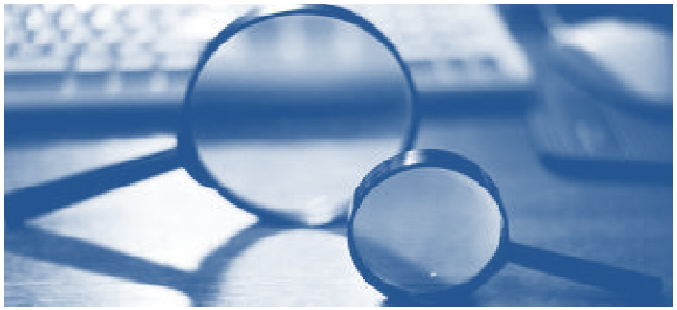
Words and phrases are not distributed randomly across a text. Instead, they are connected to its structure and associated with particular textual positions. Michael Hoey (2005) refers to this phenomenon as "textual colligation" (see Section 2 below).

Focusing on advanced student writing from different academic disciplines, this paper identifies the most common phraseological items in the Michigan Corpus of Upper-level Student

Ute Römer (uroemer@gsu.edu)

Georgia State University

Case study #1



Ute Römer (uroemer@gsu.edu)

Georgia State University

3.1 The phraseological profile of book reviews

- How can we uncover the phraseological profile of a text or text type?

  → Development of the Phraseological Profile Model (PP Model)

- PP Model summarizes the underlying procedure of text/corpus analysis and consists of 4 central steps:
  1. identification of phraseological items
  2. determination of item-internal variation
  3. examination of functions of the identified items
  4. analysis of the distribution of items across texts
- Application of the model to a 3.5 million word corpus of linguistic book reviews: Book Reviews In Linguistics Corpus (BRILC); 1,500 texts, 3.5 mio words

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

Step 1: Identification of phraseological items

- Take a corpus-driven approach (fully automatic extraction of items)
- Work with whole texts, not samples
- Use phraseological search engines to extract candidate item lists
- Tools: *Collocate* (Barlow 2004), *ConcGram* (Greaves 2009), *kfNgram* (Fletcher 2002-7)
- Extract lists of n-grams, p-frames, concgrams
- Manually filter/'weed' the lists for interesting and meaningful items

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

Step 1: Identification of phraseological items

Hits	4-gram
562	on the other hand
442	at the end of
428	on the basis of
411	as well as the
356	at the same time
330	the end of the
301	of the book is
288	in the case of
268	the fact that the
226	on the one hand
216	a wide range of
215	in the context of
214	the rest of the
207	in terms of the
206	to the study of

Figure 1. Extract of a BRILC 4-gram list (*Collocate* output)

it would be * to	101	10
it would be interesting to	44	
it would be useful to	14	
it would be nice to	11	
it would be better to	9	
it would be possible to	5	
it would be helpful to	5	
it would be fair to	4	
it would be difficult to	3	
it would be necessary to	3	
it would be good to	3	
it * be interesting to	58	3
it would be interesting to	44	
it will be interesting to	8	
it might be interesting to	6	

Figure 2. Example p-frames in BRILC, with numbers of tokens and numbers of variants (*kfNgram* output)

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

ConcGram output for it+be+interesting: constituency variation

59	children's grammars in any particular case.	It	would be	interesting to know if the children in
60	UK and US communities, all speakers of English,	it	would be	interesting to know how they interact
61	arguments that are really consistent.	It	would be	interesting to know if there are
62	. In fact,	it	might be	interesting to see if such an account
63	ldwide and	it	would be	interesting to see if, when and how it
64	s reader.	It	would be	interesting to determine if his
65	s respect	it	would be	interesting to know by which methods the
66	to me that	it	would be	interesting to examine such problems in
67	phora, and	it	would be	interesting to see how his theory can
68	calisation	it	would be very	interesting to have a survey of the
69	with respect to semantic transparency; again,	it	would be very	interesting to see this pursued in
70	French. Finally, from a theoretical standpoint,	it	would be very	interesting to expand this analysis
71	directions for future research, noting that	it	would be especially	interesting to follow the
72	explains the shift from OV to VO in English.	It	would be particularly	interesting to see if this
73	didn't find this book as exciting as I had hoped	it	might be, although Part 4 was quite	interesting,
74	"baron" is solved very elegantly in the paper,	it	would be	interesting to discuss the
75	for beginners in semantics. In my opinion,	it	would be	interesting to see how this ontological
76	However, only noun derivatives are discussed,	it	would be	interesting at least to mention verbal
77	work is felt most positively" (p. 22).	It	should be	noted that some interesting results
78	The authors also prove a pumping lemma.	It	woul! d be	interesting to see further
79	as occasionally appear in Linguist List reviews:	it	wouldn't be very	interesting, I didn't make a
80	detailed account is given on this work, though	it	seems to be very	interesting for the linguist's
81	few, oblique, and confined to the endnotes.	It	would also be	interesting to set Hornstein's view
82	a translation of a book title was omitted.	It	would also be	interesting to see if some of the
83	future generative work on corpora. Maybe	it	would also be	interesting to test the analyses in
84	restricted to the second definition. Of course,	it	would also be	interesting to find out that
85	(as entries or sub-entries, for instance). So,	it	should also be	interesting to find, among the
86	those that are involved in dictionary-making and	it	should also be	interesting to all dictionary
87	of identity between a constituent and its copy.	It	might however be	interesting to seek a connection
88	language community and their self-perception.	It	might prove to be	interesting to compare the
89	mind, Mel'cuk's criteria seem fairly reasonable.	It	would, however, be	interesting to study the
90	discourse analysis, rhetoric, semantics, etc.	It	would certainly be very	interesting to see what
91	successfully manages to carry out the action.	It	would most certainly be	interesting to look at
92	AND SYNTACTIC THEORY" by Alison Henry).	It	seems to me that it would be	interesting to
93	interesting topic for their theses. Summing up,	it	must be said that this book is indeed a very	
94	interesting (and reasonably persuasive as far as	it	goes), but it seems to be unnecessary. Given the	
95	been interesting to know. Having said that,	it	must be stressed that this criticism of too	
96	the paper is quite interesting and substantial.	It	seems to be specially valuable, first of all, due	
97	very easy and yet very interesting to read.	It	would be enlightening for both the professor or	
98	very straightforward but very interesting, and	it	would be still more interesting to see it applied	
99	is highly stimulating, original and interesting.	It	must be read by anyone interested in time, tense	
100	paper gives us an idea how interesting	it	might be if such a theory was successfully	

