Using corpus tools to investigate citation

Hilary Nesi
EAP and Corpora, Coventry University
21 June 2014
1. Presentation of findings from my investigation of citation styles in the BAWE corpus.

2. Hands-on investigation of these citations using corpus query language (CQL).
To develop **descriptors** for all the genres of British university student assignment – identifying assignment types according to their social purposes.

**Project title was 'An Investigation of Genres of Assessed Writing in British Higher Education'.**
• 6,506,995 words
• 2,896 texts
• 2,761 assignments
• 1,953 written by L1 speakers of English
• 1,251 “distinction” and 1,402 “merit”
• 1000+ modules & 300 degree courses

### Numbers of texts at each level and in each domain

<table>
<thead>
<tr>
<th>Domain</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts &amp; Humanities</td>
<td>255</td>
<td>229</td>
<td>160</td>
<td>80</td>
</tr>
<tr>
<td>Life Science</td>
<td>188</td>
<td>206</td>
<td>120</td>
<td>205</td>
</tr>
<tr>
<td>Physical Science</td>
<td>181</td>
<td>154</td>
<td>156</td>
<td>133</td>
</tr>
<tr>
<td>Social Science</td>
<td>216</td>
<td>198</td>
<td>170</td>
<td>207</td>
</tr>
</tbody>
</table>
### 30+ disciplines represented:

<table>
<thead>
<tr>
<th><strong>Arts &amp; Humanities</strong></th>
<th>Archaeology, Applied Linguistics, Classics, Comparative American Studies, English, History, Philosophy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Life Sciences</strong></td>
<td>Agriculture, Biological Sciences, Food Sciences, Health, Psychology, Medical Science</td>
</tr>
<tr>
<td><strong>Social Sciences</strong></td>
<td>Anthropology, Business, Economics, HLTM (Hospitality, Leisure and Tourism Management), Law, Politics, Publishing, Sociology</td>
</tr>
<tr>
<td><strong>Physical Sciences</strong></td>
<td>Architecture, Chemistry, Computer Science, Cybernetics &amp; Electronics, Engineering, Mathematics, Meteorology, Physics, Planning</td>
</tr>
</tbody>
</table>
13 Genre Families

1. Case Study
2. Critique
3. Design Specification
4. Empathy Writing
5. Essay
6. Exercise
7. Explanation
8. Literature Survey
9. Methodology Recount
10. Narrative Recount
11. Problem Question
12. Proposal
13. Research Report
Referencing systems: influenced by

• Discipline
• Genre
• The role of the source text

Instancing, Background, Exhibit, Argument or Method (Bizup 2008)
So, for example, we know that some disciplines favour the Vancouver (author-number) system, and others the Harvard (author-date) system.

**VANCOUVER**
For ease and graphical purposes it is sufficient to consider a duopoly. We also assume linear demand (1), constant marginal cost (2) as well as (to start with) homogeneous products that are perfect substitutes (3).

**HARVARD**
Schapiro et al. (2001) demonstrated how affiliative behaviours within socially housed groups of rhesus macaques could be manipulated.
And we can guess that genres aimed at ‘developing powers of independent reasoning’ are more likely to employ reporting verbs, to indicate the writer’s stance in relation to the source (integral Harvard).

**INTEGRAL**

Schapiro et al. (2001) demonstrated how affiliative behaviours within socially housed groups of rhesus macaques could be manipulated.

**NON-INTEGRAL**

…..they also provide key insights into standardization (Bradley 1996).
On different occasions the same source might be used:

- To establish a context (instancing)
- To provide objective information (background)
- As a primary source for analysis (exhibit)
- To provide key concepts and theories for discussion (argument)
- As a model for research procedures (method)

Citation purpose is likely to influence citation style.
Queries to try to find all the ways in which students cite sources:

1. Ben Fine (2001) argues that this concept is an oxymoron

2. As Archer (2001) says, ‘science is located in the social world’ / Radical feminists such as Andrea Dworkin (1976) have broadened the definition of violence

3. (McCracken 1990:24)

4. A meme according to Blackmore is anything passed on by imitation

5. Vancouver-style ‘author-number’ references

6. *ibid*, *op cit*, and follow-on *he argues that*
All the ways in which students cite sources?

