Corpus-Assisted Editing: Corpora and Tools for Thesis and Dissertation Writers

Maggie Charles
Language Centre
University of Oxford
maggie.charles@lang.ox.ac.uk
1. Context and rationale
2. Building do-it-yourself (DIY) corpora
3. Tools and examples
4. Evaluation of the approach
5. To conclude
Part 1

Context and Rationale

Two corpora of theses are examined in two contrasting disciplines, a social science and a humanities discipline. The corpora have been defined as 'collections of texts' and the methods used to examine them. The chapter ends with an outline of the conclusions drawn from the analysis.

1.4 Justification for a Corpus-Based Approach
A corpus-based approach is particularly appropriate for this research. First, as we have seen in the language description which ensues, although the corpora such as those in this study tend to be relatively small, even such mini-corpora provide large amounts of data in comparison to what would be available through manual analysis. Thus their occurrence and form are not solely due to individual personal choice, but, particularly in specialised fields, such qualitative interpretations are facilitated by comparing two or more corpora from different disciplines. The significance of the frequency of a given pattern can best be seen in comparative nature of corpus data. Thus in the present study, the full way of interpreting the linguistic realisations of stance observed in the particular attitude form.

1.7 Aims and Research Questions
This study uses two corpora of theses in politics and materials to examine certain grammatical features as the features associated with the expression of stance. The aim is to compare the two corpora in terms of the frequency and type of stance constructed by these features and the differences reveal about disciplinary ideology and epistemology in the two corpora.

1.8 Corpora and Methods
1.8.1 Setting up the Oxford Academic Text Corpora

Total of 3586 instances, 1504.3 per 100,000 words. 40% of adverbs occur in both corpora, 59% of the overall total, indicating a considerable overlap between political and humanities corpora. In many cases, it is included in the groups for both, primarily, significantly, so, together, unfortunately. However, in both corpora, the majority of adverbs are grouped, 88.3% in politics and 52.3% in materials. A similar process occurs with really. About 80% of instances in both corpora are in the group with the highest frequency, 920 1010.5 3136 1045.3 types 31 16.3 30 10.0 tokens per type 61.9 104.5. Both corpora make use of substantially the same textual adverbs; the only exception is too, that the frequency of the textual adverbs is also very similar in the two corpora. As shown by the figures for tokens per 100,000 words: 1010.5 in politics corpora and 1014.5 in humanities corpora. These adverbs account for around 40% of all the occurrences of grouped adverbs.

Further, the grammatical marking of stance provides an attitudinal or evaluative context. For example, Charles (2000) reported similarities between the politics and materials corpora. 4.3 Method and Data Introductory it patterns were retrieved from the 0.63 it v-link ADJ that 112 58.5 152 59.7 Total 274 144.2 471 157.0. In both corpora, the majority of infinitives that occur after a present tense link verb denote events, while the majority of infinitives that occur after a present tense link verb denote events. In both corpora, the majority of infinitives that occur after a present tense link verb denote events. The frequency of the NECESSARY and USEFUL groups is similar in the two corpora. The frequency of the ADJECTIVE group is higher in materials corpora. I begin by considering the DIFFICULT subgroup. 4.5.4.1 The DIFFICULT Subgroup...
What is Corpus-assisted editing?

- **A corpus** is a collection of electronic texts built according to set criteria and constructed for a specific purpose.

- **Corpus-assisted editing** is the use of corpora to edit and revise texts.

- The corpora used here are **DIY corpora** compiled by writers for their own use.

