Corpus and Text Analysis for Research in the Humanities

Martin Wynne
martin.wynne@oucs.ox.ac.uk
Head of the Oxford Text Archive
Oxford e-Research Centre
Oxford University Computing Services
Faculty of Linguistics, Philology and Phonetics

DH@OXSS
Monday 2nd July 2012
...alternative titles...

Corpus linguistics for non-linguists

or

How can I use the datasets, tools and methods of corpus linguistics to answer research questions of a primarily socio-historical nature?

aka

“Culturomics”
Outline of today's lecture

1. How do we do corpus linguistics?
   - Corpora
   - Tools
   - Methods
   - Research questions
   - Interpretations
2. Corpus linguistics shading into cultural theory – an example
3. Corpus Linguistics techniques for socio-cultural enquiry – some suggestions
4. Exercises
<table>
<thead>
<tr>
<th>WORD</th>
<th>W1</th>
<th>W2</th>
<th>W1/W2</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>REPORTED</td>
<td>91</td>
<td>0</td>
<td>182.0</td>
<td>192.9</td>
</tr>
<tr>
<td>DISPLAYED</td>
<td>65</td>
<td>0</td>
<td>130.0</td>
<td>137.8</td>
</tr>
<tr>
<td>DETECTED</td>
<td>54</td>
<td>0</td>
<td>108.0</td>
<td>114.4</td>
</tr>
<tr>
<td>DISPLAY</td>
<td>46</td>
<td>0</td>
<td>92.0</td>
<td>97.5</td>
</tr>
<tr>
<td>PROCESSING</td>
<td>38</td>
<td>0</td>
<td>76.0</td>
<td>80.5</td>
</tr>
<tr>
<td>ENCOUNTERED</td>
<td>32</td>
<td>0</td>
<td>64.0</td>
<td>67.8</td>
</tr>
<tr>
<td>RETURNS</td>
<td>26</td>
<td>0</td>
<td>52.0</td>
<td>55.1</td>
</tr>
<tr>
<td>SAMPLING</td>
<td>20</td>
<td>0</td>
<td>40.0</td>
<td>42.4</td>
</tr>
<tr>
<td>RE-RUN</td>
<td>18</td>
<td>0</td>
<td>36.0</td>
<td>38.1</td>
</tr>
<tr>
<td>SPELT</td>
<td>14</td>
<td>0</td>
<td>28.0</td>
<td>29.7</td>
</tr>
<tr>
<td>INCLUDES</td>
<td>12</td>
<td>0</td>
<td>24.0</td>
<td>25.4</td>
</tr>
<tr>
<td>INDICATING</td>
<td>23</td>
<td>1</td>
<td>23.0</td>
<td>24.4</td>
</tr>
<tr>
<td>REPRESENT</td>
<td>11</td>
<td>0</td>
<td>22.0</td>
<td>23.3</td>
</tr>
<tr>
<td>SUPPLIED</td>
<td>11</td>
<td>0</td>
<td>22.0</td>
<td>23.3</td>
</tr>
<tr>
<td>OCCURRED</td>
<td>183</td>
<td>9</td>
<td>20.3</td>
<td>21.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WORD</th>
<th>W2</th>
<th>W1</th>
<th>W2/W1</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUPPOSE</td>
<td>18</td>
<td>0</td>
<td>36.0</td>
<td>34.0</td>
</tr>
<tr>
<td>'M</td>
<td>47</td>
<td>2</td>
<td>23.5</td>
<td>22.2</td>
</tr>
<tr>
<td>AFFORD</td>
<td>18</td>
<td>1</td>
<td>18.0</td>
<td>17.0</td>
</tr>
<tr>
<td>LETTING</td>
<td>9</td>
<td>0</td>
<td>18.0</td>
<td>17.0</td>
</tr>
<tr>
<td>TELLING</td>
<td>9</td>
<td>0</td>
<td>18.0</td>
<td>17.0</td>
</tr>
<tr>
<td>DISMISS</td>
<td>7</td>
<td>0</td>
<td>14.0</td>
<td>13.2</td>
</tr>
<tr>
<td>WO</td>
<td>7</td>
<td>0</td>
<td>14.0</td>
<td>13.2</td>
</tr>
<tr>
<td>REGARD</td>
<td>12</td>
<td>1</td>
<td>12.0</td>
<td>11.3</td>
</tr>
<tr>
<td>BUYING</td>
<td>6</td>
<td>0</td>
<td>12.0</td>
<td>11.3</td>
</tr>
<tr>
<td>ASK</td>
<td>6</td>
<td>0</td>
<td>12.0</td>
<td>11.3</td>
</tr>
<tr>
<td>IMAGINE</td>
<td>11</td>
<td>1</td>
<td>11.0</td>
<td>10.4</td>
</tr>
<tr>
<td>GO</td>
<td>21</td>
<td>2</td>
<td>10.5</td>
<td>9.9</td>
</tr>
<tr>
<td>'VE</td>
<td>71</td>
<td>7</td>
<td>10.1</td>
<td>9.6</td>
</tr>
<tr>
<td>HEARD</td>
<td>5</td>
<td>0</td>
<td>10.0</td>
<td>9.4</td>
</tr>
<tr>
<td>FORGIVE</td>
<td>5</td>
<td>0</td>
<td>10.0</td>
<td>9.4</td>
</tr>
</tbody>
</table>