Step 1: Identification of phraseological items

- Manual 'weeding': requires to go back and forth between lists and concordances
- Important to look at wider context of items
- Procedure resulted in a database of around 8,000 phraseological items (freq of occurrence in BRILC ≥20)
- Results reported on here are based on subsets of high-frequency items

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

Step 2: Determination of item-internal variation

- This step examines how variable (or fixed) a repeatedly occurring sequence of words is
  - Where in an item does variation occur? What are the most frequent variants in a * slot in a p-frame?
  - P-frame: n-gram with one internal variable slot, e.g. A*CD, AB*D
- Step 2 also measures the degrees of variation of common word sequences: VPR (variant/p-frame ratio – a TTR for p-frames)
- Items with low VPRs (e.g. *on the * hand*, *a * range of*) have few variants per p-frame; low degree of variation

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

Step 2: Determination of item-internal variation

Span	P-frame types	Examples from BRILC
3	A*C	*the * of, a * job*
4	A*CD	*the * that the, the * of language*
	AB*D	*at the * of, on the * hand*
5	A*CDE	*the * of the book, at * end of the*
	AB*DE	*in the * of the, the first * of the*
	ABC*E	*the book is * into, it would be * to*
6	A*CDEF	*the * part of the book, the * of the book is*
	AB*DEF	*at the * of the book, the first * of the book*
	ABC*EF	*in the first * of the, the book is * into three*
	ABCD*F	*it would have been * to, book is divided into * parts*

Variant-type distribution

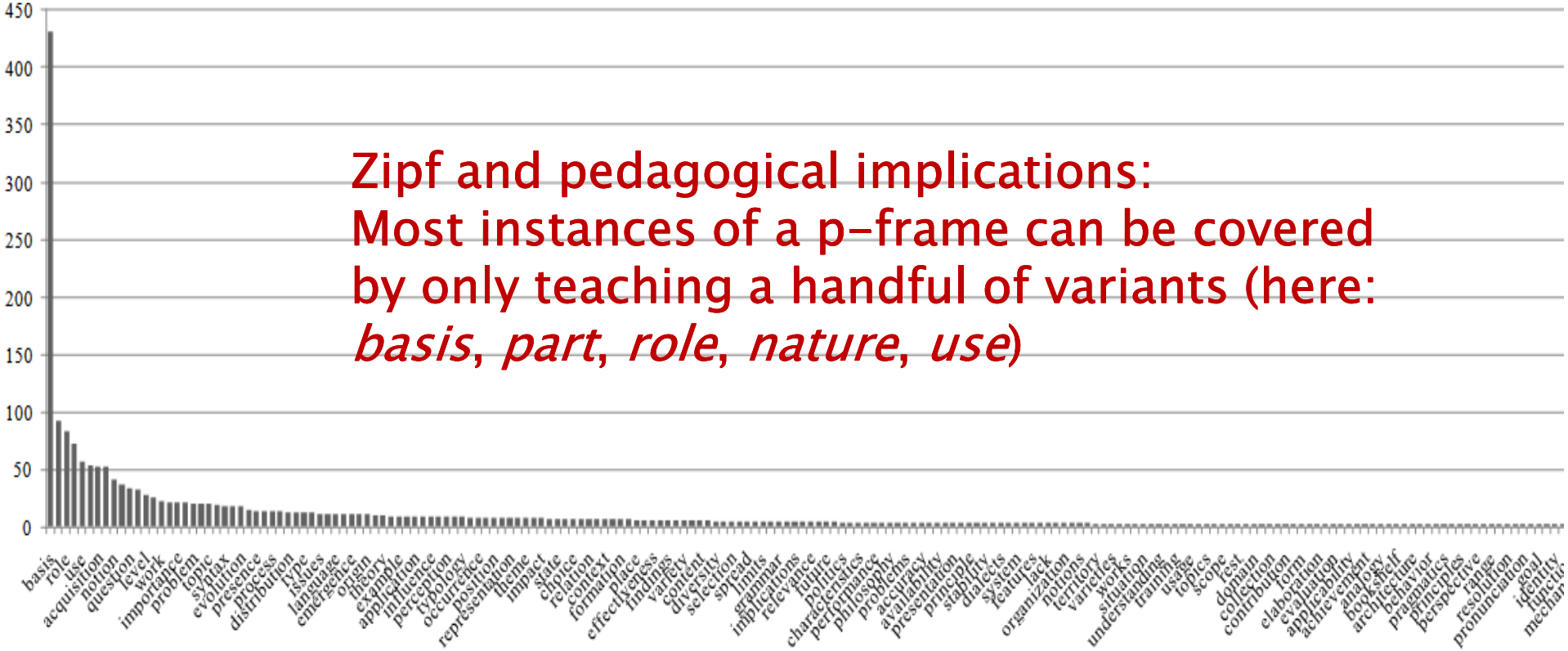
- Is the distribution of variants per p-frame Zipfian?

