

Guillaume Desagulier

Corpus Linguistics and Statistics with R

Introduction to Quantitative Methods in Linguistics

 Springer

Contents

1	Introduction	1
1.1	From Introspective to Corpus-Informed Judgments	1
1.2	Looking for Corpus Linguistics	3
1.2.1	What Counts as a Corpus	3
1.2.2	What Linguists Do with the Corpus	6
1.2.3	How Central the Corpus Is to a Linguist’s Work	8
	References	10

Part I Methods in Corpus Linguistics

2	R Fundamentals	15
2.1	Introduction	15
2.2	Downloads and Installs	15
2.2.1	Downloading and Installing R	16
2.2.2	Downloading and Installing RStudio	16
2.2.3	Downloading the Book Materials	17
2.3	Setting the Working Directory	17
2.4	R Scripts	17
2.5	Packages	18
2.5.1	Downloading Packages	18
2.5.2	Loading Packages	19
2.6	Simple Commands	19
2.7	Variables and Assignment	20
2.8	Functions and Arguments	21
2.8.1	Ready-Made Functions	21
2.8.2	User-Defined Functions	22
2.9	R Objects	24
2.9.1	Vectors	24
2.9.2	Lists	33
2.9.3	Matrices	34
2.9.4	Data Frames (and Factors)	36
2.10	for Loops	41
2.11	if and if...else Statements	43

2.11.1	if Statements	43
2.11.2	if . . . else Statements	44
2.12	Cleanup	45
2.13	Common Mistakes and How to Avoid Them	46
2.14	Further Reading	47
	Exercises	47
	References	49
3	Digital Corpora	51
3.1	A Short Typology	51
3.2	Corpus Compilation: Kennedy's Five Steps	52
3.3	Unannotated Corpora	54
3.3.1	Collecting Textual Data	54
3.3.2	Character Encoding Issues	55
3.3.3	Creating an Unannotated Corpus	57
3.4	Annotated Corpora	58
3.4.1	Markup	58
3.4.2	POS-Tagging	58
3.4.3	POS-Tagging in R	59
3.4.4	Semantic Tagging	63
3.5	Obtaining Corpora	65
	Exercise	65
	References	66
4	Processing and Manipulating Character Strings	69
4.1	Introduction	69
4.2	Character Strings	69
4.2.1	Definition	70
4.2.2	Loading Several Text Files	70
4.3	First Forays into Character String Processing	71
4.3.1	Splitting	71
4.3.2	Matching	72
4.3.3	Replacing and Deleting	72
4.3.4	Limitations	73
4.4	Regular Expressions	73
4.4.1	Overview	73
4.4.2	Literals vs. Metacharacters	74
4.4.3	Line Anchors	74
4.4.4	Quantifiers	75
4.4.5	Alternations and Groupings	76
4.4.6	Character Classes	77
4.4.7	Lazy vs. Greedy Matching	79
4.4.8	Backreference	80
4.4.9	Exact Matching with <code>strapply()</code>	81
4.4.10	Lookaround	82
	Exercises	85

5	Applied Character String Processing	87
5.1	Introduction	87
5.2	Concordances	87
5.2.1	A Concordance Based on an Unannotated Corpus	87
5.2.2	A Concordance Based on an Annotated Corpus	95
5.3	Making a Data Frame from an Annotated Corpus	104
5.3.1	Planning the Data Frame	104
5.3.2	Compiling the Data Frame	104
5.3.3	The Full Script	106
5.4	Frequency Lists	108
5.4.1	A Frequency List of a Raw Text File	108
5.4.2	A Frequency List of an Annotated File	110
	Exercises	113
	References	114
6	Summary Graphics for Frequency Data	115
6.1	Introduction	115
6.2	Plots, Barplots, and Histograms	115
6.3	Word Clouds	118
6.4	Dispersion Plots	122
6.5	Strip Charts	125
6.6	Reshaping Tabulated Data	127
6.7	Motion Charts	132
	Exercises	133
	References	135
 Part II Statistics for Corpus Linguistics		
7	Descriptive Statistics	139
7.1	Variables	139
7.2	Central Tendency	140
7.2.1	The Mean	140
7.2.2	The Median	142
7.2.3	The Mode	143
7.3	Dispersion	145
7.3.1	Quantiles	145
7.3.2	Boxplots	146
7.3.3	Variance and Standard Deviation	147
	Exercises	148
8	Notions of Statistical Testing	151
8.1	Introduction	151
8.2	Probabilities	151
8.2.1	Definition	151
8.2.2	Simple Probabilities	152
8.2.3	Joint and Marginal Probabilities	153
8.2.4	Union vs. Intersection	155

