

Normalisation in Discourse Analysis

Gabriella Rundblad
King's College London

Introduction

There are many different ways of analysing discourse (whether written or spoken) and there are also a fair amount of good books on Discourse Analysis. This paper is not intended to introduce or debate the qualitative aspects of Discourse Analysis, but will instead focus on how discourses and discourse features can be quantified and compared using the nowadays standard technique of normalisation.

Selecting Discourses to Compare

The first task is to choose the discourses to be compared. There is an abundance of areas to explore and which discourses are chosen naturally depend on the study in question. Regardless of which type of discourse is chosen, the two (or more) discourses need to be comparable on some levels. In short, it might be too complex, difficult and time consuming to compare discourses that are miles and miles apart; if they are so different on so many levels, how can we determine the reasons why they are different?! For example, it is not such a good idea to compare a short research article on physics written by a female last year with a long novel (i.e. fiction) written by a man 100 years ago. More suitable comparisons include articles on the same topic or from the same time period but in different newspapers, magazines or journals, chapters from the beginning, the middle and the end of the same novel, extracts from different novels, essays written by people of different backgrounds (e.g. men vs women), or, in the case of spoken discourse, excerpts from monologues (e.g. oral presentations in class) or dialogues.

Qualitative Analysis

The next step is to analyse the chosen discourses qualitatively. As mentioned above, this paper will not focus on qualitative discourse analysis, so instead we will move on to the next task.

Qualitative or Quantitative?

Not all types of discourse analysis lend themselves to quantification and not all linguists mix qualitative and quantitative methods. However, anyone intending to do "qualitative only" discourse analysis should be very wary of carrying out and including quantitative statements when they write up their results. Whereas it is perfectly possible and correct to state "text A contained 10 instances" and "text B used 5 instances", the reader might assume that there is a difference between texts A and B because of these numbers even

though the writer (hopefully!) never intended such a comparative statement. Comparisons such as “text A used more instances than text B” can, on the other hand, be unwarranted and even completely untrue; comparisons such as “more”, “majority”, and so on, are quantitative - not qualitative - and in our example (i.e. 10 vs 5 instances), there is in fact no statistical difference.

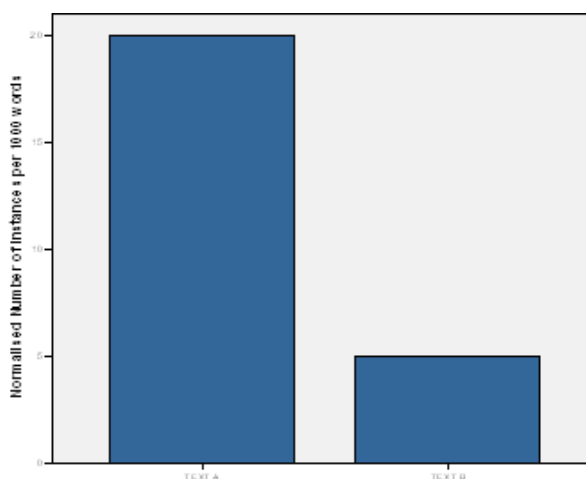
Normalising Discourse

The fact that there is no significant difference between 10 and 5 instances is not the only problem we are facing here; the first question we need to ask before we begin our quantitative comparison is “can we compare these two discourses?” If we look in detail at our two texts, we might find that text A is 500 words long but text B is 1000 words long. If text A is twice as long and has twice as many instances of a particular feature, then that is very different from both texts being equally long.

The solution to our predicament is to normalise the two discourses. A favoured way of normalising is to recalculate the number of instances for both discourses as if the texts were 1000 words; of course, we could choose to recalculate them to 100 words, 500 words, 1000 words, etc. but 1000 words is the common number. First we need to find out how long each of the discourses really is. Next we need to divide our target length (i.e. 1000) with the actual length. This way we get 2 for text A (i.e. 1000 divided by 500) and 1 for text B (i.e. 1000 divided by 1000). We can refer to 2 and 1 as the two texts’ “1000 words factors”. Finally, we use the factors to recalculate the number of instances of a feature for both texts. The normalised number of instances for text A is 20 (i.e. 10 actual instances multiplied with the factor 2) and for text B we get 5 (i.e. 5 actual instances multiplied with the factor 1).

Describing the Data

After we have normalised the data, we can proceed to describe the data and our results to the readers - perhaps through a graph. This is an essential step, and it is also the topic of another paper of mine.



