Significant Sequences in World Englishes:

A Data-driven Approach to Variationist Models

INAUGURAL-DISSERTATION

zur

Erlangung des Grades eines Doktors der Philosophie (Dr. phil.)

dem

Fachbereich Fremdsprachliche Philologien

der

Philipps-Universität Marburg

vorgelegt von

Christopher Koch

aus Hanau

Gutachter:

Prof. Dr. Rolf Kreyer

Prof. Dr. Claudia Lange

Einreichungstermin: 16.12.2020

Prüfungstermin: 27.04.2021

Marburg, 2022

Hochschulkennziffer: 1180

Vom Fachbereich Fremdsprachliche Philologien der Philipps-Universität Marburg als Dissertation angenommen am: 14.04.2020

Gutachter: Prof. Dr. Rolf Kreyer (Philipps-Universität Marburg)
Prof. Dr. Claudia Lange (Technische Universität Dresden)

Acknowledgments

This project has occupied a larger portion of my life than I would like to admit. Yet, it has also given me the opportunity to get to know many wonderful people who I might otherwise not have had the pleasure of meeting. First and foremost, I would like to express my deep and sincere gratitude to my advisors, who have been pivotal in allowing me to achieve my goals and complete my research. I will forever be grateful to Prof. Dr. Rolf Kreyer for his profound insight, valuable constructive criticism as well as supportive feedback, but certainly no less for his extraordinary patience. I believe this project would not have seen its fruition without your unending support and encouragement! I am also deeply indebted and grateful to Prof. Dr. Claudia Lange for her continued support at crucial moments and for offering great inspiration and invaluable guidance. Both of you have not only extended your wisdom and knowledge towards me but also your friendship, and either has been and will continue to be a great privilege!

I would also like to thank my fellow Ph.D. students and colleagues at the universities of Marburg, Dresden and Giessen: Tim Hoffmann (the reality-checking pessimist who knows no fear), Dr. Barbara Güldenring (deer esteemed colleague™, master motivator and uplifter), Anna Heidrich (deep-thinking, critical mensa buddy with the clarity I sometimes lacked), Dr. Sven Leuckert (perfect conference and travel partner and a connoisseur of fine phonology), and Dr. Tobias Bernaisch (who always hits the right chords within and outside of linguistics). All of you have helped me navigate the challenges of academic life as well as been great people to work and spend time with, and I am happy to count you amongst my friends!

Speaking of friends: Tascha and Alex, I could not even begin to estimate the number of times you have freely offered your support, lent an ear and generally been wonderful people to be around! Andi, it has always been a particular pleasure to be able to discuss the general state of both the world as well as our lives with you, and I'm looking forward to more! Ben and Sarah, thanks for always being there for me, and for building me up in times of crisis! My heartfelt thanks to all of you! Friends are indeed the family you choose! Dad, mom, Gertraud, Karlheinz, Tobi: You have often

been more enthusiastic about my work than I could manage to be; thank you for believing in me as well as for showing me that I could always count on you if I needed to!

Angela, you have experienced the best and the worst of times of this project, and I would not have made it this far without you! Thank you for persisting through the tough and long "Endspurt", always managing to see the bright side, unendingly believing in me and for having my back during some of the toughest times. I love you so much!

Contents

List of Abbreviations ..... **iv**

List of Tables ..... **v**

List of Figures ..... **x**

1 Introduction **1**

1.1 Motivation 1

1.2 Aims and Goals 3

1.3 Structure 7

2 Modeling World Englishes **9**

2.1 Monolithic and Pluricentric Models 9

2.2 Sources and Processes of Diversification 17

2.3 Proximity Effects in World Englishes 20

2.4 Evolutionary Dynamics of World Englishes 24

3 Collocational Sequences as a Discriminatory Measure **32**

3.1 Aspects of Co-occurrence 32

3.1.1 Foundations of Co-occurrence Research 32

3.1.2 Categorizing Co-occurrence 35

3.2 Co-occurrence and World Englishes 41

3.2.1 Patterned Language in Varieties of English 41

3.2.2 Using Co-occurrence to Differentiate Varieties 45

3.3 Operationalizing Collocational Sequences 49

3.3.1 Bigram and *n*-gram statistics 49

3.3.2 Association Measures 55

4 Methodology 64

4.1 Applying the International Corpus of English to Large-scale Studies of Co-occurrence 64

4.1.1 Studying Co-occurrence with ICE: Benefits and Caveats 64

4.1.2 Heterogeneity within the International Corpus of English 70

4.2 Data preparation and normalization 75

4.2.1 Homogenizing the ICE components 75

4.2.2 POS-annotating the Data for Grammatical *N*-grams 81

4.3 Extracting Collocational Patterns 85

4.3.1 Bigram Extraction and Statistics 85

4.3.2 From Bigrams to *N*-grams 88

4.4 Evaluating the Data 92

4.4.1 Interpreting Co-occurrence Data with Clustering Methods 93

4.4.2 Questions, Assumptions and Hypotheses 108

5 Significant Sequences in World Englishes 113

5.1 Dynamic-length Lexical *N*-grams 113

5.1.1 MI-score 117

5.1.2 T-score 122

5.1.3 Log Likelihood 127

5.1.4 Lexical Gravity 132

5.1.5 Delta $P_{2|1}$ 137

5.2 Dynamic-length POS-grams 142

5.2.1 MI-score 146

5.2.2 T-score 151

5.2.3 Log Likelihood 156

5.2.4 Lexical Gravity 161

5.2.5 Delta $P_{2|1}$ 166

5.3 Static-length Lexical *N*-grams 171
5.3.1 MI-score 174
5.3.2 T-score 182
5.3.3 Log likelihood 190
5.3.4 Lexical-gravity 198
5.3.5 Delta $P_{2|1}$ 207
5.4 Static-length POS-grams 216
5.4.1 MI-score 219
5.4.2 T-score 227
5.4.3 Log Likelihood 235
5.4.4 Lexical Gravity 244
5.4.5 Delta $P_{2|1}$ 252
6 Discussion and Evaluation 261
6.1 Clusters of World Englishes 262
6.1.1 Binary splits and Outliers 266
6.1.2 Regions and Regional Imbalances 268
6.1.3 Evolutionary Perspectives 273
6.2 Factors of Cluster Variation 275
6.2.1 Cluster Variation across Types of Base Data 275
6.2.2 Clusters Variation across Speech and Writing 278
6.2.3 Cluster Variation across Measures 280
6.2.4 Cluster Variation across Sequence Lengths 281
7 Conclusion and Outlook 283
References 290
Appendix A: Digital Material 309
Appendix B: Summary of the Study in German 310

List of Abbreviations

ICE	International Corpus of English Components are addressed by shorthand forms: EA (East Africa), KY (Kenya), TZ (Tanzania), GB (Great Britain), GH (Ghana), HK (Hong Kong), IND (India), IRL (Ireland), JA (Jamaica), NIG (Nigeria), NZ (New Zealand), PHI (Philippines), SIN (Singapore), SL (Sri Lanka), UG (Uganda), USA (United States of America).
SPK / WRT	The spoken or written *n*-gram data shared between varieties, i.e. the intersect of all variety-specific sequences.
ALL	The merger of the SPK and WRT data, and thus their intersect.
HCA	Hierarchical Agglomerative Cluster Analysis

Shorthand Forms for Cluster Descriptions

IC / IC_{NA} / IC_{GB}	Shorthand forms for 'Inner Circle' clusters: All varieties within a dataset or the North-American or British-epicentral subsets.
'+' (plus)	Mergers of two (groups of) varieties within a joint cluster, e.g. GB+IRL+NZ.
'-' (minus)	Exclusion of a variety from an otherwise consistent descriptor, e.g. Asia-IND in case an Asian cluster only lacks the Indian data.
'()' (brackets)	Lower confidence in the bracketed cluster or merger, e.g. HK+SIN (+PHI) implying that the latter merger is only indicated by some methods.

Further codes only employed in Chapter 6:

'/' (forward slash)	Alternatives if two clusters are equally supported across methods, e.g. EA+IND/NIG implying EA+IND or EA+NIG.
',' (comma)	Related subclusters (e.g. by region) presented within a single line as a shorthand form.
' ' (space)	Indicates higher distances than contrasting groups within the same expression, e.g. KY+TZ + UG + GH+NIG implying some dissimilarity of UG to both KY+TZ as well as GH+NIG.

List of Tables

Table 3.1. Contingency table underlying most collocational statistics 50
Table 3.2. Expected item frequencies and their ways of calculation from observed frequencies..... 50
Table 4.1: ICE components in the present study. 76
Table 4.2: Markup removed together with corpus text in all corpus components. ... 78
Table 4.3: Markup and encoding changes specific to some corpus components..... 79
Table 4.4: Markup (and annotated text) used for the segmentation of the corpus text 81
Table 4.5: Original corpus files and results of the cleanup and homogenization procedure in two corpus samples 82
Table 4.6: Cleaned and homogenized corpus data, C7-annotated version and final edited version for POS-gram analysis..... 84
Table 4.7. Expected item frequencies and their calculation from observed frequencies..... 88
Table 4.8: Basic observed and expected co-occurrence values for the first 26 bigrams in ICE-HK W1A-001 89
Table 4.9: Association scores calculated for the first 26 bigrams in ICE-HK W1A-001. 90
Table 4.10: Threshold values for the selection of bigrams for dynamic-length *n*-grams. 91
Table 4.11: Sample results produced by Greenacre's permutation test for clusteredness 103
Table 4.12: ICE varieties and respective phases in the Dynamic Model..... 111
Table 5.1: Frequencies of lexical bigram types and tokens, plus TTRs, extracted from the ICE data 113
Table 5.2: Threshold values for the selection of bigrams for dynamic-length *n*-grams (association scores need to be greater than these values). 114
Table 5.3: Spoken lexical bigram token and type frequencies, plus TTRs, after the application of threshold values 115

Table 5.4: Written lexical bigram token and type frequencies, plus TTRs, after the application of threshold values 116
Table 5.5: Lexical *MI n*-gram type frequencies by length in the intersects of the variety-specific datasets..... 119
Table 5.6: Lexical *MI n*-grams with highest and lowest association scores..... 119
Table 5.7: K-means clustering results for specific values of *k* for lexical *MI n*-grams 121
Table 5.8: Lexical *t n*-gram type frequencies by length in the intersects of the variety-specific datasets..... 124
Table 5.9: Lexical *t n*-grams with highest and lowest association scores..... 124
Table 5.10: K-means clustering results for specific values of *k* lexical *t n*-grams.... 126
Table 5.11: Lexical G^2 *n*-gram type frequencies by length in the intersects of the variety-specific datasets 129
Table 5.12: Lexical G^2 *n*-grams with highest and lowest association scores 129
Table 5.13: K-means clustering results for specific values of *k* for lexical G^2 *n*-grams 131
Table 5.14: Lexical *g n*-gram type frequencies by length in the intersects of the variety-specific datasets..... 134
Table 5.15: Lexical *g n*-grams with highest and lowest association scores..... 134
Table 5.16: K-means clustering results for specific values of *k* for lexical *g n*-grams 136
Table 5.17: Lexical ΔP *n*-gram type frequencies by length in the intersects of the variety-specific datasets 139
Table 5.18: Lexical ΔP *n*-grams with highest and lowest association scores..... 139
Table 5.19: K-means clustering results for specific values of *k* for lexical ΔP *n*-grams 141
Table 5.20: Frequencies of POS bigram types and tokens, plus TTRs, extracted from the ICE data..... 142
Table 5.21: Spoken POS bigram token and type frequencies, plus TTRs, after the application of threshold values 144
Table 5.22: Written POS bigram token and type frequencies, plus TTRs, after the application of threshold values 145

Table 5.23: POS *MI n*-gram type frequencies by length in the intersects of the variety-specific datasets 148
Table 5.24: POS *MI n*-grams with highest and lowest association scores 148
Table 5.25: K-means clustering results for specific values of *k* for POS *MI n*-grams 150
Table 5.26: POS *t n*-gram type frequencies by length in the intersects of the variety-specific datasets 153
Table 5.27: POS *t n*-grams with highest and lowest association scores 153
Table 5.28: K-means clustering results for specific values of *k* for POS *t n*-grams. 155
Table 5.29: POS G^2 *n*-gram type frequencies by length in the intersects of the variety-specific datasets 158
Table 5.30: POS G^2 *n*-grams with highest and lowest association scores 158
Table 5.31: K-means clustering results for specific values of *k* for POS G^2 *n*-grams 160
Table 5.32: POS *g n*-gram type frequencies by length in the intersects of the variety-specific datasets 163
Table 5.33: POS *g n*-grams with highest and lowest association scores 163
Table 5.34: K-means clustering results for specific values of *k* for POS *g n*-grams 165
Table 5.35: POS ΔP *n*-gram type frequencies by length in the intersects of the variety-specific datasets 168
Table 5.36: POS ΔP *n*-grams with highest and lowest association scores 168
Table 5.37: K-means clustering results for specific values of *k* for POS ΔP *n*-grams 170
Table 5.38: Static-length lexical *n*-gram average frequencies and association values 172
Table 5.39: Lexical *n*-gram type frequencies by length in the intersects of the variety-specific datasets 172
Table 5.40: Lexical *MI n*-grams with highest and lowest association scores 174
Table 5.41: K-means clustering results for specific values of *k* for lexical *MI* 2-grams 177
Table 5.42: K-means clustering results for specific values of *k* for lexical *MI* 3-grams 179

Table 5.43: K-means clustering results for specific values of *k* for lexical *MI* 4-grams 182
Table 5.44: Lexical *t n*-grams with highest and lowest association scores 183
Table 5.45: K-means clustering results for specific values of *k* for lexical *t* 2-grams 185
Table 5.46: K-means clustering results for specific values of *k* for lexical *t* 3-grams 188
Table 5.47: K-means clustering results for specific values of *k* for lexical *t* 4-grams 190
Table 5.48: Lexical G^2 *n*-grams with highest and lowest association scores 191
Table 5.49: K-means clustering results for specific values of *k* for lexical G^2 2-grams 193
Table 5.50: K-means clustering results for specific values of *k* for lexical G^2 3-grams 196
Table 5.51: K-means clustering results for specific values of *k* for lexical G^2 4-grams 198
Table 5.52: Lexical *g n*-grams with highest and lowest association scores 199
Table 5.53: K-means clustering results for specific values of *k* for lexical *g* 2-grams 202
Table 5.54: K-means clustering results for specific values of *k* for lexical *g* 3-grams 204
Table 5.55: K-means clustering results for specific values of *k* for lexical *g* 4-grams 207
Table 5.56: Lexical ΔP *n*-grams with highest and lowest association scores 208
Table 5.57: K-means clustering results for specific values of *k* for lexical ΔP 2-grams 210
Table 5.58: K-means clustering results for specific values of *k* for lexical ΔP 3-grams 213
Table 5.59: K-means clustering results for specific values of *k* for lexical ΔP 4-grams 215
Table 5.60: Static-length POS *n*-gram average frequencies and association values 217
Table 5.61: POS *n*-gram type frequencies by length in the intersects of the variety-specific datasets 218

Table 5.62: POS *MI n*-grams with highest and lowest association scores 219
Table 5.63: K-means clustering results for specific values of *k* for POS *MI* 2-grams 222
Table 5.64: K-means clustering results for specific values of *k* for POS *MI* 3-grams 224
Table 5.65: K-means clustering results for specific values of *k* for POS *MI* 4-grams 227
Table 5.66: POS *t n*-grams with highest and lowest association scores 228
Table 5.67: K-means clustering results for specific values of *k* for POS *t* 2-grams. 230
Table 5.68: K-means clustering results for specific values of *k* for POS *t* 3-grams. 233
Table 5.69: K-means clustering results for specific values of *k* for POS *t* 4-grams. 235
Table 5.70: POS G^2 *n*-grams with highest and lowest association scores 236
Table 5.71: K-means clustering results for specific values of *k* for POS G^2 2-grams 238
Table 5.72: K-means clustering results for specific values of *k* for POS G^2 3-grams 241
Table 5.73: K-means clustering results for specific values of *k* for POS G^2 4-grams 243
Table 5.74: POS *g n*-grams with highest and lowest association scores..... 244
Table 5.75: K-means clustering results for specific values of *k* for POS *g* 2-grams 247
Table 5.76: K-means clustering results for specific values of *k* for POS *g* 3-grams 249
Table 5.77: K-means clustering results for specific values of *k* for POS *g* 4-grams 251
Table 5.78: POS ΔP *n*-grams with highest and lowest association scores 252
Table 5.79: K-means clustering results for specific values of *k* for POS ΔP 2-grams 255
Table 5.80: K-means clustering results for specific values of *k* for POS ΔP 3-grams 257
Table 5.81: K-means clustering results for specific values of *k* for POS ΔP 4-grams 260
Table 6.1: Clusters in the lexical data by agreement of clustering methods 264
Table 6.2: Clusters in the POS data by agreement of clustering methods 265

List of Figures

Figure 2.1: Modified version of Kachru's model to account for functional nativeness (Yano 2001: 123) 13
Figure 2.2: Sources and processes of the formation of post-colonial Englishes (Schneider 2007: 100) 19
Figure 2.3: 3-D framework of language variation (Mahboob 2017: 17) 20
Figure 2.4: Strevens's (1992: 33) world map of English 21
Figure 2.5: Görlach's (1988) Circle model of English 21
Figure 2.6: McArthur's (1987) Circle of World English..... 22
Figure 4.1: Regular expression used to extract bigrams..... 86
Figure 4.2: Methodological steps in the extraction of static- and dynamic-length *n*-grams 92
Figure 4.3: Dendrogram of a fictitious cluster analysis of English consonants 95
Figure 4.4: Three fictitious datasets and their (dis-)similarities 98
Figure 4.5: Dendrogram of the fictitious consonant data with dashed black lines indicating clusters achieving values of AU≥95..... 101
Figure 4.6: Cluster dendrogram of some base data next to a random permutation.102
Figure 4.7: Visual representation of the results of Greenacre's permutation test next to an application of the 5-cluster segmentation to the fictitious consonant data.104
Figure 4.8: Percent variability explained (black) and within-cluster variation (gray) plotted against the number of k-means clusters 106
Figure 4.9: Sample NeighborNet visualization 107
Figure 5.1: Distribution of lexical *MI n*-gram lengths across the varietal datasets .. 118
Figure 5.2: Number of shared lexical *MI n*-grams between any two datasets..... 119
Figure 5.3: Hierarchical clustering results for lexical *MI n*-grams..... 120
Figure 5.4: Jumps in node heights and respective *p*-values for lexical *MI n*-grams 120
Figure 5.5: NeighborNets of the spoken and written data for lexical *MI n*-grams... 121
Figure 5.6: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *MI n*-grams 121
Figure 5.7: Distribution of for lexical *t n*-gram lengths across the varietal datasets 123
Figure 5.8: Number of shared lexical *t n*-grams between any two datasets..... 124

Figure 5.9: Hierarchical clustering results for lexical *t n*-grams 125
Figure 5.10: Jumps in node heights and respective *p*-values lexical *t n*-grams 125
Figure 5.11: NeighborNets of the spoken and written data lexical *t n*-grams 126
Figure 5.12: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters lexical *t n*-grams 126
Figure 5.13: Distribution of lexical G^2 *n*-gram lengths across the varietal datasets. 128
Figure 5.14: Number of shared lexical G^2 *n*-grams between any two datasets 129
Figure 5.15: Hierarchical clustering results for lexical G^2 *n*-grams 130
Figure 5.16: Jumps in node heights and respective *p*-values for lexical G^2 *n*-grams130
Figure 5.17: NeighborNets of the spoken and written data for lexical G^2 *n*-grams . 131
Figure 5.18: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical G^2 *n*-grams 131
Figure 5.19: Distribution of lexical *g n*-gram lengths across the varietal datasets .. 133
Figure 5.20: Number of shared lexical *g n*-grams between any two datasets 134
Figure 5.21: Hierarchical clustering results for lexical *g n*-grams 135
Figure 5.22: Jumps in node heights and respective *p*-values for lexical *g n*-grams 135
Figure 5.23: NeighborNets of the spoken and written data for lexical *g n*-grams ... 136
Figure 5.24: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *g n*-grams 136
Figure 5.25: Distribution of lexical ΔP *n*-gram lengths across the varietal datasets 138
Figure 5.26: Number of shared lexical ΔP *n*-grams between any two datasets 139
Figure 5.27: Hierarchical clustering results for lexical ΔP *n*-grams 140
Figure 5.28: Jumps in node heights and respective *p*-values for lexical ΔP *n*-grams 140
Figure 5.29: NeighborNets of the spoken and written data for lexical ΔP *n*-grams. 141
Figure 5.30: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical ΔP *n*-grams 141
Figure 5.31: Distribution of POS *MI n*-gram lengths across the varietal datasets... 147
Figure 5.32: Number of shared POS *MI n*-grams between any two datasets 148
Figure 5.33: Hierarchical clustering results for POS *MI n*-grams 149
Figure 5.34: Jumps in node heights and respective *p*-values for POS *MI n*-grams 149
Figure 5.35: NeighborNets of the spoken and written data for POS *MI n*-grams ... 150

Figure 5.36: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *MI n*-grams 150
Figure 5.37: Distribution of POS *t n*-gram lengths across the varietal datasets 152
Figure 5.38: Number of shared POS *t n*-grams between any two datasets 153
Figure 5.39: Hierarchical clustering results for POS *t n*-grams 154
Figure 5.40: Jumps in node heights and respective *p*-values for POS *t n*-grams 154
Figure 5.41: NeighborNets of the spoken and written data for POS *t n*-grams 155
Figure 5.42: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *t n*-grams 155
Figure 5.43: Distribution of POS G^2 *n*-gram lengths across the varietal datasets ... 157
Figure 5.44: Number of shared POS G^2 *n*-grams between any two datasets 158
Figure 5.45: Hierarchical clustering results for POS G^2 *n*-grams 159
Figure 5.46: Jumps in node heights and respective *p*-values for POS G^2 *n*-grams . 159
Figure 5.47: NeighborNets of the spoken and written data for POS G^2 *n*-grams 160
Figure 5.48: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS G^2 *n*-grams 160
Figure 5.49: Distribution of POS *g n*-gram lengths across the varietal datasets 162
Figure 5.50: Number of shared POS *g n*-grams between any two datasets..... 163
Figure 5.51: Hierarchical clustering results for POS *g n*-grams..... 164
Figure 5.52: Jumps in node heights and respective *p*-values for POS *g n*-grams ... 164
Figure 5.53: NeighborNets of the spoken and written data for POS *g n*-grams..... 165
Figure 5.54: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *g n*-grams 165
Figure 5.55: Distribution of POS ΔP *n*-gram lengths across the varietal datasets... 167
Figure 5.56: Number of shared POS ΔP *n*-grams between any two datasets 168
Figure 5.57: Hierarchical clustering results for POS ΔP *n*-grams 169
Figure 5.58: Jumps in node heights and respective *p*-values for POS ΔP *n*-grams. 169
Figure 5.59: NeighborNets of the spoken and written data for POS ΔP *n*-grams ... 170
Figure 5.60: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS ΔP *n*-grams 170
Figure 5.61: Number of shared lexical 2-, 3- and 4-grams between any two datasets 173
Figure 5.62: Hierarchical clustering results for lexical *MI* 2-grams..... 176

Figure 5.63: Jumps in node heights and respective *p*-values for lexical *MI* 2-grams 176
Figure 5.64: NeighborNets of the spoken and written data for lexical *MI* 2-grams. 176
Figure 5.65: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *MI* 2-grams 177
Figure 5.66: Hierarchical clustering results for lexical *MI* 3-grams..... 178
Figure 5.67: Jumps in node heights and respective *p*-values for lexical *MI* 3-grams 178
Figure 5.68: NeighborNets of the spoken and written data for lexical *MI* 3-grams. 178
Figure 5.69: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *MI* 3-grams 179
Figure 5.70: Hierarchical clustering results for lexical *MI* 4-grams..... 180
Figure 5.71: Jumps in node heights and respective *p*-values for lexical *MI* 4-grams 180
Figure 5.72: NeighborNets of the spoken and written data for lexical *MI* 4-grams. 180
Figure 5.73: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *MI* 4-grams 182
Figure 5.74: Hierarchical clustering results for lexical *t* 2-grams..... 184
Figure 5.75: Jumps in node heights and respective *p*-values for lexical *t* 2-grams. 184
Figure 5.76: NeighborNets of the spoken and written data for lexical *t* 2-grams.... 184
Figure 5.77: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *t* 2-grams 185
Figure 5.78: Hierarchical clustering results for lexical *t* 3-grams..... 187
Figure 5.79: Jumps in node heights and respective *p*-values for lexical *t* 3-grams. 187
Figure 5.80: NeighborNets of the spoken and written data for lexical *t* 3-grams.... 187
Figure 5.81: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *t* 3-grams 188
Figure 5.82: Hierarchical clustering results for lexical *t* 4-grams..... 189
Figure 5.83: Jumps in node heights and respective *p*-values for lexical *t* 4-grams. 189
Figure 5.84: NeighborNets of the spoken and written data for lexical *t* 4-grams.... 189
Figure 5.85: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *t* 4-grams 190
Figure 5.86: Hierarchical clustering results for lexical G^2 2-grams 192

Figure 5.87: Jumps in node heights and respective *p*-values for lexical G^2 2-grams192
Figure 5.88: NeighborNets of the spoken and written data for lexical G^2 2-grams . 192
Figure 5.89: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical G^2 2-grams..... 193
Figure 5.90: Hierarchical clustering results for lexical G^2 3-grams 195
Figure 5.91: Jumps in node heights and respective *p*-values for lexical G^2 3-grams195
Figure 5.92: NeighborNets of the spoken and written data for lexical G^2 3-grams . 195
Figure 5.93: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical G^2 3-grams..... 196
Figure 5.94: Hierarchical clustering results for lexical G^2 4-grams 197
Figure 5.95: Jumps in node heights and respective *p*-values for lexical G^2 4-grams197
Figure 5.96: NeighborNets of the spoken and written data for lexical G^2 4-grams . 197
Figure 5.97: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical G^2 4-grams..... 198
Figure 5.98: Hierarchical clustering results for lexical *g* 2-grams 201
Figure 5.99: Jumps in node heights and respective *p*-values for lexical *g* 2-grams 201
Figure 5.100: NeighborNets of the spoken and written data for lexical *g* 2-grams . 201
Figure 5.101: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *g* 2-grams..... 202
Figure 5.102: Hierarchical clustering results for lexical *g* 3-grams 203
Figure 5.103: Jumps in node heights and respective *p*-values for lexical *g* 3-grams 203
Figure 5.104: NeighborNets of the spoken and written data for lexical *g* 3-grams . 203
Figure 5.105: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *g* 3-grams..... 204
Figure 5.106: Hierarchical clustering results for lexical *g* 4-grams 206
Figure 5.107: Jumps in node heights and respective *p*-values for lexical *g* 4-grams 206
Figure 5.108: NeighborNets of the spoken and written data for lexical *g* 4-grams . 206
Figure 5.109: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *g* 4-grams..... 207
Figure 5.110: Hierarchical clustering results for lexical ΔP 2-grams..... 209