- Schapiro et al. (2001) demonstrated .......
- Workers experience little autonomy to try and fail, as Leadbeater (1999: 83) suggests.
- Howe (1998) cites the work of Murry
- (Crosby 1984, cited in Lockwood 1996)
- Piaget (as cited in Rubin, LeMare and Lollis, 1990)
- As Archer (2001) says, ‘science is located in the social world’
- Radical feminists such as Andrea Dworkin (1976) have broadened the definition of violence
- They also provide key insights into standardization (Bradley 1996).
- Vancouver-style ‘author-number’ references
- Ibid; op cit
Recent criticism has argued that the Cold War split has caused intellectuals to make an over-simplified distinction between ‘individualistic liberalism and state collectivism’.

‘follow-on’ reporting clauses (Shaw 1992; Charles 2006)

The researchers found that more Black patients admitted to wards were not registered in primary care than other ethnic groups (Koffman et al, 1997). They also found that a high proportion of the Black population were admitted to a psychiatric unit.

‘implicit attributions’ (Williams 2010)

Perhaps the connection with this aridity of the psyche develops the blood metaphor identified by Sinclair.
Query 1: integral citations

proper noun + number within brackets + lexical verb or modal verb.

[tag = "NP.?"] [word = "et"]? [word = "."]? [word = "al"]? [word = "."]? [word = "\("] [tag = "MC"] [word = "\)"] [tag = "VV|VO|VVD|VVZ|VM"]
Allows for the options

- *et al*
- *et. al*
- *at al.*
- *et. al.*
Query 2: non-integral citations

proper noun + number and up to 5 words
additional text, all within brackets

[word = "\(\)" [tag = "NP.?" ] [tag = "MC.?"] [] {0,5} [word = "\)" ]]
Excludes references in bibliographies, footnotes or endnotes in Queries 1 and 2.
Manual filtering still necessary for:

**Internal references**
- Downs syndrome *(Trisomy 21)* is an example of a trisomy condition affecting an autosome

**Formulae, equations, etc.**
- Crystalline quartz *(SiO$_2$)* PZT ceramic Barium titanate Zinc oxide Lithium tantalate *(LiTaO$_3$)* Lithium niobate *(LiNbO$_3$)*
Queries 3 and 4

*ibid* and *op. cit.*

3. [word ="ibid"]
4. [word ="op"] [word = "\."]?[word ="cit"]
Query 5

Vancouver-style ‘author-number’ entries in bibliography sections

<p> [textpart = "bibliography" & word = "1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23|24|25|26|27|28|29|30|31|32|33|34|35|36|37|38|39|40"]
Query 6

the use of *cited in* in the main body of assignments, as an indicator of the quantity of secondary citation in the corpus.

[word = "cited"] [word = "in" & textpart != "bibliography|note"]
## Results per million words

<table>
<thead>
<tr>
<th>Query</th>
<th>AH</th>
<th>LS</th>
<th>PS</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral citations</td>
<td>119.0</td>
<td>253.6</td>
<td>17.4</td>
<td>217.4</td>
</tr>
<tr>
<td>Non-integral citations</td>
<td>544.7</td>
<td>369.9</td>
<td>31.7</td>
<td>605.8</td>
</tr>
<tr>
<td><em>ibid</em></td>
<td>91.4</td>
<td>11.4</td>
<td>-</td>
<td>75.9</td>
</tr>
<tr>
<td><em>op.cit</em></td>
<td>14.7</td>
<td>-</td>
<td>-</td>
<td>30.1</td>
</tr>
<tr>
<td>Vancouver numbering</td>
<td>12.0</td>
<td>138.5</td>
<td>219.7</td>
<td>67.5</td>
</tr>
<tr>
<td>cited in</td>
<td>35.7</td>
<td>196.1</td>
<td>3.1</td>
<td>134.2</td>
</tr>
</tbody>
</table>
## Vancouver referencing

<table>
<thead>
<tr>
<th>Query</th>
<th>AH</th>
<th>LS</th>
<th>PS</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral citations</td>
<td>119.0</td>
<td>253.6</td>
<td>17.4</td>
<td>217.4</td>
</tr>
<tr>
<td>Non-integral citations</td>
<td>544.7</td>
<td>369.9</td>
<td>31.7</td>
<td>605.8</td>
</tr>
<tr>
<td><em>ibid</em></td>
<td></td>
<td></td>
<td>-</td>
<td>75.9</td>
</tr>
<tr>
<td><em>op.cit</em></td>
<td></td>
<td></td>
<td>-</td>
<td>30.1</td>
</tr>
<tr>
<td>Vancouver numbering</td>
<td>12.0</td>
<td>138.5</td>
<td>219.7</td>
<td>67.5</td>
</tr>
<tr>
<td>cited in</td>
<td>35.7</td>
<td>196.1</td>
<td>3.1</td>
<td>134.2</td>
</tr>
</tbody>
</table>