- i.e. does a small number of variants (types) account for a large share of p-frame tokens?

Ute Römer (uroemer@gsu.edu)

Georgia State University

Step 2: Determination of item-internal variation



Zipf and pedagogical implications:
Most instances of a p-frame can be covered by only teaching a handful of variants (here: *basis*, *part*, *role*, *nature*, *use*)

Figure 4. Variant type distribution for the p-frame *on the * of*, following Zipf's law

basis 431, *part* 92, *role* 83, *nature* 73, *use* 57

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

Step 3: Examination of functions of the items

- What meanings are expressed by the most frequent phraseological items in our text collection?
- Need to look at items in context (concordance analysis)
- Assigned one meaning to each item (while not all items were monofunctional, one meaning always dominated)
- P-frames were only assigned a function if variants were semantically related
- The examined items convey 4 functions:
  1. express EVALUATION
  2. refer to the STRUCTURE of the book under review
  3. refer to the CONTENT of a book
  4. organize the DISCOURSE

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

Step 3: Examination of functions of the items

Examples…

- Expressing evaluation:
  *it would be * to, it is * that, a wide range of, it is not clear*
- Referring to the structure of a book:
  *in the * chapter, in the first part, the * of the book*
- Referring to a book's content:
  *the history of, the * of English, the relationship between * and*
- Organizing the discourse:
  *in order to, with respect to, with * to the*

Ute Römer (uroemer@gsu.edu)

Step 4: Analysis of the distribution of items across texts

- Where in a text (here a book review) does an item most commonly occur?
- Are there any relations between phraseol. items and text structure? (cf. "textual colligation", Hoey 2005)
- Knowing where in a text an item most commonly occurs and which position(s) it avoids, facilitates text processing and production
- Textual distribution of items can be observed in corpus tools (WordSmith Tools "dispersion plot", AntConc "concordance plot", MonoConc Pro "distribution of hits")
- But: hard to evaluate results systematically (other than eyeballing bar code lists)

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

Step 4: Analysis of the distribution of items across texts

- Division of each file into four parts of equal size (Q1/Q2/Q3/Q4); THANKS to Matthew B. O'Donnell!
- Quarters correspond with structural elements (introduction, summary of contents, critical evaluation)
- Retrieval of n-gram/p-frame lists from 'quartered' version of BRILC
- Compute shares of item occurrence across quarters (for top-300 phraseological items, >200 hits)
- Each text quarter prefers different n-grams
- Selected high-frequency items show interesting textual distributions...