Figure 5.111: Jumps in node heights and respective *p*-values for lexical *ΔP* 2-grams 209
Figure 5.112: NeighborNets of the spoken and written data for lexical *ΔP* 2-grams209
Figure 5.113: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *ΔP* 2-grams 210
Figure 5.114: Hierarchical clustering results for lexical *ΔP* 3-grams..... 212
Figure 5.115: Jumps in node heights and respective *p*-values for lexical *ΔP* 3-grams 212
Figure 5.116: NeighborNets of the spoken and written data for lexical *ΔP* 3-grams212
Figure 5.117: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *ΔP* 3-grams 213
Figure 5.118: Hierarchical clustering results for lexical *ΔP* 4-grams..... 214
Figure 5.119: Jumps in node heights and respective *p*-values for lexical *ΔP* 4-grams 214
Figure 5.120: NeighborNets of the spoken and written data for lexical *ΔP* 4-grams214
Figure 5.121: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *ΔP* 4-grams 215
Figure 5.122: Number of shared POS 2-, 3- and 4-grams between any two datasets 218
Figure 5.123: Hierarchical clustering results for POS *MI* 2-grams 221
Figure 5.124: Jumps in node heights and respective *p*-values for POS *MI* 2-grams221
Figure 5.125: NeighborNets of the spoken and written data for POS *MI* 2-grams . 221
Figure 5.126: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *MI* 2-grams..... 222
Figure 5.127: Hierarchical clustering results for POS *MI* 3-grams 223
Figure 5.128: Jumps in node heights and respective *p*-values for POS *MI* 3-grams223
Figure 5.129: NeighborNets of the spoken and written data for POS *MI* 3-grams . 223
Figure 5.130: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *MI* 3-grams..... 224
Figure 5.131: Hierarchical clustering results for POS *MI* 4-grams 226
Figure 5.132: Jumps in node heights and respective *p*-values for POS *MI* 4-grams226
Figure 5.133: NeighborNets of the spoken and written data for POS *MI* 4-grams . 226

Figure 5.134: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *MI* 4-grams 227
Figure 5.135: Hierarchical clustering results for POS *t* 2-grams 229
Figure 5.136: Jumps in node heights and respective *p*-values for POS *t* 2-grams.. 229
Figure 5.137: NeighborNets of the spoken and written data for POS *t* 2-grams 229
Figure 5.138: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *t* 2-grams 230
Figure 5.139: Hierarchical clustering results for POS *t* 3-grams 232
Figure 5.140: Jumps in node heights and respective *p*-values for POS *t* 3-grams.. 232
Figure 5.141: NeighborNets of the spoken and written data for POS *t* 3-grams 232
Figure 5.142: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *t* 3-grams 233
Figure 5.143: Hierarchical clustering results for POS *t* 4-grams 234
Figure 5.144: Jumps in node heights and respective *p*-values for POS *t* 4-grams.. 234
Figure 5.145: NeighborNets of the spoken and written data for POS *t* 4-grams 234
Figure 5.146: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *t* 4-grams 235
Figure 5.147: Hierarchical clustering results for POS G^2 2-grams 237
Figure 5.148: Jumps in node heights and respective *p*-values for POS G^2 2-grams 237
Figure 5.149: NeighborNets of the spoken and written data for POS G^2 2-grams.. 237
Figure 5.150: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS G^2 2-grams 238
Figure 5.151: Hierarchical clustering results for POS G^2 3-grams 240
Figure 5.152: Jumps in node heights and respective *p*-values for POS G^2 3-grams 240
Figure 5.153: NeighborNets of the spoken and written data for POS G^2 3-grams.. 240
Figure 5.154: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS G^2 3-grams 241
Figure 5.155: Hierarchical clustering results for POS G^2 4-grams 242
Figure 5.156: Jumps in node heights and respective *p*-values for POS G^2 4-grams 242
Figure 5.157: NeighborNets of the spoken and written data for POS G^2 4-grams.. 242
Figure 5.158: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS G^2 4-grams 243
Figure 5.159: Hierarchical clustering results for POS *g* 2-grams 246

Figure 5.160: Jumps in node heights and respective *p*-values for POS *g* 2-grams. 246
Figure 5.161: NeighborNets of the spoken and written data for POS *g* 2-grams.... 246
Figure 5.162: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *g* 2-grams 247
Figure 5.163: Hierarchical clustering results for POS *g* 3-grams..... 248
Figure 5.164: Jumps in node heights and respective *p*-values for POS *g* 3-grams. 248
Figure 5.165: NeighborNets of the spoken and written data for POS *g* 3-grams.... 248
Figure 5.166: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *g* 3-grams 249
Figure 5.167: Hierarchical clustering results for POS *g* 4-grams..... 250
Figure 5.168: Jumps in node heights and respective *p*-values for POS *g* 4-grams. 250
Figure 5.169: NeighborNets of the spoken and written data for POS *g* 4-grams.... 250
Figure 5.170: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *g* 4-grams 251
Figure 5.171: Hierarchical clustering results for POS *ΔP* 2-grams 254
Figure 5.172: Jumps in node heights and respective *p*-values for POS *ΔP* 2-grams254
Figure 5.173: NeighborNets of the spoken and written data for POS *ΔP* 2-grams . 254
Figure 5.174: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *ΔP* 2-grams..... 255
Figure 5.175: Hierarchical clustering results for POS *ΔP* 3-grams 256
Figure 5.176: Jumps in node heights and respective *p*-values for POS *ΔP* 3-grams256
Figure 5.177: NeighborNets of the spoken and written data for POS *ΔP* 3-grams . 256
Figure 5.178: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *ΔP* 3-grams..... 257
Figure 5.179: Hierarchical clustering results for POS *ΔP* 4-grams 259
Figure 5.180: Jumps in node heights and respective *p*-values for POS *ΔP* 4-grams259
Figure 5.181: NeighborNets of the spoken and written data for POS *ΔP* 4-grams . 259
Figure 5.182: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *ΔP* 4-grams..... 260

1 Introduction

1.1 Motivation

There was a time when the landscape of English worldwide must have been simpler. Up to the mid-20th century, there was really only one proper English. The local roots this language sprouted in the remote corners of the British Empire were certainly taken note of but otherwise not awarded any greater relevance. This changed in the post-colonial world, when English was, to some surprise perhaps, retained in many of the newly independent nations and gradually claimed as a cultural resource by people not genetically 'native' in the language. When reports from post-independence nations began to illustrate the worldwide linguistic diversity of local Englishes, and progressively-minded researchers started to advocate for their autonomy, this was quickly met with severe backlash and resistance to what was felt by some to be a glorification of "inherently unstable" forms of English (Quirk 1990: 5). Lacking any systematic data on the new localized Englishes, the ensuing heated debate about the mono- or pluricentric character of English can be regarded as essentially revolving around competing world-views, and as such around internalized models of the post-colonial situation of English.

The split between language-external modeling and language-internal description appears to persist to this day, and approaches to either one only rarely intersect. Certainly, they frequently complement one another, but some of the far-reaching theories within the field and the predictions they entail are not supported by tests on sufficiently expansive sets of data to gauge their actual potential. On the other hand, if descriptive approaches attempt to model linguistic processes, usually only microscopic models are the result. The macroscopic situation is thus mostly approached via political or cultural factors and not on the basis of actual language data. What is more, if studies attempt to evaluate models on truly linguistic grounds, they restrict themselves in other regards, especially by revolving around only a handful of varieties, usually in regional proximity to each other. Conceptually, this disregards the truly global scale of World Englishes, but methodologically it may be even worse: If groups of only three or four varieties are scrutinized, as is common practice, even a single variety can

disrupt a consistent interpretation if it does not comply with the predictions made on the grounds of the specific model under focus. Worst of all, it appears that any individual study at most attempts to evaluate predictions made on the basis of a single model. It may be that this is a result of a predominant tendency towards corpus-based methodologies, which work in a top-down fashion and thus require an a priori choice of a particular theory. It could, however, just as well be a form of the survivorship bias, in that a best-fitting explanatory approach may be adopted after the linguistic outcomes are determined, thus dropping any less impactful models from a study. No matter the precise reasons, World Englishes research still lacks a fundamental understanding of the relative linguistic impacts of the diverse range of models as well as a consistent examination of their predictions on an appropriately large scale.

For a truly sufficient evaluation of the current modeling of World Englishes, a larger range of varieties seems desperately needed. A large varietal scope not only compensates for any individual outliers but furthermore allows for the discernment of the relative linguistic effects of different models. This entails the adoption of a true bird's-eye perspective, and as such, in contrast to the predominant corpus-based approach, a strictly data-driven methodology is indicated. The rationale for this goes beyond the more immediate concern of feasibility of a study resting on vast amounts of data: Instead of the common top-down corpus-based approach, a methodology truly driven by the corpus data makes it possible to discover groups of varieties within a bottom-up, theory-neutral process. Only after clusters of maximally similar varieties are substantiated through objective means should the fit of different language-external models be considered, thus addressing the level of uncertainty left by the fragmentary model evaluations of previous approaches. This will make it possible to assess the linguistic effects of different models side-by-side: Is there an actual worldwide patterning of varieties according to prescribed evolutionary stages as Schneider's (2007) Dynamic Model suggests? Could varieties of predominantly 'non-native' or 'functionally native' speakers, against all egalitarian efforts in terms of their acceptance, emerge as a group clearly distinct from (genetically) 'native' varieties, as a multitude of

apparently descriptive approaches imply?[1] Or does proximity in the sense of regional centers (cf. Görlach 1988, McArthur 1987) or epicentral effects (cf. e.g. Hundt 2013) supersede any of these concerns? It might be none of the above or all in equal measure, but in order to assess this global situation and the worldwide applicability of models, it becomes necessary to systematically map out the language-internal differentiation between all World Englishes and attempt to correlate this with our predominant theories.

1.2 Aims and Goals

The present study seeks to address these questions by establishing groups of varieties from the ground up following a strictly data-driven methodological approach. Only after groups of varieties are obtained on the basis of actual linguistic similarity will established models be evaluated in terms of their fit onto the linguistic situation. Unlike most other approaches, language-external categories are not expected to find reflection in the language-internal data, and diverse explanatory models can be contrasted against the emerging varietal groups. The central question to be followed concerns whether the groups emerging from the data can be consistently mapped onto models of World Englishes, particularly from either a static, regional perspective based on proximity, regional centers and aspects of epicentricity – or otherwise a dynamic, evolutionary angle rooted in cultural developments and processes of identity formation as proposed within the Dynamic Model (or a mixture of the two). All the while, the analysis sets the 'historical input variety' of British or US English not as a focal point of the analysis but rather as one within the larger picture only, describing and mapping similarities and differences within the whole spectrum of 'national' varieties of English. The linguistic object of study best suited for this analytical purpose can be found in lexical and grammatical sequences in the form of *n*-grams. In the present study, these are chosen as the analytical focus on the basis of at least three major inherent benefits

[1] The division lies at the heart of many outwardly neutral models, such as the separation of speech communities by predominantly genetic natives, second-language and foreign-language users. The same holds for the apparently temporal classification of the spread of English into two diasporas, which however mainly reflects the same division into settler- vs. indigenous-predominated varieties. It even applies to Kachru's Three Circles (1992b) – if only in the underlying linguistic makeup of the variety and not in terms of its political agenda. See Chapter 2 for a more comprehensive discussion.

– linguistic relevance, neutrality and consistency, feasibility and sophistication – which will be discussed in turn.

From the perspective of language theory, linguistic sequences, preference patterns and habitual language use has come to be understood as one of the prime factors of regional variation. This stands in stark contrast to some of the earliest reports on the characteristics of local Englishes, which relied heavily on qualitative and intuitive accounts fueled from personal experience and reception of the media. Corpus-based efforts helped to alleviate the inherent unsystematicity of these approaches, but it is also true that ever more peripheral, low-frequency phenomena came into the focus of corpus research. Their overemphasis fragmented the analysis of World Englishes, and as late as 2007, Schneider observes a disregard for this larger picture, in that it still was "customary to view individual [post-colonial Englishes, CK] in isolation, independently of each other, as unique cases shaped by idiosyncratic historical conditions and contact situations." (Schneider 4). In contrast, it may be more fruitful to address patterns of higher prevalence across the entirety of World Englishes, since qualitatively distinctive features of individual varieties, while clearly noticeable and highly unique, only go so far in describing the overall situation:

> [I]ndividual varieties differ from each other first and foremost in their combinatory preferences, in their constructions, in the frequencies of their lexicogrammatical choices, collocations, word uses, and so on. It is, not only, and perhaps not even primarily, the occasional occurrences of well-known 'distinctive features' that attribute its uniqueness to a variety; it is the subconscious set of conventions regulating the norm level of speech habits, of what is normally done and uttered, the 'way things are said' in a community. (Schneider 2007: 92)

The lion's share of the linguistic variation regularly encountered thus occurs on the level of preference patterns for the same types of sequences instead of entirely distinct ways of expression (which is why the root Englishes still applies). Unlike highly conspicuous qualitative variation, quantitative preference cannot easily be described intuitively even by competent speakers of a language. Preference patterns rather express the local way of idiomatic language use, i.e. the capacity of a competent speaker to know, in addition to "what combinations are *possible*, […] which particular combinations are *conventional* in a language community although other combinations are conceivable" (Warren 2005: 40). If the bulk of English is actually qualitatively similar across varieties, language patterns much more closely approximate this common core

(Quirk et al. 1985: 16) of English than individual idiosyncrasies. Different degrees of association can hold between qualitatively identical items across Englishes, stemming from varying usage preferences and resulting in variety-specific association profiles. Variation across association profiles thus presents itself as an ideal tool to differentiate varieties by their locally characteristic ways to habitually employ the language.

Beyond their relevance for language theory, quantitative patterns also appear as one of the most neutral approaches to observing variation. Co-occurrence in language is unavoidable and finds consistent reflection in quantitative profiles irrespective of the underlying structural constraints or personal, idiosyncratic choices which led to the use of a particular sequence: In either case, it will become available as data for the distinction of varieties. This is only restricted by the necessity to trim any list of items to those shared across the entirety of the data, which excludes sequences not found in all varieties. However, association values for shared sequences are still impacted by all competing choices within a variety, even if qualitatively not all sequences remain within the analysis. It is a rare case indeed that a new feature fully replaces another, more common-core item. Rather, the new feature complements existing patterns, with which it starts to compete for usage preference:

> [I]n its early stages this indigenization of language structure mostly occurs at the interface between grammar and lexis, affecting the syntactic behavior of certain lexical elements. Individual words, typically high-frequency items, adopt characteristic but marked usage and complementation patterns. […] Hence, grammatical nativization in [post-colonial Englishes, CK] typically sets out with a specific set of patterns which appear to occur more frequently than others. (Schneider 2007: 46)

It is thus lexis and lexical co-occurrence which accounts for the majority of early changes within a local English, providing the springboard by which new structural elements enter the variety on its road to nativization: Lexical changes are those that manifest earliest in the process of variety formation (phase 1 of Schneider's 2007 Dynamic Model), while still remaining relevant in varieties that have reached the end of the cycle. Even in case of variety-specific items, their use occurs at the expense of competing common-core patterns, causing the latter to find less stable association. As such, even variety-specific sequences become available for study by proxy. In addition to the typically lexis-based description of patterned language, the present approach extends the analysis to also include grammatical (POS) sequences. This not only abstracts away from concrete linguistic instantiations, but furthermore allows for a more

specific scrutiny of potential structural changes in World Englishes and degrees of similarities between varieties.

Finally, in a study of a large varietal scope, the choice of a suitable feature necessarily also concerns aspects of feasibility across a diverse host of varieties. Certainly, lexical as well as grammatical sequences have a special appeal for the analysis of vast amounts of data, which stems from the fact that they lend themselves particularly well to automatic extraction and evaluation. This comes at a price, however: Transforming linguistic behavior and qualitative items into purely quantitative data requires a certain degree of faith in the statistical technique. This is all the more challenging since the mathematical model that lies at the heart of each specific measure can be more or less applicable to any given linguistic context. Common wisdom is thus to employ more than a single association measure and to contrast findings from relatively well-understood techniques. The present study aims at increasing this level of reliability as much as possible, incorporating five different measures of both conventional as well as more innovative designs and triangulating their results. This makes it possible to draw onto both the ease of computation as well as the methodological sophistication of co-occurrence-based research without any cost in terms of reliability of the results obtained.

Lexical and grammatical sequences are extracted in the form of *n*-grams (/POS-grams) following either a dynamic definition of the best sequence lengths based on data-driven measures or otherwise a more traditional technique based on static lengths (2-, 3-, 4-grams). The choice of the static lengths is, however, based on the lengths preferred within the initial dynamic approach. This analytical approach is carried out with five complementary association measures, each favoring distinct types of co-occurrence and thus allowing inspection of preference patterns in World Englishes from a diverse set of angles. In addition to this underlying linguistic analysis, the data-driven approach is also extended to the statistical toolset, which embraces cluster analysis as the framework of choice. For the present application, clustering methods emerge as powerful tools for dealing with large amounts of heterogenous data, while furthermore being designed for the very purpose of mapping out complex interrelatedness between objects, in this case groups of varieties. Any groups which emerge from the forty distinct analyses of four clustering methods each will be

evaluated in a stepwise triangulative fashion, first within each separate dataset and finally across all analyses. While individual description of findings and detailed evaluation of items is mostly disregarded within the consistent step-by-step analytical process embraced by the present analysis, it is a fundamental assumption of this study that even abstract, quantitative patterns of similarity and difference will be able to systematically reflect our current theoretical understanding of World Englishes on a truly global scale.

The database of choice for the present endeavor lies in the *International Corpus of English* (ICE), which constitutes the major resource for evaluations of 'national varieties' to this day. In particular, the analysis employs all ICE components which have either been fully released or for which the initial, written parts are finalized (15 at the time of writing). Certainly, larger corpora have become available, particularly with the GlowbE database (Davies & Fuchs 2015). Still, the ICE data is deemed superior on the grounds of three inherent benefits of its design: It offers a clear systematicity in its parallel design across varieties, follows a balanced approach to data collection which results in higher reliability and representativity of the data, and furthermore includes a substantial amount of spoken language reflecting more dynamic interactions between local languages. On the point of its limited size, it appears that this is rather felt to be an issue in studies focusing on relatively peripheral, low-frequency phenomena. There is thus arguably still a lot of potential concerning patterns of higher prevalence even within the ICE data. Within the analytical framework laid out above, the ICE corpus presents a highly reliable choice of data to this day, allowing a consistent comparison across the large number of World Englishes sampled therein.

1.3 Structure

In order to account for and systematically map out patterns of (dis-)similarity within World Englishes, the present analysis will proceed through the following consecutive steps: Following this introduction, Chapter 2 provides an overview of relevant theoretical notions, concepts and models which have accompanied the formation of the analytical field of World Englishes – and which today still resonate within it. This concerns both more traditional concepts still providing prevalent terminology today (such as the ENL/ESL distinction) as well as more recent developments, both of which are included

as long as they apply to the study of 'national' standard(izing) varieties of English. Chapter 3, in turn, explores processes of linguistic co-selection, focusing on lexical and grammatical sequences in the form of *n*-grams. It will lay out statistical approaches for their extraction and evaluation, discussing potential challenges for the analysis different types and lengths of sequences, as well as present the association measures applied in the present study. Chapter 4, then, provides the methodological framework for the consecutive analysis. It presents the *International Corpus of English* as the data of choice but also addresses issues brought about by heterogeneity within the data before addressing ways of solving these issues within the present context. Furthermore, it will lay out the specific details of the extraction procedure for both lexical and grammatical sequences as well as the precise nature of the methods applied within the analytical approach and the way in which they will be applied as part of a coherent methodology. Chapter 5, finally, approaches the analysis of preference patterns across all varieties under scrutiny, discussing 40 separate sets of data overall comprising *n*-grams and POS-grams of both dynamic lengths as well as static-length 2-, 3-, and 4-grams. While preliminary findings will be described in each individual step, discussion of the findings will be deferred until the evaluation in Chapter 6, which summarizes and triangulates the findings and attempts to systematically map them onto the language-external models laid out before. Lastly, in Chapter 7, the study provides a final conclusion as well as prospects for further research.

2 Modeling World Englishes

English is a truly international language. However, the path towards its current state has been analyzed and modeled in different ways, which the present chapter attempts to retrace. It focuses on major forms of conceptualizing the spread of English, so as to provide the framework for the interpretation of the data in Chapter 5. The chapter proceeds by the analysis of influential concepts and models of the colonial expansion of English. Readers are, however, advised to consider that models can also be at work on a more implicit level – just as there are almost no theory-free approaches, so are models everywhere. Consequently, Schneider (2007: 19) cautions that "many seemingly descriptive statements […] entail culturally biased value judgments".

2.1 Monolithic and Pluricentric Models

Not all that long ago, one of the words decorating the cover of this book would have been considered a misuse of the English language: Englishes! The prevalent (internalized) model was that there existed but one English, which had been transplanted to new territories. While it inevitably acquired some local oddities along the process, English was still a monolithic entity and by no means pluralized.[2] A single 'Standard English' was taken to form the root of the genealogical tree of the language, and 'deviation' from the respected norm was not taken kindly to. It may appear unimaginable today that only a hundred years ago, most US citizens were still convinced that "no such thing as an American variety of English existed – that the differences […] constantly encountered […] were chiefly imaginary" (Mencken [1919] 1921; cf. also Bolton 2006a: 306). But even today, negative attitudes towards American English can be attested within some societal groups in Great Britain, e.g. university staff (Jenkins 2014), and stances of people in positions of power and prestige could long be considered as looking down upon the historically younger American variety. This is vividly

[2] While the term 'Englishes' was popularized by Kachru & Smith (1985: 210) after taking over editorship of the journal *World Language English* and renaming it to *World Englishes* (Bolton 2006b: 187), Strevens (1980: 90) already acknowledges a "marvelously flexible and adaptable galaxy of 'Englishes' which constitute the English language".

expressed in a *Times* report of a speech by Prince Charles in 1995 (quoted in Jenkins 2015: 5):

> The Prince of Wales highlighted the threat to "proper" English from the spread of American vernacular yesterday as he launched a campaign to preserve the language as world leader. He described American English as "very corrupting" and emphasised the need to maintain the quality of language […] Prince Charles elaborated on his view of the American influence. "People tend to invent all sorts of nouns and verbs, and make words that shouldn't be. I think we have to be a bit careful, otherwise the whole thing can get rather a mess."
>
> (The Times, 24 March 1995)

It seems astounding that the variety which for a long time provided the most numerous speech community of English, only gained acceptance through the economic and cultural dominance of the United States over the course of the 20th century. Given this fact, it should come as no surprise that the many other territories to which the language has been transplanted during colonial times have faced even stronger opposition. In Australia, for instance, the first dictionary printed within the country was published as late as 1976, representing a trend towards linguistic independence from Great Britain and ending the 'colonial cringe' (Jenkins 2015: 26). But even several years later, Quirk (1990: 6) still regards the respective standards as "rather informally established", and it has only been recently that tests of English as an international language have started to incorporate aspects of Australian English (cf. Jenkins 2014: 51). For all varieties not within "the traditional cultural bases of English" (Kachru 1992b: 356), i.e. the "New Englishes" (Platt et al. 1984)[3], even less readiness to accept the local standards has been the norm.[4]

Certainly, the emergence of local varieties did not occur spontaneously after the independence of the former colonies. Instead, it appears that users in all emerging or established varieties attempted to uphold the monolithic model even against evidence to the contrary. Conceptions of a unified, monolithic language have in fact accompanied the entire expansion of English from a relatively unimportant, nation-bound

[3] Platt et al. (1984: 2–3) define New Englishes along the four criteria of (a) having developed through the education system as a result of the fact that (b) English was not the language spoken by most of the population, but being nevertheless (c) used for a range of functions between speakers (and not only for international purposes), all the while (d) displaying features of nativization, i.e. new but stable local linguistic features.

[4] This applies both internationally as well as intranationally, as, for instance, the Indian struggles for the future role of English demonstrate.

language to a somewhat "unintended world language" (Schneider 2007: 1–2),[5] and the (perceived) 'dissemination' of English across the globe was often met with a "fervent triumphalism" (Pennycook 1998: 134). In this world-view, English brought enlightenment to "the savage", whose languages bore the "impress of degradation" (Trench 1891, quoted in Bailey 1991: 278), and which observers saw the "grossly impure structure of heathenism wrought into […], that the bare study of them often proves injurious to the mind of the European" (The *London Missionary Society*, 1826, quoted in Bailey 1991: 135–136). The major difference pre- and post-independence thus rather lies in who was seen to control the language. After the end of colonial rule, English was no longer under the purview of the colonial rulers, but instead belonged to the indegenous population. Given the age-old understanding of a single English disseminated throughout the world, this posed a major threat to the monolithic model.

After the loss of colonial influence, it was the genetics of the predominant speaker population which became the yardstick by which to assess the quality and reliability of a local English: Varieties influenced mostly by settler-driven colonial expansion were generally more readily accepted into the monolithic core of English under the roof of 'global English' or 'English as an International Language'. The New Englishes, however, were often (implicitly or explicitly) regarded as tainted by multilingualism, making them the 'illegitimate offspring' "not fully descended from Europeans" (Mufwene 2001: 108).

> The first-diaspora varieties [settler-driven colonialism, CK] of America, Australia, and New Zealand have often been regarded (explicitly or implicitly) as branches of a "Greater British" family of English dialects organically and naturalistically related to each other and the wider Germanic family. The "new" Englishes of Asia and Africa have been less comfortably placed at the family table; not least because such varieties are used by speakers of non-Germanic ethnicities in complex multilingual settings and have often had contentious colonial histories. (Bolton 2006a: 303)[6]

[5] Since no greater language planning was involved.

[6] Please note that the apparently neutal distinction between a 'first' and 'second' diaspora does not strictly hold. The temporal order it implies is somewhat misleading in that, for instance, the contact of the English language with South Asia predates the first-diaspora settings of Australia and New Zealand, while Western Africa faced speaker migrations even before North America (cf. Jenkins 2015: 7 for an overview and Bolton 2006a: 296 for precise dates of the second diaspora). The model thus essentially rests upon the distinction between predominant settler-driven colonialism and majority multilingualism and repackages it in apparently neutral terminology. Furthermore, distinguishing only two diasporas disregards the much earlier colonial expansion of English within the British Isles, the "first steps" as King (2006a) calls it. Consequently, Kachru et al. (2006: 3) divides the process into three diasporas.

The dichotomy between predominant settler-driven colonialism or multilingual language contact has resurfaced under several guises. Most prominently, it is reflected within the tripartite separation of countries into predominant native, second- or foreign-language use of English (ENL/ESL/EFL; Quirk et al. 1972). Outwardly neutral, this division has usually been understood to imply decreasing orders of proficiency, which glosses over substantial amounts of internal variation and disenfranchises the language use of highly competent ESL speakers by relegating the actual core of English to the ENL countries:[7]

> In many statements on global Englishes there is an inherent hidden tendency to regard and portray Britain and other ENL countries as the 'centers', thus entitled to establishing norms of correctness, and conversely, [post-colonial Englishes, CK] as peripheral, thus in some sense deviating from these norms and, consequently, evaluated negatively. (Schneider 2007: 19)

Several authors have attempted to arrive at more neutral terms, such as Gupta's (1997) five-way classification system: 'monolingual ancestral' (e.g. Britain, USA, Australia, New Zealand), 'monolingual contact' (e.g. Jamaica), 'monolingual scholastic' (e.g. India), 'multilingual contact' (e.g. Singapore, Nigeria, Ghana), 'multilingual ancestral' (e.g. Canada, South Africa). Another redefinition of central terms was proposed by Kachru (1998), who expanded the notion of nativeness two a binary system of 'genetic' and 'functional nativeness'. While these contributions have been well-received, they still failed to affect substantial modifications to the perceived difference between settler-driven language spread and multilingual contact, and the ENL-ESL-EFL model remained in active use.

