A LEVEL

Teacher Guide

ENGLISH LANGUAGE

H470
For first teaching in 2015

How to use quantitative analysis in a language investigation

Version 1

How to use quantitative analysis in a language investigation

This guide contains a series of activities that could be given to a student, either as separate lesson activities or as a complete booklet, for use in quantitative analysis when working on their independent language investigation. Some form of quantitative analysis is required in this investigation (see pages 13 and 18 of the specification).

The first lesson activity is on percentages and charts –it could be used as a single activity or broken down into the two constituent parts.

The second lesson activity looks at useful maths calculations – mean, median, mode range. Again it could be used as an overarching single activity or broken down into the constituent parts.

The final lesson activity looks at different web tools that could be used for analysing text – Wordle, Tweetcloud, Concordle, Google Ngram Viewer and Readability tests. As with the other two lesson activities, these could be used as an overarching single activity or broken down into the constituent parts.

Percentages and charts

You cannot make valid comparisons of amounts of language features in two texts or data sets if your texts/data sets are very different in length.

Saying that there are 25 deontic modal verbs ('must', 'should' etc.) in Text A but only 5 in Text B means nothing if text A is 1000 words long and text B is only 100.

Therefore, you must **convert raw numbers into percentages** to make valid comparisons.

A reminder of how to do this is here:

http://www.bbc.co.uk/schools/gcsebitesize/maths/number/percentagesrev1.shtml

In the example above, 25/1000 when compared to 5/100 when converted to percentages is:

2.5% compared to 5% - so the **density** of deontic modal verbs in Text B is twice that of Text A

Better still, these numbers can be presented as a visual graph which makes it easier to understand and will reduce your word count.

Use the following process:

1. The **significance** of the statistics on a graph should be introduced simply in words:
   eg: *There are twice as many deontic modal verbs in Text B than Text A when expressed as a proportion of all the words in each text.*

2. Next, include your chart at that point (keep charts embedded within the relevant part of the analysis). Make sure you label your chart, is axes and the units of measurement clearly.
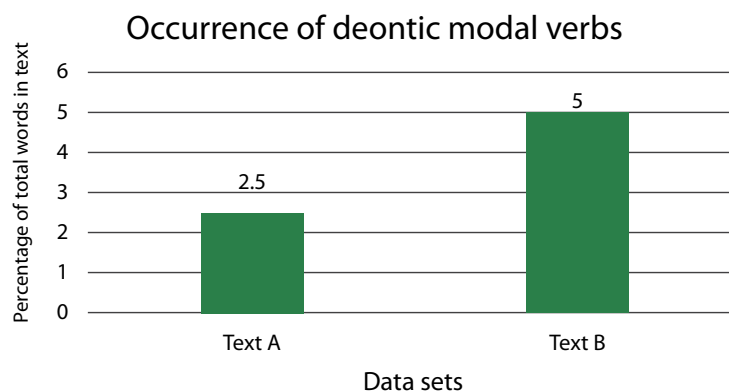
Graphs and associated data sources have been fabricated for illustrative purposes.

3.   Quantitative analysis, however, is NEVER enough on its own. The graph should then be followed up by detailed **qualitative** analysis of textual **examples** of the deontic modal verbs in the text. You should quote evidence of this **pattern of language**, link it to how it is used in its context and evaluate the likely **effect** of these language features on the reader (AO3 contextual evaluation.

In this example, as you can see, Text B has a higher proportion of these authoritative verbs than Text A. So you would need to 'dig deeper' into the texts themselves to discuss *whether* the impact or effect of bossiness created by these verbs in text B is more or less than text A.

To do this 'digging', you need to look at *what other* language features add to or soften the effect of the authoritarian tone conveyed by the deontic modal verb pattern. One language feature (or even one *pattern* of language features) alone hardly ever has such an impact/effect on the reader. So go off and find more evidence (eg of imperative verbs or other face threatening acts) to see if it is a **cumulation** of features which conveys the tone.

Charts

When you are **comparing** different data sets, bar charts are appropriate as you can see in the example above (as long as you have converted raw data into percentages!)

Excel can create these charts for you and the guide below can guide you through the process (although you have probably been taught this at school).

https://support.office.com/en-us/article/Create-a-chart-0baf399e-dd61-4e18-8a73-b3fd5d5680c2

However, when you want to calculate how much something is as **a proportion of a whole**, use a pie chart.

