Chapter 4: Exercises

1) Look at the topics and research questions in Table 4.14. Decide if the linguistic feature research design is appropriate (YES/NO). If yes, define the appropriate lexico-grammatical frame. The first row was completed for you as an example.

Table 4.14 Appropriate research design

TOPIC: Research question	Linguistic features (examples)	Linguistic feature research design?	Lexico-grammatical frame
DATIVE ALTERNATION: What factors have an effect on dative alternation in English?	*She handed the student the book.* *She handed the book to the student.*	YES	All dative constructions.
A/AN ALTERNATION: When is a non-standard version of the indefinite article (*a* before a word beginning with a vowel) used in spoken English?	*an apple, a apple*		
SWEARWORDS: Do speakers use more strong or weak swearwords?	*fuck, cunt, motherfucker* etc. vs. *damn, crap, hell etc.*		
GENITIVE ALTERNATION: What factors influence the choice between *s*- and *of*-genitive?	president's speech, *the speech of the president*		
EPISTEMIC MARKERS: Does corpus data support the hypothesis that 'we only say we are certain when we are not' (Halliday)?	*This is certainly the case.* *This is maybe the case.*		
ATTENDED/UNATTENDED *THIS:* What factors influence the presence of a noun after *this*?	*This is an example.* *This sentence is an example.*		

2) Analyse the following cross-tab table: Add row and column totals as well as the grand total. Is there a difference in the use of *must*, *have to* and *need to* between British and American English? Add the percentages that will help you answer this question and calculate the chi-squared test (with raw frequencies).

Variety	Modal			Total
	must	*have to*	*need to*	
American	352	355	201	
British	448	405	190	
Total				

3) Interpret the following mosaic plot. It displays the use of three modal expressions of strong obligation (*must*, *have to* and *need to*) in BE06, a one-million-word corpus of written British English. The genres displayed are academic writing ('Acad'), fiction ('Fiction'), general prose ('General') and newspapers ('Press').

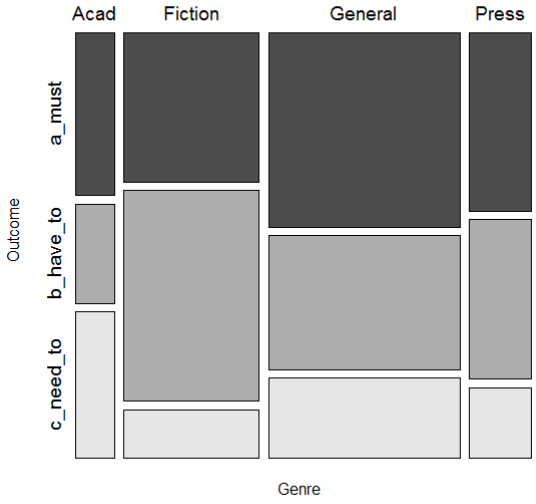


Figure 4.8 *Must, have to* and *need to* in British English (BE06)

4) Interpret the following models, outcomes of logistic regression, that come from a study of modal expressions of strong obligation. The study was guided by the following research question: In what contexts do speakers use *must* as opposed to semi-modals of strong obligation (*have to* and *need to*)?

The following variables were coded in the dataset:

- Outcome variable: *must* vs. *have to* and *need to* combined (baseline).
- Predictor 1 (Variety): British vs. American (baseline).
- Predictor 2 (Genre): fiction vs. general prose vs. press vs. academic writing (baseline).
- Predictor 3 (Subject): *I* vs. *you* vs. other subject (baseline).