The most influential tripartite model, and also the most inclusive approach, can be found in Kachru's (1992b: 356) 'Three-Circles model'. While the categorization into Inner, Outer and Expanding Circle varieties of Englishes[8] worldwide produces identical groupings as the ENL/ESL/EFL distinction, Kachru's concern was a political one, placing "greater emphasis on the Outer Circle, and also on the Expanding Circle"

[7] Major inaccuracies concern, for instance, that ENL communities exist within ESL territories (and vice versa) are ignored (1998), and similar caveats apply for EFL. Furthermore, (Jenkins 2015: 16–17) ENL is not as uniform as implied by the system but has different standards (McArthur 1998, Jenkins 2015: 16–17). More generally, monolingual nativeness is actually the global exception, which reduced the meaning of the overall distinction.

[8] Distinguishing varieties into the respective circles on the basis of (a) the types of spread of English, (b) the patterns of acquisition and (c) the functional domains of English.

(Schneider 2007: 13) and arguing for the functional "range" (administrative and judicial systems, business but also cohesive functions between family and friends) and "depth" of permeation of English in a community (social strata with access to English). Through the increasing sizes of the circles farther out from the 'center', attention is moreover drawn to the fact of the relative numerical superiority of Outer and Expanding Circle use of English. Yano (2001: 123; Figure 2.1) provides a version of the traditional representation (without speaker numbers) modified to include the dichotomous forms of nativeness.

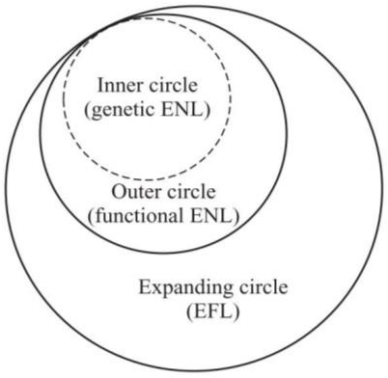


Figure 2.1: Modified version of Kachru's model to account for functional nativeness (Yano 2001: 123)

The more egalitarian approach spearheaded by Kachru advocated for a pluricentric perspective onto the landscape of English, in which regional centers and standards of the language exhibit varying degrees of mutual similarity. This soon clashed with the monolithic view, adherents of which expressed fears about language purity, lamented falling standards and saw the impending fragmentation of English.[9] At the heart of this lie concerns about 'Standard English' or 'the standard', which can be seen as another expression of the monolithic perspective. Actually defining this single standard presents more challenges than laying out "what it isn't": not a language, accent, style, register or set of prescriptive rules (Trudgill 1999). Several authors have attempted to explain the desire to defend this 'standard', arguing, for instance, that it might be "not simply a means of communication but the symbolic possession of a particular

[9] Worries about falling standards, however, are nothing new and have been expressed for over 300 years, cf. Mugglestone (2003) or Crystal (1997). Often, language purity is equated with the idea of order in general, as in the case of the conservative MP Norman Tebbit on BBC Radio 4 in 1985:

> If you allow standards to slip to the stage where good English is no better than bad English, where people turn up filthy at school […] all these things tend to cause people to have no standards at all, and once you lose standards then there's no imperative to stay out of crime. (Cameron 1995: 94, quoted in Jenkins 2015: 79)

community" and a symbol of unification (Hudson 1996: 33) Thus, Widdowson (1994: 381) surmises that "to undermine standard English is to undermine what it stands for: the security of this community and institutions." Essentially, this means that "standard English has almost come to have a life and power of its own" (Mesthrie & Bhatt 2008: 14–15), manifesting the prestige of the speakers said to represent it (Milroy 2001: 532).

It took some time and one of the most notorious controversies in the field – the Quirk-Kachru debate in the journal *English Today* (Quirk 1990, Kachru 1991) – to establish that 'indigenization' of the English language and ensuing pluricentricity did not necessarily have to entail fragmentation of English and mutual unintelligibility in international contexts.[10] In essence, the controversy was sparked by Quirk (1990), who denounced all accounts of pluricentric Englishes as "half-baked quackery" undermining the importance of Standard English. Quirk instead argued for the importance of institutionalization, while heavily implying that non-native varieties cannot achieve this state. To his understanding, "radically different internalizations" of English between first- and second-language users were a cause for alarm, since they threatened the stability of English as a tool for international communication. The only eventual effect of acceptance would thus be a loss of respect for Standard English.[11] Kachru (1991), in his rebuttal (and later expansions in Kachru 1992b), points out "fallacies about the uses and users of English" in assuming that all (second-language) learning of English were geared towards access to and immersion in American and British culture and language and that it were the goal of second-language speakers to approximate respective standards in their everyday communication. Instead, he argues that most speakers only cared for an outside norm when international communication is concerned, and otherwise, intra-national concerns were rated higher. This is also why he refuses Quirk's demand for English education to be performed through native

[10] Disregarding the fact that (partial) unintelligibility within the English-speaking world is a historical fact: "[H]istorically or in terms of the present day situation, [..] it has always been the case that some English speakers have been at least to some degree unintelligible to other English speakers." (Kachru et al. 2006: 6).

[11] Several years before this, Quirk (1962: 17–18) still had a more liberal perspective on correctness: "English is not the prerogative or 'possession' of the English […] Certainly, we must realize that there is no single 'correct' English".

speakers (cf. also Seidlhofer 1999 for a discussion of the disadvantages of native-speaker teachers).

For all the controversy, however – and certainly it has shifted linguistic reasoning – actual developments in this regard rest on the users of a particular variety claiming their right to ownership of their standard(izing) English. As late as 2008, Mesthrie & Bhatt see the eventual resolution of the debate as a future concern, to be decided case by case by the users of English within a particular territory:

> Ultimately the Kachru-Quirk controversy can only be resolved outside the ivory tower, by the attitudes and actions of parents, pupils, teachers, administrators and the like. Linguistic hegemony power can be contested, but it is seldom dismantled by reason alone. (Mesthrie & Bhatt 2008: 208)

At this point in time, it appears that the controversy has been postponed: For pluricentric Englishes, the paradigm of World Englishes established by Kachru has been fully accepted and the terminology has become "the most neutral label for the discipline and its objects in recent years." (Schneider 2017: 40) Within the World Englishes paradigm, then, every speech community is seen to have the "right […] to deviate" (Gomes Matos 1998: 15).[12] Kachru acknowledges the divide in perspectives onto English on a worldwide scale:

> Those who see the canon of English literature or of Englishness as relatively fixed, a starting point with distance measured from it to far-flung reduplications of the pattern, are those who view the spread of English with "fear and aversion," while those who see it from the world Englishes paradigm react to the same data with attitudes of "celebration and esteem." (Kachru et al. 2006: 6–7)

In contrast to the acceptance of pluricentric standards within World Englishes research, the 'international' dimension of English may still be more influenced by notions of a singular standard (cf. Williams 2007: 402 and Schneider 2017: 39 for attestations to the fact that the distinction into 'good' and 'bad' English and a single standard is still going strong). Initially, the future was believed to bring the eventual formation of a shared global standard: Trudgill & Hannah's (1982) *International English* presented several (new and New) Englishes as *Varieties of Standard English* (Bolton 2006a: 291). In a similar vain, McArthur (1987), Görlach (1988) and Crystal (1997) supposed some form of global 'common core' (Quirk et al. 1985: 16), which was expected to form through

[12] A catchy term, but immediately restricted by the author to "noncrucial areas that do not affect intelligibility or communication".

mergers of regional norms. Different terms have been proposed for this and used somewhat interchangeably, be it International English (Görlach), World Standard English (McArthur) or World Standard Spoken English (WSSE; Crystal). For all accounts, virtually none of these developments have taken place beyond the further dissemination of the two major national standards of America and Britain. Despite hopes for the contrary – "if ever possible, English for global use should be dissociated from the norm of any English-speaking society" (Yano 2001: 129) – these still almost exclusively shape any standards for global English and provide the benchmarks. While the notion of globalized learners and the description of English as an international lingua franca have gained traction (cf. Seidlhofer 2001 and the contributions in Deshors 2018a), Mufwene (2010) regards global standards as an illusion, and instead foresees further diversification of English:

> The universal trend has been for the prevailing language to diversify, especially in the spoken form […]. Worse for the wishful thinking, even Standard English itself, which is controlled by several institutions, has diversified. It seems utopian to me to conjecture that speakers of 'native Englishes' will be accommodating, midway, all those other populations speaking their language with a foreign element, and will thus contribute to the development of some WSSE, in order to guarantee mutual intelligibility. (Mufwene 2010: 46–47)

Halliday (2006) seeks to reconcile the different notions of the international and global spread, arguing for, one the one hand, international and pluralized World Englishes and otherwise a singular Global English, which "has expanded – has become 'global' – by taking over, or being taken over by, the new information technology". He presents these two paradigms as overlapping notions, with the potential for World Englishes to influence the globally dominant standards.

> Infotechnology seems still to be dominated by the English of the Inner Circle; under pressure, of course, but not seriously challenged, perhaps because the pressures have no coherent pattern or direction. If the Englishes of the Outer Circle had more impact on the global scene, those who monopolize the media would no longer automatically also monopolize the meanings. If African and Asian varieties of English are not simply vehicles for their regional cultures but also their communities' means of access to a culture that is already in effect global, those who speak and write these varieties are not constrained to be only consumers of the meanings of others; they can be creators of meanings, contributors to a global English which is also at the same time international. (Halliday 2006: 363)

Still, there is little doubt that this largely concerns an interaction between many local varieties and norms of American English (cf. e.g. Butler 1997: 107–109, Mair 2013: 260, Buschfeld & Kautzsch 2017: 115).

2.2 Sources and Processes of Diversification

World Englishes research concerns "linguistically identifiable, geographically definable" (Kachru 1992a: 67) varieties across the world. As is most common in World Englishes research, this also pertains to 'national' varieties within the present study, i.e. generalization across a state-bound form of English. Tracing the precise ways in which the varieties under scrutiny have diversified lies beyond the limitations of this publication (but cf. e.g. Platt et al. 1984, Schneider 2007, Mesthrie & Bhatt 2008, Sand 2008, Kortmann & Schneider 2008, Mukherjee & Hundt 2011, Filppula et al. 2017, as well as Deshors 2018a for recent critical approaches to the concept of 'national varieties'). Still, some general effects can be observed. In particular, the ways of transplantation of English to the colonies are prime factors for the English retained:

> Each colonization style has determined particular patterns of interaction between the colonizers and the indigenous populations as well as the particular kind of economic structure that is now in place. (Mufwene 2002: 168)

Mufwene (2001) distinguishes broadly between three 'styles' of colonial systems: Trade, exploitation and settlement. The latter subsumes a fouth 'plantation' type which Schneider (2007) regards as sufficiently distinct to inform a separate category. While some trade contexts developed, over time, into exploitation or settlement colonies, they are most strongly characterized by the limited and sporadic interaction between varieties of English and local languages. Settlement colonies are distinguished by the large-scale settlement in which migrants from a wide range of backgrounds (and dialects) mixed, but indigenous languages played only a marginal role and were displaced, resulting in the gradual production of local or regional monolingualism (Mufwene 2002: 169). Plantation colonization had similar results within a clearly demarcated settler population, but additionally saw the importation of indentured or slave labor but more severely restricted communication between the groups in contact. Within exploitation colonies, colonists formed a segregated elite within a society run to varying degrees through English, with the language mainly introduced to provide a

"managerial group sandwiched between the colonizers and the colonized." (Schneider 2007: 24)[13]

The different colonial systems with diverging contact situations in turn grew to different linguistic outcomes. While early trade colonies resulted in pidgins and plantation colonies in creolization processes, settler-driven colonization was strongly characterized by the contact, mixture and formation of dialects (Trudgill 2004: 13). Linguistic outcomes are, respectively, characterized by a higher retention of non-standard features from dialects from the British Isles. While dialect contact also existed in exploitation colonies, the introduction of English in a scholastic fashion to the indigenous population led a lower spread of non-standard features, but conversely to stronger influences of contact between languages onto the structure of the local English. The two (or more) languages involved can be seen in a H (high) and L (low) configuration, for which English almost always came out on top, as the H language and superstrate for the local substrate influences (cf. e.g. King 2006a: 30–31). It is useful to recall that the superstrate was highly diverse due to the parallel dialect contact, and thus presented a 'moving target', as Mesthrie (2006: 277) advises. On the basis of this continuation from standards and dialects, the levelling of dialects through contact and, particularly in non-settler colonies, strong language contact, further innovative changes could occur.

Additionally, for each individual variety, the backgrounds of the speakers transplanting the English may be relevant. In some shape or form, the British Empire spanned almost four centuries, during which the standards and dialects of the 'input variety' changed (cf. e.g. Heller et al. 2017 for potential effects), and colonies received influx from different speaker groups with various backgrounds (cf. King 2006b: 26–27, Mesthrie 2006: 282), all of which provided the local 'pool of linguistic features' (Mufwene 2001). These features are, in turn, available as potential features of an

[13] A system which Brutt-Griffler (2002), at least before the Advisory Committee on Education in the Colonies (1923), regards as essentially an effort to run an empire "on the cheap" (2002: 86), "a policy of limiting the spread of English to what was minimally necessary to running a colonial empire" (2002: 105). "English-medium instruction was generally favored […] only to the extent that it forstered a locally recruited civil service, or, in some instances, locally trained clerks for commerce" Bolton (2006a: 297), and consequently "[t]he English language spread to Africa and Asia by political and economic means, not demographic […] English never became the language of industry and of the major agricultural districts; instead, it was the language primarily of the colonial administration" (Brutt-Griffler 2002: 117).

emerging variety, mediated through the force of a 'founder effect' (Mufwene 1996: 84), i.e. the pull exerted by the original linguistic constellation. Over time, any local English will change due to what Mufwene (2005) calls "imperfect replication", i.e. that all learning never encompasses the entirety of a system in its present configuration. Schneider (2007) provides a schematic overview of these "sources and processes leading to the formation of post-colonial Englishes" (Figure 2.2). However, within most of these processes, Mufwene (2010: 46) reminds us that "the burden has been on speakers of 'non-native Englishes,' which are generally treated as 'deviations' […], to 'improve' their intelligibility – not the other way around."

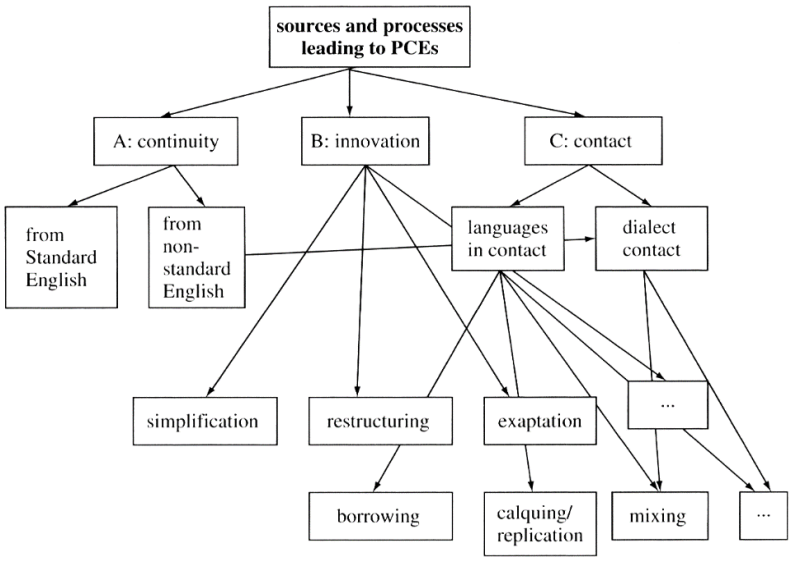


Figure 2.2: Sources and processes of the formation of post-colonial Englishes (Schneider 2007: 100)

Consistent with the notions of localized World Englishes vs. centralized Global English, diversification in varieties of English concerns speech more than writing, which "has its own conventions, some of which have little connection with features of speech." (Mesthrie & Bhatt 2008: 41) Mair (2007: 97) observes this for the two major international norms, in that "British and American standard English may differ quite considerably in speech in areas in which they resemble each other closely in writing." Taking Mahboob's (2017: 17) framework of language variation as a guide (Figure 2.3), speech, for the most part, encapsulates more local concerns of lower social distance and everyday/casual discourses while specialized/technical discourses favor writing over global distances. Of course, there are exceptions to this, for example in the pull of some popular dialects disseminated through international media, everyday concerns being covered in blog posts and written tutorials, and suchlike. On average, however,

writing attracts convergence, not least because the possibility of editing and its impact in reducing localized forms should not be underestimated.

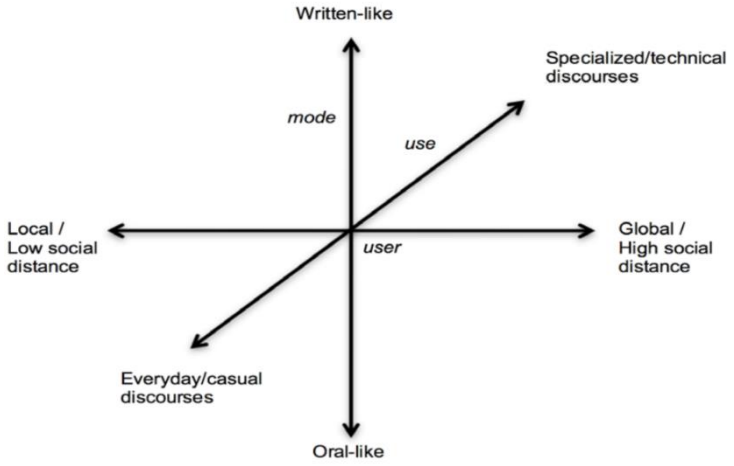


Figure 2.3: 3-D framework of language variation (Mahboob 2017: 17)

This editing process may, moreover, "be of specific importance in the case of published writing, since […] ESL-speakers are often advised to seek native-speaker counsel before publishing an English text" (Götz & Schilk 2011: 83). Similarly, Schneider (2007: 82) observes that "in formal speech production, and most characteristically in writing, a 'standard' form of language can be observed which is largely devoid of localisms."

2.3 Proximity Effects in World Englishes

Proximity of some kind has always factored into the description of varieties of English. From the conceptualization of 'brothers' of the 'English race' (Bolton 2006a: 296–297) to favoring settler-driven colonies more than the others, feelings of regional or cultural proximity have always shaped understandings of English throughout the world. In linguistic handbooks, this finds reflection in the arrangement of chapters, which historically go by world regions even if local histories and current dynamics may be different. From the earliest visualizations, proximity has been a prime descriptor, e.g. in Strevens's (1992) world map of English (Figure 2.4), in which "geography rules" (Schneider 2017: 38) by reducing differences between Englishes to two branches of the genealogical tree.

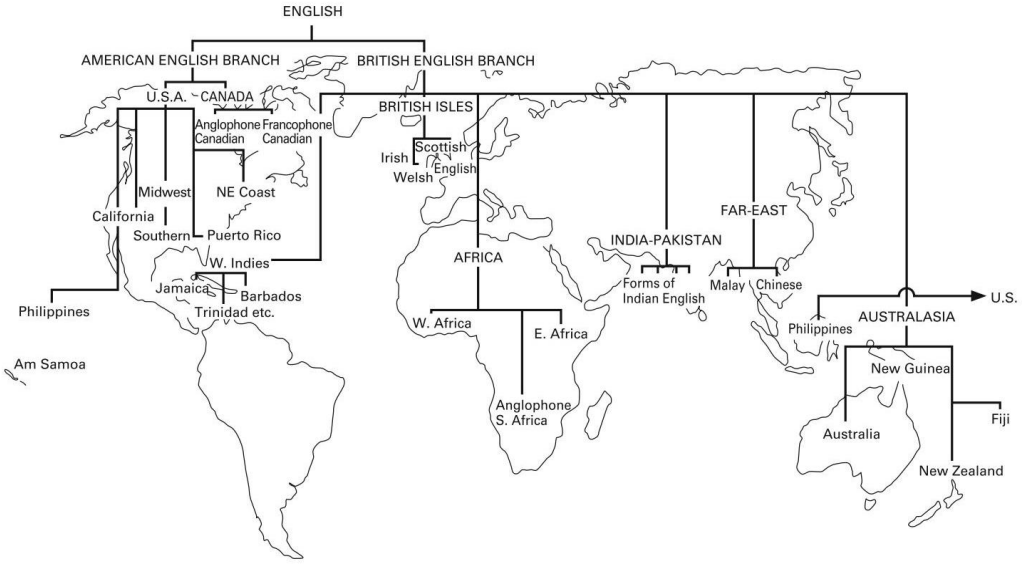


Figure 2.4: Strevens's (1992: 33) world map of English (visualization by Jenkins 2015)

Both Görlach's (1988) 'Circle model of English' (Figure 2.5) as well as McArthur's (1987) 'Circle of World Englishes' (Figure 2.6) are further influential models incorporating aspects of proximity. Similar in many regards, they revolve around some forms of global standard, which are of a more theoretical nature, and then distinguish regional and sub-regional standards (cf. Schneider 2017: 42–43 for further discussion). Proximity also features in typologically-inspired discussions of linguistic universals, particularly in the notion of 'areoversals', i.e. "features common in languages which are in geographical proximity" (Szmrecsanyi & Kortmann 2009: 33).

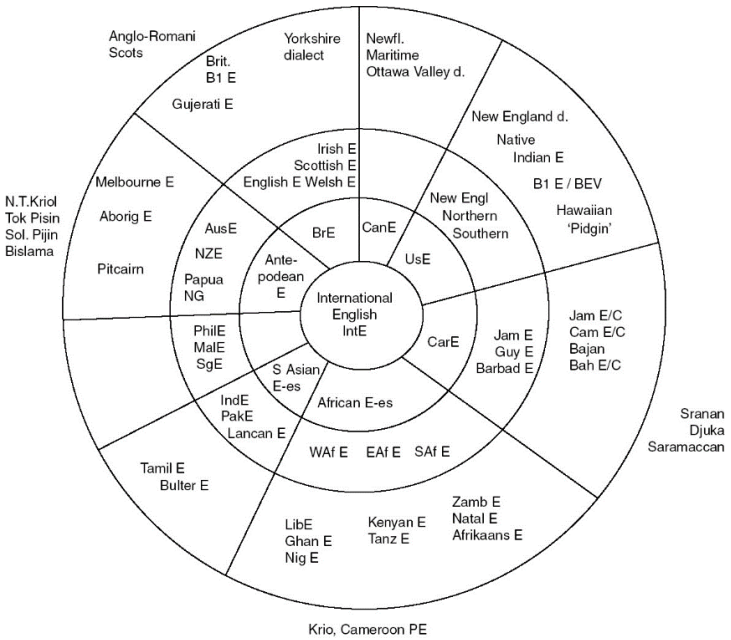


Figure 2.5: Görlach's (1988) Circle model of English

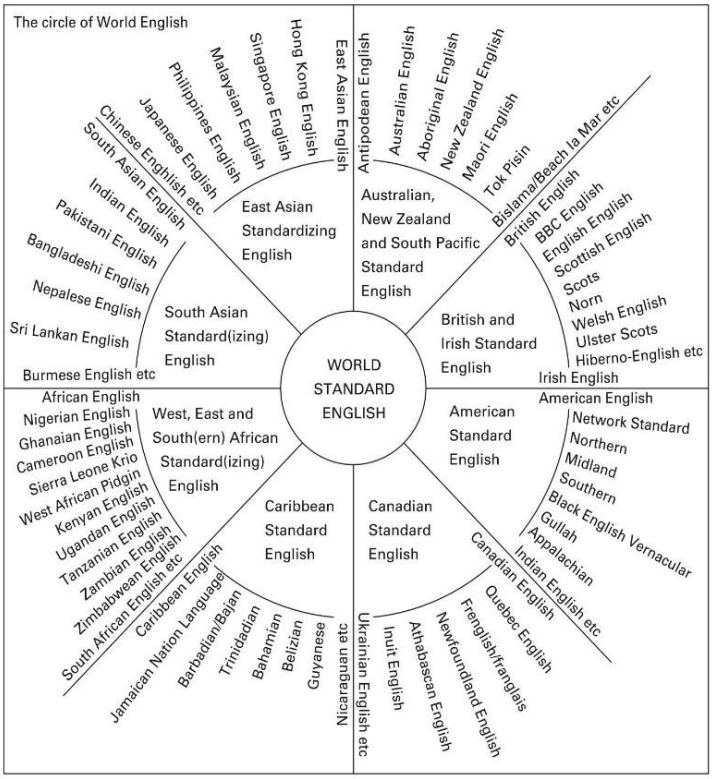


Figure 2.6: McArthur's (1987) Circle of World English

A notion which has more recently gained traction is that of epicenters, i.e. "focal points in the pluricentric constellation" (Mair 2013: 257). Suggested originally by Leitner (1992), the concept has more recently been formalized by Hundt (2013):

> The consensus definition of what an epicentre is so far involves two dimensions: a variety can be regarded as a potential epicentre if it shows endonormative stabilization (i.e. widespread use, general acceptance and codification of the local norms of English) [...] on the one hand, and the potential to serve as a model of English for (neighbouring?) countries on the other hand. (Hundt 2013: 189)

For a variety to be a potential epicenter, it has to be established and accepted as a local norm (endonormative stabilization), entailing at least some institutionalization, and it also has to exert a modeling effect on other varieties in its proximity. Epicentral influence may take place, among other aspects, through the reception of media, economic ties, frequent migration or (semi-)direct export of linguistic norms (e.g. testing schemes, dictionaries, teachers[14]). Embracing the picture of seismic shockwaves emanating from the centerpoint of an earthquake and radiating outwards, the model also assumes a weakened effect over increasing distances, similar to how "waves emanating from an earthquake epicenter have a more or less immediate (and damaging)

[14] Export of teaching norms, for instance, can be observed from India to Sri Lanka through teachers trained at Indian universitites but migrating to the neighboring country for employment, thus likely exporting Indian English norms and exerting epicentral influence.

effect on the adjacent surroundings" (Hundt 2013: 189). In turn, varieties in closer proximity will be influenced more strongly (Indian English being an epicenter for several nearby South Asian varieties; cf. Leitner 1992, Gries & Bernaisch 2016, Heller et al. 2017) than those farther away (Singaporean English being relatively spatially removed from other Englishes; Heller et al. 2017).

While metaphorically modeled on geographical aspects, the notion extends into cultural proximity as well. In this regard, a variety can have epicentral influence on others by virtue of the respective nation's economic power or cultural allure, as in the case of American English. Similarly, what is usually metaphorically labeled 'the West' in political and cultural reports is also strongly based on an epicentral understanding of American and European ideas and values. Nations within this epicentral sphere invariably have closer connections to each other than with many other cultural spheres, and cultural contact furthers linguistic contact and approximation. Within postcolonial Englishes, this mutual contact and the epicentral influence of one localized norm onto another may accelerate structural changes, producing "postcolonial Englishes squared" (Bernaisch & Lange 2012: 13). Moreover, a variety may be under the effect of several epicentral influences at once, as Mair (2013) describes:

> As many New Englishes have historically developed from British input and remained under British influence for long stretches of their historical development, Britain and its norm are considered central and thus the yard-stick […]. Today, the American standard has a global reach and the potential to affect all other (standard and non-standard) varieties of English. The British standard has a similarly global reach (with the not insignificant rider that, although widely known, it is largely irrelevant in the US), but true complexity arises because in addition there are now several transnational epicentres, for example in Australia […] or in India. (Mair 2013: 259)

Incorporating an aspect of cultural influence and media presence into the concept of epicenters is also regarded as crucial from the perspective of Mair (2013), whose 'World System of Englishes' aims at highlighting power differentials between varieties of Englishes worldwide. In this system, he distinguishes a 'hyper-central variety' or 'hub' variety of Standard American English (similarly to the hub-and-spokes models by Görlach and McArthur, but with an actually existing global variety) from consecutively less influential but ever more numerous Englishes: British English stands out from a number of around ten 'super-central' standard Englishes (e.g. Nigerian, Indian or Australian English) of 'transnational relevance'. The majority of Englishes are, however, of limited regional influence and mostly distinguished by their degrees of

institutionalization and speaker numbers. They consequently fall either into the 'central' category (Irish, Jamaican, Ghanaian, Kenyan, Sri Lankan, New Zealand English, etc.) or as seen as 'peripheral' varieties (Maltese or Cameroon English, etc.). Linguistic effects will more commonly be 'downward', in that the more central a variety, the stronger its effects onto those varieties further within the periphery. While it appears counterintuitive, adoption of actual linguistic structures from a potentially epicentral variety does not clearly correlate with the degree of positive attitudes expressed towards it: Heller et al. (2017: 137) observe epicentral influences of Indian onto Sri Lankan English even though speakers of both varieties do not express overly positive attitudes towards the respective other variety (Bernaisch 2012, Bernaisch & Koch 2016).

