Table of contents

CHAPTER 1
Introduction 1

CHAPTER 2
A review: Corpus statistics, the 'usual' approaches 12

2.1 Frequency 13
2.1.1 How frequency is measured 13
2.1.2 What frequency is important for 19
2.2 Dispersion 23
2.2.1 How dispersion is measured 24
2.2.2 What dispersion is important for (corpus-linguistically) 32
2.2.3 What dispersion is important for (conceptually) 39
2.2.4 A (very brief!) first reaction to Nelson (2023) 43
2.3 Association 46
2.3.1 How association is measured 47
2.3.2 What association is important for 66
2.4 Keywords 68
2.4.1 How keyness is measured 69
2.4.2 What keyness is important for 78

CHAPTER 3
Unification of measures 80

3.1 Introduction: The Kullback-Leibler divergence 80
3.2 Characteristics of the *KLD* 87
3.2.1 Directionality 88
3.2.2 Independence of sample size 89
3.2.3 Normalization 90
3.2.4 Smoothing 92
3.3 Background and interpretation 99
3.4 Application to dispersion 104
3.4.1 A first simple example 104
3.4.2 A more realistic example 108
3.4.3 A (very brief!) second reaction to Nelson (2023) 113

3.5 Application to association 115
3.5.1 A first simple example 115
3.5.1.1 The direction row-to-column 116
3.5.1.2 The direction column-to-row 117
3.5.1.3 Quick discussion 118
3.5.2 A second simple example 119
3.5.2.1 The direction row/verb→column/construction 119
3.5.2.2 The direction column/construction→row/verb 120
3.5.3 A more realistic example 120
3.5.3.1 The direction row/verb→column/ditransitive 124
3.5.3.2 The direction column/ditransitive→row/verb 127
3.6 Application to key words 131
3.6.1 A first simple example 131
3.6.1.1 The direction row/word→column/corpus 133
3.6.1.2 The direction column/corpus→row/word 134
3.6.2 A more realistic example 135
3.6.2.1 The direction row/word→column/corpus 137
3.6.2.2 The direction column/corpus→row/word 140
3.7 A brief excursus on contributions to KLD 142
3.7.1 Contributions to *KLD* for keyness 142
3.7.2 Contributions to *KLD* for association 146
3.8 Application to concordancing 149
3.8.1 A first simple example 151
3.8.2 A more realistic example 153
3.8.3 Implications for frequency 158
3.9 Interesting extensions 160
3.9.1 Psycholinguistic applications (incl. prototypicality) 160
3.9.2 Synchronic and diachronic corpus homogeneity/comparisons 164
3.10 Interim summary 168

CHAPTER 4
The role, and the 'partialing out', of frequency 170
4.1 Dispersion and its correlation with frequency 171
4.2 Association and its correlation with frequency 177
4.3 Keyness and its correlation with frequency 193
4.4 Partialing out frequency 195
4.4.1 Dispersion without frequency 196
4.4.2 Association without frequency 208
4.5 Interim summary (and ominous implications?) 219

CHAPTER 5
Tupleization 229
5.1 Frequency and dispersion 231
5.2 Frequency and association 241
5.3 Frequency and association and dispersion 245
5.4 Keyness as frequency, association, and dispersion 251

CHAPTER 6
What should be next 269
6.1 Quantifying uncertainty 270
6.1.1 An example of bootstrapping 270
6.1.2 Excursus: On significance 273
6.2 Scaling things up 275
6.2.1 Speed: Scaling up with parallelization 276
6.2.2 Speed: Scaling up with Rcpp 279
6.2.2.1 Dispersion 281
6.2.2.2 Association 283
6.2.2.3 Keyness 284
6.2.3 Size: Scaling up with base 285
6.2.4 Size *and* speed: Scaling up with data.table 293
6.3 The dimensions to tupleize 297
6.3.1 Dimensions of information: Type frequencies & distributions 297
6.3.2 What are our tokens? 300

CHAPTER 7
Conclusion 304

References 308

Index 319