Let's imagine you are looking at a woman's magazine and you want to know whether it contains content stereotypically aimed at women. You have counted all the editorial pages (not including the adverts) and there are 134. You have identified 12 categories of content related to the semantic fields in these pages and want to show what proportion of the whole magazine is devoted to each semantic field.

This time, you don't have to worry about converting raw numbers to percentages because the chart will do it for you (134 pages will be 100% of the magazine)

Excel will convert your raw **tally chart**…

Content Categories	Number of pages
Fashion	25
Money	19
Technology	18
Sex	16
Beauty	10
Politics	10
Work	10
Nutrition	6
Sport	6
Family	5
Travel	5
Relationships	4
Total pages	**134**

…into a visually more accessible pie chart:

## Content analysis of magazine
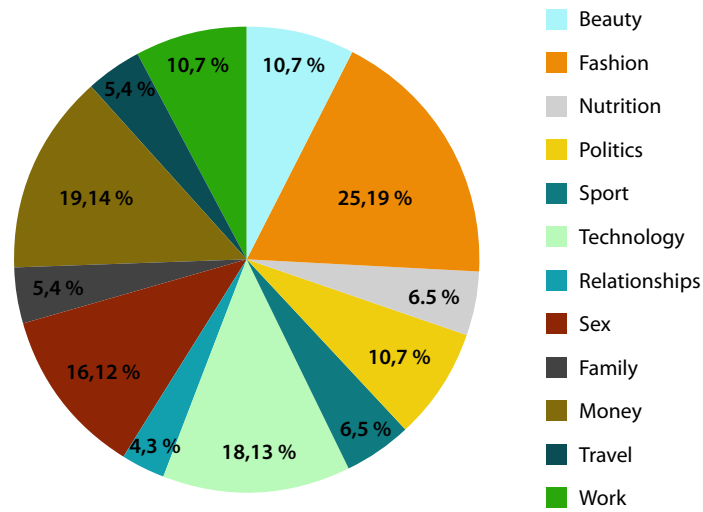
Graphs and associated data sources have been fabricated for illustrative purposes.

However, this is still difficult to read. It is not easy to tell which category had the most pages devoted to it and which had the least. This is because the amounts of pages have not been **rank ordered**.

Rank ordering of charts

The convention on a pie chart is to see it as a clock and, reading clockwise from noon, display the proportions in reducing size order. This can be achieved by **rank ordering** your data on the Excel spreadsheet before creating the chart. When you do it will look like this:

Content analysis of magazine
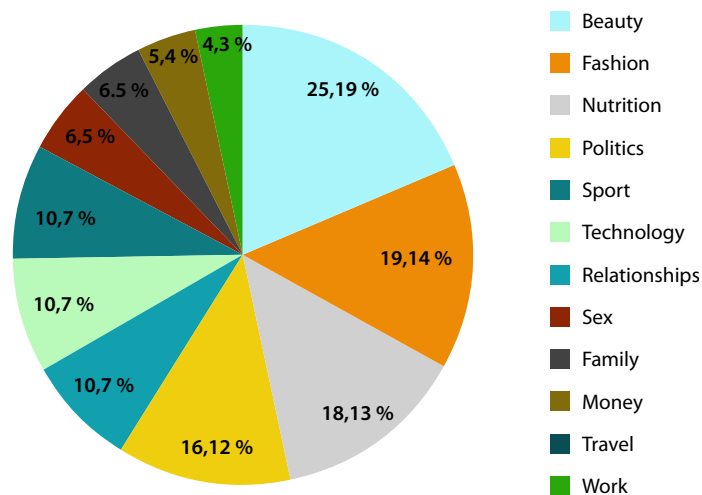
Graphs and associated data sources have been fabricated for illustrative purposes.

Now, at a glance, you can see that although, as stereotype would suggest, fashion is the largest section of the magazine, the next largest sections are money, technology and sex which we might associate more with magazines aimed at men. Indeed 39% of the content of the magazine is devoted to these issues which is certainly worth looking closer at in your analysis. Again the quantitative work here is no more than a starting point. It is useless on its own.

As a result of this initial content analysis, you might decide where to focus your qualitative analysis in more depth. You might decide that you will analyse extracts from the fashion pages to see if a less stereotypical approach to fashion journalism is being adopted there alongside some analysis of extracts from the money and technology sections. As these 3 sections make up nearly half of the magazine, they surely will contribute greatly to the impact upon the reader of what are the priorities of that particular magazine.