2.4 Evolutionary Dynamics of World Englishes

Increased use of English within a local speech community fosters the development of local norms and potential standards. Depending on a complex and interrelated mixture of factors, primarily "language policies and language attitudes, globalization and 'acceptance' of globalization, foreign policies, and the effect of the sociodemographic background of a country, and of course colonization and attitudes towards the colonizing power" (Buschfeld et al. 2018: 23), a local speech community may drift more towards the global domain or the locally characteristic. Some authors have attempted to model this evolutionary process, mapping increased attitudinal distance to the colonial situation and the culture and language of the colonizers to differentiation from their respective norms (as supplied by the colonial situation and present influences). Among these are Trudgill's (2004) deterministic model, which is, however, mostly disregarded within World Englishes research,[15] as well as Moag's 1992 "life cycle of non-native Englishes". Most influential, however, of all these models has clearly been Schneider's (2007, 2010) Dynamic Model. Schneider's model rests on to the notion of

[15] Trudgill (2004) deterministic model only truly applies to 'tabula rasa' situations, and as such appears largely hypothetical in nature. Its disregard for identity formation has sparked a lively debate in the *Language in Society* journal in 2008 (issue 2) between Trudgill and various commentators. Within a series of eight 'discussion' articles, most responses underscored the importance of identity and accomodation over a deterministic perspective. However, there is disagreement on when this affects the emerging variety most strongly.

'social identity', i.e. "the systematic establishment and signification, between individuals, between collectives, and between individuals and collectives, of relationships of similarity and difference", which entails a "construction and reconstruction by symbolic linguistic means" (Schneider 2007: 26). Thus, the Dynamic Model centrally rests on the importance of mutual accommodation and subsequent identity rewritings of what Schneider broadly distinguishes as settler and indigenous speech communities evolving into settler (STL) and indigenous (IDG) strands of the shared contact situation: It is thus characterized by

> the assumption that […] speakers keep redefining and expressing their linguistic and social identities, constantly aligning themselves with other individuals and thereby accommodating their speech behavior to those they wish to associate and be associated with." (Schneider 2007: 21)

The force behind the evolutionary process is "the reconstruction of the group identities as to who constitutes 'us' or the 'other'" (Schneider 2007: 29). The model "postulates a […] monodirectional causal relationship" in that "*sociopolitical* and *historical background*" shape the "*identity* constructions", which in turn "are decisive for the *sociolinguistic conditions* which shape the communicative settings, and on these, in turn, the resulting *linguistic effects* […] are dependent." (Schneider 2014: 11) Most importantly, this process is seen as largely uniform between different contact settings:

> [D]espite all obvious dissimilarities, a fundamentally uniform developmental process, shaped by consistent socio-linguistic and language-contact conditions, has operated in the individual instances of relocating and re-rooting the English language in another territory, and therefore it is possible to present the individual histories of PCEs [Post-colonial Englishes, CK] as instantiations of the same underlying process. (Schneider 2007: 5)

In so doing, the model posits attitudes and identity as the overarching answer to de Klerk's (1999: 315) question of "[w]hen does a substratal feature assert itself sufficiently to overcome the fear that if deviations are allowed, the rules will be abandoned and chaos will ensue?" While de Klerk asks whether this will be "when speakers use it often enough to silence or exhaust the prescriptors", Schneider draws on a broader context of societal factors.

The process of identity (re-)definitions is modeled by Schneider in terms of five successive phases, to be briefly discussed in turn below: The initial (foundation) phase marks the actual transplantation of English to a new territory and thus entails the earliest contacts between dialects of English with indigenous languages. Feelings within

both social strands are clearly those of belonging to one's original nation, with the STL group regarding themselves as British colonists, usually on temporary mission. Linguistically, apart from a few contacts, most changes occur within the STL strand in the form of koinéization, i.e. the forming of a 'middle-of-the-road variety', an effect "largely confined to informal, oral contexts, the spoken vernacular, and it is strongest in settlement colonies, where large numbers of speakers predominantly from the lower social strata are involved" (Schneider 2007: 35). The IDG strand has no particular influence on any linguistic effects other than providing toponyms (place names) as loan words, and there are only few indigenous people who see a need to learn the STL strand's language.

The second phase is outwardly characterized by the colony being firmly established and attracting a growing number of English settlers/speakers to the new territory. These bring with them English forms, and while non-standard ones might be integrated into the local koiné, the standard forms serve as a (potentially updated) model for how English is supposed to be used properly. Norm orientation is thus usually clearly directed towards British English (exonormative stabilization):

> In teaching matters, and to the extent that reflection is spent upon questions of language correctness at all, they share a conservative and unaltered, though increasingly distant cultural and linguistic norm orientation, unsupported by local realities (Schneider 2007: 38)

However, the self-perception of the STL residents is slowly changing to incorporate the local in addition to their English heritage. This is mirrored by the IDG group, who incorporate English elements into their identity, with bilingualism becoming more prevalent if restricted to a social elite. The increased contact between the two strands results in stronger borrowing reflecting the 'English-plus' mentality and particularly leads to an incorporation of terms for flora, fauna and cultural practices. These terms might become recognized both inside and outside the colony as Americanisms, Indianisms, etc. (cf. Schneider 2007: 39). A few changes may occur on the morphological and syntactic levels, which, however, "even if they are consistent and systematic, are likely to pass largely unnoticed and unrecorded, being restricted to the spoken vernaculars in the beginning" (Schneider 2007: 40).

The third phase presents "the central phase of both cultural and linguistic transformation" (Schneider 2007: 40) within the Dynamic Model. For the STL strand, the

feeling of being at home in both worlds simultaneously decreases in favor of a strong association with the colonial territory. This furthers their wishes for greater autonomy, leading in many cases to a strive for political independence, which in turn is a "precursor to linguistic independence" (Greenbaum 1996b: 11). Linguistically, this phase is the most vibrant one, with structural changes increasing greatly thanks to the predominant feeling of being separate from the former mother country. Crucially, the local English undergoes structural nativization, i.e. changes on "levels of organization which do not carry referential meaning, namely morphology and syntax", thereby "developing constructions peculiar to the respective country" (Schneider 2007: 44). While contrasts between STL and IDG are usually reduced to a sociolinguistic distinction, the "labor of approximating each other tends to rest predominantly upon the IDG strand group" (Schneider 2007: 45). A particular influence lies in structural calquing, a process by means of which

> [g]rammatical or conceptual material is transferred from a "model language" to a "replica language. In this process speakers seek equivalence relations between both languages and thus transfer both patterns (recurrent discourse pieces) and functional categories. What happens in most cases [...] is that a minor pattern of the replica language gets refunctionalized under the impact of some element or pattern of the model language. This process frequently follows the principles of grammaticalization and results in a new grammatical system in the replica language. (Schneider 2007: 108)

Within the STL strand, both conservative and innovative speakers are found, and discussions about the status of the new (and now very visible) variety abound, with a strong complaint tradition arguing for the retention of the British standard and further exonormative orientation. Schneider (2007: 43) points out, however that this may well represent "class struggles in disguise", since the discussion is mostly held within the upper classes and about the written norm, while "adoption of IDG strand features by STL strand speakers is more likely to occur in the lower social strata and in informal communication" (Schneider 2007: 42).

Whether a variety enters the consecutive phase of endonormative stabilization is largely a result of the outcome of the aforementioned political and social argument about changing norms and the strength of the complaint tradition. If traditionalists win the upper hand and assert the British dominance over formal usage, a variety is likely to remain in phase 3. Embracing the innovations, on the other hand, and thus owning the local variety will lead to it entering this stage, with the complaint tradition

becoming a minority position. It also "typically follows and presupposes political independence", but cultural self-reliance is the more decisive factor (Schneider 2007: 48). Culturally, this phase is characterized by the local population feeling as members of a newly born nation, with rigid ethnic distinctions becoming less important (Schneider 2007: 49). Literary creativity in the local English may be sparked by this new identity, addressing the cultural hybridity and the use of English (Schneider 2007: 50). Local linguistic forms are evaluated positively and are likely to be found in increasing use, reinforcing the feeling of political autonomy (Schneider 2007: 52), and codification of the local English is likely to ensue. This is usually expressed by 'English in X' becoming 'X English', showing "different conceptualizations of the status of the language" (Schneider 2007: 50). This may even lead to actual internal differences being downplayed in order to fulfill "a young nation's desire to imagine 'national singularity and homogeneity'" (Schneider 2007: 51). While this stage can be entered gradually through the processes within the previous stage, there is commonly an 'Event X', a quasi-catastrophic occurrence at least as far as the feeling of connection to Great Britain and its authority over English is concerned. This may take various shapes, as the feeling of abandonment prevalent in Australia during WWII or the Southern Indian protests to retain English as an official language beyond its intended abolishment 15 years after independence.

The final stage (differentiation), as of yet only fully observable in Inner Circle Englishes, sets in once it becomes apparent that the fictional homogeneity of stage 4 was just that. This shines a spotlight on internal variation and brings out a "composite of subgroups" and social networks using what was conceived of as a single variety in a multitude of different ways (Schneider 2007: 53). New regional variants are recognized, but with a central standard variety in place, and differences between STL and IDG "are likely to resurface as ethnic dialect markers" (Schneider 2007: 54). This reduction of homogeneity is only normal, since "degree of uniformity [is] in inverse proportion to historical depth" (Trudgill 1986: 145). Difference is tolerated, however, since "[b]y this time, the still somewhat shaky, slightly questioned independence [...] has given way to the secure existence of a stable young country" (Schneider 2007: 52).

Testing the Dynamic Model

Over the last years, the Dynamic Model has become one of the most influential perspectives on the evolution of World Englishes. Beyond the factor that it models a diverse range of contexts as one relatively uniform process, one main attraction of the model is that it claims "a causal relationship between historical conditions, socio- psychological consequences, and linguistic effects", so that the extralinguistic context will "eventually find reflection in linguistic/structural effects" (Schneider 2017: 45). In turn, this implies testability of the evolutionary phases along linguistic features, in that "linguistic structures are indicators of varietal progress in [the] evolutionary cycle." (Gries et al. 2018b: 249) It might thus be expected that varieties further along the cline will display a greater amount of structurally local characteristics than those oriented more closely towards outside norms. However, this may be too substantial an abstraction, since divergence from the 'input' is a diachronic matter building on locally particular initial setups. While a general mapping of difference to evolutionary stage might be found diachronically, this does not necessarily entail that several varieties of similar evolutionary status display the same amount of structural change.

Notwithstanding these caveats, many studies have attempted to test whether the universal process described in the Dynamic Model also leads to systematic linguistic outcomes across varieties. The wealth of studies performed with reference to the model cannot be replicated here (cf. in particular Schneider 2014 for reflections by the original author), but brief mentions of some studies with large-scale, comparative approaches similar to the present analysis are in order. These previous studies appear to mostly confirm predictions made on the basis of the Dynamic Model. For instance, Mukherjee & Gries (2009) find evolutionary stages to correlate to increased divergence in terms of the complementational behavior of ditransitive verbs. By contrast, however, Edwards & Laporte (2015: 135) discover an inversely proportional relationship between the degree of institutionalization and similarity to ENL varieties with respect to the patterning of the *into* preposition. While these results are conflicting in terms of convergence vs. divergence over the stages of the Dynamic Model, they still show a directed process to apply through consecutive stages. Correlations between evolutionary stages and patternings of varieties along linguistic features have also been observed by Collins (2012; singular existential *there*), Schneider (2012; complement

clause patterns), Lunkenheimer (2013; morphosyntactic features of the WAVE data, cf. Kortmann & Lunkenheimer 2013) and Werner (2013; preference for past over present tense in time adverbials).

Turning to lexical co-occurrence, however, Gries & Mukherjee (2010) fail to find a clear correspondence between stages and (dis-)preferences for *n*-grams between British English and three Asian varieties, prompting them to ask whether features on this level may be too topic-dependent to provide consistent results within the Dynamic Model. Schneider (2014: 10) acknowledges that "as in any case of model-making, comparison and abstraction the question is one of granularity, i.e. of how closely one wishes to look, how much attention is to be directed to similarities or differences, respectively". The efficacy of the Dynamic Model in predicting linguistic outcomes may consequently be dependent on the granularity of the feature under scrutiny: Features on a more abstract level might produce results that are more consistent with the model, while lexical analyses may be influenced more strongly by text/topic-dependent factors (cf. also Bernaisch et al. 2014). Schneider (2017: 51) similarly expects that "some [features] are more variety-specific, whereas others lend themselves more strongly to generalization and can be accounted for in a wider, possibly areal, cognitive or functional perspective."

The present Chapter has presented the major ways of categorizing World Englishes on a national level. It initially addressed models of a more implicit nature, such as the age-old monolithic perspective onto English. It discussed the debates and controversies that sparked from these foundational models and helped to lay the foundations of modern World Englishes research. The chapter then discussed two major perspectives onto the post-colonial landscape of national varieties of English, and identified an approach based on regional and cultural proximity as well as an evolutionary perspective. While the latter is explicitly formulated in terms of Schneider's (2007) Dynamic Model, the former somewhat more implicitly underlies models of the formation of regional standards. Both models have been used to inform predictions about patterns of similarities between World Englishes, which the present study aims to address in a data-driven fashion. The next chapter will introduce *n*-grams as the linguistic object which only enable the large-scale, bird's-eye differentiation of World Englishes attempted within the present study.

3 Collocational Sequences as a Discriminatory Measure

Items in language tend to occur in company, and some combinations are more likely than others. The present chapter discusses the linguistic object of study, sequences of both word forms and parts of speech in the form of (lexical) *n*-grams and (grammatical) POS-grams. These are conceptually strongly intertwined with the notion of collocation, which had been an interest of language description even before the advent of modern linguistics but has experienced a re-interpretation and great methodological sophistication through corpus research. In Section 3.1, the present chapter will contextualize the linguistic phenomena under scrutiny within the larger domain of lexicogrammar and provide definitional criteria. Section 3.2 will then discuss the relevance of patterned language as a discriminatory tool within the context of varieties of English. Finally, Section 3.3 lays out the concrete operationalization of *n*-grams and POS-grams within the present study in anticipation of the larger methodological approach to be discussed in Chapter 4, with a particular focus being placed on the statistical evaluation of relevant sequences.

3.1 Aspects of Co-occurrence

3.1.1 Foundations of Co-occurrence Research

Almost all language use consists of more than single words. For substantial lengths of time, however, any potential for syntagmatic dependency between items has usually been confined to the description of syntactic structures which lay down the slots into which individual lexical elements may be inserted (the 'slot and filler model'; Sinclair 1991: 109–110). Within this "traditional view that knowing a language involves two types of knowledge: rules and lexical items – period" (Warren 2005: 35), "many – perhaps most – linguists […] regarded the lexicon as a marginal part attached to grammar" (Johansson 2011: 19).

> [T]here is always this basic distinction, of a component which produces patterns of organization and a component which produces items that fill places in the patterns; the items

> tend to be chosen individually, and with little reference to the surrounding text. (Sinclair 2000: 191)

Given this mindset, connections between lexical choices were for the most part merely modeled on the basis of some form of selection restriction governing the use of lexical items in context (i.e. *drink* + a fluid) or otherwise relegated to highly fixed sequences in the form of "opaque and 'funny' idioms" (Lindquist 2009: 91), i.e. structurally, semantically or pragmatically relevant, discrete and usually idiosyncratic units. Certainly, even some earlier studies showed a relatively modern understanding of the relevance of the lexical co-text,[16] such as Cruden's 1737 concordance and collocationanalysis of the bible (Kennedy 1998: 91–92).[17] The typical perception of patterned language, however, relied on a combination's opaqueness, as can be seen on the cover of Palmer's (1933) *Second Interim Report on English Collocations*: "A collocation is a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts".

In contrast to this long-standing disregard for more fine-grained interactions along the interface between lexis and grammar, modern research has recognized "the tension caused by generating a strict distinction between the lexical and grammatical, and distinguishing between the syntagmatic and paradigmatic dimensions of language" (McEnery & Gabrielatos 2006: 40). This resulted in the establishment of an intermediary perspective under the paradigm of lexicogrammar (e.g. Halliday 1991), which is considered "a unique contribution made by corpus linguists to linguistic theory" (McEnery & Gabrielatos 2006: 40).[18] The origins of this modern approach are usually seen in John Rupert Firth's papers, in which he paved the way for a more general

[16] Co-text' referring to the textual surroundings of a word, with 'context' more generally also pertaining to what is relevant for pragmatic inferencing, i.e. background or world knowledge, communication styles, general features of the discourse (genre) and the communicative situation.

[17] The earliest attestation of 'collocation' as a technical term was, however, not until 13 years later, in 1750 (Bartsch (2004: 28–30)).

[18] Sinclair's (2000) term 'lexical grammar', which more directly focused on the pivotal nature of lexis, is no longer in common use. Halliday's (1991: 31–32) perspetive instead regards "lexicogrammar as a unified phenomenon, a single level of 'wording', of which lexis is the 'most delicate' resolution." Sinclair (2000: 191) remained very skeptical of lexicogrammar in that "it does not integrate the two types of pattern as its name might suggest – it is fundamentally grammar with a certain amount of attention to lexical patterns within the grammatical frameworks; it is not in any sense an attempt to build together a grammar and lexis on an equal basis."

approach to interrelations within language on the lexical level, i.e. a co-textual theory of meaning:

> You shall know a word by the company it keeps! […] The habitual collocations in which words under study appear are quite simply the mere word accompaniment, the other word-material in which they are most commonly or most characteristically embedded. (Firth 1957/1968: 179–180)

In addition to forms of lexical co-occurrence, Firth's papers extended this fledgling co-text-based linguistic theory to also account for grammar through the notion of 'colligation'.[19] Originally, this only pertained to "the co-occurrence of **grammatical** choices" (Sinclair 1991: 85, boldface CK), "but later researchers often had a less restricted meaning" (McEnery & Gabrielatos 2006: 42), who often regarded colligation as a "[s]econd type of extended **lexical** units" (Lindquist 2009: 87, boldface CK):

> Grammatical relations should not be regarded as relations between words as such–between *watched* and *him* in 'I watched him'–but between a personal pronoun, first person singular nominative, the past tense of a transitive verb and the third person pronoun singular in the oblique or objective form. (Firth 1957/1968: 181)

This change in perspective made it possible to conceptualize syntax as being centered around the word and its patterns instead of isolating grammatical structures and lexical choices.[20] Following the work carried out by Firth and his successors, research into patterned language began to include "less striking strings of words which are semantically and grammatically more 'normal'" (Lindquist 2009: 91). Along with this came an expansion of the central term of 'collocation' into a more empirical and data-driven direction under the 'neo-Firthian' tradition most closely associated with John Sinclair. These continually diverged further from the traditional, intensional definition of collocation, usually within a phraseological framework. Evert (2004) recommends to "reserve the term *collocation* for an intensionally defined concept that does not depend on corpus frequency information" (Evert 2004: 17; italics his), while discussing the distributional characteristics of linguistic items under the notion of *co-occurrence*.[21]

---

[19] The term was, however, actually first introduced by Simon (1953: 327), who Firth (1957/1968) acknowledges in footnote 49. Simon devised the term as a parallel to Firth's 'collocation', but also mentions that the term itself had been suggested by a Dr. S. A. Birnbaum, providing, however, no additional information about this source. (Thanks to the anonymous authors on http://www.bfsu-corpus.org/content/j-r-firth-not-collocated-collocation-or-colligation for pointing this out.)

[20] And consecutively extended into the areas of semantics and pragmatics through the notions of semantic preference and semantic prosody (cf. Sinclair 2004: 34, Louw 1993).

[21] Evert (2004: 17) recommends to reserve the term 'collocation' for discussions of intensional sense and using 'cooccurrence' for all other cases describing distributional characteristics of linguistic items.

In addition to the conceptual expansion of essential notions, the emerging field of lexicogrammar also began to diversify in terms of the operationalization of co-occurrence (cf. e.g. Wray 2002 for a historical account, Cortes 2004: 398–399 for a summary starting in 1873, or Bartsch 2004: 27–64 and Kreyer 2013: 205–212 for an overview of approaches to co-occurrence). For the most part, these methodological and conceptual expansions were only made possible by technological advances and the accessibility of computers accompanying the advent of corpus linguistics. Xiao (2015: 106) surmises that "corpus linguistics has not only redefined collocation but has also foregrounded collocation as a focus of research by neo-Firthian linguists as well as those of other traditions". In fact, Sinclair (2004: 165) goes so far as to suggest that the sharp divide between lexis and grammar presupposed in earlier theories of language might primarily be seen as "a consequence of the inadequacy of the means of studying language in the pre-computer age".

3.1.2 Categorizing Co-occurrence

The growth of the field of lexicogrammar came with the major caveat that the astoundingly diverse set of methods and associated terminology all trace their origins back to a small set of unfortunately vaguely-defined notions: For instance, Firth may have sparked an empirical re-definition of 'collocation' through his postulation of a "*mutual expectancy* […] The words are mutually expectant and mutually prehended" (Firth 1957/1968: 181). Simultaneously, however, he also foregrounded the semantic component of collocation when he proposed "to bring forward as a technical term, meaning by collocation, and apply the test of collocability." (Firth 1957: 194, cf. also Xiao 2015: 107). Furthermore, his statistical ideals were also unclear and potentially in conflict, since the synonymous use of 'most common' and 'most characteristic' word accompaniment (Firth 1957/1968: 179–180) equates potentially meaningless high-frequency items with low-frequency but highly characteristic ones (cf. Kreyer 2013: 214–215 for a discussion). Based on these ambiguities in terms of a pivotal concept, heterogeneity in terms and approaches may not come as much of surprise, and it may be telling that a mere 20 years ago, Manning & Schütze (1999: 121) only managed to narrow down Altenberg's (1991: 127) "recurrent word combinations of some kind" to

"[s]ome conventional way of saying things".[22] Bartsch (2004: 27) consequently calls collocations "notoriously hard to capture except by means of multi-variate criteria building on properties at different levels of linguistic organization" and attests to the term's various competing uses:

> The term collocation itself has been employed indiscriminately as a cover term for different types of word combinations that are too obvious and frequent to be ignored, yet defy explanation based on currently accepted paradigms of linguistic description. Collocation thus denotes a much more heterogeneous set of lexical co-occurrences than the single term suggests [...]. This diversity of structures subsumed under the term 'collocation' has led to a proliferation of definitions in the literature. (Bartsch 2004: 27)

Criticism like this has prompted researchers to attempt to formalize the criteria which otherwise often remained at least partially implicit in previous studies and subsumed under allegedly clear terms, so as to avoid that "the same facts are clad in ever-new representational formats" (Römer & Schulze 2009: 2). Concerning the necessity of a particularly clear definition of the linguistic structure under scrutiny in lexicogrammatical research, Gries (2008a: 10) argues that

> it is essential that we, who are interested in something as flexible as patterns of co-occurrence, always make our choice of parameter settings maximally explicit to facilitate both the understanding and communication of our work.

A comprehensive yet neutral classification system for this purpose is proposed by Kreyer (2013: 206), who distinguishes item strings by five parameters, to be discussed in turn below.[23] The presentation will focus on *n*-grams and POS (part-of-speech)-grams as the form of co-occurrence to be studied in the volume at hand, but it will cast its net wide enough to contextualize these particular sequences in the larger picture of co-occurrence research. Yet, it should be noted that even within these relatively clear-cut parameters, approaches are better seen to lie "within a continuum rather than in discrete conceptual spaces" (Bartsch 2004: 33).

[22] A step up from van der Wouden (1997: 53–54), who displays a very negative perspective on collocation as the "junkyard of linguistics and reckons that "the first aim of collocation research should be to reduce the collocational behaviour or fixed combinations to a more respectable level or module of grammar, be it phonology, semantics or syntax."

[23] A similar typology is offered by Gries (2008a, 2013), albeit from a more phraseological perspective which expects at least one element of the combined form to be a word form or lemma (cf. the first parameter). Kennedy (1998: 110–121) also provides a detailed account from a time when parameters were not as worked out as they are today.

Grammatical status of constituent items

The first criterion concerns the type of elements involved in the co-occurrence, the "*nature* of the elements involved" (Gries 2008a: 5). Most frequently, this pertains to words, and particularly to word forms. This can be seen as adhering to Sinclair's (e.g. 2004: 17) understanding that lemmatization may conflate divergent patterns (and thus meanings) of different word forms.[24] Studies of *n*-grams and lexical bundles usually follow this system. On the other hand, lemmatization is carried out freely in studies along Halliday's approach, and according to Stubbs (2002: 223), the practical impact of lemmatization on the collocational patterns identified may not always be as drastic as it is sometimes made out to be.

Exceeding the potential abstraction introduced through word lemmatization, elements within a co-occurrence can also be conceived in more abstract terms, particularly in the form of placeholder items for parts of speech, syntactic constructions, meaning facets and pragmatic implications.[25] The latter two concern the concepts of "semantic preference" and "semantic prosody" (Sinclair 2004: 34 and Louw 1993, respectively), which describe either a semantic aspect common to all collocates (e.g. the field of 'visibility' before *the naked eye*) or otherwise an evaluative component (e.g. 'difficulty experienced' in connection with *the naked eye*; Sinclair 1991, 2004). Grammatical and syntactic co-occurrence, on the other hand, extends co-occurrence phenomena into the field of 'colligation', as established in Section 3.1 above.

The concept of colligation furthermore informed the notion of Part-of-Speech grams (POS-grams) by Stubbs (2007), i.e. sequences of POS-tags instead of lexical items extracted from a corpus.[26] The choice of POS tags used within these sequences necessarily results from the particular corpus annotation scheme: Within the British National Corpus, for instance, the sequence "'PRP AT0 NN1 PRF AT0' can be attested, which represents the CLAWS C5 markup for the sequence 'preposition other than *of* + article + singular common noun + preposition *of* + article' (e.g. *at the end of the*)."

[24] Sinclair (1991: 41) strongly rejects lemmatization as the basis for corpus-based analysis, "as *a priori* lemmatization is seen as introducing the analyst's subjective intuitions".
[25] Gries's (2008a) definition only addresses "lexical items and grammatical patterns", and while it "does not commit to a particular level of granularity regarding the lexical elements involved", it leaves out the broader aspect of semantic and pragmatic co-occurrence incorporated by Kreyer (2013: 206–208).
[26] The term may have actually been used before that date within the PIE database (Fletcher 2003-2010).

(Kreyer 2013: 207–208) In the expanded CLAWS C7 tagset, this sequence would be differently encoded as "II AT NN1 IO AT" (UCREL 2007).

