

Using corpus tools to investigate citation



Hilary Nesi
EAP and Corpora, Coventry University
21 June 2014

A red jagged outline, resembling a lightning bolt or a starburst, surrounds the first text block.

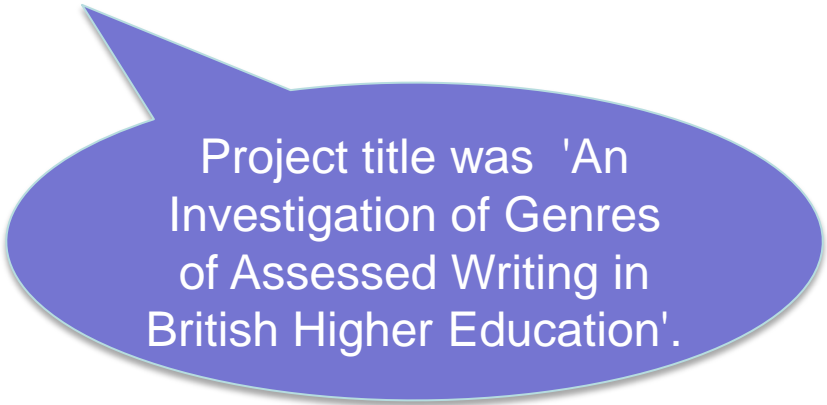
1. Presentation of findings from my investigation of citation styles in the BAWE corpus.

A green jagged outline, resembling a lightning bolt or a starburst, surrounds the second text block.

2. Hands-on investigation of these citations using corpus query language (CQL).

The aim of the 2004-7 project:

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

BAWE corpus
contents

Numbers of texts
at each level
and in each
domain

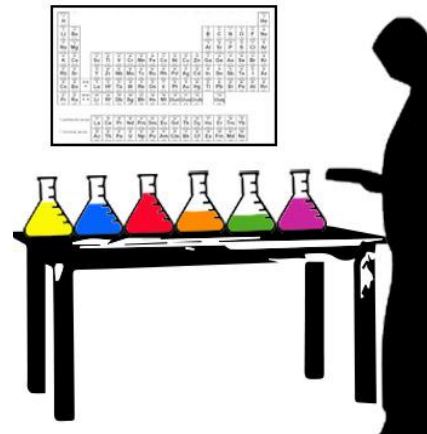
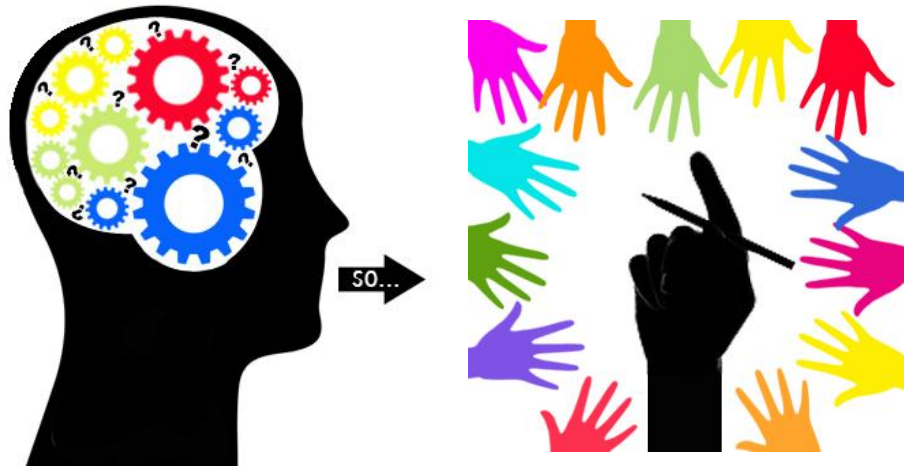
	Year 1	Year 2	Year 3	Year 4
Arts & Humanities	255	229	160	80
Life Science	188	206	120	205
Physical Science	181	154	156	133
Social Science	216	198	170	207

30+ disciplines represented:

<p><i>Arts & Humanities</i></p> 	<p>Archaeology, Applied Linguistics, Classics, Comparative American Studies, English, History, Philosophy</p>
<p><i>Life Sciences</i></p> 	<p>Agriculture, Biological Sciences, Food Sciences, Health, Psychology, Medical Science</p>
<p><i>Social Sciences</i></p> 	<p>Anthropology, Business, Economics, HLTM (Hospitality, Leisure and Tourism Management), Law, Politics, Publishing, Sociology</p>
<p><i>Physical Sciences</i></p> 	<p>Architecture, Chemistry, Computer Science, Cybernetics & Electronics, Engineering, Mathematics, Meteorology, Physics, Planning</p>