[COMPARE BNC AND COCA]

This website allows you to quickly and easily search the 100 million word British National Corpus (1970s-1993). The BNC was originally created by Oxford University Press in the 1980s - early 1990s, and now exists in various versions on the web. If you find this version useful, you may also be interested in other corpora that we have created, including the 410+ million word Corpus of Contemporary American English (1990-2010) and the 400+ million word Corpus of Historical American English (1810-2009).

As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (colligates) within a certain word window (e.g., all nouns, some common prepositional phrases, all adjectives near...
What do people do in corpus linguistics?
- datasets
- tools
- methods
- research questions and outcomes
cause

naked eye

Explore at http://corpus.byu.edu/bnc/
“I'm just going out to commit certain deeds.”
"Summer has set in with its usual severity."

Samuel Taylor Coleridge (1772-1834)
What can't you use a corpus for?

• Automatic analysis
• Quantitative analysis without checking meaning in the texts
• Negative evidence (proving something doesn't ever occur)
• Generalisations beyond the corpus design parameters
• Comparison of text types not sampled to be representative
“The electronic analysis of large text collections has led to a new level of language description that is based on the lexico-grammatical patterns from concordance data.” (Adolphs, 2006)
Lexis and grammar

New techniques and theories:

- Concordance
- Collocation
- Colligation
- Semantic preference
- Semantic prosody

Corpus data can reveal patterns of co-occurrence, co-selection and co-dependence
An example of doing corpus linguistics with a socio-historical bent

In Risk, Lupton summarizes the scholarship on the several historical semantic shifts of “the word ‘risk’ ” (1999: 5–13), arguing that at the advent of modernity, risk was “scientized” and treated in terms of calculable probability, as a means of evaluating knowable outcomes both desirable and undesirable: “‘risk’ could be both ‘good’ and ‘bad’ ” (1999: 6–7). However, by the end of the twentieth century, the meaning was again shifting, with both the notion of calculability and most associations of the positive falling away. Lupton’s conclusions regarding contemporary uses of risk accord, in the main, with Beck’s: “In everyday lay people’s language, risk tends to be used to refer almost exclusively to a threat, hazard, danger or harm” (Lupton 1999: 8).

A genre, collocational, and constructional analysis of RISK
Donald E. Hardy, University of Nevada, Reno
Crystal Broch Colombini
A tentative suggestion of further socio-historical research with corpus linguistic resources

Searching text collections for:
- word forms
- senses
- lemmas
- phrases
- terms
- sets of related words
- concepts?
The holocaust in history

“I will argue that “The Holocaust” is an ideological representation of the Nazi holocaust...Until recently, however, the Nazi holocaust barely figured in American life. Between the end of World War II and the late 60s, only a handful of books and films touched on the subject.”

Types of findings

1. Emergence of new word forms, postdating
   a: Emergence of new phrases

2. Emergence of new senses, changes in meaning
   a: Changes in collocations, semantic preferences
   b: emergence of multi-word expressions

3. Changes in reference

4. Changes in frequency

5. Derivations

6. Emergence and decline of related words
Interpretation

How do we interpret the results? We need to understand:

- Dataset
- Sampling
- Representativeness
- Tools
- Tagging, tokenization, annotation
- Metadata
- Frequencies
- Statistical significance
Interpretation 2

How do we interpret the results? We need to ask the questions:

• What's in my dataset? What's missing?
• What will the sampling procedure miss?
• What population of texts in the world can I make claims about by searching this dataset?
• What is the right tool for the job?
• Will I successfully retrieve all occurrences of the word forms which I am looking for?
• How can I make my search term more sophisticated?
• What claims can I make about the significance of the frequencies?
• How can I improve the process and refine the results?
• What do I need to investigate further?
What's available?

Bank of English: large monitor corpus of modern English
BNC: large snapshot of British English c. 1994
Corpus of Old English: complete collection of texts
EEBO: large collection of early published material
ECCO: eighteenth century texts
Google: search the web!
Google Books: huge amounts of digitized texts from libraries
JSTOR: historic journals in electronic form
LexisNexis: access to newspaper archives
WebCorp: search the web as a corpus
Voyant Tools: online analysis tools for texts
What's wrong with our current resources?