Permissible distance between constituents

While the previous section concerned itself with potential variability within a particular 'slot' of a co-occurrence (the 'granularity' of description), a second distinction needs to be made in terms of variable distance between the elements under scrutiny, i.e. between continuous and discontinuous sequences. While most understandings of collocation allow for some degree of discontinuity (but cf. e.g. Lindquist's 2009: 78 concept of 'adjacent collocations'), many other concepts of co-occurrence include only immediately adjacent elements, such as *n*-grams, lexical bundles (Biber et al. 1999) or POS-grams. This might be regarded as a severe limitation of these approaches since, as Evert (2009: 1222) shows, increasing the allowed distance ('span size') may be tempting: "For a span size of 3, *throw a birthday party* would be accepted as a cooccurrence of *(throw, party),* but *throw a huge birthday party* would not." But it also stands to reason that consecutive span increases could theoretically repeatedly be advocated for (e.g. throw a really, really huge birthday party), while lower spans are often preferable for precision, as Bartsch & Evert (2014: 57) demonstrate: "the larger the contexts become, the lower the precision drops."

Some forms of co-occurrence, such as skip-grams (*n*-grams in which tokens can be 'skipped'; cf. e.g. Guthrie et al. 2006) or collocational frameworks (Renouf & Sinclair 1991), can be discontinous or continuous depending on perspective: In these cases, the abstract representation (e.g. '*a* + ? + *of*' in case of the collocational framework) describes "a discontinuous sequence of two words, positioned at one word remove from each other" (Renouf & Sinclair 1991: 128). Thus, only the abstract representation including placeholders is truly discontinuous. Actual linguistic instantiations instead result in continuous sequences of items (Kreyer 2013: 213).

(Positional and syntactic) Flexibility

In addition to the types of elements involved in a co-occurrence and the distance between them, the internal modifiability of the entire co-occurrence provides a third source of variability. Primarily, this concerns the positional structure of the elements concerned, i.e. whether they follow a certain order, like e.g. *n*-grams, POS-grams or

lexical bundles (and which may additionally be discontinuous in case of e.g. collocational frameworks or skip-grams) – or whether they can appear in any order (like most definitions of collocation).

Beyond positional variability, syntactical modifiability is also often attributed to the aspect of flexibility, such as in the case of many idioms (e.g. *kick the bucket*, which can be modified for tense). But while this area of modifiability may provide a sensible aspect of post hoc description of a particular co-occurrence (i.e. the degree to which an item can be adapted to a context), it should be noted that this overlaps considerably with the first descriptive dimension, i.e. the degree of grammatical abstraction (with the flexibility of *kick the bucket* explainable through a degree of lemmatization of the verb). Thus, including this type of variation under the heading of 'flexibility' may conflate an operational definition of a co-occurrence with its analysis and description.

Significance of the combined sequence

As laid out before, Firth's argument to regard collocation as "the mere word accompaniment" of "actual words in habitual company" (Firth 1957/1968: 182) and his, admittedly conflicting, definitions of this 'habituality' paved the way for a revised perspective onto co-occurrence. Firth seems to regard a word's frequent and characteristic environment as essentially identical and synonymous, but a distinction should better be drawn between the two: While highly frequent combinations might be characteristic of a word (e.g. highly frequent verb+preposition combinations), a high frequency of co-occurrences like *my* + *mother/father* could be regarded as rather limited in terms of the noteworthiness of such items (Kreyer 2013: 214–215). In contrast, the reverse of low frequency of an expression but high degrees of familiarity can be observed for idioms (e.g. *kick the bucket*).

Thus, if relevant patterns of co-occurrence to be analyzed are not available a priori, a method is required for the identification of items worthy of further study. This fourth aspect of co-occurrence may well be the one by which modern approaches distinguish themselves most categorically from traditional perspectives on co-occurrence: While early approaches could only rely on introspection, often affected by the markedness of a combination (cf. Schilk 2006: 313), modern research into co-occurrence is informed through corpus research and involves some degree of statistical sophistication

for the empirical substantiation of items. This can include raw or relative frequencies, but since frequency can result from structurally required or otherwise relatively unremarkable item combinations, some quantification of the strength of the connection between items involved in a co-occurrence is required. The scope of theoretical models and potentially available measures does, however, diverge strongly from what is actually applied on a regular basis, as Gries (2008a: 20) observes: While he finds "the most comprehensive identification procedures" in the area of corpus linguistics, he criticizes that "[s]everal levels of sophistication are discernible [but] the most basic approaches are, it seems, also the most widely used ones." Section 3.3 will delve further into aspects of statistical description and selection within the focus of the current study.

Structural and semantic unity and non-compositionality

A final distinction to be made concerns whether the co-occurrences identified also need to form "units of a grammatical, semantic or pragmatic kind" (Kreyer 2013: 211). Some authors only accept complete grammatical units as relevant forms of co-occurrence, and require semantic unity and non-compositionality (/non-predictability) as prerequisites (cf. e.g. Kjellmer 1991, Simpson 2004, and also van Lancker-Sidtis & Rallon 2004: 208). This is a particular concern for studies approaching the field from a psycholinguistic perspective, which presupposes holistic storage and retrieval and includes some form of functional aspect of the co-occurrence (cf. e.g. Götz & Schilk 2011: 85–86).

The majority of strings, however, do not overlap neatly with grammatical, semantic or pragmatic unit boundaries, as Kreyer (2013: 211) observes:

> It is very rare that the recurrent item string qualifies as a grammatical unit. [… In] general, recurrent item strings may create grammatical units, but they may also fail to do so. Similarly, most recurrent item strings neither form a semantic or pragmatic unit.

This can be seen to apply for lexical bundles (and the more general notion of *n*-grams and POS-grams; cf. also Stubbs 2002): These may certainly be characteristic aspects of a text, but still are solely defined on the grounds of frequency independent from unit bounderies as "the most frequent recurring lexical sequences in a register" (Biber et al. 2004b: 376).

> Lexical bundles are recurrent expressions, regardless of their idiomaticity, and regardless of their structural status. That is, lexical bundles are simply sequences of word forms that commonly go together in natural discourse. (Biber et al. 1999: 990)
>
> [M]ost lexical bundles do not represent a complete structural unit. For example, only 15 per cent of the lexical bundles in conversation can be regarded as complete phrases or clauses, while less than 5 per cent of the lexical bundles in academic prose represent complete structural units (Biber et al. 2004b: 377)

In contrast to defining relevance through structural unity, it can be regarded as a feature of particularly data-driven methods to disregard other levels of description in a deliberate forgoing of information ('bewußter Informationsverzicht', Lehr 1996: 50), and instead to employ statistical measures as the primary definitional feature for the relevance of a co-occurrence.

3.2 Co-occurrence and World Englishes

3.2.1 Patterned Language in Varieties of English

The previous section has brought to the fore the diverse range of approaches to co-occurrence, which can all trace their origins back to the concept of collocation. While this was long viewed as a single phenomenon, the prevalence of modern approaches can be regarded as a testament to the increased recognition of diverse forms of co-occurrence patterns in language. Although terms differ, the relevance of patterned language is today acknowledged widely. Precise frequency estimates for patterned language are, however, strongly dependent on the particular methodological approach – since co-occurrence touches on "the boundary between habit and rule [..., which] has never been clear" (Nattinger & DeCarrico 1992: 35).

Quantitatively, several estimates of the frequency of patterned language have been attempted, which can be regarded as coarse reference values given the caveats mentioned above. Altenberg (1990) is often cited, who estimates that up to 70% of adult language may be formulaic in some way. While he later even increases this figure to 80% for 2-grams within the London-Lund Corpus of Spoken English (Altenberg 1998), other authors have reported considerably lower frequencies. Within a corpus partially comprising the same texts as Altenberg (1990), Erman & Warren (2000) find only 59% of all material to contain prefabricated language ('prefabs'). Even lower frequencies are observed by van Lancker-Sidtis & Rallon (2004: 213), who only attest to 25% of their phrases under scrutiny to be the result of conventionalized speech (accounting

for 16% of all words in their study). The numbers obtained by Daudaravičius & Marcinkevičienė (2004: 341) can be seen to reconcile the varying frequencies above: While they find only about one eighth of their BNC data to be comprised of highly frequent language patterns (n>10 occurrences), about 80% of the corpus material can indeed be accounted for by sequences of lower frequency.

Beyond empirical benchmarks, the relative importance of language patterns can be deduced on a more theoretical level: Following Sinclair (1991), a language user has, for every communicative act (or part thereof), a choice between fully creative use (the open-choice principle) and conventionalized expression (the idiom principle).

> [The open-choice principle] is a way of seeing language text as the result of a very large number of complex choices. At each point where a unit is completed (a word or a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness. (Sinclair 1991: 109)
>
> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. (Sinclair 1991: 110)

Of course, none of the two principles is ever singularly active, "actual language use is rarely a choice between such polar opposites but is on a cline between these extremes" (Kennedy 1998: 109) and, in practice, "the language user alternates between the open choice principle and the idiom principle" (Erman & Warren 2000: 30). Total creativity and true novelty are, however, a rare phenomenon – "poetic" in Wray & Perkins's (2000: 12) terms – making "the use of a 'purely' analytic strategy […] a peripheral activity [… and not] a major element of normal language processing". On the other hand, however, "phraseology alone cannot account for how sentences or utterances are made up" (Hunston 2002: 148), thus establishing an upper bound for purely conventionalized expression. Wray (1998) summarizes the necessity of both principles in everyday interaction:

> Without the rule-based system, language would be limited in repertoire, clichéd, and, whilst suitable for certain types of interaction, lacking imagination and novelty. In contrast, with only a rule-based system, language would sound pedantic, unidiomatic and pedestrian. (Wray 1998: 64–65)

While contextualizing the relative frequencies of use of both of Sinclair's (1991) principles, Wray (1998) thus links conventional linguistic behavior following the idiom principle to the subject of typical and idiomatic linguistic behavior: In addition to knowing "what combinations are *possible,* [...] idiomaticity [...] involves knowing which

particular combinations are *conventional* in a language community although other combinations are conceivable." (Warren 2005: 40, italics hers) Idiomaticity and conventionality consequently concern efficient language production within a particular context, and reciprocally speedy reception and ease of communication for all parties involved in a communicative event. They can conversely be modeled as entrenched statistics and intuitions about probabilities in language in the sense that an individual's "baseline strategy in everyday language processing, both production and reception, 'relies not on the *potential for the unexpected* […] but upon *the statistical likelihood of the expected*" (Wray & Perkins 2000: 13, italics theirs).

For varieties of English, this knowledge of the likelihood of certain choices exerts itself as "nativelike selection" (Pawley & Syder 1983), which leads to "nativelike fluency" only through an adherence to the conventional:

> The problem we are addressing is that native speakers do not exercise the creative potential of syntactic rules to anything like their full extent, and that, indeed, if they did do so they would not be accepted as exhibiting nativelike control of the language. The fact is that only a small proportion of the total set of grammatical sentences are nativelike in form – in the sense of being readily acceptable to native informants as ordinary, natural forms of expression, in contrast to expressions that are grammatical but are judged to be 'unidiomatic', 'odd' or 'foreignisms'. (Pawley & Syder 1983: 93)

A speaker acquiring English in a particular context would thus be exposed to that variety's conventions through other speakers, in turn shaping what is *possible* in English into what is *likely and normal* within that specific speech community. From a cognitive perspective, this can also be viewed as a consequence of the 'pattern-finding' ability of humans (Ellis 2003: 68), which 'primes' and consecutively loads an item with its combinatorial behavior (Hoey 2005: 8), reinforcing conventionality through repeated activation of the network of the mental lexicon (Kreyer 2013: 212–227). In turn, it could be expected that linguistic behavior would betray a text's origin through conventionalized expressions and patterned language, i.e. the 'from-corpus-to-cognition-principle' (Schmid 2000: 39):

> in cognitive linguistics, the frequency with which a speaker/hearer encounters a particular symbolic unit is assumed to be positively correlated with the degree of cognitive entrenchment of that unit in the speaker's cognitive system (Gries 2008a: 18)

The idea that the conventional and idiomatic finds its reflection in corpus data has consequently been argued for by Schneider (2007), who makes a strong point for the distinctive potential of combinatorial differences between varieties of English:

> Of course, practically every text betrays its regional origin by the place names and personal names in it, and perhaps by some cultural or distinctly local terms. But if you strip it of those, what remains? Not very much, but something inconspicuous but nevertheless powerful and consistent: preferences". (Schneider 2007: 91)

> [I]ndividual varieties differ from each other first and foremost in their combinatory preferences, in their constructions, in the frequencies of their lexicogrammatical choices, collocations, word uses, and so on. It is, not only, and perhaps not even primarily, the occasional occurrences of well-known 'distinctive features' that attribute its uniqueness to a variety; it is the subconscious set of conventions regulating the norm level of speech habits, of what is normally done and uttered, the 'way things are said' in a community. (Schneider 2007: 92)

In contrast to lists of 'distinctive features', which are frequently brought up when discussing the characteristics of World Englishes, Schneider (2007: 86–87) stresses the importance of differences in the "co-occurrence potential" of words and structures, and that as such, "preferences are important, not only absolutely novel uses can be expected" (cf. Lunkenheimer 2013: 870 for a similar evaluation). Certainly, "[s]urface similarities across New Englishes can be skin deep, diverging dramatically upon closer examination, due to substrate systems or substrate-superstrate interaction." (Sharma 2009: 190) However, the overall frequency of direct structural transfer may be relatively limited, as Gut (2011: 118) argues, and instead "the avoidance or the overproduction of certain structures" may be more quantitatively relevant. Direct transfer presents itself as "a phenomenon that appears to be restricted to the early stages of language acquisition", making it "less likely to underlie language productions by the advanced 'learners' of English that make up the population of postcolonial countries." Schneider (2007: 81) instead regards collocational preferences and gradual differences as "extremely characteristic" of local Englishes: "Individual speech communities are free to develop habits as to which words tend to co-occur, and so this is an element of language organization which readily perrnits the emergence of inconspicuous regional features." They are thus central to the process of grammatical nativization which lies at the heart of his Dynamic Model:

> [I]n its early stages this indigenization of language structure mostly occurs at the interface between grammar and lexis, affecting the syntactic behavior of certain lexical elements. Individual words, typically high-frequency items, adopt characteristic but marked usage and complementation patterns. […] Hence, grammatical nativization in PCEs typically sets out with a specific set of patterns which appear to occur more frequently than others. (Schneider 2007: 46)

Accordingly, locally characteristic collocations begin to diverge from other varieties already during transplantation of English to a new surrounding: They emerge early in a

variety's development but continue to shape its preferences, conventions and its speakers' perceptions of typical and preferred linguistic behavior. Being gradual in nature, they allow to distinguish varieties by what is shared between them instead of what is different:

> In corpus-based research into regional varieties of English, the focus has traditionally been on the description and analysis of features which may help to **distinguish** one particular variety from another. In contrast, a similar focus on linguistic features **shared** by a range of varieties of English has not yet been established in corpus-based research into World Englishes, notwithstanding some laudable exceptions (Bernaisch et al. 2014: 8; boldface CK)

3.2.2 Using Co-occurrence to Differentiate Varieties

The present study approaches co-occurrence patterns in varieties of English from a strictly data-driven perspective, estimating the relevance of a particular sequence from a purely empirical perspective. In this way, it approaches co-occurrence as an 'epiphenomenon' resulting from a multitude of "hidden causes" (Evert 2009: 1218): Indeed, many phenomena in language not primarily regarded as aspects of co-occurrence (compounds, semantic restrictions, cultural stereotypes and even conceptual knowledge) find their "surface reflections" in collocational structures (Evert 2009: 1242) and depending on a specific theory, different aspects of the co-occurrence potential within language may be addressed and highlighted.[27] Unlike "the explanatory perspective of theoretical linguistics", the study thus adopts a data-driven perspective on collocations as primary data, in which they "can be interpreted as empirical predictions about the neighbourhood of a word" (Evert 2009: 1218). The approach thus represents one following a clearly 'corpus-driven' (/bottom-up) paradigm in contrast to the more prominent 'corpus-based' (/top-down) procedure (Tognini-Bonelli 2001): With regard to collocations, a corpus can either be searched for instantiations of a type of co-occurrence previously defined, or it can be mined for any kind of sufficiently "mutually expectant" (Firth 1957/1968: 181) combinations of items which in turn form the basis for further analysis and potential categorization (an approach which resulted in e.g. Pattern Grammar; cf. Hunston & Francis 2000, Hunston 2014). The present study consequently expects that even outside clearly defined theoretical phenomena, quantitative language patterns (*n*-grams, POS-grams) in varieties of English (to the extent that

[27] Compare, e.g. Aitchison (2012: 102), who discusses *salt water* as a case of collocation, even if most people might intuitively describe this occurrence as a compound noun.

they become observable in corpus data) can be used to systematically distinguish varieties of English from one another even on this relatively abstract level.

Analogous applications of lexical sequences *n*-grams as a differential tool can be attested in several earlier studies. While usually no categorical distinctions can be identified between varieties on this level of analysis, quantitative differences emerge as the main indicator. Potentially the most prominent application can be found in the lexical bundles approach (*n*-grams of typically length 4) spearheaded by Biber et al. (1999) and applied to the study of academic English in e.g. Biber et al. (2004b), Biber & Barbieri (2007), Cortes (2004) and Simpson-Vlach & Ellis (2010). Moreover, bigrams have been employed for register classification through the application of multi-dimensional (Crossley & Louwerse 2007) and clustering approaches (Gries et al. 2011) as well as for evaluations of corpus homogeneity (Gries 2010a). *N*-grams have also been a methodological focus in the related disciplines of computational and cognitive linguistics, as well as for machine learning and beyond (cf. e.g. Manning & Schütze 1999, Crossley & Louwerse 2007, Jurafsky & Martin 2009, Wahl & Gries 2018: 87).

However, despite the wide-spread use of co-occurrence data in other fields, studies in this vain were still a scarcity at the time of Schneider's (2007) call for more attention towards co-occurrence patterns in World Englishes. Gries & Mukherjee (2010: 525–526) attribute this fact to the quantitative nature of variation on this level, which eludes intuitive description and requires the analysis of large amounts of data (which are not necessarily available as required). Schilk (2006: 277) still attests to the scarcity of collocational studies in general within the field, and Gries & Mukherjee (2010: 523) call the entire "lexis-grammar interface […] a blind spot in research into many postcolonial varieties of English". While a number of studies on the lexis-grammar interface have since been published (e.g. Schneider 2004 and Nesselhauf 2009 on particle and phrasal verbs, Schilk 2006, 2011 on collocational profiles of selected verbs, Mittmann 2011 on variation in formulas between American and British English, Bernaisch 2015 on light verbs), studies focusing on *n*-grams in World Englishes seem to only have emerged within the last decade (Gries & Mukherjee 2010, Götz & Schilk 2011) and should still be regarded as a rare occurrence.

While the neglect of the lexis-grammar interface in general has thus been partially remedied within the last few years, it can be observed that corpus-based rather than corpus-driven studies prevail in the field of World Englishes just as much as overall in corpus linguistics. The reason for this may lie in the fact that corpus-driven analyses come with an even stronger reliance on quantitative data – and including further varieties ever more compounds the issue. In addition to these methodological problems in managing the potentially vast amounts of data, corpus-driven studies can also become challenging from a linguistic perspective: Since they approach collocation from the side of its epiphenomenal nature, they may conflate potentially relevant theoretic distinctions and accordingly dilute boundaries between categories of co-occurrence. However, instead of regarding this as a drawback, the decision to remain theory-light should rather be seen as a deliberate forgoing of information in the sense advocated by Lehr (1996: 50, see above), neglecting the 'why' of linguistic behavior and instead retaining a descriptive approach for as long as possible. This extensional description disregards attempts at precisely modelling human cognition in co-occurrence and instead hypothesizes that all linguistic behavior within a speech community may be a result of relevant preferences and particularities. Following Lehr (1996: 55), only because intensional description is more meaningful to humans and intuitively understandable, that does not imply that there can be no systematic variation within an extensional account that might otherwise remain unnoticed. What is often disregarded is that intensional explanation may be just as incomplete by design, albeit in a different form. Certainly, in top-down, corpus-based approaches, a priori categories may lead to ignoring relevant co-occurrences or differences in language patternings between texts. Erman & Warren (2000: 33) consequently note the difficulties involved in a human researcher scanning a text for prefabs, in that "they can easily be overlooked. Some of them appear at first sight to be completely transparent combinations of words". Consequently, the authors caution against the impression that deciding on an item's prefab status was always an error-proof undertaking, since "there is no absolute proof that a particular combination of orthographic words is ready-made. It is quite possible that, say, *I am afraid* (implying 'I regret to have to inform you') was composed word for word by a speaker who simply happened to make the idiomatically correct choice of words." (Erman & Warren 2000: 33)

The present study addresses the lack of strongly corpus-driven approaches to variation between World Englishes by employing *n*-grams and POS-grams as an exclusively distinctive tool and thus relying entirely on a set of means of calculating the degree of attraction between words and POS which highlight different types of co-occurrence. Instead of having the selection of relevant types of co-occurrence be primarily informed through theory, the present study compares co-occurrence preferences between all varieties under scrutiny and gives primacy to statistical and frequency effects to distinguish them. The main hypothesis followed will thus be that even co-occurrence as an epiphenomenon will be a satisfactory tool to systematically distinguish varieties of World Englishes. It thus disregards categorization of the sequences identified and may conflate the collocation of text- or genre-specific items (such as cultural or technical terminology) with those determined by the grammar of the language (e.g. adjective+noun combinations). Some of the latter type may even appear to be entirely self-evident (e.g. *and a*) but can still represent aspects of the variety at hand, such as aspects of grammatical determination (cf. Lenko-Szymanska 2012). What appears like the insignificant result of a rule may still diverge strongly in terms of habits of speakers or speech groups. It is therefore not the intent of this analysis to identify, for each variety, the most typical collocations and selecting these for further description. Rather, in a bottom-up, data-driven approach, the measure of collocational strength serves as a means of distinguishing varieties of English by their respective (dis-)preferences for the same set of shared sequences.

While analyses with a similar aim of distinguishing World Englishes through *n*-grams are already virtually non-existent, the scope of the present study exceeds previous analyses in that multiple measures of attraction are applied in an effort to triangulate the findings. Their presentation and the description of their characteristics will be the core of Section 3.3.

3.3 Operationalizing Collocational Sequences

3.3.1 Bigram and *n*-gram statistics

The previous sections have discussed the prevalence of co-occurrence features in language and their relevance for the description of varieties of English. Furthermore, the quantitative nature of co-occurrence has been shown to be far more pervasive than acknowledged for a substantial length of time in linguistic research. Since, however, precise empirical relationships elude human perception, both in language use and analysis, several measures have been devised to quantify the degree of association between linguistic items.

Association may be measured in many ways, and certainly the simplest form can be seen in absolute frequency of co-occurrence.[28] However, while frequency addresses the likelihood of two items occurring together in contrast to all other combinations in a corpus, it does not convey any information as to the probability of that occurrence against chance. As such, it will assign the highest scores to combinations of highly frequent items (i.e. function words) that do not actually attract one another beyond the likelihood of their individual occurrence, i.e. "the fact that very frequent words will frequently occur near any word by sheer chance" (Lindquist 2009: 74).[29] This is why "frequencies must be normalized and checked against chance levels before they can be interpreted" (Gries et al. 2010b: 71), which leads to 'statistical association' (Sinclair 1966: 418). Ellis & Ferreira-Junior (2009) provide an illustrative examples from category learning:

> Consider how, in the learning of the category of birds, while eyes and wings are equally frequently experienced features in the exemplars, it is wings which are distinctive in differentiating birds from other animals. Wings are important features to learning the category of birds because they are reliably associated with class membership, eyes are neither.

[28] Several values from one to higher numbers have been used, often selected in relation to the size of the corpora employed. Altenberg (1998) chooses a frequency of 2 as his threshold value to consider an *n*-gram as recurrent. This threshold is also employed by Hunston (2014: 101) in her discussion of recurrence in pattern grammar: "Because a pattern of any kind exists only if it is repeated, it is obvious that any instance […] must be observed at least twice to be considered". That this number is mostly chosen arbitrarily can be seen in e.g. Götz & Schilk (2011: 85–86), who, in a relatively small dataset (4 sets of c. 87,000 words each), require that "at least five occurrences of any 3-gram have to be attested in the corpus data."

[29] Frequency alone can, however, be informative once a strong combination has already been identified, in that frequent exposure may reinforce the collocation (but cf. the comparison of frequency vs. attraction information in Gries et al. (2010b: 71).

> Raw frequency of occurrence is less important than the contingency between the cue and interpretation. (Ellis & Ferreira-Junior 2009: 194)

Frequency information is thus one of the few measures which is not contingency-based (cf. e.g. Levshina 2015: 225-226, 234–235), while most association measures include contingency information, and calculate it on the basis of a 2x2 co-occurrence matrix as presented in Table 3.1. In such contingency tables, the co-occurrence frequency of two items (cell a) is contrasted against the number of times that any of the two constituents co-occurs with some other lexical item (cells b and c) and/or against the total number of co-occurrence of any other two items (d). While cell a is always of interest, not every measure uses all the remaining cells provided in Table 3.1, and any combination of the frequencies can form the basis of calculation depending on the specifics of the measure.

Table 3.1. Contingency table underlying most collocational statistics (cf. e.g. Levshina 2015, Wahl & Gries 2018: 91). The ¬ symbol indicates negation.

	word₂	**¬ word₂**	**Totals**
word₁	a	b	R_1
¬ word₁	c	d	R_2
Totals	C_1	C_2	N

In addition to observed frequencies, most measures also relate these to expected frequencies for some or all cell values above. Expected frequencies are based on the supposition of random distribution of the linguistic data, and are thus calculated solely on the basis of the item frequencies (i.e. row and column frequencies).[30] For any cell for which an expected frequency is to be derived, the respective row and column frequencies are multiplied and then divided by the overall number of items in the data (N), which results in Table 3.2:

Table 3.2. Expected item frequencies and their ways of calculation from observed frequencies.

	word₂	**¬ word₂**
word₁	$a_{exp} = {R_1 * C_1}/{N}$	$b_{exp} = {R_1 * C_2}/{N}$
¬ word₁	$c_{exp} = {R_2 * C_1}/{N}$	$d_{exp} = {R_2 * C_2}/{N}$

A second distinctive criterion of association measures concerns the direction of attraction: As before, the majority of measures falls into the same group in that almost all frequently-used methods calculate bi-directional attraction scores, which indicate

[30] Cf. e.g. Stubbs (1995: 31), who comments critically on the assumption of 'complete independence' and 'expected frequencies' in language.

how closely connected the constituents are overall, in contrast to other co-occurrences. This comes with a drawback in that bi-directional measures cannot distinguish average, mutual attraction from cases in which it is more strongly exerted from only one constituent, i.e.

> – instances [...] where *X* attracts *E* but *E* does not attract *X* (or at least much less so);
> – instances where *E* attracts *X* but *X* does not attract *E* (or at least much less so);
> – instances where both elements attract each other (strongly). (Gries & Durrant to appear).