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*

Not included

‘general’ references

- 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" ]
```

[word = "et"]? [word = "\."]?
[word = "al"]? [word = "\."]?

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 = "\")"]
```


[textpart != "bibliography|note"]

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 (**SiO₂**) PZT ceramic Barium titanate Zinc oxide Lithium tantalate (**LiTaO₃**) Lithium niobate (**LiNbO₃**)

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|3  
1|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

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

Vancouver referencing

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

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

Query	AH	LS	PS	SS
Integral citations	119.0	253		7.4
Non-integral		369		5.8
		11.4	-	75.9
<i>op.cit</i>	14.7	-	-	30.1
Vancouver numbering	12.0	138.5	219.7	67.5
cited in	35.7	196.1	3.1	134.2

In the Arts & Humanities *op.cit* is only used in footnotes and endnotes

No *op.cit* or *ibid* in the Physical Sciences

Results for Life Sciences

Query	AH	LS		
Integral citations	119.0	253.6		
Non-integral citations	544.7	369.9		
	4	11.4	-	75.9
	7	-		
Vancouver numbering	12.0	138.5		
cited in	35.7	196.1		

Food Sciences and Medicine do not cite much at all - using either the Harvard or the Vancouver method.

LS has the most instances of 'cited in' - largely due to use of secondary sources in Health studies.

Psychology prefers integral citations - Agriculture, Biological Sciences and Health prefer non-integral.

Results for Social Sciences

Query	A	PS	SS
Integral citations	119	17.4	217.4
Non-integral citations	54		605.8
<i>ibid</i>	91.4	11.4	75.9
<i>op.cit</i>		-	30.1
Vancouver numbering		219.7	67.5
cited in		3.1	134.2

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"]

user: anonymous corpus: British Academic Written English Corpus (BAWE)

Concordance
Word List

Corpus: British Academic Written English Corpus (BAWE)
Hits: 298 (35.7 per million)

Save

View options

KWIC

Sentence

Sort

Left

Right

Node

References

Shuffle

Sample

Filter

Frequency

Node

Node

Doc ID

Text Types

Collocations

ConcDesc

?

Menu position

Agriculture,Methodology recount	and the average composition	can be seen	in Table 2. Diets were made isonitrogenous by calculating
Agriculture,Methodology recount	analysis for the compound feed	can be seen	in Table 3. </p> Table 2: Mean nutrient composition
Archaeology,Methodology recount	ected to definite butchery. This	can be seen	when there is a clean cut, often with an area sticking
Archaeology,Methodology recount	. The re-crystallised salts observed in the sample	can only be seen	in cross-polarised light and appear as coatings on
Biological Sciences,Methodology recount	has on its rate of reproduction. </p><p> Some clones	can also be seen	to be more affected by the varying plant quality
Biological Sciences,Methodology recount	moved P element carrying in the w+ gene. </p><p> As	can be seen	, there were approximately 10 times more white eyed
Biological Sciences,Methodology recount	little p53 growth at all, but on what there is blue	can be seen	. There is much more lamin growth, showing more diploids
Biological Sciences,Methodology recount	of oxygen evolution in isolated chloroplasts. <p> As	can be seen	from the results above, when DCMU, a herbicide, is
Biological Sciences,Methodology recount	growth, showing more diploids were produced and blue	can be seen	here also. This implies that the reporter gene MEL-1
Biological Sciences,Methodology recount	and e	can be seen	in figures 5 and 6. </p> Figure 5 Staining of 3rd instar
Biological Sciences,Methodology recount	show	can be seen	in graph 1. below, when pH increases, the enzyme
Biological Sciences,Methodology recount	Spain	can be seen	in the amount the shell has reduced in thickness.
Biological Sciences,Methodology recount	eggs, as a slow recovery to the eggshells thickness	can be seen	in the results. Eventually both species eggshells
Biological Sciences,Methodology recount	diagram below shows the generalised amino acid: </p><p> It	can be seen	that apart from having an amino group (A terminus
Biological Sciences,Methodology recount	cells. </p><p> Referring to Appendix 1, Table 2, it	can be seen	that in samples two to four for the X antibody that
Biological Sciences,Methodology recount	AH109-lamin = 1716 cfu/µg DNA, Y187-SV40 = 10 cfu/µg DNA. It	can be seen	that the most efficient was p53 and the least efficient
Biological Sciences,Methodology recount	after DDT was introduced. Through this experiment it	can be seen	that the Peregrine Falcons eggs had suffered the
Biological Sciences,Methodology recount	<p> Figure 2 shows a photograph of an X-a plate. It	can be seen	that there was very little p53 growth at all, but
Biological Sciences,Methodology recount	with substrate. Although from the reaction: </p><p> it	can be seen	that water is also a reactant, the reaction is first
Biological Sciences,Methodology recount	proportion of dropped aphids on Poor quality plants; clones	can be seen	to act in a number of ways and each to a different
Biological Sciences,Methodology recount	blood cells (RBC) to form haemagglutination, which	can easily be seen	with naked eye. Another method of virus identification
Chemistry,Methodology recount	reaction. </p><p> Effect of Changing the Monomer: This	can be seen	by comparison of the rate of the reaction for Polymer
Chemistry,Methodology recount	Effect of time of addition: </p><p> The effect of this	can be seen	from comparing the rate of reactions of Polymer 5C
Chemistry,Methodology recount	d: - </p><p> As	can be seen	from the table derivatization of the starting material
Chemistry,Methodology recount	e C-H stretches	can be seen	however the O-H stretch is lost. There were some
Chemistry,Methodology recount	s. <p> This plot	can be seen	in results - "Plot of [F-] vs. Ionic radius", the
Chemistry,Methodology recount	demonstrate the common-ion, "salting out" effect, this	can be seen	in solution C, KNO3, The Solubility of KIO4 is greatly
Chemistry,Methodology recount	further simplified: </p><p> or </p><p> From equation 6 it	can be seen	that a plot of logb against logA (see Figure 1) will
Chemistry,Methodology recount	pressure against temperature was plotted (graph 1), it	can be seen	that as the pressure increases; the temperature of
Chemistry,Methodology recount	2v point group. </p><p> From the character table, it	can be seen	that only the three vibrations, A 1, B 1 and B 2
Chemistry,Methodology recount	molecular orbital energy diagram for the allyl anion, it	can be seen	that there are two occupied MO's and one vacant. </p>
Chemistry,Methodology recount	pie bonds with oxygen. The effects of this bonding	can be seen	when comparing their physical properties to their