Chi Square

After going through our normalised data, we might spot a few cases where a graph, a table or the mere numbers themselves seem to suggest a difference worth investigating further. Our next step is to do an inferential statistics test - here I would recommend Chi Square. We can use a statistics program such as SPSS, EXCEL or even the good old fashion pen, paper and calculator technique to this end.

If we use SPSS (see Appendix A on how to do this) to see if we indeed have a statistically significant difference between texts A (20 normalised instances) and B (5 normalised instances), we get the following result:

	TEXTs
Chi-Square ^a	9.000
df	1
Asymp. Sig.	.003

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 12.5.

If we had calculated Chi Square by hand we would have to look up the Chi Square value (here: 9.000) in a Chi Square table in order to find the p-value, but here SPSS has calculated it for us: 0.003. This means that we have a highly significant difference between the two texts. I would report this result in the following way:

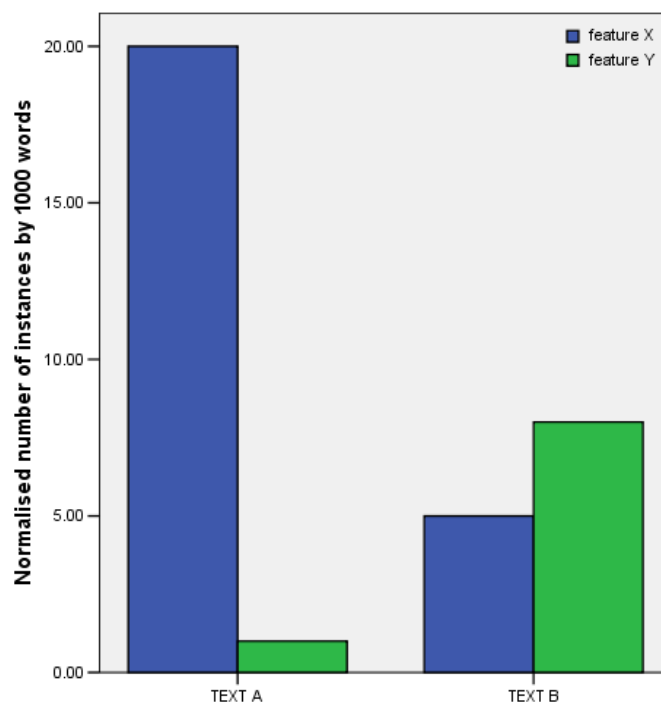
There is a highly significant difference between text A and text B ($\chi^2_{(1)} = 9.0$, $p = 0.003$).

Please note that the degree of freedom (df) is reported as well (i.e. the number 1 within the little parenthesis); the degree of freedom confirms that we are comparing two texts (i.e. the degree of freedom is the number of samples (here: texts) minus 1).

Parametric Correlation (Pearson)

There are more patterns than “more in A” versus “less in B” (or the other way around) that we might find in discourse. A linguistic features, literary techniques and conventions might occur more or less often in a text because a second feature occurs more or less often in that text – correlation. There are two types of correlation that we might find: positive (i.e. if feature X occurs a lot then feature Y does too – if feature X occurs hardly at all then feature Y does too) and negative (i.e. if feature X occurs a lot then feature Y will not – if feature X occurs hardly at all then feature Y will be used a lot).

After we have normalised our data and described it using graphs and tables, we can combine two linguistic features in a graph, like here:



This graph seems to suggest that we have a negative correlation between the two linguistic features – text A has lots of normalised instances of X but very few of Y (only 1 instance), whereas the pattern is the opposite for text B (8 instances of Y). This graph suggests that this discrepancy might be worthy of further investigation.

The inferential statistics test we need to do is a correlation test. There are two tests to choose from: Pearson and Spearman. The difference between the two is that Pearson is parametric (i.e. when comparing scalar data such as number of metres, number of minutes, number of mistakes, and number of instances), whereas Spearman is non-parametric (e.g. the difference between choosing a high versus a low grade on a questionnaire question – please see my paper on Analysing Non-Parametric Data). Here I will use Pearson since our data is scalar.

Once again I will use SPSS to see if we have a negative correlation between the two linguistic features X and Y in the texts A and B. Because we are assuming (i.e. hypothesising) a particular outcome, I will do a one-tailed test (see Appendix B on tails and how to do this). This is the result:

Correlations

		VAR00001	VAR00002
VAR00001	Pearson Correlation	1	-1.000**
	Sig. (1-tailed)		.
	N	2	2
VAR00002	Pearson Correlation	-1.000**	1
	Sig. (1-tailed)	.	
	N	2	2

** . Correlation is significant at the 0.01 level (1-tailed).

The first things we notice are the asterisks. There are two. This means two things: first, that we have found a correlation, and second that this correlation is highly significant. Had there only been one asterisk, the correlation would have been significant (but without “highly”) – also a good result. If there had not been any asterisks at all, we would not have found any correlation at all.