If, for instance, *in spite of* were concerned, then within each of the two bigrams involved, *spite* presents a better predictor than either *in* or *of*, leading to two different directions of association. Thus, uni-directional association measures may be better suited to explaining such bigrams than bi-directional ones, which disregard this distinction or conflate directional attraction to mutual attraction. While such 'left- and right-predictive' bigrams (in Kjellmer's 1991 terms; cf. also Michelbacher et al. 2011) are often described post-hoc, they are rarely calculated:

> Although such asymmetries are often reflected in skewed marginal frequencies (the collocation being more important for the less frequent word), hardly any of the known association measures make use of this information (Evert 2009: 1245)

A final important distinction, and one in which most common measures vary the most, lies in the statistical way of quantifying the attraction. In this regard, measures can be divided into two general types: probability-based and significance-based measures. Probabilistic methods indicate the effect size of the co-occurrence, i.e. the degree to which the two items attract each other beyond chance levels (a corpus in random order), which is commonly called the *strength* of association (Hunston 2002: 71). By contrast, statistical significance-testing methods analyze the amount of evidence for the co-occurrence, no matter the effect size, and consequently address the *certainty* of association (Hunston 2002: 72). The two perspectives are, however, not entirely unrelated, as Evert (2009: 1228) acknowledges, in that "a word pair with a large 'true' effect size is also more likely to show significant evidence against the null hypothesis in a sample". Both methods do, however, have their specific advantages and drawbacks: Since probabilistic measures only focus on the degree to which observed frequencies exceed expectations, a co-occurrence frequency of 2 against an expectation of 1 leads to the same probability as 20 observed co-occurrences against

10 expected, even if the latter type clearly has more statistical support. Significance-testing methods, while highlighting such latter cases as being more substantiated, also find ever more evidence the more data is involved. They thus indicate higher attraction even if all values involved increased proportionally: "The larger the corpus is, the more significant a large number of co-occurrences is." (Hunston 2002: 73)[31] Thus, both types of measures run the risk of misrepresenting collocativity: "Effect-size measures typically fail to account for sampling variation and are prone to a low-frequency bias […], while significance measures are often prone to a high-frequency bias." (Evert 2009: 1228)

With differences between association measures being thus already very pronounced in terms of these three dimensions, it comes as no surprise that also practically "different association measures may lead to entirely different rankings of the word pairs" (Evert 2009: 1216). This situation, in turn, raises the question which method might be the 'best' one currently available. The respective discussions may focus on the mathematical properties (cf. Stubbs 1995 for an intuitive approach to the characteristics of some well-known measures) and the derivation of a measure from a statistical test, but these are only one side of the coin and actual analyses of the 'geometric' properties of the measure, i.e. which types of co-occurrences it foregrounds, may provide very different results. Evert (2009), for instance, argues against the commonly used *t*-score on the grounds that it "has been derived from an inappropriate hypothesis test" (Evert 2009: 1229). Yet, he finds that "despite its mathematical shortcomings", it empirically outperforms the other commonly-used measures in his study in case of PP-verb co-occurrences, which "illustrates the limitations of a purely theoretical discussion" (Evert 2009: 1238).[32] Similarly, Gries & Durrant (to appear) surmise that "the discussion of how to best approach the quantification and exploration of co-occurrence is likely to continue for the foreseeable future." Consequently, there is no single best measure, but rather different measures are variously well-suited to specific tasks. While the behavior of some frequently-used measures is relatively well-known, "[a]t this point, no definitive recommendation can be made. It is perhaps better to apply several measures with well-understood and distinct properties than attempt

[31] Making them inapplicable to differently-sized corpora.
[32] In contrast, Evert (2004: 145) finds the t-score slightly less powerful than both the log-likelihood and chi-square measure.

to find a single optimal choice." (Evert 2009: 1243) The present study follows this recommendation by approaching the same sets of data from the perspective of a combination of well-known and experimental measures to be presented in Section 3.3.2.

Association between >2 elements

The discussion above only pertained to the co-occurrence of two items. This is not an issue in case of collocations, even within a larger window, since these still only include two items and can thus still be analyzed within the frame of the 2x2 contingency tables above. For *n*-grams, however, this becomes problematic for any lengths above *n*=2, since these essentially contain two or more overlapping bigrams, each of which has its own contingency information. Thus, the question presents itself as to how the calculation of association scores can be extended to more than two elements, and indeed this is not an uncontroversial task (cf. Wahl & Gries 2018, Gries 2018a). Many common concordancers like *AntConc* actually avoid the issue entirely, offering only absolute frequencies or transitional probabilities between the first and *all other* items (which are thus treated like a single item). Fortunately, the issue can be addressed if *n*-grams are operationalized as sequences of bigrams. This perspective was introduced by Jelinek (1990), who proposes an iterative approach (cf. also Gries 2013, 2018a):

1. define a minimum value *m* of association strength,
2. extract all bigrams from the data,
3. select all bigrams exceeding the threshold *m* and merge them into longer sequences.

Implementations of this approach have been attempted in a few studies. While some are based on absolute frequencies (e.g. Stubbs's 2002 'collocational chains' or O'Donnel's 2011 Adjusted Frequency List), these come with the caveats of relying on frequency information discussed above. In terms of association scores, beyond Jelinek's (1990) MI-based study, attempts have also been made with the log-likelihood measure (the MERGE algorithm in Wahl & Gries 2018) and the rarely-seen lexical gravity measure (Daudaravičius & Marcinkevičienė 2004, Gries & Mukherjee 2010).

The iterative approach comes with the additional benefit that it provides a possible solution to another common problem, which lies in determining the length *n* for the sequences to be extracted. Usually, the issue is resolved by setting a value for *n* a

priori, for which "[c]urrently, $n = 4$ appears to be most fashionable." (Gries & Mukherjee 2010: 522) This can be seen to conform to the frequent recommendation of a collocation window span of 4 units, which was expressed in Jones & Sinclair (1974), Sinclair (1991) or Kennedy (1998). It originally appears to stem from Sinclair et al.'s (2004 [1970]) *OSTI Report*, reaffirmed in the reprint on the basis of a larger corpus (Sinclair et al. 2004 [1970]: xix).[33,34] Stubbs & Barth (2003), however, find 3- and 4-grams to provide better text-type-discriminating capabilities than both shorter and longer sequences. Different values are again recommended by Gries et al.'s (2011: 10) analysis, which yields the best discriminatory power at lengths 3 and 5 and finds 2- and 4-grams to be occasionally off. However, longer sequences appear to be relatively rare, with Altenberg (1998) finding the majority of *n*-grams to be of lengths between 2 and 4. Erman & Warren (2000), moreover, show that typical lengths depend on the mode of interaction and the function of the sequence: While three out of four types of prefabs in their analysis display means lengths only slightly over 2 words, lexical types average around 3. Since writing greatly prefers the latter category, average sequence length in the written mode is higher than for speech (2.8 vs. 2.61).

The realization that no single previously-defined *n*-gram length can equally do justice to all kinds of sequences leads to the question whether it might be more beneficial to dynamically determine the best length *n* for any specific bigram under scrutiny. The present analysis accommodates both perspectives in that it processes the corpus data in two separate forms: In a first step, the analysis follows a dynamic approach to determining the length *n* of any individual sequence within each corpus component. This allows to retrieve more variety-specific *n*-grams, since sequence lengths within a component depend entirely on collocational preferences between consecutive pairs of items. However, this comes at the cost of increased heterogeneity between corpus parts, since qualitatively different sequences can be produced within each component. This in turn reduces the frequency of *n*-gram types available for further study, since fewer types will be shared between varieties. To compensate for this, a second

[33] Of course, a 4-gram and a span of 4 refer to quite different structural circumstances: A 4-gram, by design, only takes into account three items to either the left or right of a particular node (and requires two separate *n*-grams to reflect either), while a collocation with span of 4 takes into account eight items altogether and at once.

[34] In contrast, Bartsch & Evert (2014: 57), in a study of precision and recall of various spans of collocations, discover that "the smaller L3 / R3 span is better than the commonly used L5 / R5 span".

methodological approach relies on more traditional sequences of fixed lengths. The lengths under scrutiny within this second step are, however, chosen on the basis of the lengths preferred by the data within the first, dynamic-length analytical step.[35] Without taking away from the results obtained in Chapter 5, it can be said that this leads to an analysis sequences of 2, 3 and 4 items in length, i.e. traditional 2-, 3- and 4-grams. The particulars of the current approach will be presented in Chapter 4, after the discussion of association measures applied by the present study in the next section.

3.3.2 Association Measures

For the purposes of the present study, rather than trying to build all discriminatory work on a single measure, the recommendation of triangulating between several measures will be followed. Towards this goal, the analysis combines what probably are the three most frequently applied and well-understood measures (MI-score, t-score, log-likelihood; cf. Gries 2018a: 227) with more innovative approaches which have so far shown promising results (lexical gravity, Delta P). This combined approach makes the findings of the study compatible with areas of linguistic research in which "some measures have been established as de facto standards, e. g. log-likelihood in computational linguistics, t-score and MI in computational lexicography" (Evert 2009: 1236), while simultaneously allowing for comparisons between these conventionalized measures as well as newer approaches. Side-by-side comparisons of different association metrics may be an ever more urgent task given that Gries (2008a: 20) as well as Evert (2009) appear to insinuate that there may be a certain degree of convention without substantiation. Both authors present the selection of a statistical approach as something more commonly made on the basis of ease of computation rather than on explanatory potential, with choices largely made on the basis in what typical software offers. While the present analysis will compare results obtained with different association measures, it should be noted that it remains practical in scope, applying measures to the task of attempting to discriminate varieties through combinatorial preferences. For detailed analyses of the mathematical and geometric properties of association measures, better sources can be found in Evert (2004) and Evert (2009),

[35] Calculation of association strengths still follows the above procedure of consecutively averaging association scores for each bigram involved.

which represent comprehensive studies to this day, but valuable information on collocation measures and statistics is also contained in e.g. Bartsch & Evert (2014: 56–59), Gries & Durrant (to appear), Daudaravičius & Marcinkevičienė (2004) and also Levshina (2015).

Mutual Information (MI)

The first metric included in the present study may well be the most commonly-used association measure in corpus linguistics in general. The extensive use that Pointwise Mutual Information (*MI*; Church & Hanks 1990) has found over the years may mostly be a result of its prominence or otherwise its relatively well-known behavior. It might just as much only be the fact that it is included as a statistic (and indeed the default setting) in two of the most widely used concordancers, *WordSmith Tools* and *AntConc*, that cause this measure to remain in constant use. In any case, "Mutual Information is gradually taking a central position in corpus linguistics" (Daudaravičius & Marcinkevičienė 2004: 324), has been "widely used in computational studies with very many target words and contextual features" (Levshina 2015: 238), and no introduction to collocation is complete without addressing *MI*. Merely due to its prominence and familiarity it serves the function of a benchmark measure, and is therefore also relevant in the context of the present study. MI is calculated as the base-2 logarithm of observed frequencies divided by expected frequencies:

$$MI = \log_2 \frac{a}{a_{exp}}$$

As a probabilistic/effect-size based rather than a statistical measure (it merely contrasts observed and expected frequencies), the MI-score is not dependent on corpus size and can therefore be applied to corpora of different sizes without relatively overestimating collocational attraction in the larger dataset. It does, however, suffer from the low-frequency bias discussed above, in that items of very low frequency can quickly attain association scores disproportionate to their relevance. Generally, "pointwise *MI* is known to return low-frequency but perfectly predictive collocations" (Gries & Durrant to appear: 14). Geometric analysis of MI shows that it appears to focus on pairs of items with roughly equal frequencies, of which those with overall high frequencies will receive low association scores while those with overall low frequencies are awarded higher values (Daudaravičius & Marcinkevičienė 2004: 325).

Therefore, MI retrieves mostly "quotations in foreign languages, specific noun phrases, first names and surnames preceded by titles, names of institutions and organisations" (Daudaravičius & Marcinkevičienė 2004: 335, cf. also Evert 2009: 1230). While some highly specific terms retrieved by *MI* may consequently be useful, the measure has a tendency of highlighting irrelevant combinations (which are likely to include hapaxes; Wahl & Gries 2018: 91). Correspondingly, Evert (2009) finds the measure to only fractionally outperform a random selection of item pairs:

> MI performs worse than all other measures and is close to the *baseline precision* […] corresponding to a random selection of candidates among all recurrent word pairs. Evaluation results always have to be interpreted in comparison to the baseline, and an association measure can only be considered useful if it achieves substantially better precision. (Evert 2009: 1239)

Several corrections have been offered for the weaknesses of the measure, which may include changes to the formula. Such approaches aim at increasing the relative weight of the observed frequencies in order to counterbalance the low-frequency bias of the measure. This is (partially) achieved by calculating some *k*-th power of the observed frequency (i.e. a^k; cf. Bartsch & Evert 2014: 55 for a comprehensive evaluation), but these changes also result in a very different measure from the original method (the MI^k family). The more straightforward (and more commonly applied) solution introduces a frequency threshold which can be variously decided (cf. Evert 2009: 1229–1230 for comparisons of MI with frequency thresholds of 5, 10, and 50). Within the current study, this threshold is required for the extraction of relevant items within the dynamic-length approach (cf. Chapter 4 for details). It is set at a moderate number of five occurrences, which is supposed to reflect the limited size of the ICE corpora. Pairs of items are furthermore only accepted as significant is they reach or surpass an MI-score of 3, and thus $MI \geq 3$ is set as a second threshold (cf. Hunston 2002: 71, McEnery et al. 2006: 56, Wood 2010, Biber & Reppen 2015).

T-score (t)

The t-score (Church et al. 1991) provides yet another standard measure of collocational attraction discussed in every introductory text and offered by virtually every software. It is often portrayed as a complement to *MI*, since it provides a statistical (instead of probabilistic) perspective on the data and compensates *MI*'s tendency of highlighting rare and specific sequences with an emphasis on frequent, established patterns. However, this also comes with the high-frequency bias discussed before, leading to t-score

awarding higher association to the same item combinations in a larger dataset and potentially disregarding relevant low-frequency cases (Levshina 2015: 238).

Typical t-score collocations reveal the measure's "focus on grammatical patterns like *of the* or *to be*" (Evert 2009: 1230), but it still finds high association for items of "a great variety, e.g. specific noun phrases, proper nouns, idioms, verb phrases." (Daudaravičius & Marcinkevičienė 2004: 335) In geometric analyses, the measure performs best for grammatical phenomena like the identification of PP-verb collocations in Evert (2009) but somewhat less so for e.g. figurative expressions (Evert 2004: 145).

The formula applied for the calculation of t-scores in the present analysis takes the following form:

$$t = \frac{a - a_{exp}}{\sqrt{a}}$$

Given the prominence of the observed frequencies for the final score, Stubbs (1995) analyzes the t-score as essentially approximating $t = \sqrt{a}$, but with a bias against combinations of two individual high-frequency items resulting from the subtraction of a_{exp} (with two items of high individual frequency resulting in a high value for their joint expected frequency). Daudaravičius & Marcinkevičienė (2004) accordingly observe that

> *T*-score highlights frequent word pairs which have high sums of frequencies [... F]or the same word frequency sums, the *T*-score increases when the respective word frequencies in a pair are considerably different, and decreases when they are similar. Therefore only those words forming pairs with individual high frequencies register high on the curve (Daudaravičius & Marcinkevičienė 2004: 326)

Concerning the implementation of the measure, Evert (2004: 83) criticizes the t-score for its essentially flawed derivation from Student's test and the approximations used therein. However, contrary to what might be expected based on this premise, both Evert (2004) and Evert (2009) find the precision of the t-score to be surprisingly good, even outperforming other measures in some datasets. A reason for this may lie in the measure implicitly setting a frequency threshold, so that "no pair type with $o \leq 22$ can achieve a significance of pv = 10^{-6} , regardless of its expected frequency." (Evert 2004: 114) Within the current study, it can be assumed that the focus on high-frequency items and grammatical sequences in addition to this implicit frequency

threshold may lead to more sequences being shared between the varietal datasets for variable-length *n*-grams. (Daudaravičius & Marcinkevičienė 2004: 336)

Threshold values to identify strongly collocated items are often set at $t > 2$ (cf. Hunston 2002: 70, Fitzpatrick & Barfield 2009: 136, Wood 2010) or $t \geq 2.576$ (cf. Bartsch 2004: 101, Xiao 2015: 110, Glass 2019: 123). The present study employs the second, more restrictive threshold (which results in a significance level of 0.01), since more reliable sequences are deemed more relevant within the dynamic-length approach than a larger number of items.

Log-likelihood (G^2)

A final classic measure is the log-likelihood statistic (G^2; Dunning 1993), which is often portrayed as an in-between measure of *MI* on the one hand and *t* on the other (Lindquist 2009: 76–78). When contrasted against other measures, G^2 has been found to extract figurative expressions better than the previous measures and to show relatively stable precision values for different linguistic variables (Evert 2004: 145), even if it is not always the best-performing measure. It thus represents a solid all-purpose choice, while exceeding only rarely within a specific taask.

Like the t-score, it constitutes a significance-based measure and is thus likely to rank more frequent items higher (Gries & Durrant to appear: 14). But while the former underestimates significance in contrast to Fisher's exact test, log-likelihood shows an excellent approximation to this statistic (cf. Evert 2009: 1237, Gries & Ellis 2015: 237), making it a mathematically more sound measure than *t*. Calculation of G^2 is, however, somewhat more involved, which may explain its comparatively lower extent of use. The formula employed in the present study takes all cell values of the co-occurrence table into account, focusing mainly on a contrast of observed vs. expected frequencies for all cells (much like an extension of MI), but multiplying each fraction with the respective observed frequency before summation:

$$G^2 = 2 * \left(a * \log_2 \frac{a}{a_{exp}} + b * \log_2 \frac{b}{b_{exp}} + c * \log_2 \frac{c}{c_{exp}} + d * \log_2 \frac{d}{d_{exp}} \right)$$

In contrast to the other measures discussed before, this formula does, however, not result in negative values for O<E, i.e. cases of repulsion between the elements

involved. In order to compensate for this, values are multiplied by -1 in case the observed frequency of a bigram is less than the expected.

Only few studies appear to discuss typical association thresholds for the log-likelihood measure (e.g. Xiao 2015: 111, who uses a relatively low $G^2 \geq 3.84$). Instead, critical values are usually chosen on the basis of probabilities (significance levels) of a certain log-likelihood score (cf. McEnery et al. 2006: 55). Since best practice accounts are rare, the association threshold for G^2 within the present study is set so that it results in a significance level of 0.01 (as in case of the t-score). This results in a threshold of $G^2 \geq 6.64$.

Lexical Gravity g

With several established measures thus included in the present study and serving the function of a familiar benchmark, the remaining metrics represent newer and less familiar approaches. The first of these comes in the form of the lexical gravity measure (*g*) originally proposed by Mason (1997, 1999) and applied to corpus linguistics in the form of *gravity counts* by Daudaravičius and Marcinkevičienė (2004). In contrast to the previously-discussed measures, lexical gravity offers a unique perspective on co-occurrence data in that it takes not only token but also type frequencies within a multiword unit into account and thus introduces a 'coefficient of diversity': The statistic shows a higher degree of collocational attraction for any two items if the absolute number of absences of any one constituent (cells b, c in Table 3.1, above) is caused by a large number of different types (and indeed it appears to be the only measure to involve type frequencies; cf. Gries & Durrant to appear: 12).

The design of the measure thus introduces competition of types over a given slot in addition to frequent co-occurrence of the two types involved. As such, it promotes 'promiscuity', i.e. types chosen over competitors for co-occurrence with the other constituent:

> At its heart, *LG* is based on the sum of the forward and backward transitional probabilities (TPs) of a two-way co-occurrence. However, each TP is weighted by the type frequency (i.e., the number of different word types) that can occupy its outcome slot, given its cue. Thus, for a given (forward or backward) TP, there is a reward for promiscuity in possible outcomes and a punishment for faithfulness (this is because a high TP is more impressive when it occurs in the context of many possible outcomes). (Wahl & Gries 2018: 92)

If, for example, *Starship* co-occurred with *Enterprise* 100 times in a given dataset, while in 50 cases it did not, lexical gravity of *Starship Enterprise* would be recognized as lower in case these 50 cases were comprised of only one other type than if there were many different types (*Voyager*, *Defiant*, etc.[36]) 'competing' for co-occurring with *Starship* in the dataset (cf. Gries & Mukherjee 2010: 529 for visualizations of the impact of manipulating individual values). Consequently, the formula for lexical gravity goes beyond the factors in Table 3.1: For the frequencies of types after the first constituent $O_{types\ after\ 1}$ and before the second constituent $O_{types\ before\ 2}$, g is calculated as follows:

$$g = \log_2\left(\frac{a * types_{after\ w1}}{R_1}\right) + \log_2\left(\frac{a * types_{before\ w2}}{C_1}\right)$$

In this sense, lexical gravity transcends more traditional measures with their simpler 2x2 (co-)occurrence tables. While it certainly appears sensible to take the type frequencies of competing co-occurrence items into account, it is not necessarily true that the metric also provides an improvement over others computationally: As Gries (2013: 161) discusses, co-occurrence contingencies are Zipf'ian distributed just as all corpus data, which means few high-frequency types but many of lower frequency. Thus, if there are only few tokens after word$_1$, they are likely to be of only few types, and if there are many tokens, they will usually be of proportionally more types. Thus, while theoretically, lexical gravity offers an improvement over measures only using token frequencies, the results need not be dramatically different.

In terms of the typical sequences extracted, one of the few studies yet performed on the basis of the measure (Daudaravičius & Marcinkevičienė 2004) observes a wider range of *n*-grams to be identifiable by *g* than by *t*. The authors regard this as a consequence of the measure being less dependent on the frequencies of individual words in the sequence, and thus producing more balanced results than other measures (cf. also the comprehensive evaluation in Gries 2010a as well as Ferraresi & Gries 2011). Still, the most reliable results are obtained for word pairs with combined constituent frequencies of n>10 (Daudaravičius & Marcinkevičienė 2004: 331). Gries (2010a) also

[36] Which are predominantly referred to by their designations only, i.e. USS Defiant, while *Starship Enterprise* recurs in the opening thematic of the respective show.

observes a better reflection of the corpus sampling scheme in the results of *g*-based bigrams than those observed for t-scores. Within the analysis of variable-length *n*-grams, however, the measures tendency towards producing many unique sequences (Daudaravičius & Marcinkevičienė 2004: 336) may lead to strongly diverging datasets and consequently to relatively fewer shared data between varieties.

Association thresholds for the lexical gravity measure have hardly been discussed. The original authors recommend a critical value of $g \geq 5.5$ (Daudaravičius & Marcinkevičienė 2004: 333), which has not been challenged. As such, the few studies which employ the measure apply this value or disregard thresholds entirely. Wahl & Gries (2018: 92) further comment that this threshold corresponds to a "*p*-value of approximately 0.02". This makes it slighly less restrictive than those discussed above, but in lieu of clear recommendations, it appears sensible to employ what has been suggested by the original authors.

Delta P (ΔP)

Another very different route to co-occurrence is again taken by the Delta P (*ΔP*) measure (Gries 2013), which, unlike all other included in the present study, includes directionality of the collocation into its statistic (cf. Michelbacher et al. 2011 for an overview of further asymmetric association measures). As such, it attempts to differentiate the two values of collocational strength that are actually conflated in bi-directional scores by calculating the transitional probabilities of one word co-selecting the other. Within the present study, only a prospective version of directionality is analyzed with the help of Delta P, since this mirrors the direction of sequence generation within the dynamic-length approach. Direction of association can be designated by lowercase numbers, indicating the probabilities of 'item 2 selected by item 1' ($\Delta P_{2|1}$):

$$\Delta P_{2|1} = \frac{a}{(a+b)} - \frac{c}{(c+d)}$$

Within this formula, the sums in the denominators correspond to the marginal frequencies R_1 and R_2, respectively, which represent the total frequencies of word$_1$ occurring or not occurring. As such, the formula can be loosely rephrased as

$$\Delta P_{2|1} = \frac{a}{word_1} - \frac{c}{\neg word_1}$$

Delta $P_{2|1}$ thus quantifies the probability of finding the bigram (cell a) in case $word_1$ is found in the initial position, minus the chance of finding $word_2$ (cell c) in case of any other item in the first position. Given the larger number of items in the second denominator, the second part of the equation will always be minute in comparison to the first part. As such, "ΔP is transitional probability minus a small adjustment, which 'punishes' pairs whose second word also frequently occurs in other combinations." (Schneider 2018: 7)

By including the Delta P measure, the present study attempts to find out whether directional measures harmonize better with the consecutive merging of adjacent bigrams. In theory, this should apply, since a directional measure suits the directional process of consecutive bigram joining. Whether in actuality better results can be obtained with this process and the application of ΔP has so far not been a topic of research.

Threshold values for ΔP for use within the dynamic-length approach to *n*-gram generation have not yet been established, neither by the original author nor within later-studies (cf. Gries 2013, Wahl 2015, Schneider 2018). In lieu of empirical values to identify mutually attracting items, the dynamic-length method will employ a comparatively loose criterion of $\Delta P_{2|1} \geq 0$. This only removes bigrams in which the first item disprefers the second, but leaves all other items within the data. This is likely to result in the largest retention of base bigrams, and it will be interesting to determine how this affects sequence generation and clustering results within the dynamic-length approach. Please recall that any threshold values only come into effect in the dynamic determination of sequence length but not within the static-length ('traditional') approach.

4 Methodology

The following chapter sketches out the methodological framework and issues specific to this study. For this purpose, the corpus employed for the analysis will be discussed first, since the question needs to be answered why this particular data, the *International Corpus of English* (ICE), was selected. Section 4.1 will focus on the advantages of the corpus as well as potential caveats of applying it to the study of collocational patterns, also discussing relevant aspects of internal inconsistencies between the various components. Section 4.2 addresses the question of how these issues can be resolved for the current purposes, and how we can arrive at the two types of data employed by this study, i.e. a set of homogenized corpus text in lexical form as well as a POS-annotated version. Building on this, Section 4.3 details the extraction procedure for both lexical as well as grammatical bigrams and the generation of two types of longer sequences (*n*-grams of both dynamic and static lengths), together with their association statistics. Section 4.4, then, addresses the evaluation of the data thus generated, discussing clustering techniques as the statistical approach chosen for the detection of patterns in the data as well as how to interpret these. It will also lay out the research questions and hypotheses which form the framework analysis in Chapter 5.[37]

4.1 Applying the International Corpus of English to Large-scale Studies of Co-occurrence

4.1.1 Studying Co-occurrence with ICE: Benefits and Caveats

When work on the *International Corpus of English* project (Greenbaum 1988, Greenbaum 1996a) started, the goal was "to provide the resources for comparative studies of the English used in countries where it is either a majority first language (ENL) (for example, Canada and Australia) or an official additional language (ESL) (for example, India and Nigeria)." (Greenbaum 1996b: 3) With this aim, the project surpassed previous ambitions in corpus linguistics, which had mostly been limited to comparisons of

[37] With the major exception of part-of-speech annotation, all methodological steps (extraction and analysis) were carried out using *R* (versions 3.5.3 and 3.6.3 for data retrieval/cleanup and analysis, respectively; R Core Team (2019, 2020)).

the two native speaker varieties of British and American English: "For the last decade we have had two corpora that have stimulated scholars to make comparisons between American and British English." (Greenbaum 1988: 315) What Greenbaum refers to are, of course, the first two members of what is known today as the Brown family of corpora[38], i.e. the Brown and LOB corpus. These two collections of linguistic data represented the largest set of comparable corpora, i.e. such that were compiled under the same framework and general guidelines to achieve maximal similarity in corpus design and thus provide the best data for comparison of the varieties of English sampled therein.

While a few other varieties of English had also received some attention with regards to corpus compilation – most prominently Indian English with the Kolhapur Corpus (Shastri et al. 1986) – major drawbacks lay in low numbers of varieties covered and a usual disregard for spoken language (not sampled in the underlying Brown compilation scheme). It was only with the ICE project that comparisons of a large number of both native and non-native Englishes would become a corpus-linguistic reality, and that not only written but also spoken language could be analyzed comparatively on similarly structured data. In this setup, written texts display the more stable and established characteristics, since they "inevitably […] approximate to local varieties of standardized English" (Kirk & Nelson 2018: 698). Spoken language, by contrast, often provides the source of innovations which may eventually become nativized within a variety, and thus the inclusion of a substantial amount of spoken data (60% of tokens) can be regarded as a major benefit of the ICE corpus over other data. Since the publication of its earliest components (starting with ICE-GB in 1998) roughly 20 years ago, a number of studies "now all too copious to encapsulate within a single ICE bibliography" (Kirk & Nelson 2018: 699) have been employing the diverse spectrum of regional components collected as part of the ICE project, which certainly attests to the success and relevance of this enterprise.