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

Step 4: Analysis of the distribution of items across texts

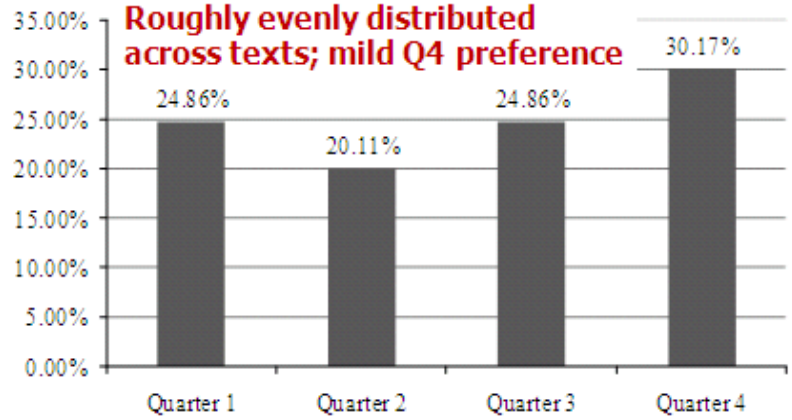


Figure 5. *At the same time* across texts

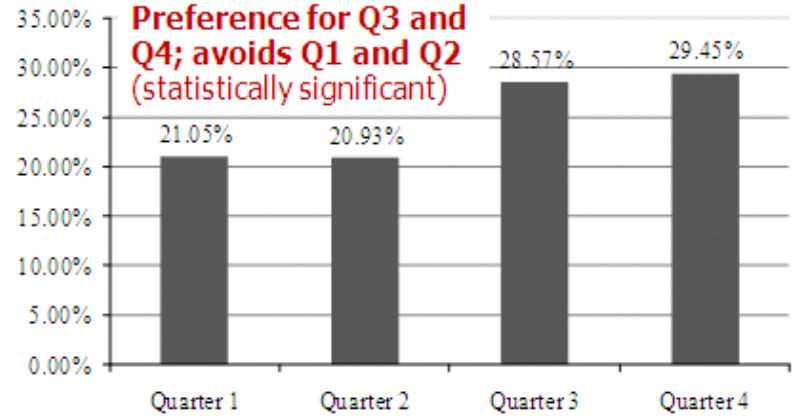


Figure 6. *On the * hand* across texts

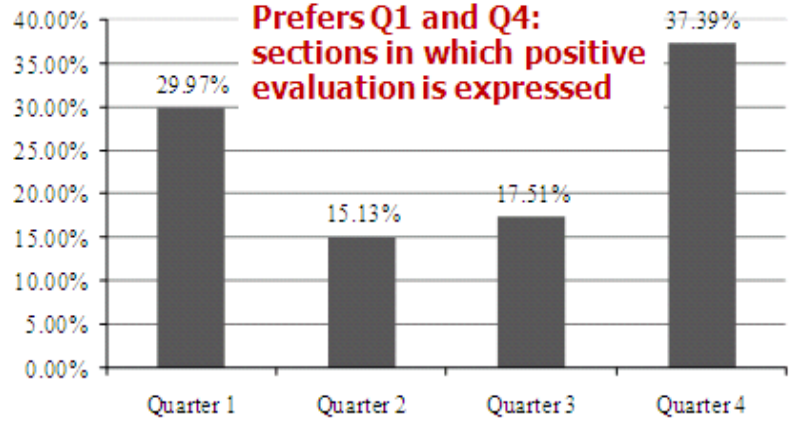


Figure 7. *A * range of* across texts

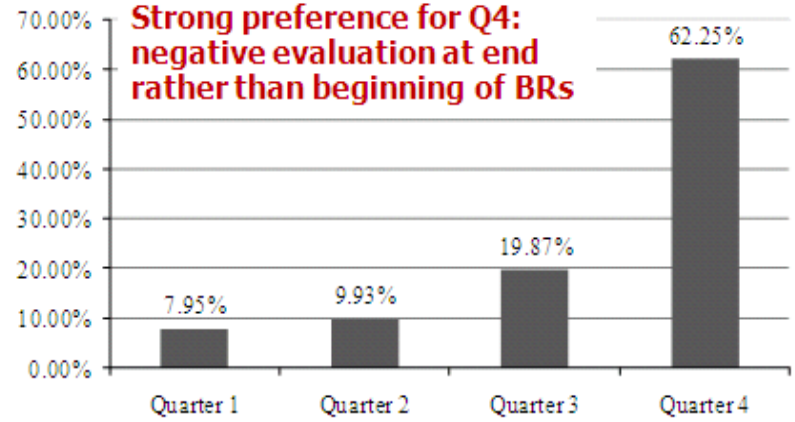


Figure 8. *It would have been* across texts

Ute Römer (uroemer@gsu.edu)

Georgia State University

3.1 The phraseological profile of book reviews

- Study has put forward a new model for text and corpus analysis
- 4 steps (I–V–F–D) help establish the phraseological profile of a text/text collection:
  - What are the central phraseological items?
  - How variable are they? What type(s) of variation do they allow?
  - What functions do they most commonly express?
  - How are items distributed across texts? How does their occurrence relate to text structure?
- Analyses show: a lot of important information about the co-selection and textual distribution of words/ phrases has not yet been captured

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

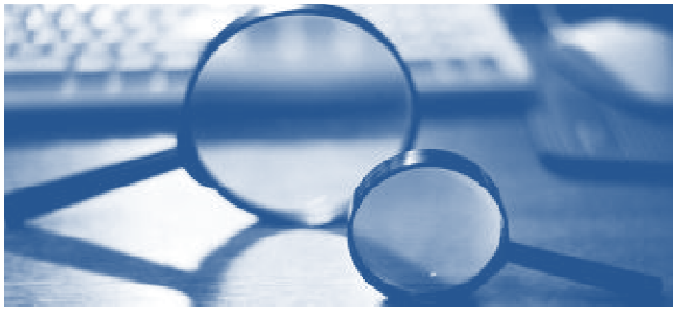
3.1 The phraseological profile of book reviews

- Outcome: a text-type specific inventory of items, their variation, functions and textual distribution
- Findings have implications for creation of text-type or genre specific reference works
- Also: implications for pedagogical practice – What do learners (or in this case novice academic writers) need to know about the use of common phrases in a particular text type?
- Still: need to find out a lot more about patterns in the language of book reviews

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

Case study #2



Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

3.2 The textual distribution of phraseological items

- Focus on phraseology of student academic writing
- Paper identifies common phraseological items in MICUSP and relates them with text structure
- Analysis is based on Hoey's (2005) observations on textual colligation
  - Words and phrases may carry with them associations for occurrence at a specific location in a text or textual unit
  - Preference for or avoidance of textual positions
- Which items do typically occur at the beginning or end of a text, paragraph, or sentence?

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

3.2 The textual distribution of phraseological items

MICUSP

- Michigan Corpus of Upper-level Student Papers
- 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
- See http://micusp.elicorpora.info

Ute Römer (uroemer@gsu.edu)

Georgia State University

3.2 The textual distribution of phraseological items

For each paragraph in every MICUSP XML document:

1. tokenize text into sentences and words (NLTK)
2. process each word, recording:
   - paper ID, discipline & student level
   - up to 8-word right context
   - word # within sentence & sentence length
   - word # within paragraph & paragraph length
   - word # within text & text length
   - sentence # & # of sentences within paragraph
   - paragraph # & # of paragraphs in text

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

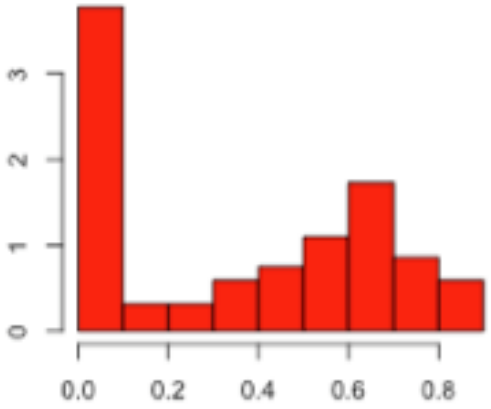
3.2 The textual distribution of phraseological items

