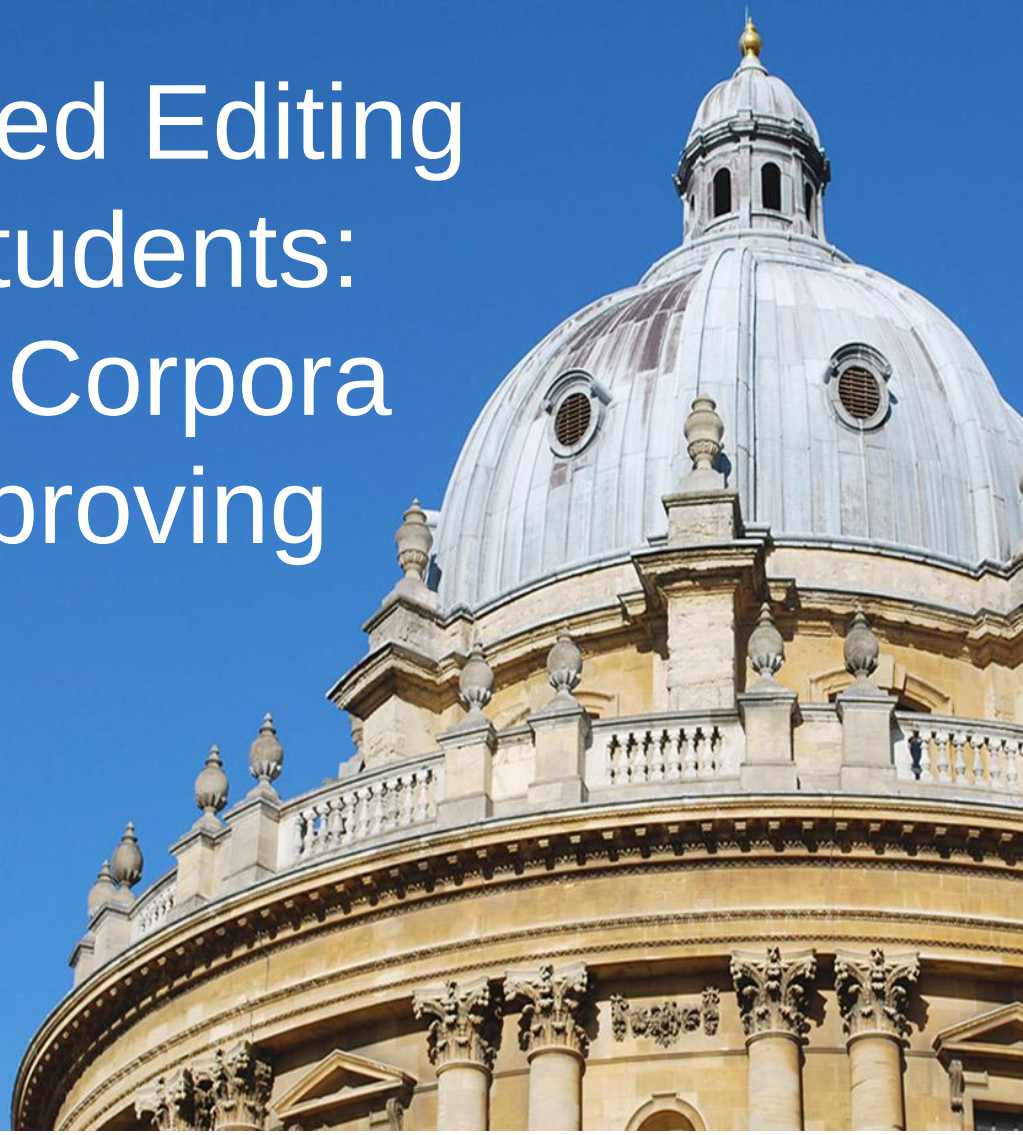
BAAL Corpus SIG Symposium, Using Corpora in EAP
31st March 2017, Durham University

UNIVERSITY OF OXFORD

Corpus-Assisted Editing for Doctoral Students: Do-It-Yourself Corpora & Tools for Improving Writing

Maggie Charles
Language Centre
University of Oxford



Editing your Thesis with Corpora: Course Details

Aim:	enhance graduates' editing skills prior to thesis submission
Frequency:	2-3 times per year (10 in total)
Timing:	1 2-hour session/week for 6 weeks
Venue:	computer laboratory
Class size:	7 – 12 (maximum 12)
Composition:	multi-disciplinary
Software:	AntConc (Anthony 2014) AntFileConverter (Anthony 2015)

AntConc Tools

Concordance: usage of search term; frequency, context

Clusters: groups of words which include the search term

Collocates: a list of words that co-occur with the search term

Keyword List: words which are unusually frequent or infrequent in one corpus when compared to a reference corpus

N-grams: all groups of words of size *n* in the corpus

Concordance Plot: a graphic display of the search term distribution

Word List: a list of all words in the corpus with frequencies

Research Questions

- How useful are the individual corpus tools for editing purposes?
- What are the affordances of each tool that make it useful for editing purposes?