[38] The so-called Brown family of corpora refers to the initial Brown Corpus (Francis & Kucera 1964), its British English equivalent LOB (Lancaster-Oslo/Bergen Corpus; Johansson et al. 1978), as well as their 'updated' counterparts containing samples of language collected 30 years after these initial two corpora, FROWN (Freiburg Brown Corpus; Hundt et al. 1999) and FLOB (Freiburg LOB Corpus; Hundt et al. 1998).

However, three decades after its initial conception, it comes as no surprise that the ICE corpus begins to show some signs of ageing: Most of the actual data at the time of writing is sourced from material at least 20 years of age. This presents obvious problems with the ICE data for the study of current linguistic settings, since in many contexts (and at different speeds) the ways people speak will have changed. While contextualizing the exact degree of change would require an updated set of data (and indeed calls for this have been made), it is clear that 20 years represents a substantial 'lag' to current language use. Compounding this problem, there is the issue that, while most (earlier) ICE components source data from the early 1990s, material for some other ICE components is of more recent date,[39] which introduces a (however slight) diachronic perspective into the set of ICE components (for a current overview of the state of the ICE components, cf. Kirk & Nelson 2018). Both issues are, however, usually more the norm than the exception in (contrastive) corpus studies and might just have to be accepted and acknowledged in the interpretation of the data.

A third type of criticism concerns the size of corpora like ICE, which may appear tiny in comparison to recent mega corpora potentially incorporating billions of words, and the related question whether they can actually be fruitfully applied to the analysis of a particular linguistic feature. Gerald Nelson, coordinator of the ICE project until 2017, acknowledges these limitations for some types of studies:

> It is certainly true that corpora of this size are very limited for those who are interested in phraseology, or in collocational studies, or in "rare" linguistic phenomena. […] I must point out that ICE corpora were not designed to be "all-purpose" corpora. Instead, they were designed primarily for the study of the grammar of English worldwide. (Nelson 2015: 38)

Some researchers have attempted to complement the ICE components with additional data to compensate for their limited size (e.g. Sedlatschek 2009), while more commonly calls for larger corpora have been made. On the subject of lexicogrammatical variation, Schilk (2006: 313) presents a skeptical case in point, stating that "a token-size of one million is simply not enough for lexical analyses". To this end, he later (Schilk 2011) develops a mega-corpus of c. 100m words of acrolectal Indian newspaper English and contrasts it with findings from ICE-India. On the basis of this comparison, he weakens his previous claim about the inadequacy of smaller corpora,

[39] E.g. ICE-Nigeria, which sources 2000s data for the written part and early 2010s for speech, or ICE-Sri Lanka's timeframe of 2003-2009 for writing and the consecutive period 2010-2018 for spoken data.

distinguishing between central uses (for which ICE can be used) and the "peripheral use of linguistic terms" (Schilk 2011: 153), for which larger datasets are usually required. In essence, this supports the use of ICE data for the description of more general trends within a variety and their degrees of (dis-)similarity. Truly tracing the roots of a particular (innovative) feature will usually not be feasible, since they start out on the periphery and may only enter into central use at a later stage. Beyond their overall size, data sparsity in ICE may furthermore interfere with more reliable inferences about text-type specific characteristics or the social constraints of language use. The non-availability of socio-biographic information for most ICE components often limits the corpus to an analysis of 'educated' types of English on the level of 'national standard(izing) varieties'.

Despite the potential issues of smaller corpora in general and the ICE corpus in particular, size is not the only important corpus feature when it comes to describing a variety. A particularly prominent case can be found in the *Corpus of Global Web-based English* (GloWbE; Davies 2013, Davies & Fuchs 2015), which comprises 1.9 billion words and may have presented a viable alternative to the ICE corpus in the context of the current study. However, two recent studies on this data demonstrate that bigger is not necessarily better: On the one hand, Loureiro-Porto (2017) evaluates the GloWbE data for Great Britain against more carefully compiled corpora in the form of ICE, FLOB and the *British National Corpus* (BNC). She finds that the traditional corpora, even if they exhibit different internal structures, provide relatively mutually consistent results, which contrast distinctly against those of GloWbE. While the web-based nature of GloWbE may produce some of that effect, she still assesses that "[t]he strong degrees of correlation between ICE, BNC and FLOB cannot […] be considered a random result, and therefore, these corpora must be considered to represent the GB variety very thoroughly." (Loureiro-Porto 2017: 467) She further assesses the (vague) internal structure of GloWbE (into blogs and 'general') to not hold up to empirical evaluation. She thus cautions against regarding the corpus as a true alternative to more carefully compiled corpora, asserting that "the differences between GloWbE and ICE are too pervasive to consider these two corpora equivalent alternatives for any linguistic study." (Loureiro-Porto 2017: 468) To make matters worse, the classification into varieties based on regional identifiers in the web documents appears to be a highly

error-prone subject, even beyond the initial cautionary notes provided by Nelson (2015: 39): In her attempt at authenticating authorship of the GloWbE texts, Güldenring (2020) discovers that the attribution of speakers to national varieties as performed by GloWbE should not be relied upon, finding substantial part of the data to consist of quotations from texts from other varieties. In some cases, texts from outside the indicated variety actually constituted the majority of her data, with 57% of her Singaporean English data provided from external sources (Güldenring 2020: 94). These findings should raise serious doubts as to Davies & Fuchs's (2015: 5) claim that "all subcorpora constitute representative samples of how these national varieties of English are used in web-based communication."

All that said, the need for larger corpora than ICE shall not be rejected here. It can, however, be seen that smaller and well-assembled corpora may still be the more viable option for many studies even given today's mega-corpus trend. Particularly for studies such as the present one, which focus on general tendencies within a variety, even smaller corpora can still be regarded as being sufficient in size. The relevance of smaller but well-structured corpora is also stressed by Kirk & Nelson (2018) in their discussion of a potential second generation ICE corpus (based on feedback provided by current ICE teams):

> Whatever decisions regarding second generation corpora are finally taken, the value of small-scale, carefully structured, well annotated (especially with comprehensive biodata) corpora continued to be preferred for many research purposes—not least comparable studies of national varieties—over rapidly-compiled, anonymous, indiscriminate, web-derived mega-corpora such as GloWbE (Kirk & Nelson 2018: 707)

What is more, the argument about the limited applicability of ICE to the study of low-frequency features could also be flipped on its head, in that it might be worth questioning whether these less frequent forms truly shape the character of a variety of English: From that perspective, high-frequency items and central trends (i.e. the typical and idiomatic, cf. Chapter 3) can be regarded as being of greater importance in determining the general character of a variety than low-frequency and peripheral uses (i.e. rare, potentially highly marked features, possibly entering a variety). Empirical findings of this kind can, for instance, be inferred from Koch & Bernaisch (2013), who zoom in on 'new ditransitives' in South Asian Englishes. While they discover a general trend for deriving 'new ditransitives', only a single verb can be attested at somewhat elevated frequencies. Similarly, Koch et al. (2016) retrieve a clear distinction between

established and peripheral uses of the 'intrusive *as*'-construction in complex-transitive verbs (e.g. "They called that as nonsense"), with only six out of 60 verb lemmas contributing the vast majority of occurrences. The large number of fringe cases reported by these studies may certainly open up interesting avenues for more specialized research focusing on the processes by which varieties of English develop their own norms, but they can certainly not be said to represent the quantitatively distinguishing characteristics of these varieties.

Blanket statements on corpus size are certainly inadequate, as can be seen in relevant findings from Lange's (2012) in-depth analysis of Spoken Indian English on the ground of only c. 20% of the ICE component or Biber's (1990) assessment that even lexical sequences can differentiate genres in samples as small as 5,000–10,000 words of text. Usually, fine-grained objects are said to require more data to account for larger degrees of diversity of the linguistic items while coarse-grained studies (e.g. syntactic structures) are necessarily found in less diverse configurations and thus produce more coherent data in less linguistic material. Feature granularity is, however, a two-faced aspect. While at a first glance, it should be expected that collocational analyses run the risk of not having enough data for a contrast of detailed variety-specific patterns, it has been found that they indeed more easily display variety-specific preferences, in contrast to items on the other end of the spectrum:

> First empirical studies […] point to the fact that the degree to which variety-specific […] structures can be found is related to the level of descriptive granularity. While fine-grained objects of investigation (such as collocational or complementational patterns, cf. Schilk 2011) may be more likely to yield variety-specific structures, more coarse-grained studies […] tend to show cross-varietal similarities (Bernaisch & Koch 2016: 118–119)

In summary, it thus appears that patterned language use within a variety can well be described with the help of the ICE corpus, but only on the level of general difference and/or similarity intended by the present study. Comparisons need to focus on a more limited set of high-frequency items. It may even show that fine-grained lexical structures produce distinction between varieties more easily, i.e. with less data, while similarities may be more common on the grammatical end of the lexicogrammatical spectrum even if more data were available.

4.1.2 Heterogeneity within the International Corpus of English

The conventional way of referring to the *International Corpus of English* in general and its variety-specific parts is to call the variety-specific collections of texts 'components' which together form the entirety of the 'corpus'.[40] This is commonly meant to strengthen the claim that ICE is the larger whole while the variety-specific components supply individual facets to the greater picture. Conceptually, there is certainly some beauty to this approach – practically, however, general similarity of the components quickly gives way to divergent individual realizations of each component, a point conceded by Kirk & Nelson (2018):

> Although, as ICE Coordinator, the second named author of this paper has strenuously pursued the creation of proverbial unity amongst diversity to prevent fragmentation, he has found it increasingly challenging to create overall a prevailing, unifying ethos. (Kirk & Nelson 2018: 698)

The unfortunate consequence of this process is that, for a corpus as dedicated to facilitating cross-varietal comparisons as the *International Corpus of English*, researchers delving into the corpus components are faced with a considerable degree of internal variation, the reasons of which are often attributed to challenges specific to a regional setting and different compilation time frames of the various. Issues may thus arise in terms of corpus design and structure in general, i.e. the material which forms the corpus, but also and much less documented in terms of the homogenous application of markup to the texts contained in the corpus. In all these cases, ICE-East Africa presents a particular drastic case: In terms of content, the component samples material from two varieties of English, reaching the usual spoken token count only by merging the Kenyan and Tanzanian data, and deviating from the usual genre/text type stratification.[41] The component furthermore deviates from the normal genre/text type stratification, representing several written and all but one spoken text category (class

[40] Regular deviations from this pattern can, however, be found, e.g. in the nomenclature used by the previous ICE coordinator Nelson in Kirk & Nelson(2018), in the passage from Nelson (2015: 38) quoted earlier in Section 4.1.1 or as expressed in the focus of the 34th *ICAME Journal* on "ICE corpora of New Englishes in the making" (underscore CK).

[41] Cf. Hudson-Ettle & Schmied (1999: 5) for a discussion of the reasons that led to the eventual creation of "two parallel East African written corpora" under the roof of a single ICE component, with the spoken part mixing texts from both settings.

lessons) with the normal number of texts (Hudson-Ettle & Schmied 1999: 52–53).[42] Three further components currently lack (completed) spoken parts but are still included in the present study with only their written data. This concerns ICE-USA as well as the Ghanaian and Ugandan ICE components, for which the spoken parts are still under development.[43]

A second major reason for internal inconsistency stems from styles of text annotation and markup application. Slight deviations from the guidelines put forward in the ICE markup manuals (Nelson 2002a, Nelson 2002b) are to be expected given different project teams and the necessity of adapting general guidelines to a concrete linguistic scenario. A few components, however, also diverge in relatively major ways from the norm. For qualitatively-oriented analyses of a restricted dataset, these differences may be relatively easy to handle through appropriate judgement on any individual inconsistency. With the inclusion of further components and the application of a quantitative approach to the data, less immediate interaction with and partially automated handling of the data are the consequence. Within the scope of the present project, this requires a separate discussion of corpus heterogeneity and the potential for homogenization.

The ICE manuals distinguish markup broadly into three categories: 'Essential' markup includes the annotation of text units, speaker IDs, extra-corpus material or uncertain transcription and unclear words, 'recommended' markup concerns, for instance, incomplete or foreign or indigenous words, pauses, quotations and overlaps, while markup for text normalization, discontinuous words and any typeface are considered 'optional' (cf. Nelson 2002a: 14 and Nelson 2002b: 18). Additionally, the ICE markup manuals provide a list of special character SGML encodings to be shared by

[42] Several other components differ from the norm in less drastic ways, such as the Nigerian component not compiling texts below the 2,000 word count into a single file for a text category, thus totaling 902 files instead of the usual 500, of which only 92 are in excess of the 300-text norm for speech. An ICE-Philippines component with only 278 instead of 300 spoken texts, without an explicit notice included in either version, has also come to the author's notice. ICE-Ghana (under development but finished in the traditional sense of other ICE components according to the compilers) currently also offers 12 files with written material beyond the usual 200, included in a non-standard "W1C" text category comprising business and private e-mails. This category is excluded from the analysis in Chapter 5, since the category is incomplete and in excess of the regular corpus contents.

[43] For ICE-USA, the *Santa Barbara Corpus of Spoken American English* constitutes the spoken part. While parts of this corpus totaling 249,000 words are currently available, the difference in size to the otherwise 600,000 words of other components makes inclusion of this corpus not feasible for this study.

all components (Nelson 2002b: 17)[44]. Using all components largely restricts the usable markup to the 'greatest common denominator', i.e. the 'essential' category. However, in case further markup is applied, it may be necessary to extract relevant information out of these tags.

Beyond the limitations in terms of the 'functional' scope of markup application, divergences can also extend to the concrete forms by which a type is expressed in the corpus text. Some of these differences appear more motivated than others, such as the use of Unicode-based XML markup in some newer components (Nigeria and Ghana), but even these introduce divergent tag names for the same general type and purpose. Smaller deviations concern, for instance, ICE-Nigeria's version of speaker markup ("`Transcription 1/2/3`", meaning speaker 1, 2 or 3) in the plain text version of the spoken data, [45] ICE-Ireland remaining at the reduced text-unit markers (`<#>`), and ICE-East Africa furthermore using these in divergent forms (`<#/>`)[46] in writing (only in the RTF version[47]) or not at all in speech. An example of the result of the latter decision is shown in example (1), which presents a particularly non-standard form of annotation at best suggesting speech units through the use of capital letters. This may already be only barely parsable even by a human researcher, let alone reliably by a computer.

(1) <$A> I am Goro Kamau Welcome Well to begin with I would like to ask Dr Gikenye to tell us something in brief what <./>ma medicine is all about (ICE-EA S1B021K / br-discussionK.rtf)

Further surprising (and unnecessary) deviations of ICE-East Africa from the ICE norm concern, for instance, mismatching pairs of opening and closing labels (e.g. the extra-corpus text marker `<X_>…</X>` instead of `<X>…</X>`), the insertion of corpus text identifiers (e.g. "`W1B-BT1`") inside the corpus text (i.e. not as a tag), or displacement of

[44] While both manuals (for spoken and written markup) contain lists of SGML characters, the one presented in the manual for written texts lists those used in the spoken corpus parts in addition to those for written parts.

[45] The latest release as of the time of writing (Version 'Nov 3, 2015') introduces a separate plain-text version without any text unit markup, presenting each utterance in a separate line.

[46] Empty-element tags in ICE-EA, i.e. those without a closing label (compare, e.g. the text unit markup `<#/>` against the subtext marker `<I>…</I>`) appear to follow more modern annotation styles by always ending in a forward slash. In ICE-Ireland (retaining the reduced marker even after release), ICE-East Africa's `<#/>` would be represented as `<#>`.

[47] There are two TXT versions of the corpus: The version edited for use with *WordSmith Tools* changes the markup drastically without respective documentation in the manual, while the "text2000" version is missing other markup such as the subtext and extra-corpus tags as well as the speech unit marker.

normalized corpus text (erroneous use for which putative corrections are suggested by the editors) into the normalization tag itself (cf. the manual's example of the correction of superfluous <m> in "<-/tommorrow>", which removes the original spelling from the corpus text; Hudson-Ettle & Schmied 1999: 14).[48] Some of the inconsistencies between components might eventually be resolved through the release of ICEonline, but this undertaking is, at the current moment, neither completed nor does it encompass all ICE components (ICE-EA being a case in point).

These differences between corpus components may still be relatively straightforward to identify, even if it may already be necessary to delve into the data to identify and understand them properly. What is worse, not all components clearly spell out their design strategies (and consequent deviations from the norm), with the lack of systematic documentation thus unsurprisingly emerging as a major weak spot of the ICE components from the questionnaires presented in Kirk & Nelson (2018). However, there is even more variation on a less immediately visible level for which documentation is almost universally absent, and which only become apparent after close and long inspection of the different datasets in comparison. One of these areas concerns the annotation of special characters ('SGML codes'), with many components adding SGML characters codes beyond the specifications in the ICE manuals (e.g. **&obrack;** and **&cbrack;** for an opening and closing bracket, respectively, or **&ampersand;** instead of **&** for <&>; cf. also Nelson 2002b: 17). Again, this may not pose a problem for the human researcher, who can quickly understand the intent of the code, but an automatic process will need to account for these encodings. This is, however, more easily said than done, since these idiosyncrasies are only rarely documented in the accompanying manuals, while other manuals spell out their encoding schemes (e.g. ICE-Ireland; Kallen & Kirk 2008) but simultaneously introduce entirely new sets of (non-SGML) codes specific to only a single component (e.g. "**/e**" for the letter <é>).[49]

[48] It should be stated that the normal ICE practice also somewhat confounds the distinction between corpus layers by including the putative corrections as corpus text marked up as corrections ("<+>…</+>").While it could be argued that what constitutes final corpus text is dependent on corpus compilers' perspectives (thus influencing markup application), the reversal of regular ICE markup practice in ICE-East Africa still appears unwarranted.

[49] This even varies within the component, since the same character is coded as "**&/e**" in the written component, presumably to avoid ambiguity with a word-initial <e> after a forward slash (to separate two words). All of these changes probably follow the intent of making the text file more easily readable, but sacrifice direct comparability in the process.

Finally, even the same type of markup can be applied in different ways. A particularly surprising example was discovered in the annotation of line breaks in ICE-Canada and ICE-Jamaica, where whitespaces surround all markup. This can even extend to the line-break tag (<l>), which results in the introduction of a whitespace into a continuous word. Mere removal of the tag would consequently split this word into two parts. This is not even a scarce phenomenon, since line breaks can be found in 3,879 and 2,384 instances in these components, respectively. Similarly, some components split up contracted forms into the (pro-)noun and (reduced) verb (e.g. ICE-Hong Kong, ICE-Philippines) while others remain at one continuous form overall (e.g. ICE-New Zealand, ICE-Singapore).

It becomes clear that a consistent semi-automatic approach to the ICE data needs to account for the high levels of (unexpected) heterogeneity within the varietal components, so that differences in corpus design do not become a major contributor of difference between components. For researchers focusing on relatively little corpus material, extensive manual inspection, correction of errors, application of additional markup as well as eventual familiarity with their data presents a feasible, if painstaking, process. For studies building on a multitude of components and employing at least partially automated extraction and analysis, variation in corpus annotation results in the 'largest denominator' type of approach discussed above, which may have to ignore potentially relevant information simply on the grounds that annotation is not equally represented across all corpus components. Based on the discussions above, the following subsection will lay out which of the corpus markup can still be used to make the ICE components as similar in shape as possible.

Some particularly attractive types of markup unavailable to the present analysis due to their inconsistent application across components concern the identification of incomplete words and text normalization, consistent removal of which (and reliance on the putative corrections) might allow cleaning up the data to the patterns actually intended by the speakers, e.g. in case of stutters, repetitions, crossed-out sections or other types of self-editing. That said, it also stands to reason that attempting to 'correct' this perceived 'noise' may inadequately skew the data. In particular, any corrections beyond individual letters introduce the analyst's perspective into potentially ambiguous scenarios, and the editors of both ICE-Ireland and ICE-East Africa explicitly

refrain from introducing putative targets (Kallen & Kirk 2008: 14, Hudson-Ettle & Schmied 1999: 11) with the latter also acknowledging the immense effort of such an undertaking.[50] In other components, different perspectives on correction emerge in the form of quantitative differences in error markup: ICE-Ireland and ICE-Hong Kong only contain normative insertions in c. 500 cases, while ICE-India and ICE-Nigeria (the `<error>` markup and its variants) surpass 2,000 and 3,000, respectively, and ICE-Canada approaches 3,800 cases of editorial corrections.[51] Surprisingly, for all the rejection of corrective intrusion into the data, ICE-East Africa still displays relatively a high number of 1,872 insertions, in part due to the compilers' decision to also include grammatical corrections beyond the ICE guidelines.[52] While different degrees of erroneous language use may have an effect on these numbers, they are far more likely to be a result of different perspectives on intrusion into the text as well as practical concerns, presenting them as an unreliable source in a quantitative approach.

4.2 Data preparation and normalization

4.2.1 Homogenizing the ICE components

While the 'largest denominator' effect from the high degree of heterogeneity between the ICE components discussed above can already exert a considerable effect when working with only a few select components, the problem is compounded by each additional corpus component. For the present study, building on 15 ICE components, this challenge is particularly daunting. Table 4.1 lays out all ICE components and the shorthand labels used to refer to them, and furthermore indicates which components offer spoken or written parts.[53] ICE-EA, which combines material from Kenya and

[50] Specifically commenting here on the correction of student examination essays.
[51] The components for Ireland, Hong Kong and Nigeria only contain normative insertions in writing.
[52] Additionally, it is by no means always possible to identify the extent of other types of intrusion into the corpus text. Judging from markup frequency alone, the entirety of ICE-Singapore, for instance, should not contain a single instance of incomplete words, even in the spoken texts, which seems very unlikely.
[53] All further discussions as well as the analyses in Chapter 5 pertain to the txt-based versions of the ICE components with the following restrictions in case choice between variants was possible: For ICE-East Africa, data was extracted from the RTF files and converted to TXT in order to access all markup as discussed in Section 4.1.2 (`EA_rtf2txt.r`) and split into a combined spoken part and two separate

Tanzania, contained sufficient tokens in the written part for the creation of separate Kenyan and Tanzanian sub-components. In speech, however, there is only enough material for a joint EA component.

Table 4.1: ICE components in the present study, together with shorthand codes and availability of spoken or written data.

Country/Region	Code	Spoken Data	Written Data
Canada	CAN	✓	✓
East Africa	EA	✓	✓
⌞ Kenya	KY		✓
⌞Tanzania	TZ		✓
Great Britain	GB	✓	✓
Ghana	GH	×	✓
Hong Kong	HK	✓	✓
India	IND	✓	✓
Ireland	IRL	✓	✓
Jamaica	JA	✓	✓
Nigeria	NIG	✓	✓
New Zealand	NZ	✓	✓
Philippines	PHI	✓	✓
Singapore	SIN	✓	✓
Sri Lanka	SL	✓	✓
Uganda	UG	×	✓
United States of America	USA	×	✓

Given the compounded effect of corpus heterogeneity, a considerable effort was made to homogenize and normalize the data as much as possible, and this section lays out conceptual considerations and methodological steps involved in this process. It should be noted at the outset that simple deletion of all markup would not have sufficed: For instance, confer the case of corrections discussed above, for which ICE-

written parts according to the folder structure and manual. The separate 'written to be spoken' category was included in the spoken part. A pre-release copy of ICE-Ghana, dated 2018 and finished in the sense of other ICE components (except for the additional and incomplete W1C category, which was discarded), was obtained from its compilation team under the supervision of Magnus Huber at Justus Liebig University Giessen. ICE-Nigeria was used in the Nov 2015 version available from sourceforge.net, and the TXT files were used for the spoken data (with segmentation by "`Transcription 1/2/…`" markup). ICE-Sri Lanka was obtained from its compilation team under the supervision of Joybrato Mukherjee at Justus Liebig University Giessen. ICE-Uganda (written part) was obtained from its compilation team under the supervision of Christiane Meierkord at Ruhr-Universität Bochum. The remaining ICE components were employed in the forms as available through their respective download platforms (e.g. the new home of the ICE corpora at the University of Zurich). ICE-Great Britain was not employed in its fully parsed form to be used with the ICECUP software but instead in a plain-text format. While it would have been highly relevant to include ICE-Australia, this corpus remains closed off to a wider scholarship and requests for permission remained unanswered. Overall, twelve fully-released ICE components were used in the current study (CAN, EA, GB, HK, IND, IRL, JA, NIG, NZ, PHI, SIN, SL) as well as three for which only written data was available (GH, UG, USA).

East Africa includes corpus text within markup while other components insert corrected forms into the corpus surrounded by <+>…</+> markup. Similarly, extra-corpus material (i.e. extra text for context in excess of the intended word count) needs to be removed in conjunction with its markup, as well as several other types of annotation and their respective corpus text. In addition to these basic steps, several further and more complex cases need to be handled appropriately and are discussed below. Indeed, it can be said that all efforts to homogenize the data within the current study are in excess of what is usually attempted, with most ICE-based analyses not acknowledging the vast degree of difference between the components.

The main issues addressed in turn below concern the following fields (readers are encouraged to consult the relevant script file included in the appendix):

1. Removal of markup spanning over corpus text including marked-up corpus text, e.g. extra-corpus text (<X>…</X>).
2. Handling of component-specific inconsistencies and homogenization not based on markup, e.g. single-word normative deletions in ICE-EA (<-/…> and <-_…>) or whitespaces surrounding line breaks (<l>) in ICE-CAN and ICE-JA.
3. Harmonization of diverse and partially undocumented special-character encodings, with particular regard for the requirements of POS annotation, e.g. non-standard &ampersand;.
4. Homogenization of diverse text segmentation indicators to facilitate sentence-internal extraction of bigrams, e.g. <Transcription *x*> or <#/>.

The first area of cleanup is concerned with the removal of markup spanning over corpus text which should not be included in the analysis of bigrams. The major concerns for this step lie in the removal of extra-corpus material (including editorial comments and untranscribed text) and putative corrections included by the corpus compilers, but a few finer aspects were also addressed, such as superscript and footnote references which would interfere with bigram detection or material deleted by the original author of the text. All markup addressed in this regard is listed in Table 4.2, which provides the markup tag, its function and notes on the reasons for its removal. Detailed notes on the categories of ICE markup and its usage are available within the ICE manuals (Nelson 2002a, Nelson 2002b). All replacements were performed case-

insensitively, which is why lowercase forms also represents uppercase markup in the table below.

Table 4.2: Markup removed together with corpus text in all corpus components. Note that all markup types also have a corresponding closing tag, which is omitted to improve readability.

Markup[54,55]	Markup Function	Notes
<x> <x_> (EA) <unannotated> (NIG)	Extra-corpus text	Additional material in excess of regular file size (most components follow the 2,000 word limit). May also mark up quotations.[56]
<&> <&_> (EA) <ed> (GH) <editor-comment> (NIG)	Editorial comments	Hints included by the editors to facilitate understanding of certain words and structures (e.g. abbreviations: "vid.<&>vid=video</&>").
<o>	Untranscribed text	Marks up graphics and captions which cannot be transcribed or which would introduce single words such as 'diagram' or verbal descriptions of Figures etc. into the corpus text.
<+> <+_> (EA) <plus> (GH) <error …> (NIG)[57]	Normative insertion	Insertions were handled vastly differently across components and could not be relied upon to homogenize the data.
<sp> <superscript> (NIG)	Superscript	Would obstruct word/bigram recognition if placed within a sentence.
<fnr>	Reference to footnote	ditto
<del> <deleted> (NIG)	Deleted text (self-deletions by the author)	Since some components do not contain this markup, heterogeneity would be introduced if kept in text.