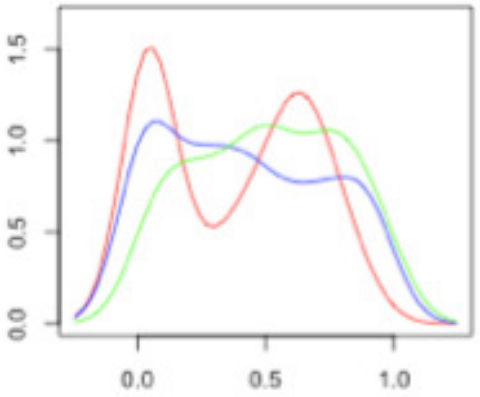
Paper ID	w1	w2	w3	w4	w5	w6	w7	w8	w9	Word in Sent	Sent length	Word in Para	Para length	Word in Text	Text length
PHY.G2.02.1	supersymmetric	theories	have	long	provided	a	theoretical	framework	which	1	19	1	99	1	1941
PHY.G2.02.1	theories	have	long	provided	a	theoretical	framework	which	can	2	19	2	99	2	1941
PHY.G2.02.1	have	long	provided	a	theoretical	framework	which	can	overcome	3	19	3	99	3	1941
PHY.G2.02.1	long	provided	a	theoretical	framework	which	can	overcome	many	4	19	4	99	4	1941
PHY.G2.02.1	provided	a	theoretical	framework	which	can	overcome	many	of	5	19	5	99	5	1941
PHY.G2.02.1	a	theoretical	framework	which	can	overcome	many	of	the	6	19	6	99	6	1941
PHY.G2.02.1	theoretical	framework	which	can	overcome	many	of	the	shortfalls	7	19	7	99	7	1941
PHY.G2.02.1	framework	which	can	overcome	many	of	the	shortfalls	of	8	19	8	99	8	1941
PHY.G2.02.1	which	can	overcome	many	of	the	shortfalls	of	the	9	19	9	99	9	1941
PHY.G2.02.1	can	overcome	many	of	the	shortfalls	of	the	standard	10	19	10	99	10	1941
PHY.G2.02.1	overcome	many	of	the	shortfalls	of	the	standard	model	11	19	11	99	11	1941
PHY.G2.02.1	many	of	the	shortfalls	of	the	standard	model		12	19	12	99	12	1941
PHY.G2.02.1	of	the	shortfalls	of	the	standard	model			13	19	13	99	13	1941
PHY.G2.02.1	the	shortfalls	of	the	standard	model				14	19	14	99	14	1941
PHY.G2.02.1	shortfalls	of	the	standard	model					15	19	15	99	15	1941
PHY.G2.02.1	of	the	standard	model						16	19	16	99	16	1941
PHY.G2.02.1	the	standard	model							17	19	17	99	17	1941
PHY.G2.02.1	standard	model								18	19	18	99	18	1941
PHY.G2.02.1	model									19	19	19	99	19	1941
PHY.G2.02.1	the	major	shortfall	of	supersymmetr	susy	however	has	been	1	18	20	99	20	1941
PHY.G2.02.1	major	shortfall	of	supersymmetr	susy	however	has	been	a	2	18	21	99	21	1941
PHY.G2.02.1	shortfall	of	supersymmetr	susy	however	has	been	a	complete	3	18	22	99	22	1941
PHY.G2.02.1	of	supersymmetr	susy	however	has	been	a	complete	lack	4	18	23	99	23	1941
PHY.G2.02.1	supersymmetry	susy	however	has	been	a	complete	lack	of	5	18	24	99	24	1941
PHY.G2.02.1	susy	however	has	been	a	complete	lack	of	direct	6	18	25	99	25	1941

Ute Römer (uroemer@gsu.edu)

3.2 The textual distribution of phraseological items

- Database enables us to extract positional data according to textual units
- By means of an R-script: Derive histograms with set number (e.g. 10) of bins and continuous density distribution for sentence/paragraph/text



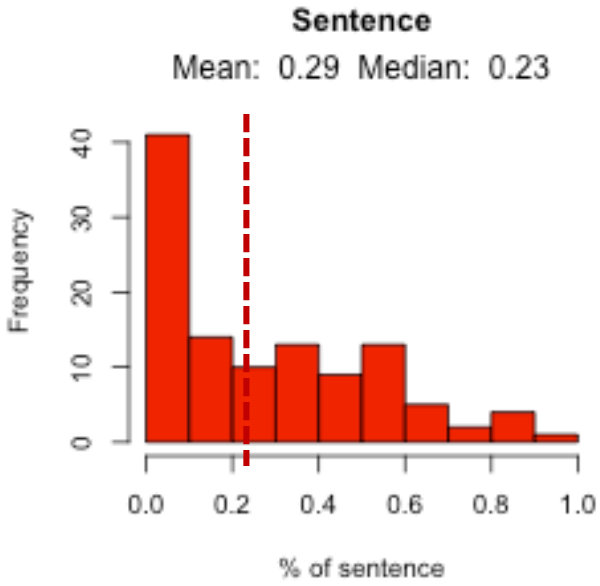


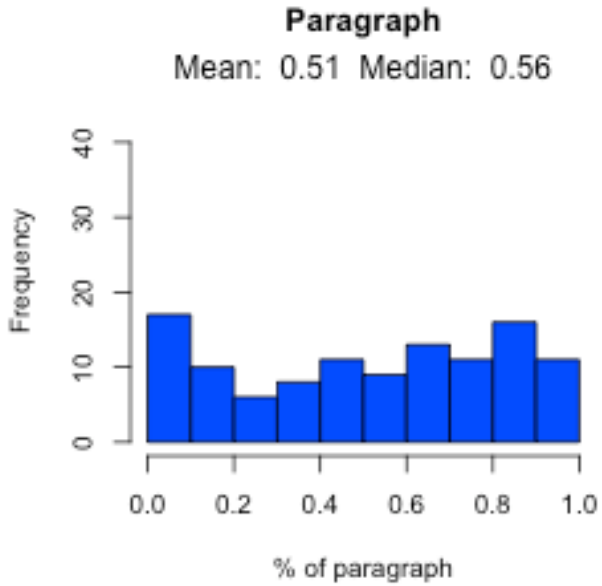
Ute Römer (uroemer@gsu.edu)

