welcome...

analysing semantic change: a corpus-driven approach

dominic smith

Background image: Entry for '*coz*', Diccionario de Autoridades, Real Academia Española, 1732

semantic change

Philologists have identified three main types of change in vocabulary:

- Borrowing (eg *'microscope'*)
- Adaption, often by analogy (eg *'carriage'*)
- and something that seems more fluid:

Latin TUBELLU(S) (swelling) > Castilian *tobillo* (ankle)

Latin TALU(S) (ankle) > Castilian *talón (*heel)

Latin CALCE(M) (heel) > Castilian *coz* (kick)

All changes seem to be unpredictable.

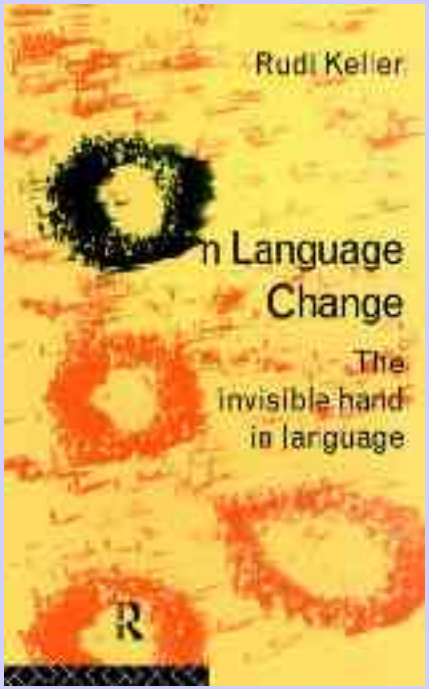
hermeneutics

- Hermeneutics says that all texts must interrelate, so we must relate a text we are interested in to texts to which it refers.
- Interpretation, and therefore the meaning assigned, is individual because any given person will have been exposed to different texts.
- But if usage of a lexical item changes over time, this change must occur through all individuals in the discourse community interpreting the meaning of that item differently.

hermeneutics

- Therefore, the change must be indicated in a number texts in the years prior to when the change took place, so that all members of the discourse community are led to a different interpretation.
- Consequently, we must consider whether semantic change can be predicted through Hermeneutic readings.
- If so, can Hermeneutics explain why such changes take place in the first place?

the 'invisible hand'



- 'The question as to how the process of change in our language takes place is therefore not an historical one, but a systematic one. The changes of tomorrow are the consequences of our acts of today.' (p.14)

Keller, Rudi (1994) On language change: the invisible hand in language (London: Routledge)

- Keller believes that an 'invisible hand' guides change. What is this hand?

the 'invisible hand'

- Micro-domain: breaking cars
- Macro-domain: traffic-jam
- 'The macro-domain is the structure generated by the micro-domain; in our case the 'traffic jam out of nowhere'. We can thus summarise that *a phenomenon of the third kind is the casual consequence of a multitude of intentional actions which serve, at least partially, similar intentions*.'

- Keller (1994) (p. 65)

the 'invisible hand' in semantic change

- Keller's sees semantic change as such a *'phenomenon of the third kind'*: it occurs as a result of an intervention (eg. purposefully using a word in a new sense) by a number of individuals.
- The result (eg. altering the meaning of a related word) is probably not intended and affects the whole discourse community.
- In the analogy, we can measure the effects of the individual applying the foot to the break pedal. Can we do the same in a corpus-driven study?

towards a corpus-driven investigation

- To attempt to investigate the individual's actions would probably be impossible: we don't have all of every individual's language use over several years recorded in a corpus.
- We can look for the effect starting to occur at the community level (the slowing-down before the traffic jam)
- If Keller is right, we should expect the changes to happen just before the noticable semantic change takes place.

the corpus

- Four sub-corpora:

1810 (1795 - 1824)

1840 (1825 - 1854)

1870 (1855 - 1884)