In order to tell whether the correlation we found was negative or positive, we look at the Pearson Correlation value in one of the cells with the asterisks in. Here the Pearson Correlation value is -1.000. The minus sign tells us that the correlation is negative – had there been no minus sign, we would have found a positive correlation.

I would report this result in the following way:

There is a highly significant negative correlation between the two features in the two texts ($r_{(2)} = -1.000$, $p < 0.01$ (one-tailed)).

The value within the little parenthesis is N which stands for the number of features compared (here we compared the features X and Y, i.e. 2). The p-value we give is also taken from the table – “significant at the 0.01 level” means that the p-value is less than 0.01.

If our correlation had been significant but not highly so, the table would have stated “significant at the 0.05 level” and consequently our report would have stated “ $p < 0.05$ ”. Similarly, if we had done a two-tailed test, the table and the report would have “two-tailed” in it.

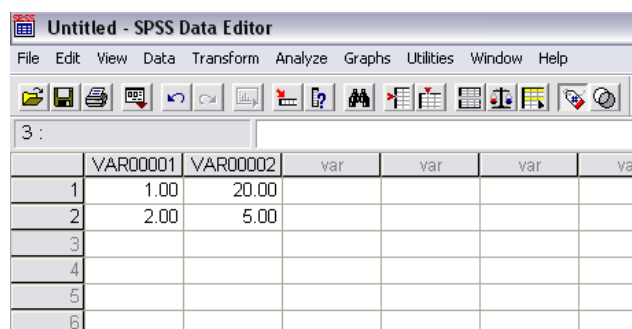
Summary

Only by normalising the discourses we are comparing, can we safely and accurately describe and inferentially prove or disprove differences between them and patterns of usage in them.

APPENDIX A

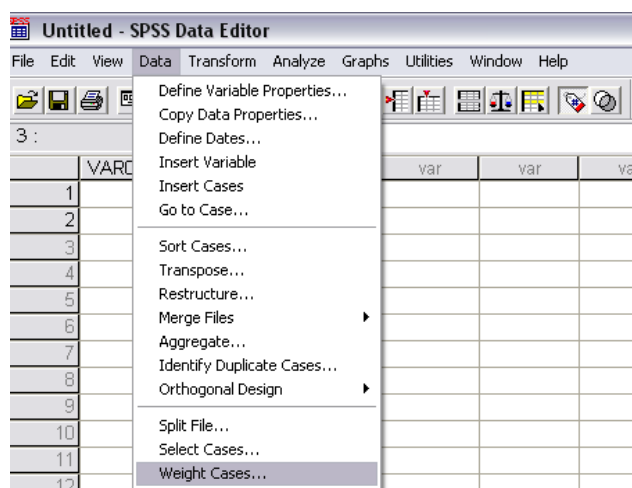
Tutorial: Chi Square using SPSS

1) Enter 1 (means text A) and 2 (means text B) in the first column; then enter the normalised number of instances for these texts in the second column.

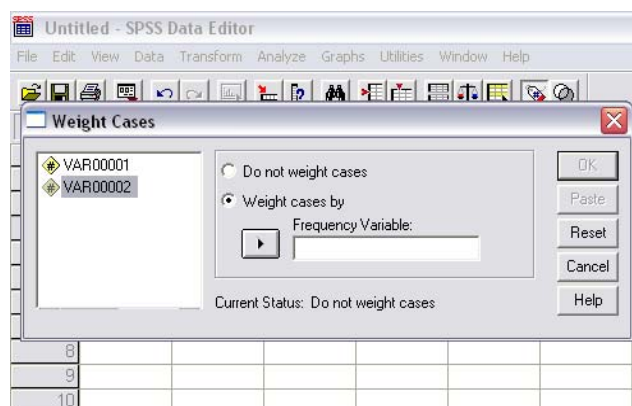


	VAR00001	VAR00002	var	var	var	va
1	1.00	20.00				
2	2.00	5.00				
3						
4						
5						
6						