8.2.5	Conditional Probabilities	155
8.2.6	Independence	156
8.3	Populations, Samples, and Individuals	157
8.4	Random Variables	158
8.5	Response/Dependent vs. Explanatory/Descriptive/Independent Variables	159
8.6	Hypotheses	160
8.7	Hypothesis Testing	162
8.8	Probability Distributions	163
8.8.1	Discrete Distributions	165
8.8.2	Continuous Distributions	169
8.9	The χ^2 Test	178
8.9.1	A Case Study: The Quotative System in British and Canadian Youth	178
8.10	The Fisher Exact Test of Independence	185
8.11	Correlation	186
8.11.1	Pearson's r	186
8.11.2	Kendall's τ	189
8.11.3	Spearman's ρ	192
8.11.4	Correlation Is Not Causation	193
	Exercises	193
	References	194
9	Association and Productivity	197
9.1	Introduction	197
9.2	Cooccurrence Phenomena	198
9.2.1	Collocation	198
9.2.2	Colligation	200
9.2.3	Collostruction	202
9.3	Association Measures	203
9.3.1	Measuring Significant Co-occurrences	203
9.3.2	The Logic of Association Measures	204
9.3.3	A Quick Inventory of Association Measures	205
9.3.4	A Loop for Association Measures	210
9.3.5	There Is No Perfect Association Measure	213
9.3.6	Collostructions	213
9.3.7	Asymmetric Association Measures	222
9.4	Lexical Richness and Productivity	226
9.4.1	Hapax-Based Measures	226
9.4.2	Types, Tokens, and Type-Token Ratio	227
9.4.3	Vocabulary Growth Curves	228
	Exercise	235
	References	235
10	Clustering Methods	239
10.1	Introduction	239
10.1.1	Multidimensional Data	239
10.1.2	Visualization	240

10.2	Principal Component Analysis	242
10.2.1	Principles of Principal Component Analysis	243
10.2.2	A Case Study: Characterizing Genres with Prosody in Spoken French	243
10.2.3	How PCA Works	245
10.3	An Alternative to PCA: t-SNE	252
10.4	Correspondence Analysis	257
10.4.1	Principles of Correspondence Analysis	257
10.4.2	Case Study: General Extenders in the Speech of English Teenagers	257
10.4.3	How CA Works	261
10.4.4	Supplementary Variables	266
10.5	Multiple Correspondence Analysis	268
10.5.1	Principles of Multiple Correspondence Analysis	269
10.5.2	Case Study: Predeterminer vs. Preadjectival Uses of <i>Quite</i> and <i>Rather</i>	270
10.5.3	Confidence Ellipses	275
10.5.4	Beyond MCA	276
10.6	Hierarchical Cluster Analysis	276
10.6.1	The Principles of Hierarchical Cluster Analysis	277
10.6.2	Case Study: Clustering English Intensifiers	278
10.6.3	Cluster Classes	279
10.6.4	Standardizing Variables	281
10.7	Networks	283
10.7.1	What Is a Graph?	283
10.7.2	The Linguistic Relevance of Graphs	285
	Exercises	290
	References	292
A	Appendix	295
A.1	Chapter 6	295
A.1.1	Dispersion Plots	295
A.2	Chapter 8	297
A.2.1	Contingency Table	297
A.2.2	Discrete Probability Distributions	298
A.2.3	A χ^2 Distribution Table	300
B	Bibliography	301
	Solutions	309
	Index	351