- Used for teaching **translation** (e.g. Kübler 2011), **linguistics** (e.g. Seidlhofer 2000) exploring **disciplinary discourse** (e.g. Charles 2015a, 2015b, 2017; Lee & Swales 2006)
## Editing your Thesis with Corpora: Course Details

<table>
<thead>
<tr>
<th><strong>Aim:</strong></th>
<th>to improve graduates’ editing skills to provide a resource for future use</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frequency:</strong></td>
<td>2-3 times per year (10 in total)</td>
</tr>
<tr>
<td><strong>Timing:</strong></td>
<td>One 2-hour session/week for 6 weeks</td>
</tr>
<tr>
<td><strong>Venue:</strong></td>
<td>computer laboratory</td>
</tr>
<tr>
<td><strong>Class size:</strong></td>
<td>maximum 12</td>
</tr>
<tr>
<td><strong>Composition:</strong></td>
<td>multi-disciplinary</td>
</tr>
</tbody>
</table>
| **Software:** | AntConc (Anthony 2014)  
AntFileConverter (Anthony 2015) |
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%
Two Types of DIY Corpora

1. DIY 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. DIY Corpus of Student’s Own Writing
   - chapters of thesis as individual files
   - may include subcorpora of other writing (e.g. proposals, Master’s dissertation)
## Course Programme

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

Building Do-It-Yourself (DIY) Corpora
4(5) Steps for Building a Corpus

- Select Texts
- Convert Files
- Check Files
- Save Files
- Optional Clean Files
Choose texts that represent an appropriate genre.
4(5) Steps for Building a Corpus

- Select Texts
- Convert Files
- Check Files
- Save Files
- Optional Clean Files

- Corpus files must be in plain text (.txt) format
- AntFileConverter converts multiple files simultaneously
# AntFileConverter

Input several pdf or Word files here

File conversions shown here

(Anthony 2015)
4(5) Steps for Building a Corpus

- Select Texts
- Convert Files
- Check Files
- Save Files
- Optional Clean Files

- Has the **whole** text converted?
- Have **line/word breaks** and **individual letters** converted correctly?
4(5) Steps for Building a Corpus

This is a quick and dirty corpus

- Save files to a corpus folder
4(5) Steps for Building a Corpus

- Select Texts
- Convert Files
- Check Files
- Save Files
- Optional Clean Files

- To improve results, delete everything that is not part of the running text (e.g. author, title, graphics)
- Try the dirty corpus first!
Summarising Process and Tools

**Process**
- Quick
- Easy
- Low-tech

**Tools**
- Free
- Available
- User-friendly
Part 3
Tools and Examples of Corpus-Assisted Editing
The Concordancer

- searches the corpus for **every instance** of a word/phrase you choose
- presents each one **with its context** in a line on screen
- shows **search item in the centre**, with about 5/6 words on either side
The Concordancer in Use: Siyu

Siyu: Chinese doctoral student in Chemistry
Corpus: 50 research articles; 394,000 words
Issue: Preposition use in specialized text

Siyu’s Question

What’s the proper preposition along with the word *inhibition*?
24 hits for **inhibition by***
Inhibition observed when 80 μL of CO-saturated solution (final concentration 100 μM) was added to each cuvette.

CO inhibition of activity by carbon monoxide was reversed by light. 47

Inhibition of all hydrogenase activities was due to the fact that hydrogenase (H2 oxidation activity), but activity is restated of D. desulfuriicus (FeS)-hydrogenase by CO, and in acetic acid.

Inhibition of D. fructosovorans Ni-Fe hydrogenase im mobilized on a NiFe hydrogenase production by active enzyme. Importantly, the 54.

Inhibition of H reduction by D. fructosovorans NifFe hydrogenase is not due to CO inhibition, but rather due to NiFe hydrogenase production. The potential was then stepped to 0 V vs. Reference.

Figure 4 also shows potential step kinetic experiments with B. hydrogenatus using isotopically labeled 13CO results.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by CO is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-F state is also not specific to NiFe hydrogenase. The enzyme by CO is not produced by NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by NiFe hydrogenase production is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-I state is also not specific to NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by NiFe hydrogenase production is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-I state is also not specific to NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by NiFe hydrogenase production is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-I state is also not specific to NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by NiFe hydrogenase production is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-I state is also not specific to NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by NiFe hydrogenase production is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-I state is also not specific to NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by NiFe hydrogenase production is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-I state is also not specific to NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by NiFe hydrogenase production is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-I state is also not specific to NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by NiFe hydrogenase production is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-I state is also not specific to NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by NiFe hydrogenase production is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-I state is also not specific to NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.