Two Corpus Types

1. **Do-It-yourself Corpus of Research Articles in student's own field/topic area**
   * based on downloaded files in own bibliography;
   * may include subcorpora of different topics/genres

2. **Do-It-Yourself Corpus of Student's Own Writing**
   * chapters of thesis as individual files;
   * may include subcorpora of other writing (e.g. proposals, Master's dissertation)

Participants

Doctoral students who have completed at least **1 substantial chapter** of their thesis

66 students (2012 – 2015)

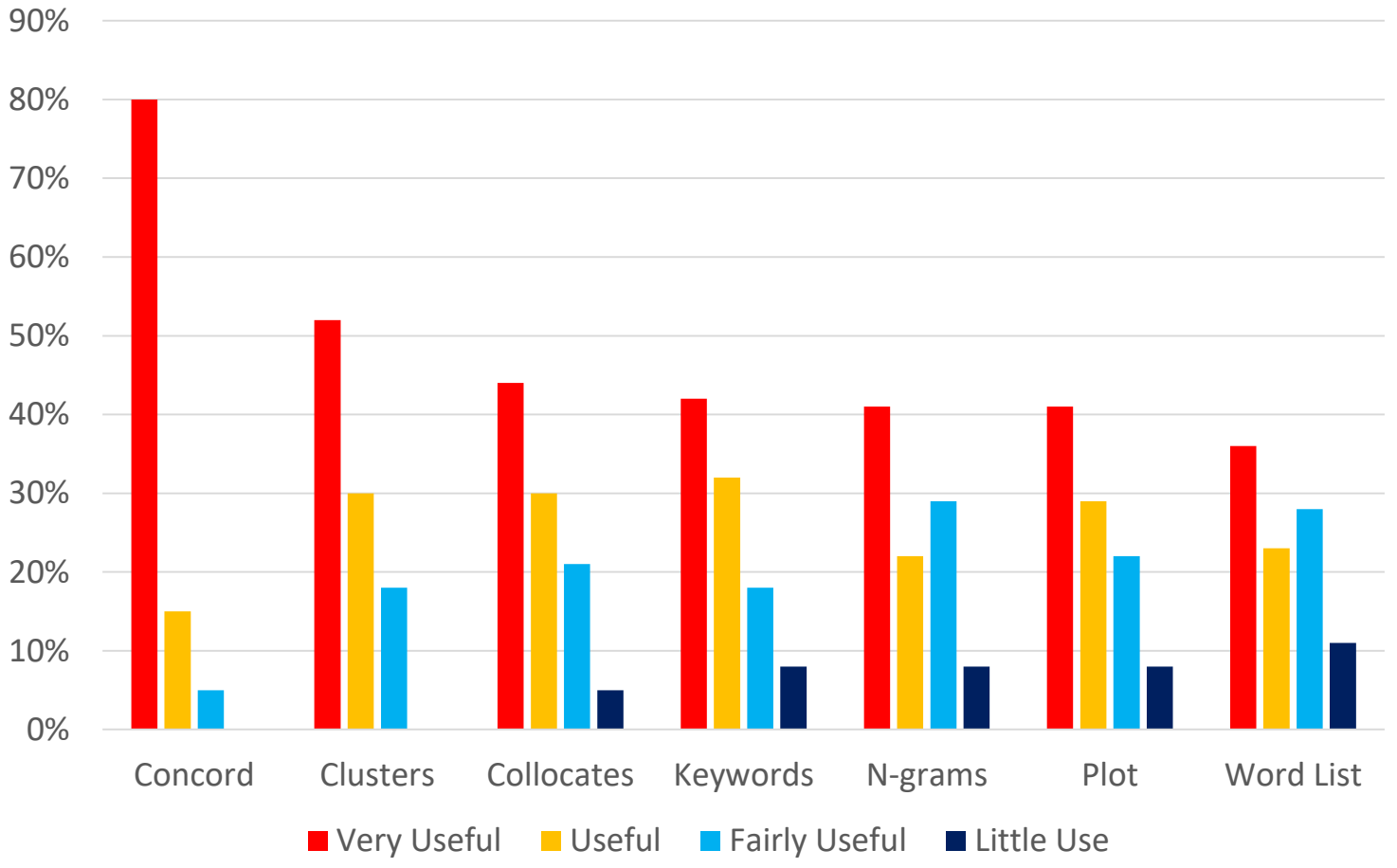
Fields

Natural Science 41%
Social Science 30%
Humanities 29%

Course Programme

Topic	Tool
1. Using concordances to answer grammar, vocabulary and usage queries	AntConc Concordance
2. Building your corpus of research articles; answering your own editing queries	AntFileConverter
3. Finding collocations and semi-fixed phrases; building a corpus of your own writing	Clusters Collocates
4. Examining the words you use; checking for consistency; comparing your own writing with expert texts	Word List N-Grams
5. Tracing content, themes, terminology, citation throughout your own writing	Concordance Plot
6. Comparing individual chapters to the whole text; comparing your own writing with expert texts	Keyword List

Student Evaluation of Tools (*n* = 66)



