Methods for Text Mining and Analysis of Text Corpora

Meaning

***Combining Words by Meaning***

*The* Historical Thesaurus of English *has been developed at Glasgow over a period of more than forty years. Since its initiation, the growth of digital text analysis has opened up avenues for using the data that were never envisaged at the time of its creation. This session will explore some of the ways in which this data can help us approach the meaning of texts.*

Look at your notes from earlier workshops and pick out some of the words which have emerged as keywords that you might want to follow up on.

Go to the *Historical Thesaurus* website: [www.gla.ac.uk/thesaurus](http://www.gla.ac.uk/thesaurus) and use the ‘Browse’ option on the top left of the screen to look at the *Thesaurus* hierarchy and to get a feel for how it works. Click on a ‘+’ symbol in the hierarchy tree to view subordinate levels. When you see a category name that interests you, click on it – the results for that category will then be displayed on the right-hand side of the screen.

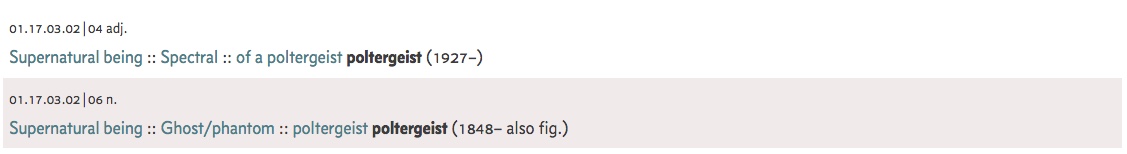
Categories contain all the words which have had the meaning in the category heading since the period of Old English (roughly 800 ce to 1100 ce) to the present day. The words are arranged in chronological order, beginning with the earliest. Dates at which the word has been found in use are given alongside each. The evidence for dating is based on the 2nd edition of the *Oxford English Dictionary*.

Use the quick search box on the top left of the page to search for one of your keywords.

How many results (i.e. different meanings) does the *Thesaurus* return for the word you entered?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….



Look through the list of results. A ‘trail’ of headings, such as that shown above, gives you an indication of the levels of the hierarchy in which your results are nested. Find the one that you think is most likely to be the meaning which matches your keyword in your text and click on it. How many different words have had the same meaning as the keyword you are searching for?

Looking at the dates, how many of these words are active in the time period of the texts in which you are interested?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

Load your text into AntConc and make sure you can see it in the ‘File View’ tab.

Switch to the ‘Concordance’ tab, and find the button marked ‘Advanced’ next to the search box. Click this and a new options window should open. Here, click the check-box next to ‘Use search term(s) from list below’. This allows you to search for multiple words rather than the single word you would enter in the normal search box.

Copy and paste or type the words from the *Thesaurus* category which mean the same as your keyword into this box. Each word should be on a new line, and there should be no punctuation between the words. When you are done, click ‘Apply’. When the options box has closed, click ‘Start’ under the normal search box. Does the search return any of the other items along with your keyword?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

Are there any of the extra items for which you feel the wrong meaning of the word is being returned?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

Try searching for some of these words in the BNC (corpus.byu.edu/bnc) and see what the results look like there. How many of the results might be useful to you?

……………………………………………………………………………………………………………………………………………….

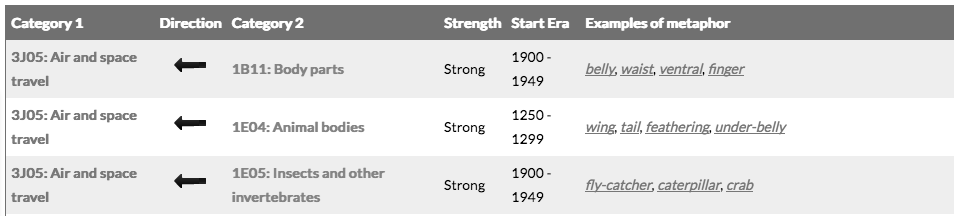
……………………………………………………………………………………………………………………………………………….

***Metaphor***

Access the Metaphor Map of English (<http://mappingmetaphor.arts.gla.ac.uk/>) and start by going to **Browse**.

You’ll see that the metaphor map is divided into three fields: *External World*, *Mental World*, and *Social World*. Under each larger field, there are smaller, more specific categories. Choose any category you are interested in and/or is relevant to your own research.

After choosing a category, the first thing you see is this **tabular view**:



This is a list of all the different categories that have a metaphorical connection with the category you have chosen and gives you examples.

1. How many categories does your category have a connection with?

……………………………………………………………………………………………………………………………………………….

2. Do any of the examples surprise you? What can you find out about the examples by clicking on them?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

*Source and target domains*

Have a look at the arrows between the categories. The direction of the arrows tells you in which direction the metaphorical connection goes.

For example, we can see that words originally used to describe fish (*fin, tail fin*) have been borrowed to describe things in theair and space travelcategory. That means that air and space travel is the **target** of the metaphor and fish is the **source**.



3. Which categories are a metaphorical source for your category?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

If the arrow points the other way, such as with air and space travel and sequence, this means that the words were originally used in air and space travel, and were later borrowed to talk about sequences in general. Therefore in this case air and space travel is the **source** of the metaphor and sequence is the **target**.



