Chapter 5: Exercises

1) Manually calculate the Pearson's and Spearman's correlations between verbs and adjectives in ten randomly selected texts from BE06. The data is provided below:

Pearson's correlation is calculated as follows:

											Mean	SD
Verbs	169.9	135.0	161.7	183.0	163.1	190.8	140.7	213.9	218.0	165.2	174.1	27.7
Adjectives	96.0	102.6	91.9	76.5	98.8	77.6	68.4	60.3	74.4	76.5	82.3	14.1
Verb - distance	-4.2	-39.1	-12.4	8.9	-11.0	16.7	-33.4	39.8	43.9	-8.9	-	-
Adj. - distance	13.7	20.3	9.6	-5.8	16.5	-4.7	-13.9	-22.0	-7.9	-5.8	-	-

$$\text{covariance} = \frac{\text{sum of multiplied distances from mean}_1 \text{ and mean}_2}{\text{total no. of cases} - 1}$$

$$= \frac{(-4.2 \times 13.7) + (-39.1 \times 20.3) + (-12.4 \times 9.6) + (8.9 \times -5.8) + (-11 \times 16.5) + (16.7 \times -4.7) + (-33.4 \times -13.9) + (39.8 \times -22) + (43.9 \times -7.9) + (-8.9 \times -5.8)}{10 - 1}$$

$$= \frac{-1988.45}{9} = -220.94$$

$$r = \frac{\text{covariance}}{\text{SD}_1 \times \text{SD}_2} = \frac{-220.94}{27.7 \times 14.1} = -0.57$$

There is a negative Pearson's correlation of -0.57 between verbs and adjectives. This can be interpreted as a large effect according to the standard interpretation (Cohen 1988: 79-80).

Spearman's correlation is calculated as follows:

Verbs	169.9	135.0	161.7	183.0	163.1	190.8	140.7	213.9	218.0	165.2
Adjectives	96.0	102.6	91.9	76.5	98.8	77.6	68.4	60.3	74.4	76.5
Verbs-rank	5	10	8	4	7	3	9	2	1	6
Adj.-rank	3	1	4	6.5	2	5	9	10	8	6.5
Rank difference	2	9	4	-2.5	5	-2	0	-8	-7	-0.5
Squared rank difference	4	81	16	6.3	25	4	0	64	49	0.3

$$r_s = 1 - \frac{6 \times \text{sum of squared rank differences}}{\text{number of cases} \times (\text{number of cases squared} - 1)} = 1 - \frac{6 \times 4+81+16+6.3+25+4+0+64+49+0.3}{10 \times 100 - 1} = -0.51$$

There is a negative Spearman's correlation of -0.51 between verbs and adjectives. This can be interpreted as a large effect according to the standard interpretation (Cohen 1988: 79-80). Both Pearson's and Spearman's

(rank) correlation show the same type of relationship: the more verbs there are the fewer adjectives and vice versa.

2) What can you tell about the relationship between the variables in the four graphs below?

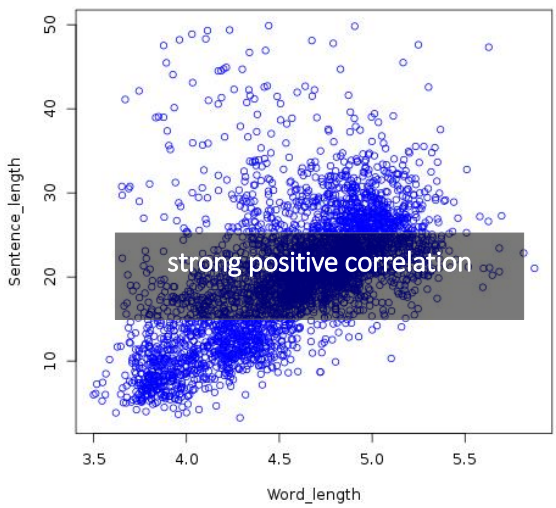


Figure 5.27 Relationship between mean word length (no. of characters) and mean sentence length (no. of words) in BNC