Editing Issues and Search Types

Editing issues

- Focus on **lexicogrammar, content, organisation**
- Aim for **accuracy** and **consistency**

Search types

1. Checking known issues

e.g. Is '*capable to do*…' correct?
Do I use terminology consistently?

2. Identifying unknown issues

e.g. What does a comparison of my text with expert texts show?

cf. 'pattern-defining' and 'pattern-hunting' (Kennedy & Miceli: 2010: 31)

Tools, Editing Issues, Search Types

	Lexicogrammar	Content	Organisation
Checking known issue	Concordance Clusters Collocates	Plot	Plot
Identifying unknown issue	N-grams Wordlist	N-grams Keywords	N-grams Keywords

- Tools with **high user input** for ***checking issues***
- Tools with **low user input** for ***identifying issues***

Example 1: Concordance Plot

Andrea: Dominican doctoral student in Modern Languages

Corpus: 4 thesis chapters; 64,000 words

Thesis title: Poetics of the urban, poetics of the self: Transience, imminence and the everyday in selected works by Jorge Luis Borges and Jacques Réda.

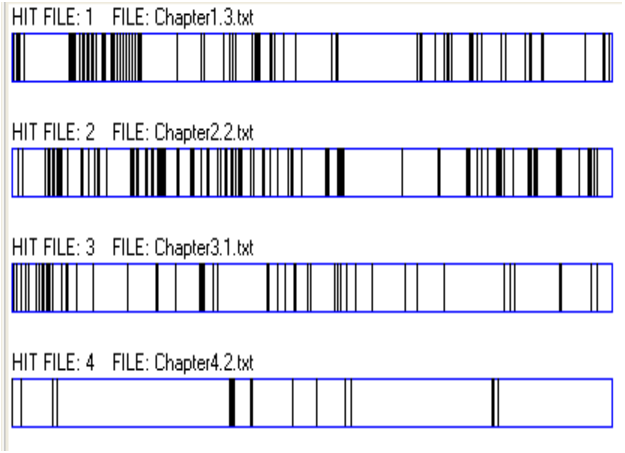
Issue: Balance of themes (checking)

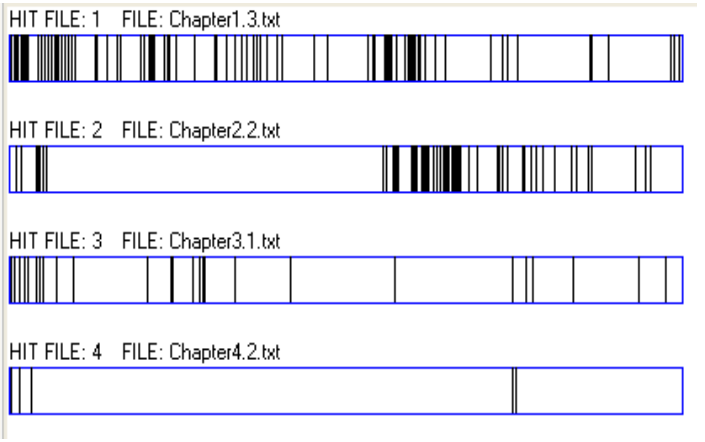
Procedure: Retrieve plots using content topics as search terms; compare distribution of topics in chapters

Andrea's Question:

'Buenos Aires and Paris: Are they balanced throughout?'