1900 (1885 – 1914)

- British fiction (sourced mainly from Project Gutenberg / Oxford Text Archive)
- Each sub-corpus: 1.5 million words; no one text / one author more than 300 000 words.

corpus content

- 1810: Ainsworth, Anne of Swansea, Austen, Hogg, Lewis, Porter, Scott, Shelley, Sleath (1,099,182 words)
- 1840: Ballantyne, Brontë, Buckstone, Bulwer-Lytton, Dickens, Gaskell, Kingsley, LeFanu, Thackeray, Trollope (1,599,749 words)
- 1870: Carrol, Dickens, Eliot, Gaskell, Hardy, James, MacDonald, Stevenson, Trollope (1,519,621 words)
- 1900: Burnett, Conan-Doyle, Grahame, Hardy, Jerome, Stoker, Wells, Wilde (766,433 words)
- Not all cleaned (yet)

methodology

- Look at most common words by frequency in each sub-corpus.
- Look at the words with the largest frequency difference (per 1000 words) between the sub-corpora.
- Look at the collocates of the word and calculate the MI-Score of the collocation.
- If the word has become unstable (cars brakeing), we would expect MI-Score to be unstable, giving rise to a change of meaning.
- ie. we assume that context will indicate the change.

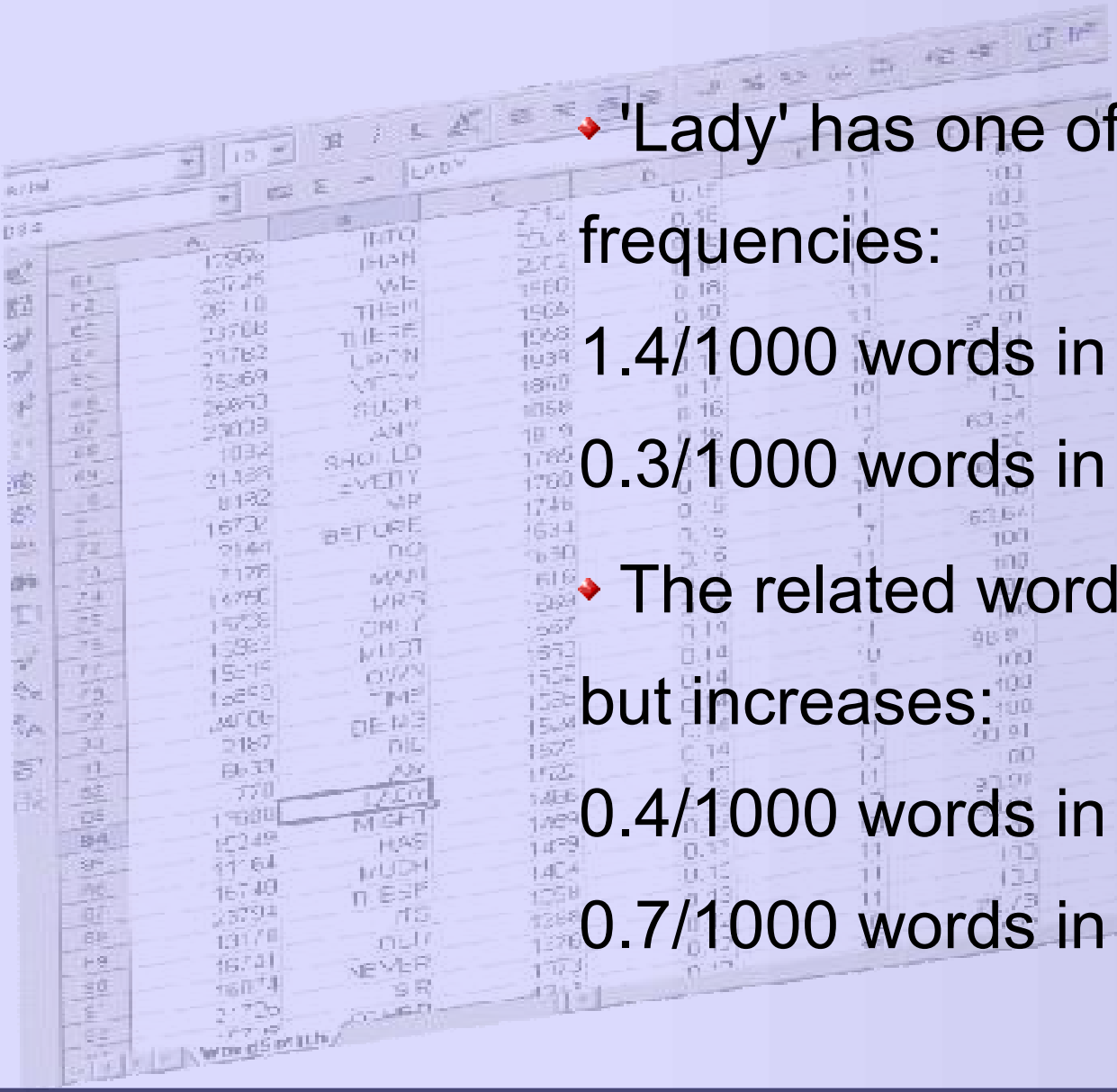
methodology

- Take the top ten collocates of the word being investigated (the *node*) by MI-Score in each sub-corpus. Get the MI-Score for each of these collocations in each of the other corpora.
- Should now be possible to graph the semantic change in terms of the change in MI over time.
- Using a 3D graph, could be possible to show the relationship between one word changing and another related word.
- At present, based on MI-Score. May change to MI3 or log-likelihood...

the software problem

- Few people use diachronic corpora because the software isn't available, but until people do studies like this, the software won't be written!