About 82% of references in the Physical Sciences and about 18% in the Life Sciences use the Vancouver system.
The use of the Vancouver numbering system in the Sciences perhaps reflects an underlying assumption that writers build on the logical and cumulative outcomes of prior research (Becher/Trowler 2001), and therefore rarely need to take issue with prior claims.
# op cit and ibid

<table>
<thead>
<tr>
<th>Query</th>
<th>AH</th>
<th>LS</th>
<th>PS</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral citations</td>
<td>119.0</td>
<td>253.6</td>
<td>17.4</td>
<td>217.4</td>
</tr>
<tr>
<td>Non-integral citations</td>
<td>369.7</td>
<td>55.8</td>
<td>-</td>
<td>75.9</td>
</tr>
</tbody>
</table>
| In the Arts & Humanities *op.cit*  
is only used in footnotes and endnotes | 14.7 | -   | -   | 30.1 |
| *op.cit*                | 14.7 | -   | -   | -   |
| Vancouver numbering    | 12.0 | 138.5 | 219.7 | 67.5 |
| cited in               | 35.7 | 196.1 | 3.1  | 134.2 |

No *op.cit* or *ibid* in the Physical Sciences.
## Results for Life Sciences

<table>
<thead>
<tr>
<th>Query</th>
<th>AH</th>
<th>LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral citations</td>
<td>119.0</td>
<td>253.6</td>
</tr>
<tr>
<td>Non-integral citations</td>
<td>544.7</td>
<td>369.9</td>
</tr>
<tr>
<td>Ibid</td>
<td>91.4</td>
<td>-</td>
</tr>
<tr>
<td>Op. cit</td>
<td>14.7</td>
<td>-</td>
</tr>
<tr>
<td>Vancouver numbering</td>
<td>12.0</td>
<td>138.5</td>
</tr>
<tr>
<td>Cited in</td>
<td>35.7</td>
<td>196.1</td>
</tr>
</tbody>
</table>

- **LS** has the most instances of ‘cited in’ - largely due to use of secondary sources in Health studies.
- Food Sciences and Medicine do not cite much at all - using either the Harvard or the Vancouver method.
- Psychology prefers integral citations - Agriculture, Biological Sciences and Health prefer non-integral.
### Results for Social Sciences

<table>
<thead>
<tr>
<th>Query</th>
<th>AH</th>
<th>LS</th>
<th>PS</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral citations</td>
<td>119.0</td>
<td>253.6</td>
<td>17.4</td>
<td>217.4</td>
</tr>
<tr>
<td>Non-integral citations</td>
<td>544.7</td>
<td>369.9</td>
<td>31.7</td>
<td>605.8</td>
</tr>
<tr>
<td><em>ibid</em></td>
<td>91.4</td>
<td>11.4</td>
<td>-</td>
<td>75.9</td>
</tr>
<tr>
<td><em>op.cit</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>30.1</td>
</tr>
<tr>
<td>Vancouver numbering</td>
<td></td>
<td></td>
<td>219.7</td>
<td>67.5</td>
</tr>
<tr>
<td>cited in</td>
<td>3.1</td>
<td>10.1</td>
<td>3.1</td>
<td>134.2</td>
</tr>
</tbody>
</table>

Non-integral citations are preferred in all the major Social Science disciplines, apart from HLTM (Hospitality, Leisure and Tourism Management).

Law students tend to refer to cases, acts and judges’ pronouncements rather than authors and dates.
Try out the queries!

- [tag = "NP.?"] [word = "et"]? [word = "."]? [word = "al"]? [word = "."]? [word = "\()] [tag = "MC"] [word = ")"] [tag = "VV|VO|VVD|VVZ|VM" & textpart != "bibliography|note"]
- [word = "\()" [tag = "NP.?"][tag = "MC."] [] {0,5} [word = ")" & textpart != "bibliography|note"]
- [word = "ibid"]
- [word = "op"][word = "."]?[word = "cit"]
- `<p>` [textpart = "bibliography" & word = "1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23|24|25|26|27|28|29|30|31|32|33|34|35|36|37|38|39|40"]
- [word = "cited"] [word = "in" & textpart != "bibliography|note"]
Law, Problem question: Relevant circumstances, the fair-minded and impartial observer would consider that there was a real possibility...

Law, Problem question: Whether the fair-minded and informed observer, having considered the facts, would conclude that...

Law, Problem question: Possibility that Geoffrey could be biased. An objective observer would most certainly not come to such a conclusion...

Thw writer of this assignment was a female student in her third year of university study. She was studying Law and Sociology.
References