Comparison: *Buenos Aires, Paris*





Buenos Aires

Chapter 1: **84** hits

Chapter 2: **133** hits

Chapter 3: **55** hits

Chapter 4: **18** hits

Paris

Chapter 1: **102** hits

Chapter 2: **65** hits

Chapter 3: **27** hits

Chapter 4: **5** hits

Outcomes: Andrea's Investigation

'Chapter 2: *Balance the Buenos Aires and Paris sections.*

Chapter 3: *Investigate the structure of the chapter.*

Chapter 4: *Very few hits for both cities. Is another theme emerging that needs to appear throughout the thesis (i.e. imminence)?'*

Why use Concordance Plot?

- to track content, ideas, terminology, citations etc. within a single chapter
- to compare usage across chapters of a thesis
- to check content issues that the student is already aware of

Example 2: Keyword List

Keiko: Japanese doctoral student in archaeological science

Corpus: 7 thesis chapters; 57,492 words

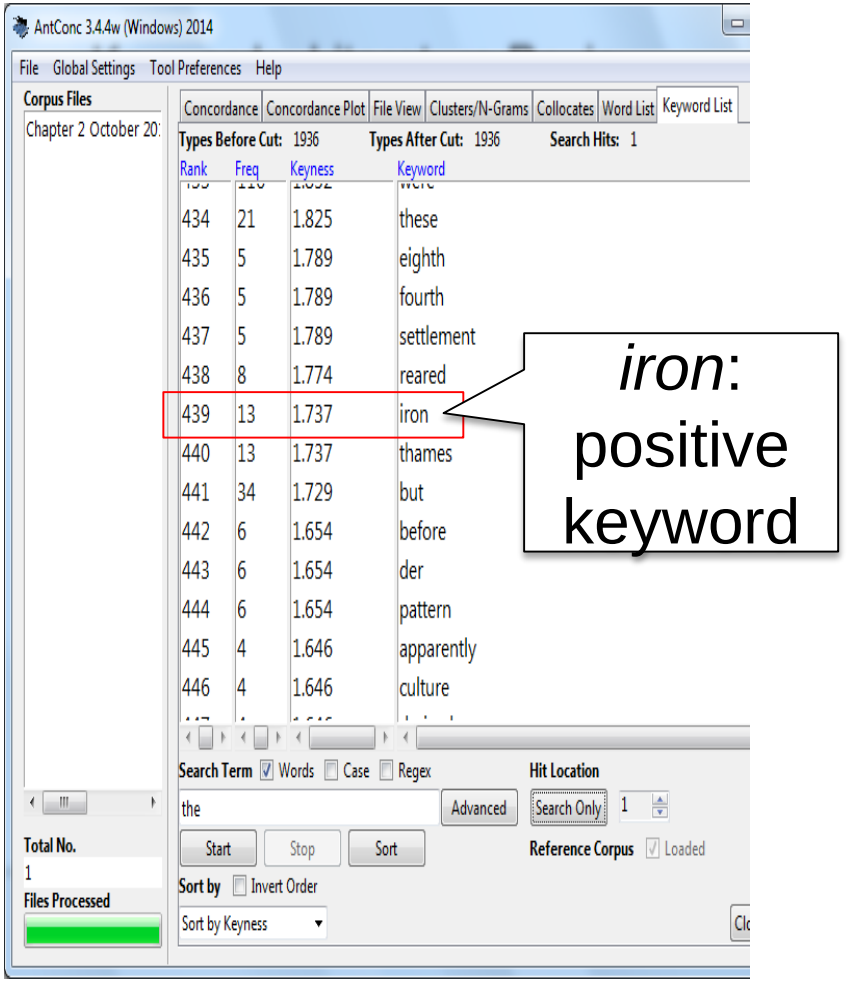
Thesis title: Transition from the Roman period to the Anglo-Saxon period in the Upper Thames Valley: Analysis using stable isotope data

Issue: content of individual chapters (identifying)

Procedure: Make keyword lists of individual chapters, using the whole thesis as reference corpus; examine keywords and negative keywords