- For in-depth analysis, we need to be able to display the exact collocations. WordSmith is very good at this synchronically, but not designed to cope with the extra variable of time.
- I'm currently designing an XML-based language to resolve this issue for my purposes, which will display results in a web-browser through an XSLT template. Code will be available when finished. For the case study you're about to see, only using WordSmith and spreadsheets.

case study: 'lady'



- 'Lady' has one of the most different frequencies:

1.4/1000 words in the 1810 sub-corpus,

0.3/1000 words in the 1900 sub-corpus

- The related word 'woman' is less different, but increases:

0.4/1000 words in the 1810 sub-corpus

0.7/1000 words in the 1900 sub-corpus

case study: 'lady'

- Problem is the number of titles (eg. *Lady Delmore*)
- Once these are ignored, the top collocates are as shown:
- Easily seen that the least significant collocates (lowest MI) are the ones found in the 1900 sub-corpus. More significant ones seem to disappear in the middle of the century.

Collocation	1810	1840	1870	1900
Lady+Natured	0	11.1	0	0
Lady+Abbess	10.5	0	0	0
Lady+Monk	0	0	10.1	0
Lady+Ladyship	0	9.6	0	0
Lady+Countess	9.5	7.7	7.1	0
Lady+Gaunt	0	9.1	0	0
Lady+Rejoined	8.9	0	0	0
Lady+Bosom	8.6	0	0	0
Lady+Mademoiselle	0	8.3	0	0
Lady+Inquired	8	0	0	0
Lady+Resumed	7.9	0	0	0
Lady+Amiable	0	7.6	0	0
Lady+Countenance	7.5	0	0	0
Lady+Gracious	0	7.3	7	0
Lady+Agreeable	7.2	0	0	0
Lady+Exclaimed	7.1	0	6.3	0
Lady+Madam	7	0	0	0
Lady+Maid	0	7	0	0
Lady+Cousin	6.4	5.4	6.9	0
Lady+Weeping	0	6.8	0	0
Lady+Quarrel	0	6.6	0	0
Lady+Cousins	0	0	6.5	0
Lady+Carriage	0	5.6	6.3	0
Lady+Madame	0	0	6.1	0
Lady+Seated	0	0	6.1	0
Lady+Matching	0	0	5.6	0
Lady+Cried	6.2	6.1	0	5
Lady+Till	4.1	0	0	3.6
Lady+Young	5	6	5.4	3.4
Lady+Dear	5.5	5.5	5.2	3
Lady+Lord	4.1	4.4	2.5	2.4
Lady+Old	2.2	4.2	2.8	2.2
Lady+Day	1	1.6	1.7	1.6
Lady+Asked	2.6	2.5	2.5	1.5
Lady+Woman	2.3	2.1	3.5	1.5
Lady+Mrs	4.9	4.6	3.6	1.4