To recap

1. Don't compare raw numbers if texts are different lengths

2. Rank order amounts in tables before creating Excel charts

3. Use bar charts for comparing 2 data sets

4. Use pie charts for calculating proportions within one text

5. Label charts:

   a. Label all axes on charts

   b. Label units of measurement

6. Introduce graphs with a topic sentence that highlights the significance statistical point in them

7. Embed charts/graphs within relevant analytical paragraph

8. Never undertake quantitative analysis without following up with qualitative analysis.

Useful maths calculations

Mean, median, mode and range

A reminder of how to calculate these are here:

http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/measuresofaveragerev2.shtml

Mean

Or in simpler terms, 'averages'. Averages are useful if you want to find out, for example:

a. why a text seems so difficult to read because you suspect it contains long complex sentences with multiple subordinate clauses, or

b. what stage of development a child has reached by calculating their MLU (mean length of utterance).

In the first example (a), you could calculate the average length of all the sentences in your two texts and then you could represent the comparison by a bar chart:
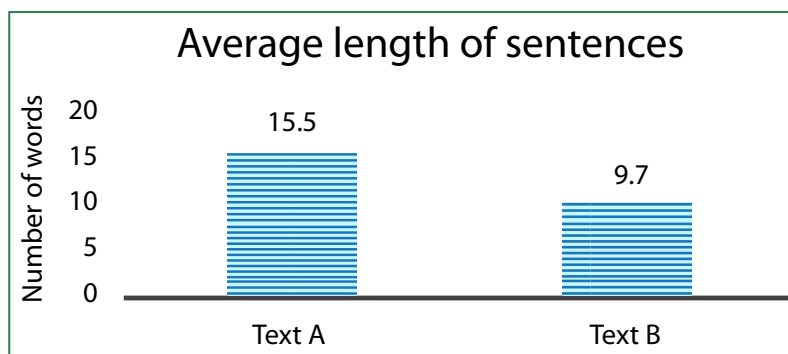
Graphs and associated data sources have been fabricated for illustrative purposes.

As always, these statistics do not tell you much on their own. But they provide an evidence-based justification for looking carefully at a few examples from each text to see if the difficulty of the reader is associated with the length of the sentence, or some other factor (obscure Latinate vocabulary or lots of obsolete words, for example)

Median

It is sometimes useful to dig deeper into raw averages, however, because very large numbers and very small numbers can distort the answer. In this case, you might want to calculate the median instead.

This is the middle number and to calculate the median of any set of numbers, you need to write the numbers in order.

To find the median number:

1.   Put all the numbers in numerical order.

2.   If there is an odd number of results, the median is the middle number.

3.   If there is an even number of results, the median will be the mean of the two central numbers.

For more detail on this see here:

http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/measuresofaveragerev3.shtml

Mode

Mode means the most frequently occurring feature and there can be more than one. You might want to find out, for example, out of the 2 modal verbs: 'must' and 'could', which is the one that occurs most frequently. This might be important if you want to find out whether the tone is bossy or more friendly.

To revise how to calculate mode, see here:

http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/measuresofaveragerev4.shtml

Range

Finally, you might want to know what the length of the longest sentence is and the length of the shortest in each text to see whether there is greater or lesser variety in each text.

Remember that to find the range of a set of numbers, you need to:

1.   Find the lowest number.

2.   Find the highest number.

3.   Subtract the lowest number from the highest number.

The difference between the highest and lowest number is the range.

For more on this see:

http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/measuresofaveragerev7.shtml

Remember:

Use the numbers discovered as starting points only to:

a.   Provide you with a rational for analysing some sections more closely than others

b.   Follow up with close textual qualitative contextual and conceptual analysis.

Useful web tools for analysing text

There are many wonderful digital, free-to-user programmes these days which enable you to 'crunch' stretches of text. Some can be very helpful in your analysis of language.

Visual Representations of Word Frequency – Wordle

http://www.wordle.net/

This programme creates a word cloud which represents visually which words occur most frequently in a text as a whole. It is helpful for identifying, at a glance, the key words and the predominant semantic fields within a text.

You paste your text (corpus) into the programme. It takes out all the grammatical words like determiners, pronouns, conjunctions and represents the rest of the (usually lexical) words visually. The more frequently used the word, the bigger it is displayed.

You can set it to display the top 25 or 50 words so it does not try and cram them all onto the cloud, and you can alter the colour, font and orientation of the text.

A drawback is that Wordle does not provide you with actual word counts, so you will still have to go back to the text and count the occurrences of words if you want accurate reliable statistical data. Also, remember: you cannot save it or print it without taking a screen shot, cropping and then pasting it into your document.