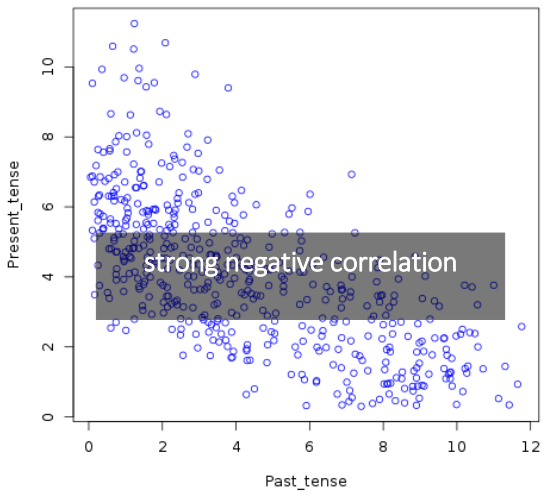


Figure 5.28 Relationship between the use of the past and the present tense in BE06

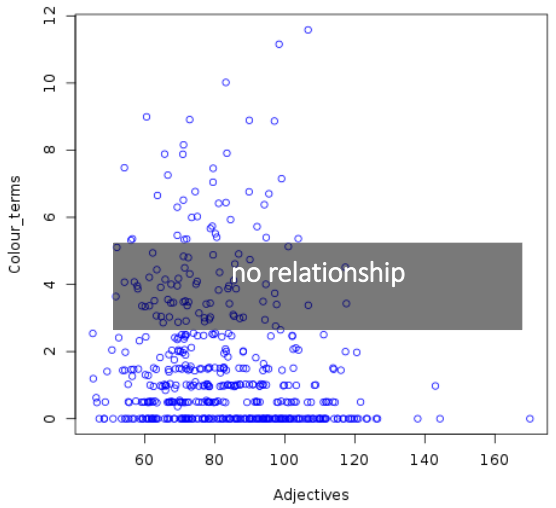


Figure 5.29 Relationship between the use of adjectives and colour terms in BE06

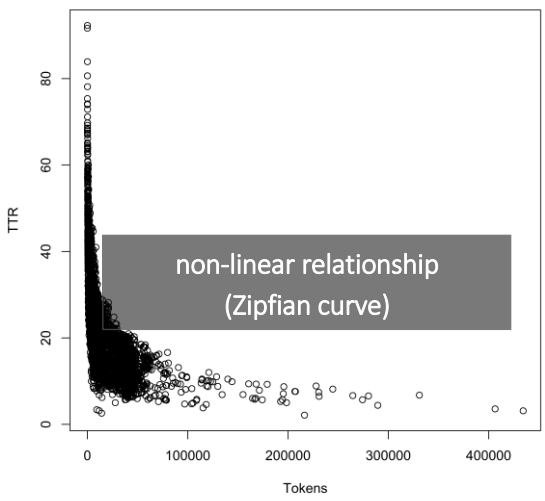


Figure 5.30 Relationship between text length (tokens) and type-token ratio (TTR) in BNC

3) Each Brown family corpus is divided into 15 different types of texts listed below (see also section 1.4).

A (Press: reportage), B (Press: editorial), C (Press: reviews), D (Religion), E (Skills, trades and hobbies), F (Popular lore), G (Belles lettres, biography, essays), H (Miscellaneous government documents, foundation reports, industry reports, college catalogue, industry house organ), J (Learned and scientific writings), K (General fiction), L (Mystery and detective fiction), M (Science fiction), N (Adventure and western fiction), P (Romance and love story), R (Humour).

Possible categoridation:

The 4 broad genres/registers	**Includes categories…**
Press	A, B & C
General prose (non-fiction)	D, E, F, G & H
Learned (academic)	J
Fiction	K, L, M, N, P & R

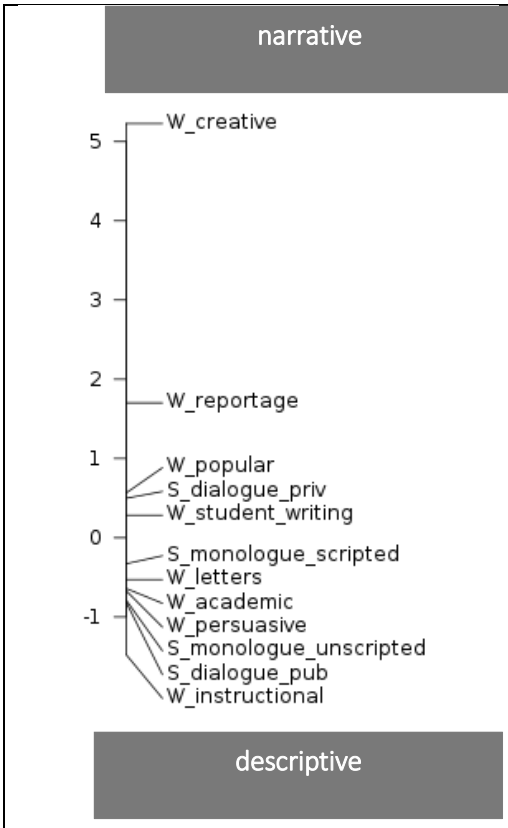
This classification is very useful; however, for some purposes it might be too detailed. Group the individual text types into larger categories based on their functional similarity. Then design a study in which you could verify your grouping.