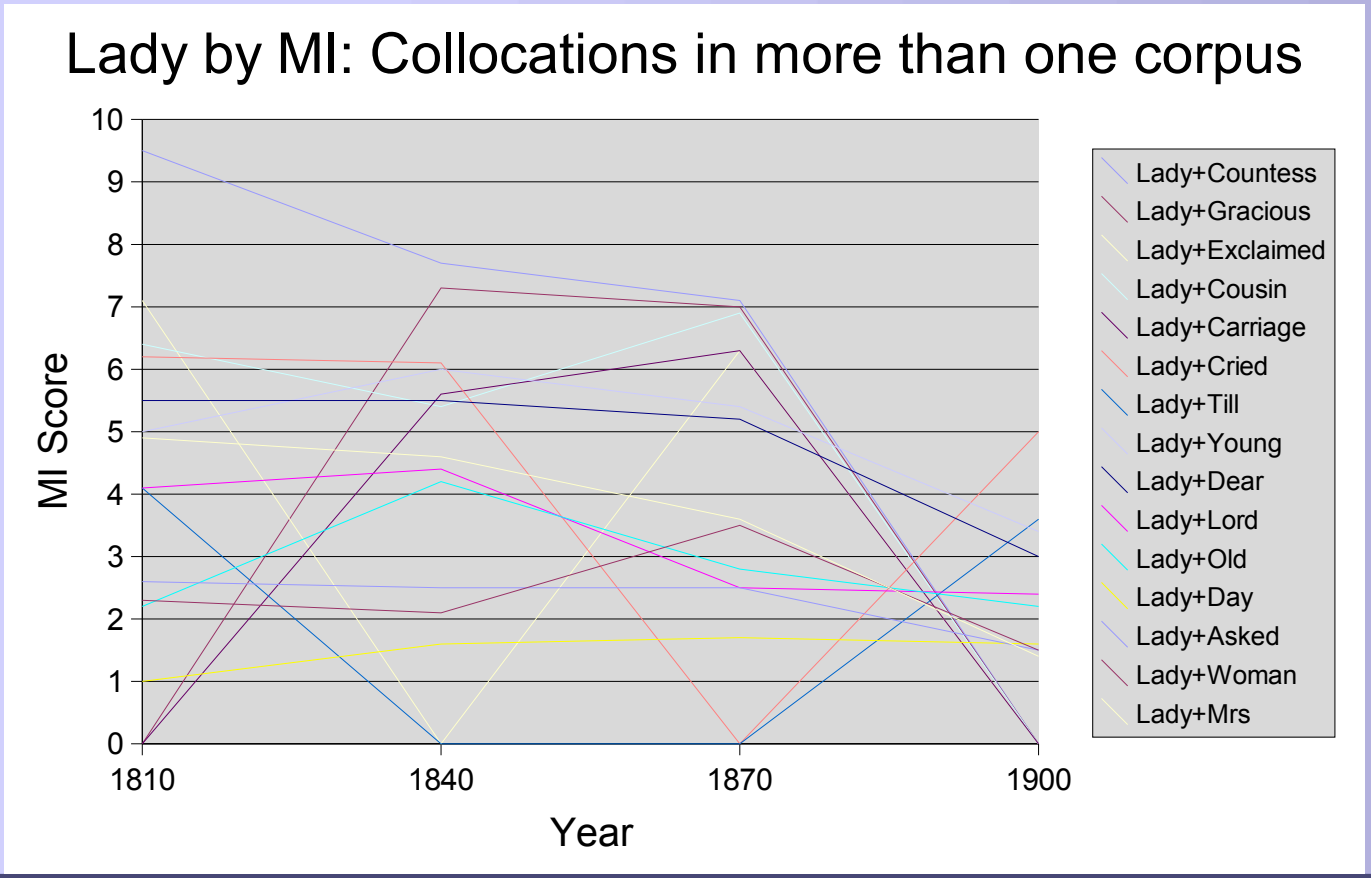
case study: 'lady'