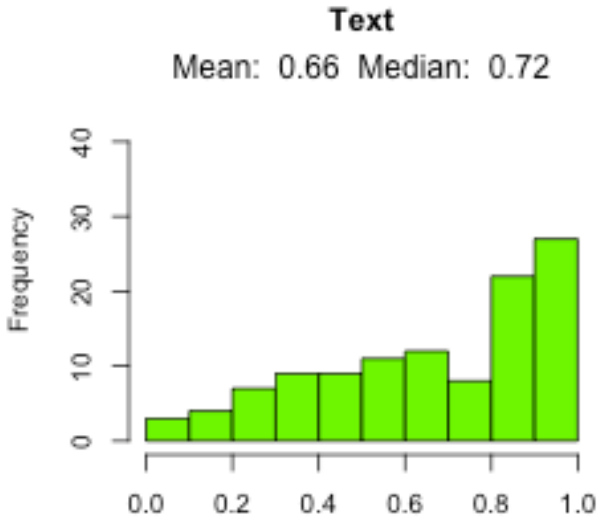
Georgia State University

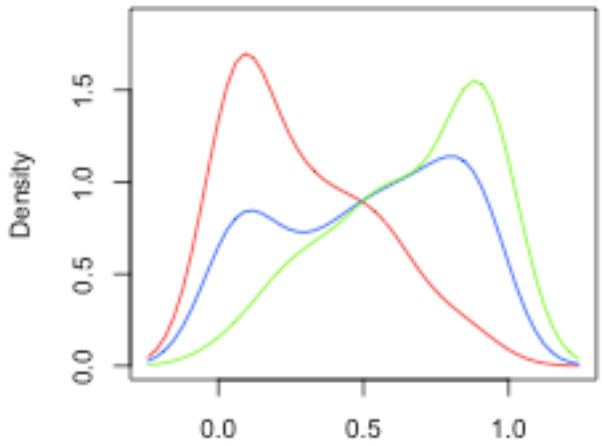
3.2 The textual distribution of...

*it would be * to*





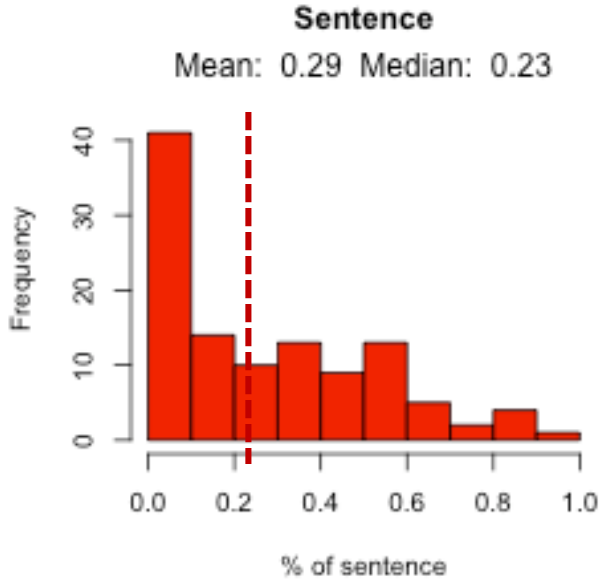


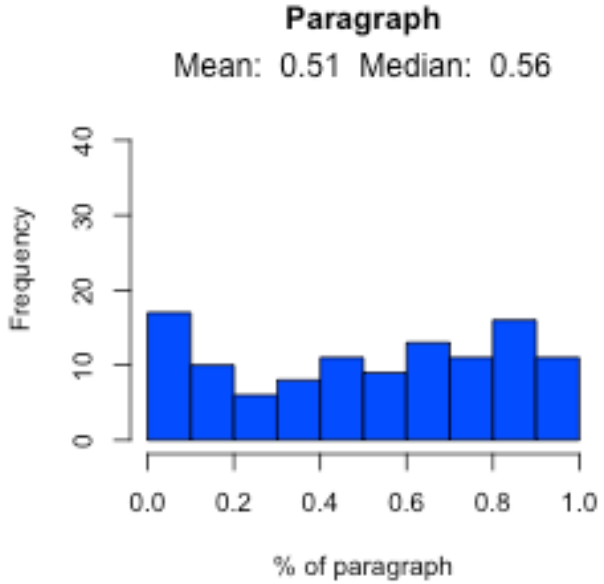


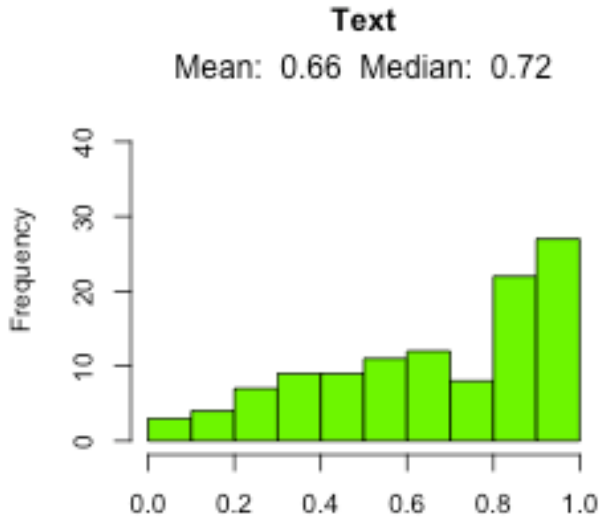
interesting	22
hard	10
difficult	9
useful	6
important	5
easy	4
helpful	4
worthwhile	3
preferable	3
easier	3
possible	3
unfair	3
false	2
appropriate	2
impossible	2
best	2
better	2
necessary	2
wise	2

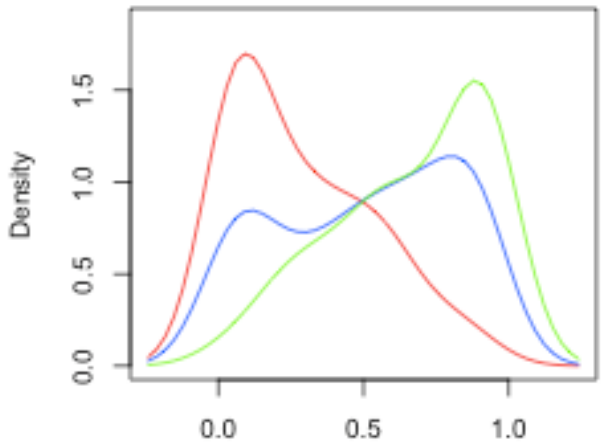
3.2 The textual distribution of...