Design of the study: 1) select a range of lexico-grammatical features that can be searched in a Brown-family corpus automatically (refer to Biber's 67 items); 2) Use #LancsBox Whelk tool to search these features and save the results to a spreadsheet; 3) Run the cluster analysis to see the grouping of the 15 text type categories and how well these match the 4 broad genres/registers from the table above.

4) The following are factor loadings of Factors 3 and 4 based on the Multidimensional analysis of New Zealand English from section 5.5. The dimension plots are also provided. Interpret each factor functionally as a dimension. Create labels for these dimensions.

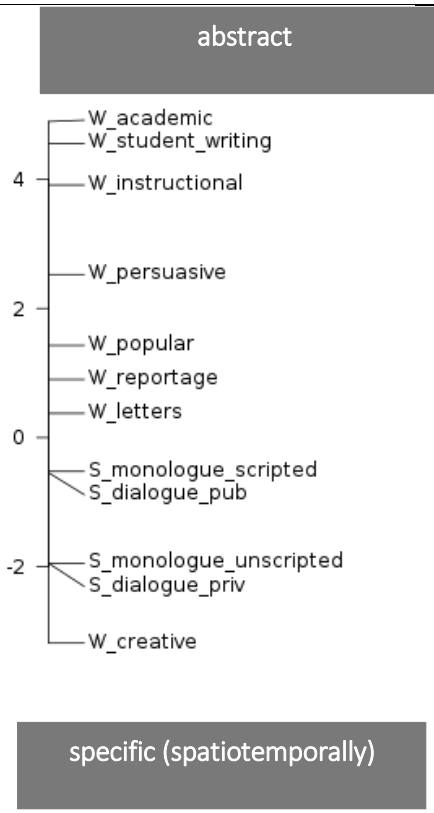
Table 5.7 Results of Factor analysis of NZ English: Factor loadings of Factors 3 and 4

Features	Factor 3 loadings	Features	Factor 4 loadings
past tense (1)	1.099	nominalizations (14)	0.618
third-person personal pronouns (8)	0.461	conjuncts (45)	0.499
attributive adjectives (40)	-0.304	agentless passives (17)	0.36
present tense (3)	-0.583	by-passives (18)	0.347
		time adverbials (5)	-0.444
		place adverbials (4)	-0.504



F (11, 488) = 25.25; *p* <.001, r^2 = 36.27%

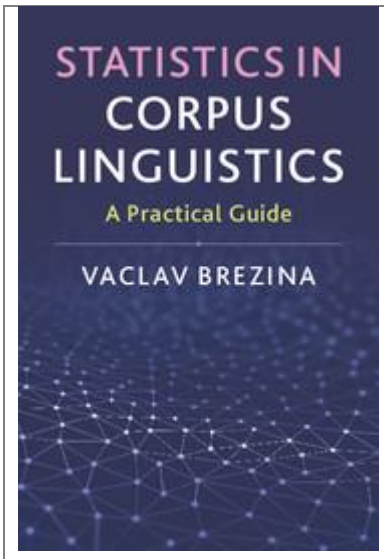
Figure 5.31 Dimension 3



F (11, 488) = 44.69; p <.001, r^2 = 50.18%

Figure 5.32 Dimension 4

5) -



Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Do you use language corpora in your research or study, but find that you struggle with statistics? This practical introduction will equip you to understand the key principles of statistical thinking and apply these concepts to your own research, without the need for prior statistical knowledge. The book gives step-by-step guidance through the process of statistical analysis and provides multiple examples of how statistical techniques can be used to analyse and visualise linguistic data. It also includes a useful selection of discussion questions and exercises which you can use to check your understanding.

The book comes with a Companion website, which provides additional materials (answers to exercises, datasets, advanced materials, teaching slides etc.) and Lancaster Stats Tools online, a free click-and-analyse statistical tool for easy calculation of the statistical measures discussed in the book.