Chapter 4: Exercises – answers

1)

TOPIC: Research question	Individual feature research design?	Lexico-grammatical frame	Reference (to find out more about the research)
DATIVE ALTERNATION	YES	All dative constructions.	Theijssen (2010)
A/AN ALTERNATION	YES	All countable nouns in singular beginning with a vowel.	Gabrielatos et al. (2010)
SWEARWORDS	NO	-	-
GENITIVE ALTERNATION	YES	All contexts where *s*- and *of*-genitive construction can in principle be used.	Leech et al. (1994) Szmrecsanyi (2010)
EPISTEMIC MARKERS	NO	-	-
ATTENDED/UNATTENDED *THIS*	YES	All cases of anaphoric *this*.	Wulff et al. (2012)

2)

Variety	Modal			Total
Var	*must*	*have to*	*need to*	
American	352 (38.8%)	355 (39.1%)	201 (22.1%)	908
British	448 (43.0%)	405 (38.8%)	190 (18.2%)	1043
Total	800	760	391	1951

Note that percentages are based on row totals.

Pearson's Chi-squared test: 5.81 (df = 2), p = 0.0548757, Cramer's V = 0.055 (95% CI: 0, 0.095), NEGLIGIBLE EFFECT

We can conclude that there is not enough evidence in the data to reject the null hypothesis (p>0.05), which says that there is no difference between American and British use of *must, have to* and *need to*. In other words, the difference is not statistically significant.

3) Overall in the dataset, we have most cases of the modal expressions in general prose and fiction – this is a reflection of the size of the subcorpora and the frequency of the use of these modal expressions in the individual genres. The largest proportion of *must* compared to *have to* and *need to* is in general prose followed by newspapers. On the other hand, the smallest proportion of *must* and the largest proportion of *have to* is in fiction. The largest proportion of *need to* is in academic writing.

4)

Overall, using AIC we can say that out of the four options the most successful model for the dataset is Model 4 with the lowest AIC value of 2515.61. The model that we can discard straightaway is Model 1, because it does not represent a significant improvement (overall p> 0.05) from the baseline (null) model; all the other three models are statistically significant. Note however, that their classification potential is fairly low (C-index below 0.7). The first two models (Model 1 and Model 2) include only main effects, while the other two models (Model3 and Model 4) include also some predictor interactions. The most successful model (Model 4) includes 'Variety', 'Genre' and 'Subject' as main effects as well as interactions between 'Variety' and 'Genre'. While in Model 2 only two estimates are statistically significant (p<0.05), in Models 3 and 4 all estimates are statistically significant. The odds ratio (including 95% CI) of each estimate needs to be interpreted against the baseline values. For instance, looking at Model 2, the odds of *must* being used in fiction are 0.419 times (95% CI: 0.303, 0.579) the odds of *must* appearing in academic writing. The models with interactions (Model3 and Model 4) are somewhat more complex to interpret. To get the relevant odds value, we need to take the odds ratios of the interaction terms and the main effects and multiply them. Alternatively, we can take log odds ratios and add these values. Let's take Model 4 as an example: if we are interested in the British variety, and the odds ratio of must in fiction as opposed to academic writing (the baseline), we take the odds ratio value 0.206 for fiction main effect and multiply it by 4.221 (interaction term: 'VarietyB_BR:GenreB_Fiction'). The result is 0.87, which is the odds ratio of must in British fiction as opposed to British academic writing. The same result would be achieved by adding 1.581 and 1.440 and then using the exponential function (e(-1.581 +1.440)). For more details see Osborne (2013: 258 – 267).

Finally, it would be advisable to collect more explanatory variables to improve the classification potential of the models (C-index) and run the analysis again.