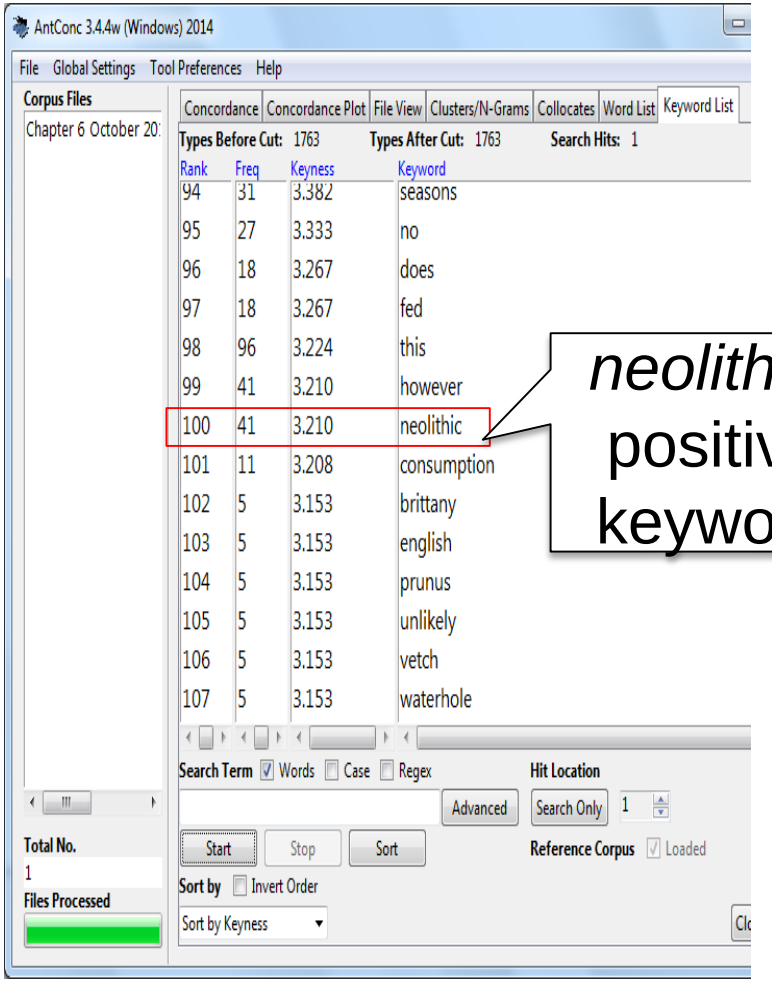
Keiko's Keywords

Literature Review



iron: positive keyword

Discussion



neolithic: positive keyword

Outcomes: Keiko's Investigation

Chapter 2 Literature Review
iron: positive keyword; *roman:* negative keyword
Chapter 6 Discussion
neolithic positive keyword; *roman* negative keyword

'I talk about Iron Age more in Chapter 2 (Literature Review) and Neolithic period more in Chapter 6 (Discussion), but my main focus is in the Roman and Anglo Saxon period. References to Iron Age and Neolithic should be reduced'.

Why use Keyword List?

- to allow aspects of content to emerge
- to identify content issues the student is not aware of

Example 3: N-Grams

Hiromi: Japanese doctoral student in sociology

Thesis topic: Integration and separation of immigrants in Japan

Corpora: 52 research articles; 523,427 words

4 thesis chapters; 18,945 words

Issue: differences between expert and student's writing (identifying)

Procedure: Make 3-gram lists of research article corpus and student's thesis corpus; compare most frequent 3-grams

Hiromi's Top Five 3-grams

Research Article Corpus	**Thesis Corpus**
1. of national identity (192)	1. of national identity (55)
2. as well as (150)	2. national identity and (46)
3. of the nation (135)	3. civic national identity (34)
4. in terms of (119)	4. ethnic national identity (31)
5. there is a (90)	5. and attitude toward (27)

Hiromi's research article corpus contains **2 referential expressions** and **1 discourse organizer** (Simpson-Vlach & Ellis (2010)

Her own writing contains only content-related 3-grams

Outcomes: Hiromi's Investigation

- '*I should check if I can write more sentences using the general expressions.*
- *It may be that I need more interpretations of the results.*
- *How is 'there is a' used in my research article corpus?*
- *It is used to reference the previous research and to explain the gap in the field of study, as well as to explain the results of the statistical analysis.'*

Why use the N-grams Tool?

- to identify frequent expressions
- to explore the difference between student's writing and expert text

Affordances of Corpus Tools for Editing

- enable ***comparisons*** of student writing e.g. with expert texts or between chapters
- facilitate ***a focus on language, content*** and ***organisation separately***
- show ***issues in language, content*** and ***organisation*** that are not evident when reading linearly
- allow both a ***bird's eye view*** from above and ***a bug's eye view*** from below
- ***de-familiarise*** an over-familiar text

References

Anthony, L., (2014). AntConc (3.4.3). [computer program] Tokyo, Japan: Waseda University. Available at: <http://www.laurenceanthony.net/>

Anthony, L., (2015). AntFileConverter (1.2.0). [computer program] Tokyo, Japan: Waseda University. Available at: <http://www.laurenceanthony.net/>

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, *32*, 301–319.

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning and Technology*, *14*(1), 28–44.

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512.

Watson Todd, R. (2001). Induction from self-selected concordances and self-correction. *System*, *29*, 91–102.