Here are 2 word clouds side by side of 2 US presidential inaugural addresses. The one on the left is Obama's and the one on the right is Bush's. They provide a good starting point for an analysis of the priorities of each of these Presidents when they took up office .

For more on these Presidential speeches, go to: http://readwrite.com/2009/01/20/tag_clouds_of_obamas_inaugural_ speech_compared_to_bushs/

NB: To save the results, you will have to take a screen shot and paste onto Word. (Shift+Print Screen or use the snipping tool in Windows Accessories, then paste in Word)

Tweetcloud

http://tweetcloud.icodeforlove.com/

Twitter has its own service where you can generate a cloud of the words most tweeted about. Here is one for the past 4 months for @ENSFCEnglish which also contains useful quantitative information at the top. You can set it for any period and change the number of words it displays

Concordle

http://folk.uib.no/nfylk/concordle/

A concordance is an alphabetical list of the principal words used in a book or body of work, with their immediate contexts. Concordances are frequently used in linguistics when studying a text, to:

- compare different usages of the same word

- analyse keywords

- analyse word frequencies

- find and analyse phrases and idioms

Concordle is a programme that will count and sort the words or phrases within a text which you can then analyse.

1. Click on the above link

2. To analyse your own text, you then clear the input and paste it in the box. It then does the rest.

3. Concordle lists the words in the text in order or frequency (as does Wordle) but it will go much further for you as an analytical tool, because:

   - It can order alphabetically the words in the pasted text ("Concordle-ordered") and show the number of times that each word occurs in the text. These numbers are useful as you might want to find and analyse all the adjectives in a text so that you can discuss their connotations. Or you might want to find all the nouns and classify how many are abstract, concrete or proper. Or you might want to identify all the verbs and identify how many are stative or dynamic in order to strengthen your analysis.

   - The programme will not help you decide which nouns are abstract, concrete or proper, but it is easier to find them in a list rather than scattered throughout a text and it also gives you the amounts so you can make statistical calculations.

   - It can also display your selected word or phrase in its immediate context in the text, allowing you to look for common collocations and examine pragmatic and connotative usage.

   - This helps you to see whether a word is being used to mean what you think it does. This might be handy for words that have the same spelling but have very different meanings. For example, the word 'bow' can mean: a hair ribbon, the front of a boat or ship, what you use to play a violin with and bending at the waist when you greet someone.

NB: To save the results, you will have to take a screen shot and paste onto Word. (Shift+Print Screen, then open Word and Paste)

Google Ngram Viewer

When you enter phrases into the Google Books Ngram Viewer, it displays a graph showing how those phrases have occurred in a corpus of books (e.g., "British English", "English Fiction", "French") over the selected years.

This tool can reveal the frequency of word use over time and is great if you are doing a language change investigation because it enables you to see visually how a word's popularity has emerged or decreased.   Be careful though, old texts printed "s" as "f" so a search for the 'F-word' can be rather misleading!

This tells you more about it and how to use it:

https://books.google.com/ngrams/info

On the Ngram below, are the occurrences of the words 'punk', 'hip hop''grunge' and 'hippy' in books since 1850. The lines are colour-coded when you see them online.

You can use this tool to see when a spelling might have changed and plot this against the context of language change (e.g. was it before or after Johnson's 1755 dictionary which helped to 'fix' English spelling?)