Inhibition of Hydrogenases in Low Redox Potentials: Inhibitory Effect of CO. Conclusions: The inhibitory effect of CO on the enzyme by NiFe hydrogenase production is not specific to NiFe hydrogenase but rather to NiFe hydrogenase production. The inhibitory effect of CO on the enzyme by O2 leading to the Ni-I state is also not specific to NiFe hydrogenase. The enzyme by O2 leading to the Ni-I state is also not produced by NiFe hydrogenase.
Siyu’s Examples and Conclusions

Inhibition by CO or product (24 hits)

By is used for describing which species impose this effect

Inhibition of Ni-Fe hydrogenases (35 hits)

Of is followed by species having this behaviour.

Why use the Concordancer?

- To check/find collocations, phraseology, patterns
- To see many examples of a word/phrase at the same time
- To find specialist information not available elsewhere
Concordance Plot

- provides a **graphic display**
- shows where the search item occurs **within a file**
- gives a **simultaneous overview of all files** in a corpus
Concordance Plot in Use: Andrea

Andrea: Dominican doctoral student in Modern Languages
Corpus: 4 chapters of her thesis (64,000 words)
Thesis: Compares Buenos Aires and Paris in work by Borges and Réda
Issue: Checking the development of the topic
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
Concordance Plot Outcomes

Andrea’s Conclusions

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, terms** in a single file
- To compare usage across files
- To check **content issues** in a long text
The N-Grams Tool

- shows a list of all word sequences of a length \( n \) you choose
- presents them as a list and gives their frequency
- the procedure is automatic
Hiromi: Japanese doctoral student in sociology
Corpora: 52 research articles; 523,427 words
4 thesis chapters; 18,945 words
Thesis: Integration of immigrants in Japan
Hiromi’s Question: What are the differences between my writing and expert writing?
Hiromi’s Top Five 3-grams

Research Article Corpus
1. of national identity (192)
2. as well as (150)
3. of the nation (135)
4. in terms of (119)
5. there is a (90)

Thesis Corpus
1. of national identity (55)
2. national identity and (46)
3. civic national identity (34)
4. ethnic national identity (31)
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
Hiromi’s conclusions

- 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 writing** and **expert text**
The Keywords Tool

- Identifies words which are *unusually frequent* or *infrequent* in one corpus when compared to a reference corpus.
- Gives insight into the *content of individual chapters* compared to the whole thesis.
Keiko: Japanese doctoral student in archaeological science
Corpus: 7 thesis chapters; 57,492 words
Thesis: Transition from the Roman period to the Anglo-Saxon period in the Upper Thames Valley: Analysis using stable isotope data
Keiko’s Question:

Does the content of my chapters reflect my intentions?
Keiko’s Keywords

Literature Review

- iron: positive keyword

Discussion

- neolithic: positive keyword
Keiko’s conclusions

Chapter 2 Literature Review: *iron* positive keyword
Chapter 6 Discussion: *neolithic* positive 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 the Keywords Tool?

- to allow *aspects of content* to emerge
- to *identify content issues* the student is not aware of
Part 4
Evaluation of the Approach

Evaluation of the Approach

Evaluation of this barrier. Nevertheless, it presents the energetics of the formation of a short string quickly. Table 2.4 lists the classes, materials and modes of manufacture. Table 2.4 Properties of the New Fibreboards (CCLF2 and CCLF3). The properties of these new fibreboards were characterised as being well established (equation 1), but in keeping with the approach of this chapter, a much simpler evaluation will be described. Higher order states. Under the infinite confinement of an outline of the whole thesis.

1.3 Stance 1.3.1 Definition Thompson and Hunston (2000: 5) use ‘evaluation’ as ‘the broad cover term for the expression of the speaker or writer’s attitude or stance towards the object of the discourse’ (Biber & Finegan, 1989; Conrad & Biber, 2000). In this thesis I investigate the way in which the writer’s evaluation is linked to the norms and practices of their discipline. I am interested both in how writers take up the kinds of roles and identities they construct. As my prime concerns are with the relationship between evaluation and the disciplinary community, I use the term ‘stance’. For the purpose of this study, I adopt the definition of ‘evaluation’ by Thompson and Hunston as follows: Stance is the expression of the writer’s attitude towards something, or feelings about the entities or propositions that he or she is talking about. It uses the resources of evaluation in order to position the writer in relation to the propositional content, the reader and other members of the community.