The blue words show you the discipline and the genre family

Click on 'Left' or 'Right' to sort in alphabetical order the words to the left or right of the search

If you click on 'Node' the red words will be listed in alphabetical order

Click on 'References' to put the blue words in alphabetical order

Click on any of the red words to see more context

Click on any of the references in blue to see more information about the assignment

Corpus: British Academic Written English Corpus (BAWE)

Hits: 13 (1.6 per million)

Law,Problem question	relevant circumstances, the fair-minded and impartial observer would consider that there was a 'real possibility
Law,Problem question	Porter v Magill applies [69]: would an independent observer would perceive bias? The Pinochet case [70] provides
Law,Problem question	consider is whether "the fair minded and informed observer ... would conclude that there was a real possibility
Law,Problem question	circumstances would lead a fair-minded and informed observer reasonably to apprehend that there was a real possibility
Law,Problem question	All ER 465 </p> "Whether the fair-minded and informed observer , having considered the facts would conclude that
Law,Problem question	account is "..... whether the fair-minded and informed observer , having considered the facts, would conclude that
Law,Problem question	concerning the question to what extent an objective observer would doubt the impartiality of a judge, whether
Law,Problem question	possibility that Geoffrey could be biased. An objective observer would most certainly not come to such a conclusion
Law,Problem question	we have to ask whether a fair-minded and informed observer knowing all the facts would come to the conclusion
Law,Problem question	circumstances would lead a fair-minded and informed observer to conclude that there was a real possibility...that
Law,Problem question	biased". In the present case, it is probable that an observer would conclude that the panel was biased, in the
Law,Problem question	presence endorsed the order tacitly so that a fair minded observer could legitimately think that he was predisposed
Law,Problem question	condemn his refusal so much so that a fair minded observer could legitimately think that he was predisposed

Click on one of the blue words to get information about the assignment

text.genre	Problem question
text.grade	unknown
text.l1	English
text.level	3
text.sex	f
text.studentage	unknown
div1.type	section
p.n	p29.32
s.n	s2.3;p29.32
text1.course	Law and Sociology
text1.date	unknown
text1.id	0374a
text1.modcode	LA201
text1.modtitle	General Principles of Constitutional and Administrative Law

The writer of this assignment was a female student in her third year of university study.

She was studying Law and Sociology

References

Becher, T. & Trowler, P. (2001) *Academic Tribes and Territories: Intellectual Enquiry and the Culture of Disciplines*. Buckingham: The Society for Research into Higher Education and Open University Press

Bizup, J. (2008) BEAM: A Rhetorical Vocabulary for Teaching Research-Based Writing. *Rhetoric Review* 27 (1) 72–86

Nesi, H. (2014) Corpus Query Techniques for Investigating Citation in Student Assignments. In: M. Gotti & D. S. Giannoni (eds.) *Corpus Analysis for Descriptive and Pedagogic Purposes: English Specialised Discourse*. Bern: Peter Lang 85-106