http://books.google.com/ngrams
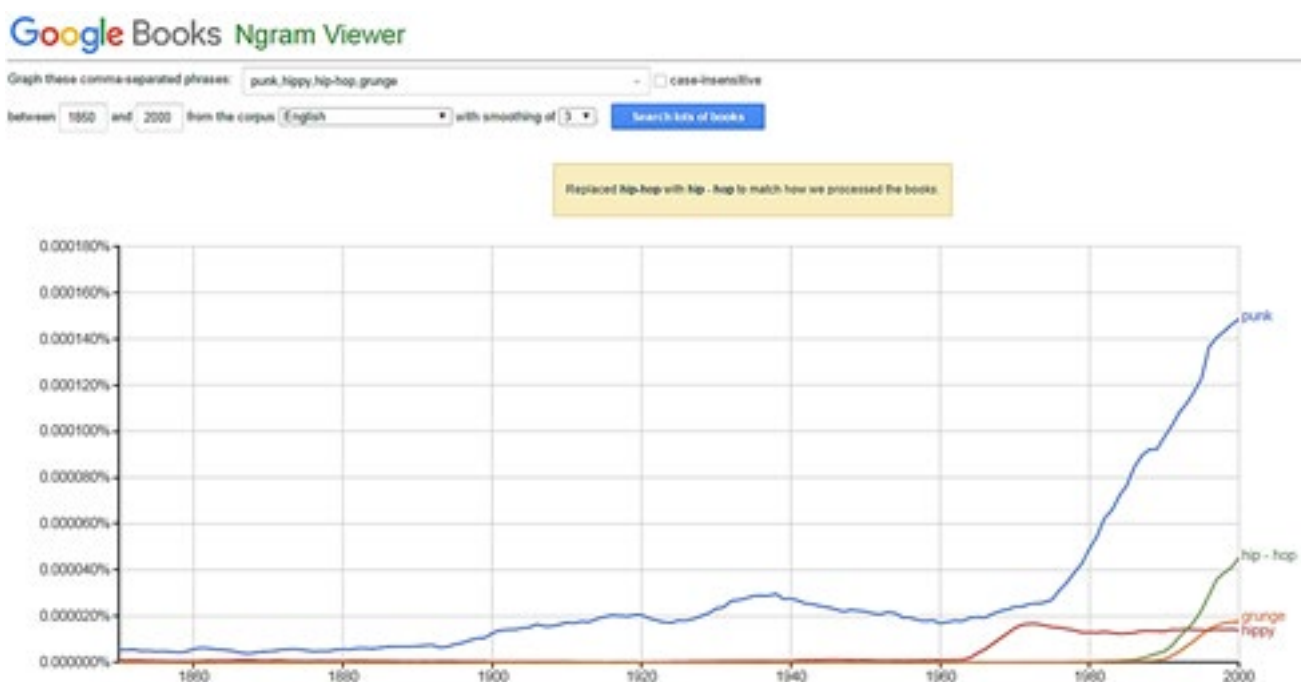
Readability

Readability tests are formulae for evaluating the readability of text.   Each readability test bases its rating on the average number of syllables per word and words per sentence. These can be useful tools for initially evaluating the complexity of a written text. The most common of these is the Flesh-Kincaid readability test.

This is how it works:

Flesch Reading Ease test

This test rates text on a 100-point scale. The higher the score, the easier it is to understand the document. For most standard files, you want the score to be between 60 and 70.

The formula for the Flesch Reading Ease score is:

206.835 – (1.015 x ASL) – (84.6 x ASW)

where:

ASL = average sentence length (the number of words divided by the number of sentences)

ASW = average number of syllables per word (the number of syllables divided by the number of words)

Although a bit technical, knowing how this is calculated gives you a way in to picking out specific examples of polysyllabic words and complex sentences to analyse in further depth after you have calculated the general readability.

Flesch-Kincaid Grade Level test

This test rates text on a U.S. school grade level. For example, a score of 8.0 means that an eighth grader can understand the document. For most documents, aim for a score of approximately 7.0 to 8.0.

The formula for the Flesch-Kincaid Grade Level score is:

(.39 x ASL) + (11.8 x ASW) – 15.59

where:

ASL = average sentence length (the number of words divided by the number of sentences)

ASW = average number of syllables per word (the number of syllables divided by the number of words)

But, thankfully, there are online tools to do the calculations for you:

https://readability-score.com/text/

Copy and paste the text you want to analyse and the tool will display the Readability Formulae and corresponding grades/score in a table.

Others

This final site tests the readability of any webpage and there are plenty more available on line or as apps for your phone or tablet these days: http://www.read-able.com/ and this one does lots of things at once: http://textalyser.net/.

A final reminder…

None of these tools is worth using on its own. A pretty graph or a numerical statistic means nothing by itself.

Instead, you always need to go beyond the statistic or the graph, to 'dig deeper' into the text itself and investigate further particular instances of language.

Graphs, charts and graphics can provide factual overviews (a sort of 'establishing shot', to use film jargon) but you should always then 'zoom in' to textual example and in-depth close analysis of language and meaning.

So…always follow quantitative statistics by quoting language examples from the text and qualitatively analysing their effect, connotations and significance in context.

Looking for a resource?

There is now a quick and easy search tool to help find **free** resources for your qualification:

www.ocr.org.uk/i-want-to/find-resources/

www.ocr.org.uk/alevelreform

OCR Customer Contact Centre

General qualifications
Telephone 01223 553998
Facsimile   01223 552627
Email general.qualifications@ocr.org.uk

OCR is part of Cambridge Assessment, a department of the University of Cambridge. *For staff training purposes and as part of our quality assurance programme your call may be recorded or monitored.*