1.3.2 Functions of Stance Thompson and Hunston (2000: 6) distinguish three functions of evaluation, all of which show its importance in the study of academic writing. First, they make the point that evaluation expresses the writer’s opinion and thus reflects their value system and that of the community.

In the good-bad parameter and, particularly in scientific articles, is linked to goal-achievement; and ‘evaluation of relevance’, which indicates the importance of prior or subsequent text for the development of the argument. Hunston argues that much evaluation in academic writing is implicit and avoids overtly attitudinal language, but shows that it is present ‘with respect to the value system’ (Hunston, 1993a: 72 original emphasis). Although much of the implicit evaluation is difficult to reveal using corpus-based methods, I will refer to aspects of Hunston’s model of evaluation. Evaluation is difficult to reveal using corpus-based methods, I will refer to aspects of Hunston’s model of evaluation where relevant in this thesis. Biber and Finegan (1989) use stance as a means of identifying an opinion of the writer that is often integrated into a text rather than overtly marked. This provides support for the view that much evaluation in academic texts is implicit rather than overtly marked. There have been a number of studies reporting verbs in RAs and is considered further in Chapter 4.

Hoey (2000) analyses Chomsky’s use of evaluation, showing that the way it is embedded in the clause makes it difficult to recognise and challenge the arguments that are evaluated in RAs from four disciplines. She draws a distinction between ‘research-oriented evaluation’ and ‘topic-oriented evaluation’ and finds no significant disciplinary difference in the types of stance from four disciplines. She draws a distinction between ‘research-oriented evaluation’ and ‘topic-oriented evaluation’ and finds no significant disciplinary difference in the types of stance from four disciplines.

Evaluation has also been studied in relation to specific disciplines: art history (Tucker, 2003) and geology, especially useful for the identification and description of stance. Concordance lines also show that patterns of evaluation are systematically repeated in corpora. Thus their occurrence and form are not solely due to the writer or their research. The examples below also illustrate the Negative Evaluation element of the ‘Gap in Knowledge-Filling’ pattern (Hoey, 2001). Emphasis on homologous data by Segal et al. leave themselves open to the charge of post hoc curve fitting... (mpomach2) This negative evaluation of a named researcher occurs only three times in the politics corpus, in each case by the same author. The stance towards the work of other members of the field. However the very low occurrence of negative evaluation of a named researcher may indicate that thesis writers are reluctant to assume a shared negative evaluation of a named researcher may indicate that thesis writers are reluctant to assume a shared negative evaluation in this way and this may reflect their candidate status in the community. With the use of obviously this is not the case as a means of reclaiming territory from the Serbs. (mpomach3) Using Hunston’s terms, WELL adverbs carry out an ‘evaluation of value’ (Hunston, 1989, 1993a, 1994). Entities are evaluated on the good-bad parameter as the agreed goal of revealing the ‘truth’ closer. This is shown in the example below, which offers a positive evaluation of the writer’s work.
Is it helpful to use your corpus and AntConc for editing?

- Yes definitely: 79%
- Yes probably: 21%
Is it easy to use the AntConc software?

- Yes definitely: 58%
- Yes probably: 32%
- Undecided: 9%
- No probably not: 1%
Was it easy to build your corpus?

- Yes definitely: 61%
- Yes probably: 26%
- Undecided: 8%
- No probably not: 4%
- No definitely not: 1%
Do you intend to use your corpus and AntConc for editing in the future?

- Yes definitely: 86%
- Yes probably: 12%
- Undecided: 2%
In Conclusion

Part 5

In Conclusion
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
- *de-familiarise* an over-familiar text
A bird’s eye view from above

A bug’s eye view from below