*it would be * to*









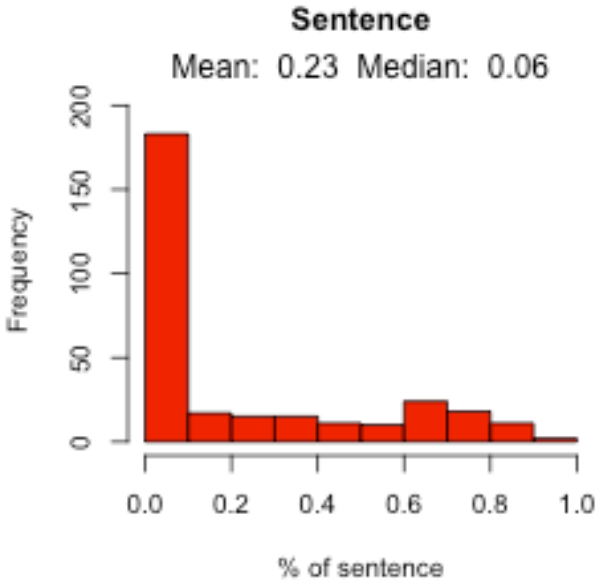
-Avoidance of sentence-final position

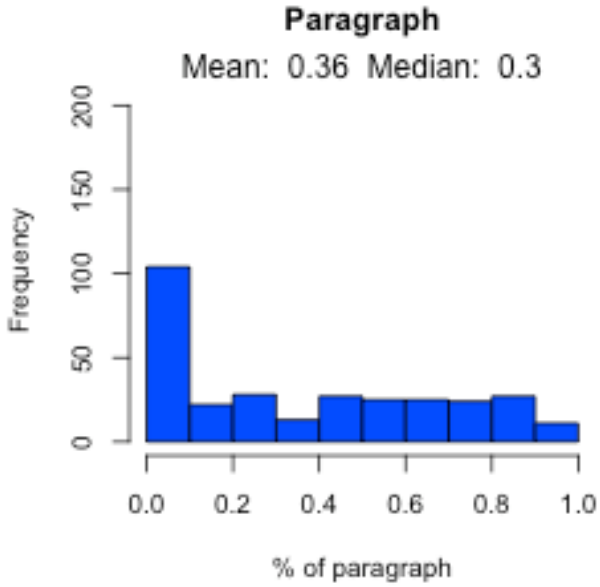
-No marked preference for particular paragraph positions

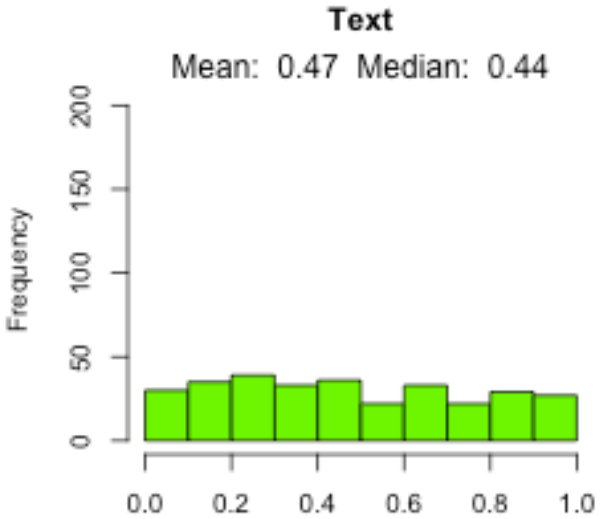
-Text-final preference

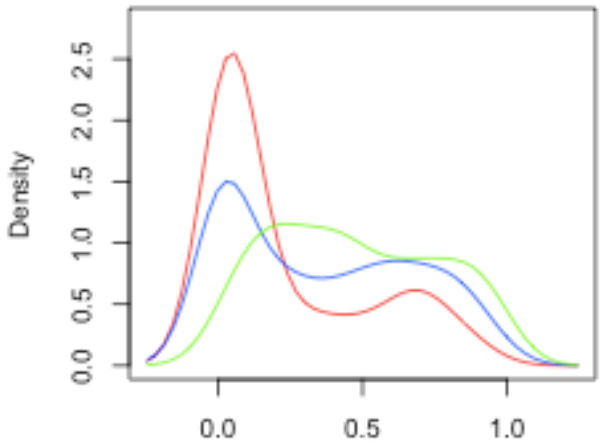
3.2 The textual distribution of...