Model1: main effect 'Variety'

Overall model statistics: Likelihood ratio test (LL): 3.52 (p = 0.061) -> NOT SIGNIFICANT; C-index: 0.52 -> NOT ACCEPTABLE; Nagelkerke R^2: 0; AIC: 2641.65

Coefficients:

	Estimate (log odds)	Standard Error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	-0.457	0.068	-6.711	0.000	0.633	0.554	0.723
VarietyB_BR	0.173	0.092	1.875	0.061	1.189	0.992	1.426

Model2: main effects 'Variety', 'Genre'

Overall model statistics: Likelihood ratio test (LL): 54.49 (p < 0.001) -> SIGNIFICANT; C-index: 0.6 -> NOT ACCEPTABLE; Nagelkerke R^2: 0.04; AIC: 2596.67

Coefficients:

	Estimate (log odds)	Standard Error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	-0.027	0.147	-0.184	0.854	0.973	0.729	1.300
VarietyB_BR	0.163	0.094	1.738	0.082	1.177	0.980	1.416
GenreB_Fiction	-0.870	0.165	-5.277	0.000	0.419	0.303	0.579
GenreC_General	-0.172	0.157	-1.092	0.275	0.842	0.618	1.146
GenreD_Press	-0.529	0.185	-2.860	0.004	0.589	0.410	0.846

Model3: main effects 'Variety', 'Genre' plus interactions

Overall model statistics: Likelihood ratio test (LL): 75.54 ($p < 0.001$) -> SIGNIFICANT; C-index: 0.61 -> NOT ACCEPTABLE; Nagelkerke R^2: 0.05; AIC: 2581.63

Coefficients:

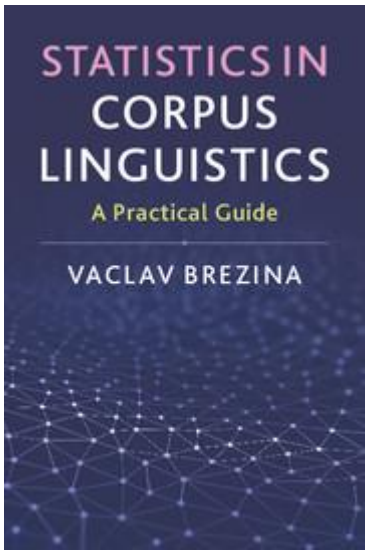
	Estimate (log odds)	Standard Error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	-0.027	0.147	-0.184	0.854	0.973	0.729	1.300
VarietyB_BR	-0.857	0.288	-2.979	0.003	0.424	0.240	0.742
GenreB_Fiction	-1.581	0.238	-6.646	0.000	0.206	0.128	0.326
GenreC_General	-0.578	0.225	-2.573	0.010	0.561	0.359	0.868
GenreD_Press	-1.166	0.266	-4.387	0.000	0.312	0.184	0.522
VarietyB_BR:GenreB_Fiction	1.440	0.337	4.275	0.000	4.221	2.190	8.214
VarietyB_BR:GenreC_General	0.893	0.321	2.785	0.005	2.443	1.308	4.605
VarietyB_BR:GenreD_Press	1.319	0.376	3.506	0.000	3.738	1.796	7.852

Model4: main effects 'Variety', 'Genre', 'Subject' plus 'Variety', 'Genre' interactions

Overall model statistics: Likelihood ratio test (LL): 145.56 ($p < 0.0001$) -> SIGNIFICANT; C-index: 0.66 -> NOT ACCEPTABLE; Nagelkerke R^2: 0.1; AIC: 2515.61

Coefficients:

	Estimate (log odds)	Standard Error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	0.542	0.200	2.714	0.007	1.720	1.168	2.561
VarietyB_BR	-0.930	0.290	-3.210	0.001	0.395	0.222	0.693
GenreB_Fiction	-1.318	0.242	-5.441	0.000	0.268	0.165	0.428
GenreC_General	-0.450	0.228	-1.974	0.048	0.638	0.406	0.993
GenreD_Press	-1.191	0.268	-4.445	0.000	0.304	0.179	0.511
SubjectB_I	-1.084	0.174	-6.232	0.000	0.338	0.239	0.472
SubjectC_you	-0.917	0.158	-5.794	0.000	0.400	0.291	0.542
VarietyB_BR:GenreB_Fiction	1.482	0.340	4.353	0.000	4.400	2.267	8.620
VarietyB_BR:GenreC_General	0.952	0.324	2.941	0.003	2.592	1.379	4.915
VarietyB_BR:GenreD_Press	1.490	0.379	3.927	0.000	4.438	2.118	9.384



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.