For example, the word *take-off* originally just referred to planes taking off, but is now also used as a metaphor for other kinds of beginnings as well (e.g. *Her singing career has just begun to take off*).

4. Which categories are a metaphorical target for your category?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

If the arrow points to both directions, that means that the connection goes both ways.



For example, if you look at the connection between air and space travel and birds**,** you find words that were originally used to describe birds but have been borrowed to also talk about planes (wing of a bird -> wing of a plane). There are also words that went the other way around: for example, the word *helicopter* was originally coined to talk about actual helicopters but has later been borrowed to describe the way some birds fly.

5. Are there any connections that go both ways in your category?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

*Timelines*

There is a dark green box on the left that you can use to change the way you view the results. Click on **Change view** and choose the **Time line view**.

6. Which metaphorical links in your category are attested earliest, and which are the most recent?

……………………………………………………………………………………………………………………………………………….

7. From which decade or century are most of the metaphorical links from? Can you speculate why?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

***Semantic Tagging***

*Manual sorting of the results of concordance searches is often essential. Depending on the words you are looking for, there may be a good number of positive as well as a good number of negative hits for the meaning you need. Semantic analysis software will be essential in speeding up this process in the future. One method for doing this is to add a layer of annotation to the text (like part of speech tagging and lemma tagging) which gives the particular meaning of that word. This software is still in its early days and there is much to be done to improve its reliability, but it is well worth experimenting with it*.

Visit the webpage for the online version of the SAMUELS project’s semantic tagger. It can be found at: phlox.lancs.ac.uk/ucrel/semtagger/English

Select some text either from your own files or a text such as those from the Project Gutenberg website we used yesterday, and enter this into the box in the tagger page. Either leave the year boundaries blank in the tagger, or set them with a margin of error of roughly twenty years on either side of the date (e.g. for 1930, enter 1910 – 1950).

When you have the result, have a look through the tags which the tagger returns. Look especially at the tags returned in the column headed ‘SEMTAG3’. The tagger lists three possible semantic categories to which each word could belong, in the order it has judged to be most likely. How often do you feel that the first tag is correct?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

Copy the table of results, open an Excel spreadsheet, and paste the results into the spreadsheet. Highlight the columns in which data is found by clicking on the ‘A’ column header and dragging the mouse pointer across to the ‘G’ column header. Then click ‘Sort and filter’ along the top menu bar and select ‘Filter’ from the drop down menu.

Click on the arrow next to the heading ‘SEMTAG3’ in the first cell in column G, and select ‘Ascending’. You will now have a list of the words in the text sorted by the first tag in the ‘SEMTAG3’ column. Have a look down the column. Are there any first (i.e. of the three in that column) which appear several times? How often are words which are given the same tags the same words, and how often are they different words given the same tag?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

*This is a very beginning of working with tagged data like this. It is possible using formulae in Excel to separate out the three tags into three separate columns, and then to search, order, or count the results in this way. The same can be done using the desktop GUI version of the tagger, which can be downloaded from* [*www.gla.ac.uk/samuels/#/projectoutputs*](http://www.gla.ac.uk/samuels/#/projectoutputs)

*The results of semantic tagging are in developmental stages and need to be used with more care than other types of tagging. However, the technology should improve in the next few years!*

Corpora which have been tagged with these meaning codes exist and are available online through the BYU website. Go to [www.hansard-corpus.org](http://www.hansard-corpus.org) to try one of these.

First try a normal word search of this corpus. In the search box on the left-hand side of the page, type ‘war’ and have a look at the results which are returned. Clicking on the word ‘WAR’ will return all the results from the corpus. The initial results display also gives you frequencies for each of the decades covered by the corpus; clicking on one of these frequency numbers will return the results for the appropriate decade.

We can also, however, search for everything that has been tagged as belonging to the semantic category ‘War’. To do this, look below the search box for the option ‘CATEGORIES’. This will bring up a list of categories in the bottom right. Find category ‘BA Armed hostility’ and click on the ‘Armed hostility’ part of the heading (rather than the ‘BA’). This will bring up a list of categories which are below ‘Armed hostility’ in the hierarchy.

In this list, find ‘BA:01 War’, and click on the ‘BA:01’ part. This will enter ‘{BA:01}’ into the search box on the left-hand side of the page. Click ‘Search’, and look at the list of results which are given this time. Click through one of the results, either the word itself or the frequency given for a decade within the corpus. This will give you the results for that word, and you can go back and look at other words which have been tagged as belonging to the same semantic category if you would like to.

Try this with a few other categories. Remember that you’ll have to delete old search terms from the text box so that they aren’t added to the new ones (unless you do want to find strings of words tagged with particular categories).

It is possible, if you already know the code for the category you want, to search for this directly without going through the menu. Just enter the code in the search box, surrounded by curly brackets – i.e. {}

The codes used by the tagger are based on the *Historical Thesaurus* but use a modified version of the hierarchy. A full list of codes in this version of the *Historical Thesaurus* can be found under the ‘Thematic Categories’ link at [www.gla.ac.uk/samuels/#/projectoutputs](http://www.gla.ac.uk/samuels/#/projectoutputs)