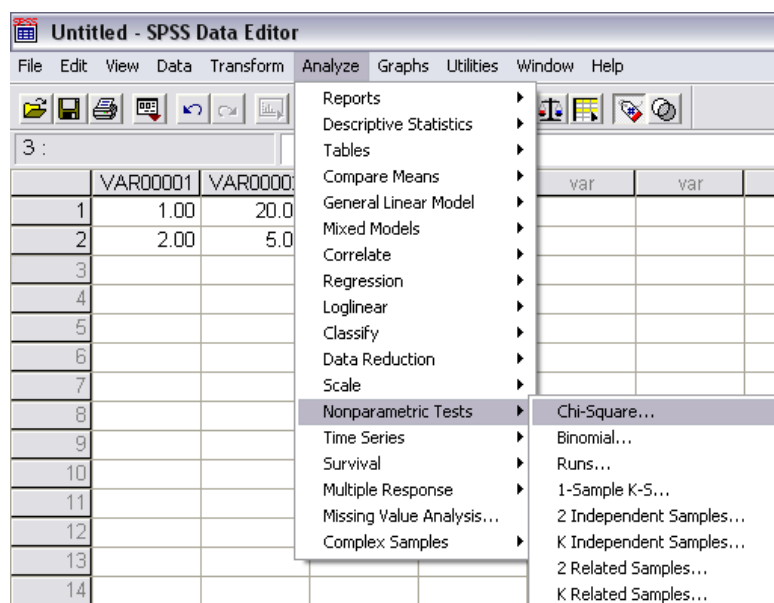
2) Click DATA, select WEIGHT CASES.



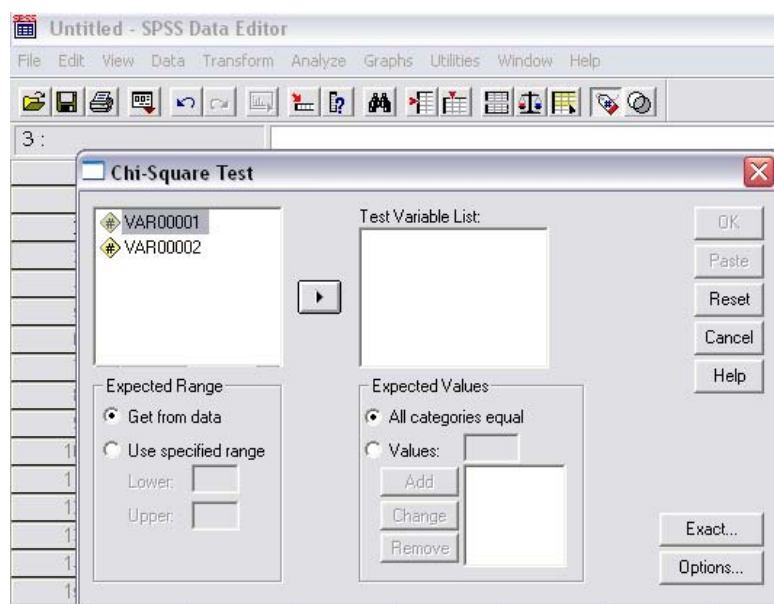
3) Tick the box that says “Weight cases by”, highlight the second column’s name (here: VAR00002), click on the arrow button. This moves VAR00002 into the little text box. Click OK.



4) Click ANALYSE, select NONPARAMETRIC TESTS, select CHI-SQUARE.



5) Highlight the second column's name (here: VAR00001), click on the arrow button. This moves VAR00001 into the little text box on the right. Click OK.

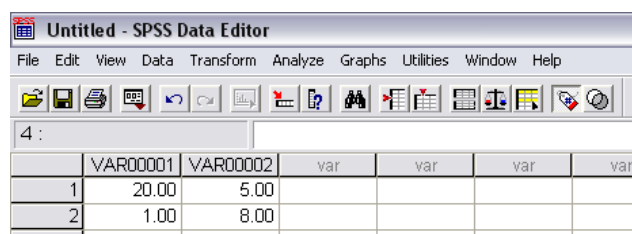


6) SPSS will now create an output window, in which the table we saw above will appear.

APPENDIX B

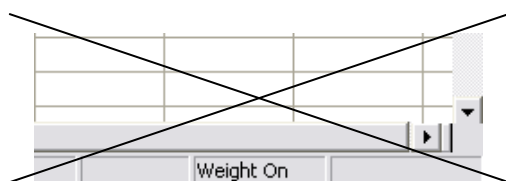
Tutorial: Parametric Correlation (Pearson) using SPSS

1) Enter the normalised instances for text A in the first column – first row for feature X and second row for feature Y; then enter the normalised instances for text B in the second column – first row for feature X and second row for feature Y.