**Bank of English**: not available online

**BNC**: one historical snapshot; out of date; only certain text types

**Corpus of Old English**: fragmentary nature of surviving texts

**EEBO**: no principled sampling, spelling variation

**ECCO**: no principled sampling?

**Google**: unknown and unstable corpus; search tools optimized for personalized document retrieval

**Google Books**: variable quality control of text and metadata

**JSTOR**: unknown sampling frame; unreliable tools

**LexisNexis**: no corpus analytic tools

**WebCorp**: based on search results, unknown corpus

**Voyant Tools**: in beta, find your own texts!
Looking forwards

• What do we need from the next generation of tools and resources?
• What new methods do we need to develop?
• What training do we need?
“warlord”
"The word "warlord" entered the English language as a translation from the German word "Kriegsherr", which was an official title of the German Emperor. Its use for Chinese military commanders who had a regional power base and ruled independently of the central government dates from the early 1920s, with Bertram Lenox Simpson being one source, according to the Oxford English Dictionary." [Wikipedia]

“ethnic cleansing”
"The practice is much older than the term, known throughout history. The term itself appears to have been popularised by the international media approximately early in 1992, following the discovery of Bosnian-Serb Concentration camps established in April 1992, and the subsequent assault on Sarajevo." [Wikipedia]

“sustainability”
What did "sustainability" mean and how was it used before its current widespread and popular usages regarding agricultural, industrial and business policies which are environmentally sound? (Also a current popular sense in terms of the potential loss and obsolescence of digital resources, especially in the humanities!)
Exercises (2)

"risk"
Explore further the claims made by Hardy and Comombini about the historical shifts in the meaning and uses of “risk”

"genocide"
The first occurrence of this word is:
“1944   R. Lemkin Axis Rule in Occupied Europe ix. 79   By ‘genocide’ we mean the destruction of a nation or of an ethnic group.” OED
Can you support or refute this assertion? Can you use this to test the accuracy of any online resources? What can you find out about usage of 'genocide'? What events is the words used to refer to at different times?

"holocaust"
Explore further the example used in the lecture: “I will argue that “The Holocaust” is an ideological representation of the Nazi holocaust...Until recently, however, the Nazi holocaust barely figured in American life. Between the end of World War II and the late 60s, only a handful of books and films touched on the subject.” Finkelstein 2000

Other that you have thought of...but your group must agree and you must write down a hypothesis about it's usage first!
Exercises (3)

And a bit harder:

The concept ENVIRONMENT

Explore words relating to the concept of the environment (referring the global natural
What words are relevant? Are there neologisms and shifts of meaning? Are there specific rhetorical and semantic strategies associated with uses of these words?

“anthropocene”

A relatively recent coinage, the word “anthropocene” seems to be topical and increasing in frequency of usage. What does it mean (are there variations)? In what sorts of texts does it occur? Think imaginatively about how to search for a word that may not appear in the historical databases.

“state”

"...no political writer before the middle of the sixteenth century used the word 'state' in anything like its modern political sense [referring to the machinery of government and social control]” Skinner (1998)

“cause”

Are the claims that Stubbs makes about the usage and meaning of "cause" true historically? Are collocations and semantic preferences different in the past?
Practical session

1. Select a word to investigate; form groups

2. Method
- discuss hypothesis, methods, resources, tools
- write down any expectations
- investigate (together, or in parallel) & discuss findings
- refine or change hypothesis
- investigate further, note results

3. Evaluation
- what do the results suggest?
- which resources, tools and techniques worked & which didn't?
- what further research is needed?
- what further tools and datasets are needed?

4. Present results
References

BYU Corpora  http://corpus.byu.edu/
CLARIN  http://www.clarin.eu/
Culturomics  http://www.culturomics.org/
EEBO  http://eebo.chadwyck.com/search
ECCO  http://quod.lib.umich.edu/e/ecco/
JSTOR  http://www.jstor.org/
LexisNexis  find via Oxlip+ at http://www.bodleian.ox.ac.uk/taylor/eresources
JISC e-Collections  http://ecollections.mimas.ac.uk/ [not available 1.7.12]


Where do we go from here?

There are some useful corpora and tools out there, but it's a fragmented, unconnected, confusing landscape.

We need to move from *ad hoc* tools to persistent, sustainable and reliable services, from academic cottage industry to specialist, professional production services. Disparate datasets and tools need to be connected together.

We can't solve these problems on our own, but we can tell the policy-makers and developers of computing infrastructure what our requirements are (e.g. join CLARIN!)