in addition to









-Strong sentence-initial preference

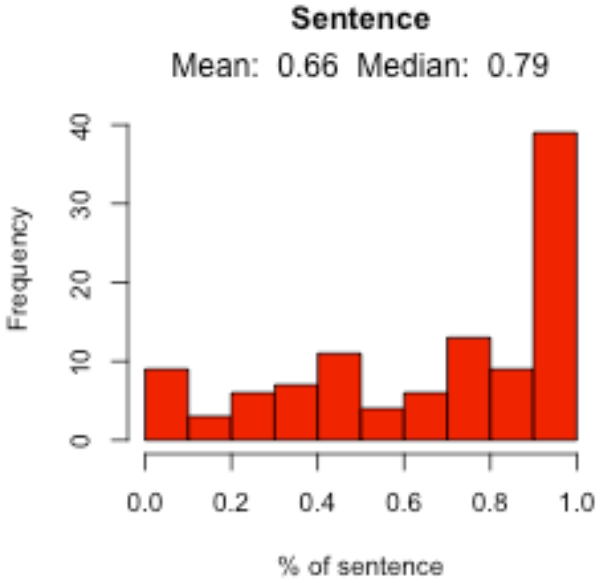
-Paragraph initial preference

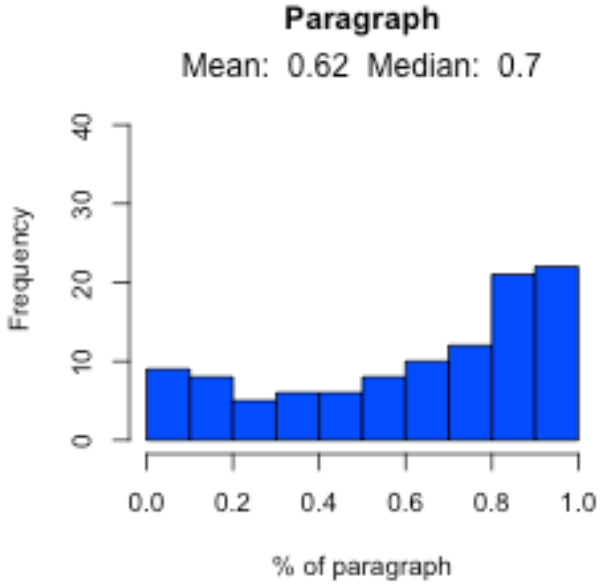
-No marked text-positional preferences

Georgia State University

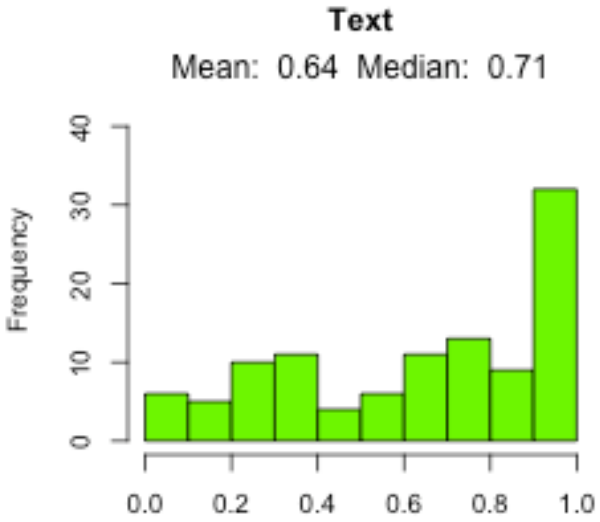
3.2 The textual distribution of...

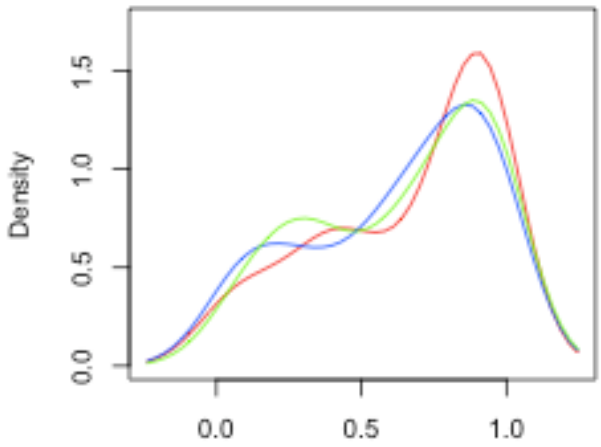
in the future





- Strong sentence and text final preference
- Tends towards the end of the paragraph





Georgia State University

3.2 The textual distribution of phraseological items

- Alternative approach: Group items by positional tendencies
  - Retrieve all 3-5 grams and 3-5 p-frames occurring 100+ in MICUSP
  - Calculate percentage distributions using 10 bins for sentences, paragraphs and text positions according to first word of each item
  - Sort items according to various positions, e.g. sentence-initial (S1+S2+S3), paragraph-final (P10+P9+P8), text-medial (T4+T5+T6) to group them by text positional behavior
- Look at results for positional tendencies of items within the paragraph

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

3.2 The textual distribution of phraseological items

Paragraph initial	*medial*	*final*
one of the most	in addition the	in this way
of this study	whether or not	in the future
of the most	he does not	the absence of
in addition to	it is possible	the possibility of
there are many	in other words	should not be
the purpose of	found that the	likely to be
a series of	at the time	as a whole
one of the	as a result of	may not be
the issue of	this is not	but it is
the concept of	would have been	needs to be

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

3.2 The textual distribution of phraseological items

Findings are pedagogically relevant:

- 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 are used where in a text
- In teaching and in materials creation it appears that textual positioning should be given greater attention

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

4. Concluding thoughts

- Corpora as powerful tools in the study of academic discourse
- Corpora and corpus-analytic techniques help highlight phraseological items in texts
  - → insights into meaning creation
  - → insights into (disciplinary) terminology
  - → insights into textual colligation
  - → insights into aboutness/topicality
  - → insights into stylistic preferences of writers
- Use new tools and techniques to explore corpora in new ways and discover new aspects of the patterned nature of language

Ute Römer (uroemer@gsu.edu)

GeorgiaStateUniversity

More 4-grams...

Freq	Collocation
927	thank you very much
218	thank you thank you
55	thank you for that
53	thank you mr chairman
52	well thank you very
51	very much thank you
42	thank you for your
41	you very much thank

uroemer@gsu.edu
www.gsu.edu/alesl