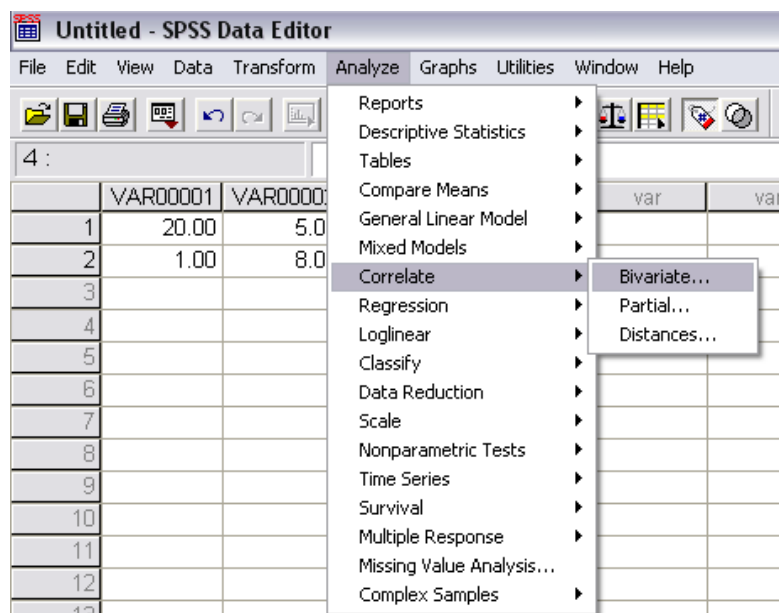


	VAR00001	VAR00002	var	var	var	var
1	20.00	5.00				
2	1.00	8.00				

2) We do NOT weigh cases for correlation tests! Make sure that the bottom right hand corner of the screen does NOT say “Weight On” – if it does click DATA, select WEIGHT CASES, tick the box that says “Do not weight cases”, click OK.



3) Click ANALYSE, then CORRELATE, the BIVARIATE.



4)

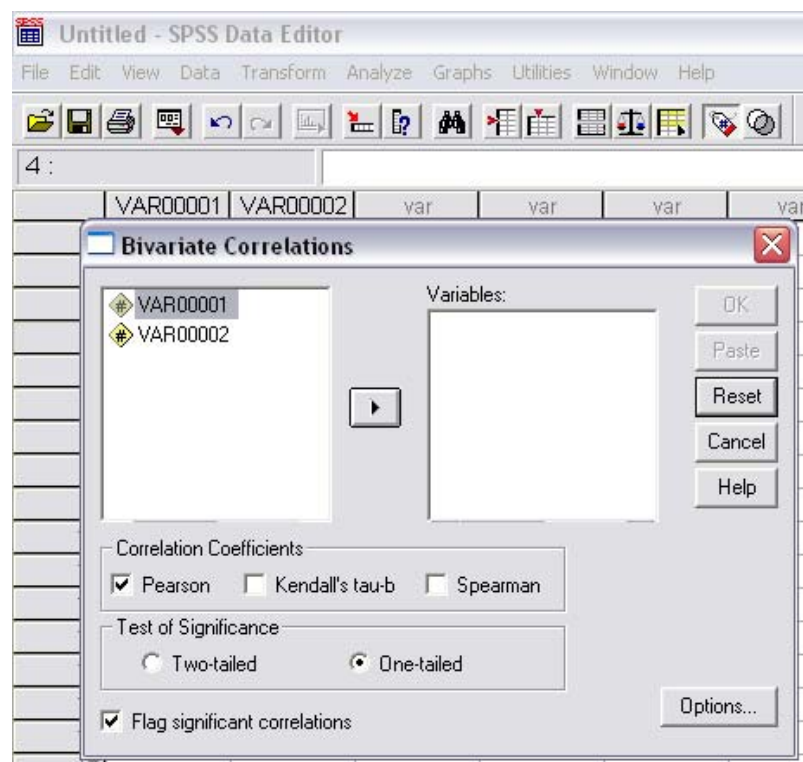
a) In turn, highlight the name of each column (first VAR00001), click on the arrow button – repeat for VAR00002. This moves VAR00001 and VAR00002 into the little text box on the right.

b) In this particular example, we hypothesised a negative correlation – therefore we need to tick the little box for one-tailed.

If we had not known whether to expect a negative or a positive correlation – or if we were not sure to expect any correlation at all but just wanted to see if there is one – we would need to take the safer option and choose a two-tailed test. Consequently, we would have checked the little box for two-tailed instead. But that is not the case in this example. Hence one-tailed it is.

c) We must also make sure that we are doing a Pearson rather than a Spearman test – make sure the correct box is ticked.

d) Click OK.



5) SPSS will now create an output window, in which the table we discussed above will appear.