- But there are a large number of collocates shown here which only appear in one sub-corpus, so confusing (even if instructive in showing that collocates may not be as stable as thought).

Collocation	1810	1840	1870	1900
Lady+Countess	9.5	7.7	7.1	0
Lady+Gracious	0	7.3	7	0
Lady+Exclaimed	7.1	0	6.3	0
Lady+Cousin	6.4	5.4	6.9	0
Lady+Carriage	0	5.6	6.3	0
Lady+Cried	6.2	6.1	0	5
Lady+Till	4.1	0	0	3.6
Lady+Young	5	6	5.4	3.4
Lady+Dear	5.5	5.5	5.2	3
Lady+Lord	4.1	4.4	2.5	2.4
Lady+Old	2.2	4.2	2.8	2.2
Lady+Day	1	1.6	1.7	1.6
Lady+Asked	2.6	2.5	2.5	1.5
Lady+Woman	2.3	2.1	3.5	1.5
Lady+Mrs	4.9	4.6	3.6	1.4
Average	**4.06**	**4.2**	**4.05**	**1.71**

case study: 'lady'

- Several collocates cease to be used (eg. *Countess, Gracious, Exclaimed*). Others continue (eg. *Young, Dear, Lord*) but general pattern is reducing MI, so collocations becoming weaker.



case study: 'lady'

- Seems *'lady'* changed from being used in idiomatic fixed phrases (eg *'gracious lady'*), probably with a notion of respect to being a near-synonym for 'woman', taking collocates such as '*young'*, '*old'* and *'Mrs.'*.
- The 1900 sub-corpus has almost no proper nouns as collocates of lady, showing that (although these were excluded here) usage as an aristocratic title also reduced, confirmed by the loss of the collocate *'countess'*.

case study: 'lady'

- 4. a. A woman of superior position in society, or to whom such a position is conventionally or by courtesy attributed. Originally, the word connoted a degree equal to that expressed by lord; but it was ... early widened in application, while the corresponding masc. term retained its restricted comprehension. In mod. use lady is the recognized fem. analogue of gentleman, and **is applied to all women above a loosely-defined and variable, but usually not very elevated, standard of social position.** Often used (esp. in 'this lady') as a more courteous synonym for 'woman', without reference to the status of the person spoken of. (OED)
- The corpus evidence indicates that the required elevation decreased in the nineteenth century but the OED (2nd edition) only has attestations from 1807, 1886 and 1888, so the key period of change is omitted.

conclusions

- Need to chart the collocates of *'lady'* vs. those of *'woman'*. I would expect a convergence to take place over the century.
- Eventually, the dissertation should contain 10 such word studies; verbs, nouns and adjectives.
- Corpus needs to be finalised and cleaned properly.
- Simple conclusion: the methodology seems to work, the results given appear more accurate than traditional descriptions and will hopefully lead to a better understanding of how languages change.

thank you

This presentation is available to download on my website:

www.domsmith.co.uk/mphil

If you have any questions or suggestions:

dom@domsmith.co.uk