In addition to the deletions presented in Table 4.2, some further markup and such more specific to single or a few components had to be addressed (Step 2 in the list above). Please note that there is a gradient between the two in that, e.g. <plus> in ICE-Ghana is a component-specific variant of the general approach to normalization and thus included in Table 4.2. Markup that necessitates component-specific changes is listed in Table 4.3. These include the extraction of original but normalized forms in ICE-East Africa, the separate treatment of line break markup including or surrounded

[54] Corpus components named by country codes within brackets (abbreviations as used by the corpus teams) denote expressions specifically designed for certain ICE components.

[55] Since several errors in markup application were discovered during initial testing, most of the deletions below only accept a maximum of 50 intervening characters between opening and closing tags (to prevent deleting massive amounts of text in case of missing closing tags). The exceptions are the <o>, <x> and <&> tags (and their variants), since these can span substantial amounts of corpus text (for instance, the first fifth of NZ-S1A is annotated as extra-corpus material).

[56] The may in turn cause quoted parts to separate corpus text.

[57] This latter type operates somewhat differently from the others but is included in the table due to its similar use. ICE-Nigeria includes corrections inside the tag, not in-between tags. Thus, the corrected form inside the tag is removed but not the text between opening and closing tags as in case of all other operations in this table.

by whitespace (cf. above) and the removal of XML metadata. It also, lastly, concerns ways of homogenizing the corpus text without material being necessarily marked up (protected whitespaces and a homogenization of abbreviated forms of address as well as contractions).

Table 4.3: Markup and encoding changes specific to some corpus components

Markup / Encoded text	Markup Function	Notes
<-/…> (EA) <-_...> (EA)	Normative deletion	Extracts original, uncorrected forms of single words or phrases from within their tags in ICE-EA.
<l> (CAN, JA) <l/> (GH) <l /> (GH)	Line breaks	While line break markup could be removed safely, ICE-CAN and ICE-JA add whitespace characters before or after, introducing superfluous spaces into coherent words.
<?xml…> <meta> <annotated…>	Various XML meta data (GH, NIG)	Specification of XML version and metadata which needed to be removed, in case of the <meta> tag including the spanned text.
<x.anonym-x …> (NIG) <object …> (NIG) <coll for …> (NIG) <W1A-001 …> (IRL)	Misc.	Various component specific types of markup which include whitespace characters inside the tag (and thus have to be handled separatedly).
<space> (GH)	Protected whitespace	Replaced by regular whitespace.
<punctuation> (NIG)	Abbreviated terms	Punctuation serving as a proper name abbreviation (cf. also below)
Mrs. / Mrs / Dr. / Prof (etc.)	Abbreviations (no markup)	Abbreviations were homogenized to not contain a period (which facilitates sentence recognition).
they're / they 're	Contractions (no markup)	Contractions are preceded by whitespaces in some components. This was homogenized to have all contracted forms immediately follow the preceding word.

After the changes presented in Tables 4.2 and 4.3, a final major inconsistency remained in the form of special character encoding (Step 3 above). While the ICE manuals specify encodings for non-ASCII characters, many components add additional encodings for characters not specified in the ICE manuals (e.g. for the ampersand character). These may, however, differ in their realizations (with the ampersand character variously encoded as "`&`" and "`&ampersand;`" and sometimes left in-text as is, particularly in the Unicode-encoded XML components). Mostly, these decisions remain undocumented. Since POS-annotation with CLAWS cannot deal with non-ASCII characters (particularly non-Latin characters such as <ɔ> or <ɛ> in ICE-GH) instead of trying to replace all possible (non-documented!) character encodings with their orthographic counterparts, the opposing route was chosen and all non-ASCII characters were replaced by a universal "`&char;`" encoding. It should be noted that this

introduces encodings into words not previously including any markup (particularly the XML-based components), but the goal of treating all data in as similar a way as possible was given precedence over this drawback for the few components concerned.

The final step of the procedure centers around the segmentation of corpus files into units sensible for the later analysis. This mostly concerns clauses or speech units, which are clearly indicated by text unit markers in the majority of components (e.g. <ICE-GB:S1A-001#12:3:A>). They can, however, also diverge across components: NIG, for instance, does not employ them and instead marks up sentences by the <p> tag; similarly, not all headings (<h>) are followed by the markup of a new text unit. Furthermore, the CLAWS system used for POS-annotation draws on uppercase characters and punctuation marks to infer sentence boundaries, which will be absent in the spoken data. Lastly, unclear words are a prominent feature of the spoken data. If these cases were merely removed from the data, non-consecutive words would be recognized as a bigram (e.g. "what I <unclear> to say"). Thus, an effort was made to segment the data into respective chunks delimited by punctuation characters. This was applied both to written and spoken data, adding periods after e.g. headings in the case of writing and similarly after every speech unit as specified by the component editors for speech (unless impossible; cf. ICE-East Africa spoken data's lack of text unit segmentation shown in example (1) above). Table 4.4 lists the types of markup which resulted in the corpus data being segmented. Note that brackets and other symbols not allowed as word characters during the later stage of bigram recognition (cf. Section 4.3) would also result in unit segmentation: The string "240 kilometers (150 mi)" and is regarded as two non-consecutive bigrams, while "150 (nautical) miles" or "Firth (1957)" retrieve no bigram.

After this segmentation of the data and all the preceding cleanup, superfluous whitespaces (double whitespaces or ones at line beginnings or ends, both usually the result of markup deletion) were removed as well as all remaining markup (only ones without internal whitespaces remaining after step 2) and, lastly 'greater' and 'lesser' signs (for which CLAWS otherwise produces errors or breaks annotation). Also, single quotation marks were removed as far as possible since text contained in these usually shows paraphrase forming part of the original utterance. It should be noted that single-quoted words ending in an <s> characters are indistinguishable from plural genitives

for a machine. Since the latter should be retained, these few cases could not be accounted for and remain in the data. Table 4.5 shows two results of the cleanup and homogenization process, with the original file contents vis-à-vis the homogenized and automatically edited text. The edited forms are then either used directly for the lexical *n*-gram analyses or fed into the CLAWS system for annotation (cf. below) and only then input into the analysis of POS-based *n*-grams.

Table 4.4: Markup (and annotated text) used for the segmentation of the corpus text

Markup / Encoded text	Markup Function	Notes
<ice…> <#> (IRL) <#/> (EA)	Text unit	Text unit markup indicates a new utterance/sentence. Appears in reduced form in EA and IRL and may contain whitespaces.
<$...> (EA) transcription … (NIG)	Speaker turn	A speaker turn can contain several text units.
</p> </h> </heading> (NIG)	End of paragraph or heading	Paragraphs are usually followed by markup for new text units, but headings do not necessarily have these.
<unclear> <unclear…> (NIG) <o_>…</o> (EA) <o/>… (EA)	Unclear word(s)	Mere deletion of unclear word markup would lead to two non-consecutive words forming a bigram.[58]

4.2.2 POS-annotating the Data for Grammatical *N*-grams

With the base lexical data prepared as explained above, the last preparatory step before the extraction and analysis of *n*-grams lies in the area of POS annotation. Annotation was carried out using the CLAWS system and C7 tagset, with some of the above changes to the data made with an eye to its applicability to this system. This particularly relates to the limited ability of the CLAWS system to deal with non-ASCII characters, and the much-increased likelihood of erroneous classification in case sentences boundaries cannot be identified, which is why full stops were inserted after spoken utterances or headlines in writing.[59] The product of this process is the same data as previously discussed but with the addition of part-of-speech annotation.

[58] Since some of this type of markup can span across corpus text, these were again restricted to a maximum of 50 characters or to a single word (EA "<o/>" is followed by a single unclear word).

[59] Similarly, while all data is handled case-insensitively by the script files, the data itself is not modified into case-insensitive format. This is yet another contribution to POS-annotation accuracy with the CLAWS system, which relies on capitalization.

Table 4.5: Original corpus files and results of the cleanup and homogenization procedure in two corpus samples (line breaks indicated by "¬").

	Original Corpus Text	Cleaned and Homogenized Text
ICE-CAN W1A-001	<I>¬ <ICE-CAN:W1A-001#1:1> <h> The Critique of Fiction and Philosophy in <}> <-> Votaire's </->¬ ¬ <+> Voltaire's </+> </}> Candide </h>¬ ¬ <p> <ICE-CAN:W1A-001#2:1> The strong personal views of Voltaire can easily be found through¬ ¬ his caricature of characters in his black comedy Candide . ¬	<I>¬ <ICE-CAN:W1A-001#1:1> <h> The Critique of Fiction and Philosophy in <}> <-> Votaire's </->¬ ¬ <+> Voltaire's </+> </}> Candide. </h>¬ ¬ <p> <ICE-CAN:W1A-001#2:1> The strong personal views of Voltaire can easily be found through¬ ¬ his caricature of characters in his black comedy Candide . ¬
ICE-EA S1B-001 K (class-lessonK)	<$A> <#/>So in other words it's more of uh the application part of the language ¬ <$B> <#/>Yeah¬ <$A> <#/>So you differ with uh basically you differ with Chomsky's idea that uh linguistic competence is basically that set of rules that is in the child you know uh let alone applying it <#/>Right <#/>Now I want to look at this <#/>I want to have uh this discussion in advanced levels ¬	<$A> <#/>So in other words it's more of uh the application part of the language. ¬ <$B> <#/>Yeah. ¬ <$A> <#/>So you differ with uh basically you differ with Chomsky's idea that uh linguistic competence is basically that set of rules that is in the child you know uh let alone applying it. ¬ <#/>Right. ¬ <#/>Now I want to look at this. ¬ <#/>I want to have uh this discussion in advanced levels. ¬

After the application of POS markup, lexical items were removed, so that the consecutive analysis of POS patterns would not require a repeated handling of these. While the default was to remove lexical items and only retain POS annotation, in a few cases lexical items were retained in case their tags belonged to closed-class items with very little lexical diversity. Only minimally more variation was introduced into the data through this decision, while at the same time the potential for erroneous classification was reduced (taggers being most reliable in the major word classes, cf. e.g. the erroneous classification of 'intrusive *as*' as a conjunction in Koch et al. 2016). Further benefits include improved readability of the resulting patterns, as well as that the overall design of the retrieved patterns is made to more closely resemble the familiar style of patterns from e.g. Pattern Grammar (Hunston & Francis 2000, Hunston 2014) which also often retains function words in their lexical forms. POS tags for items detected as a conjunction (CC, CCB, CS, CSA, CSN, CST, CSW), preposition (IF, II, IO, IW) or the infinitive marker (TO) were thus removed and the lexical realization retained. Finally, unique collocations created by the CLAWS system had to be addressed: In case CLAWS recognizes a sequence as fulfilling a joint function (e.g. *in terms of*), it applies the appropriate tag (II for preposition) to all words in the sequence and additionally assigns each item a unique numerical identifier ("in_II31 terms_II32 of_II33"). These so-called 'ditto tags' lead to the creation of unique collocations (e.g. the unique item II33), which can in turn be strongly preferred or dispreferred by individual association measures. The numerical identifiers are thus deleted, so as not to (dis-)prefer the sequences contained within the CLAWS dictionary over all others.

After the automatic editing process of the corpus data, all remaining lexical items before a tag (i.e., the word-material before an underscore followed by a combination of capital letters and potentially digits) were deleted from the corpus text, leaving patterns of data as shown in Table 4.6, representing the cleaned and homogenized corpus material from Table 4.5 viz its POS-annotated counterpart (for easier readability, the samples are aligned, even though the final output does not show line segmentation). Moreover, it should be made explicit that tags were retained (and analyzed) in their capitalized forms while lexical items were converted to lowercase in the consecutive analysis so as to avoid any potential for mistaking a POS tag for an actual word.

Table 4.6: Cleaned and homogenized corpus data, C7-annotated version and final edited version for POS-gram analysis (original line breaks indicated by "¬")[60]

	Cleaned and Homogenized Text	CLAWS C7-Annotated	POS-grams version
ICE-CAN W1A-001	The Critique of Fiction and Philosophy in Votaire's Candide. ¬The strong personal views of Voltaire can easily be found through his caricature of characters in his black comedy Candide. ¬ These characters, through their obvious ridiculousness, critique many beliefs and poke fun at many people without openly attacking them in much the same way Dante did in his Inferno. ¬	The_AT Critique_NN1 of_IO Fiction_NN1 and_CC Philosophy_NN1 in_II Votaire_NP1 's_GE Candide_NN1 ._. The_AT strong_JJ personal_JJ views_NN2 of_IO Voltaire_NP1 can_VM easily_RR be_VBI found_VVN through_II his_APPGE caricature_NN1 of_IO characters_NN2 in_II his_APPGE black_JJ comedy_NN1 Candide_NN1 ._. These_DD2 characters_NN2 ,_, through_II their_APPGE obvious_JJ ridiculousness_NN1 ,_, critique_NN1 many_DA2 beliefs_NN2 and_CC poke_VV0 fun_NN1 at_II many_DA2 people_NN without_IW openly_RR attacking_VVG them_PPHO2 in_RP much_RR the_AT same_DA way_NN1 Dante_NP1 did_VDD in_II his_APPGE Inferno_NN1 ._.	AT NN1 of NN1 and NN1 in NP1 GE NN1 . AT JJ JJ NN2 of NP1 VM RR VBI VVN through APPGE NN1 of NN2 in APPGE JJ NN1 NN1 . DD2 NN2 , through APPGE JJ NN1 , NN1 DA2 NN2 and VV0 NN1 at DA2 NN without RR VVG PPHO2 RP RR AT DA NN1 NP1 VDD in APPGE NN1 .
ICE-EA S1B-001 K (class-lessonK)	So in other words it's more of uh the application part of the language. ¬ Yeah. ¬So you differ with uh basically you differ with Chomsky's idea that uh linguistic competence is basically that set of rules that is in the child you know uh let alone applying it. ¬ Right. ¬Now I want to look at this. ¬I want to have uh this discussion in advanced levels. ¬	So_RR in_II other_JJ words_NN2 it_PPH1 's_VBZ more_DAR of_IO uh_UH the_AT application_NN1 part_NN1 of_IO the_AT language_NN1 ._. Yeah_UH ._. So_RR you_PPY differ_VV0 with_IW uh_UH basically_RR you_PPY differ_VV0 with_IW Chomsky_NP1 's_GE idea_NN1 that_CST uh_UH linguistic_JJ competence_NN1 is_VBZ basically_RR that_DD1 set_NN1 of_IO rules_NN2 that_CST is_VBZ in_II the_AT child_NN1 you_PPY know_VV0 uh_UH let_II21 alone_II22 applying_VVG it_PPH1 ._. Right_RR ._. Now_RT I_PPIS1 want_VV0 to_TO look_VVI at_II this_DD1 ._. I_PPIS1 want_VV0 to_TO have_VHI uh_UH this_DD1 discussion_NN1 in_II advanced_JJ levels_NN2 ._.	RR in JJ NN2 PPH1 VBZ DAR of UH AT NN1 NN1 of AT NN1 . UH . RR PPY VV0 with UH RR PPY VV0 with NP1 GE NN1 that UH JJ NN1 VBZ RR DD1 NN1 of NN2 that VBZ in AT NN1 PPY VV0 UH let alone VVG PPH1 . RR . RT PPIS1 VV0 to VVI at DD1 . PPIS1 VV0 to VHI UH DD1 NN1 in JJ NN2 .

[60] The final column disregards empty lines still present in the data at this point. Similarly, the middle column leaves out line breaks and the CLAWS sentence markers <s>…</s>.

4.3 Extracting Collocational Patterns

The cleaned and homogenized data prepared as laid out in Section 4.2 formed the basis for the extraction of bigrams, the calculation of association measures and the combination into longer *n*-grams. The same general approach was employed for the lexical dataset as well as for the POS-annotated version, and both variants of base data were subjected to the same methodology for extracting bigrams, calculating their association scores, and finally combining them to longer sequences.[61] The following sections lay out the process for browsing the datasets for (lexical and POS) bigrams as well as calculating their association scores by measure (Section 4.3.1) before leading on to the generation of longer sequences (*n*-grams) of both dynamic and static lengths *n* (Section 4.3.2).

4.3.1 Bigram Extraction and Statistics

Bigram extraction itself is a relatively straightforward process and hinges mostly on the definition of a word pattern. Characters accepted to represent word elements were the letters a-z, the digits 0-9 as well the apostrophe and hyphen. The latter two characters were, however, only accepted if they occurred after at least one initial letter or digit. Note that apostrophes will not be present in the POS dataset, since CLAWS assigns a separate tag to genitives or contracted verbs. If two consecutive sequences of this design could be found in the data, they were extracted as a bigram. The regular expression to represent this pattern in *R* is shown in Figure 4.1. The expression essentially contains the same word-matching twice, and furthermore expects an intervening whitespace (which leads to the retrieval of bigrams). The only difference lies in the second item being encapsuled within a so-called look-ahead criterion "`?= (…)`", which allows matches onto the second item without removing it from the data (since it could be the first constituent of a consecutive bigram). The procedure applies to all text units delimited within the clean-up and homogenization process laid out in Section 4.2. It thus typically extracts bigrams sentence- or speech-unit-internally, but some

[61] While lexical sequences were handled case-insensitively so as to avoid bigrams differentiated only by capitalization. POS sequences had the few items retained in lexical form converted to lowercase, while POS tags are consistently capitalized.

other features of the text (e.g. punctuation or unclear words, cf. above) may lead to partitions within the corpus data.

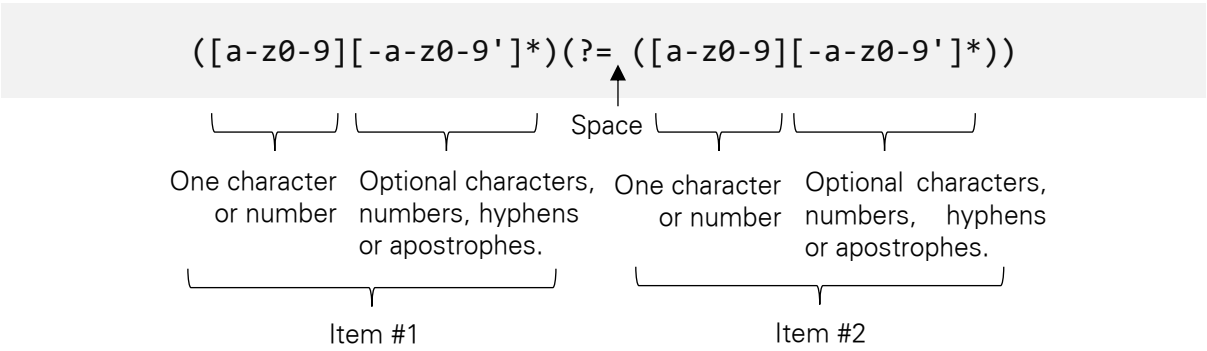


Figure 4.1: Regular expression used to extract bigrams

The procedure defined above applies to both the lexical and grammatical data in the same form. It follows the practice of a strictly orthographic word definition, which extends to the treatment of POS tags as words (e.g. NN1) in the grammatical data. However, please note that the POS annotation introduces some changes into the data which are not present in the lexical version. In particular, the CLAWS system separately tags contracted forms and genitive markers, which only represent single tokens in the lexical data. While contracted forms can be argued to contain two lexical items, thus warranting a separation of the contraction, this was not attempted in the lexical data for several reasons: Generally, an orthographic retrieval can be regarded as the default operating mode of corpus linguistics, which usually interferes as little as possible with the data at hand. While some concordancers (e.g. *AntConc*) allow to separate contractions, they perform this in a very crude way which merely splits the contraction at the apostrophe. This procedure has roughly equal chances of retrieving at least one actual word (*she'll* > *she 'll*) or to produce non-words (*isn't* > *isn 't*). During manual data inspection, it might be possible to identify such latter cases, but the unique collocations which a simple split of contractions produces clearly need to be avoided in a methodology revolving around fully automatic evaluation of the data. At first glance, it would seem that these instances can be resolved through a list-based approach which handles some negations separately (*can't* > *can 't*, but *isn't* > *is n't*), but even this fails at frequent contractions such as *won't* or also *shan't*, which result in non-words (*wo*, *sha*, *shan*) or ambiguous forms (*won*). Furthermore, a list-based approach would be blind to aspects of variation or errors during speech production which do not result in the pre-defined contractions. Even if all contractions could be

sensibly separated, ambiguity would remain, such as in case of contracted verbs after proper nouns (e.g. *Barbara's)* remaining indistinguishable from genitives. It appears that consistently dealing with contracted forms requires the adoption of a semantic word definition, which also leads to a different handling of lexicalized forms (*cannot*, *gonna* or *wanna*) or non-hyphenated compounds (which are two words under the present, orthographic framework). It is unfortunate that an extraction of 'better' orthographic words fails on so many levels and that some amount of difference between the lexical and grammatical data needs to be accepted, but any such interference with the text would introduce more problems than it helps to resolve. In an automated approach, the dangers of creating unique collocations through the separation of contractions outweigh the benefits of retrieving semantic units in those cases where it is even possible. If the actual collocations were a (qualitative) concern of the present study, the situation would certainly be different, but within the present methodology, it is ultimately irrelevant whether a semantic five-word unit (*I'll drop the mic*) is only recognized as a four-word orthographic sequence.

The procedure laid out above yielded an average count of 569,767 spoken and 360,880 written bigram tokens in the lexical data as well as 574,712 spoken and 363,136 written tokens in the POS data (cf. Chapter 5). Numbers thus diverge slightly between data types, as discussed above, since POS tags for contracted forms and genitives will themselves form bigrams in the POS data. The fact that bigram token frequencies are below individual word frequencies (600,000 and 400,000) results from the fact that some words are used in isolation (e.g. single-word sentences or utterances or single words within brackets or quotations), thus not forming bigrams. This appears to affect the written mode proportionally more strongly, i.e. relatively prevalent single-word utterances (e.g. discourse markers) seem to affect speech less than writing (e.g. bracketed or quoted text passages leading to single words or disrupting other text units).

With the extraction of bigrams complete, association values can be derived for each item. For each word in the 28 lists of bigram tokens (one for each corpus part), observed frequencies are determined, as well as type frequencies before and after each item (for the lexical gravity measure), which are then used to derive expected

frequencies as per Table 4.7 (which repeats Table 3.2).[62] These, in turn, form the basis for the calculation of the five association measures (cf. Chapter 3).

Table 4.7. Expected item frequencies and their calculation from observed frequencies.

	word$_2$	**¬ word$_2$**
word$_1$	$a_{exp} = {R_1 * C_1}/{N}$	$b_{exp} = {R_1 * C_2}/{N}$
¬ word$_1$	$c_{exp} = {R_2 * C_1}/{N}$	$d_{exp} = {R_2 * C_2}/{N}$

Table 4.8 presents an excerpt from the base statistics (observed and expected token frequencies and type frequencies) on the example of the first 26 bigrams in ICE-Hong Kong W1A-001.[63] Table 4.9, then, shows the resulting scores for the five association measures, with mean values across the corpus part in the final row indicating each measure's central tendency.[64] The list of association scores formed the basis for the consecutive *n*-gram generation and their analysis (cf. next section).

4.3.2 From Bigrams to *N*-grams

The bigrams generated as laid out above will, in their own right, become an aspect of the analysis in Chapter 5, but they furthermore serve as the basis for the generation of consecutively longer sequences of words. Section 3.3.1 already laid out theoretical and practical issues concerning association score calculation for sequences longer than two items, arguing for an iterative merging process of consecutive bigrams (and their association statistics) to avoid these. Within the scope of the current study, this will be implemented as follows:

[62] As a concrete example, consider *the greatness*: This co-occurs twice (a), but *the* by itself has a frequency of 29,842 (b) whereas *greatness* only occurs once otherwise (c), with 382,698 other tokens in the corpus part (d). Observed row (R_1, R_2) and column (C_1, C_2) totals are merely sums of either two of these values. Expected frequencies are derived by multiplication of the respective row and column totals and their joint division by the number of overall tokens (N), which itself is the summation of either row or column totals. Thus, e.g. a_{exp} = 29,844*3/(29,844+382,699) = 89,532/412,543 = 0.217.

[63] Note the sentence break and unit boundary (comma) expressed by non-overlapping bigrams in the fourth- and third-to-last lines of the table.

[64] Rounding to two fractional digits for all measures except *ΔP*, which requires further digits to be distinguishable due to it being scaled only between -1 and +1.

Table 4.8: Basic observed and expected co-occurrence values for the first 26 bigrams in ICE-HK W1A-001

Bigram	a	b	c	d	R_1	R_2	C_1	C_2	a_{exp}	b_{exp}	c_{exp}	d_{exp}	Types post 2nd	Types pre 1st
the greatness	2	29,842	1	382,698	29,844	382,699	3	412,540	0.22	29,843.78	2.78	382,696	6,999	2
greatness in	2	1	9,636	402,904	3	412,540	9,638	402,905	0.07	2.93	9,637.93	402,902	1	3,384
in the	2,439	7,199	27,405	375,500	9,638	402,905	29,844	382,699	697.23	8,940.77	29,146.77	373,758	1,940	2,293
the number	88	29,756	172	382,527	29,844	382,699	260	412,283	18.81	29,825.19	241.19	382,457	6,999	66
number of	204	56	14,160	398,123	260	412,283	14,364	398,179	9.05	250.95	14,354.95	397,928	28	3,358
of vocabulary	4	14,360	26	398,153	14,364	398,179	30	412,513	1.04	14,362.96	28.96	398,150	3,764	8
vocabulary in	3	27	9,635	402,878	30	412,513	9,638	402,905	0.70	29.30	9,637.30	402,875	12	3,384
in the	2,439	7,199	27,405	375,500	9,638	402,905	29,844	382,699	697.23	8,940.77	29,146.77	373,758	1,940	2,293
the english	51	29,793	223	382,476	29,844	382,699	274	412,269	19.82	29,824.18	254.18	382,444	6,999	95
english language	30	244	95	412,174	274	412,269	125	412,418	0.08	273.92	124.92	412,144	87	46
language is	5	120	6,526	405,892	125	412,418	6,531	406,012	1.98	123.02	6,529.02	405,888	60	1,942
is amazing	2	6,529	7	406,005	6,531	406,012	9	412,534	0.14	6,530.86	8.86	406,003	1,448	7
amazing to	1	8	13,467	399,067	9	412,534	13,468	399,075	0.29	8.71	13,467.71	399,066	7	3,195
to most	3	13,465	545	398,530	13,468	399,075	548	411,995	17.89	13,450.11	530.11	398,544	2,602	111
most learners	1	547	19	411,976	548	411,995	20	412,523	0.03	547.97	19.97	411,975	263	9
learners which	1	19	1,263	411,260	20	412,523	1,264	411,279	0.06	19.94	1,263.94	411,259	10	453
which are	79	1,185	2,918	408,361	1,264	411,279	2,997	409,546	9.18	1,254.82	2,987.82	408,291	470	999
are foreign	1	2,996	74	409,472	2,997	409,546	75	412,468	0.54	2,996.46	74.46	409,471	1,107	35
foreign to	1	74	13,467	399,001	75	412,468	13,468	399,075	2.45	72.55	13,465.55	399,002	52	3,195
to the	1,416	12,052	28,428	370,647	13,468	399,075	29,844	382,699	974.30	12,493.70	28,869.70	370,205	2,602	2,293
the language	24	29,820	101	382,598	29,844	382,699	125	412,418	9.04	29,834.96	115.96	382,583	6,999	46
language itself	1	124	65	412,353	125	412,418	66	412,477	0.02	124.98	65.98	412,352	60	54
in fact	85	9,553	72	402,833	9,638	402,905	157	412,386	3.67	9,634.33	153.33	402,751	1,940	9
the greatness	2	29,842	1	382,698	29,844	382,699	3	412,540	0.22	29,843.78	2.78	382,696	6,999	2
greatness in	2	1	9,636	402,904	3	412,540	9,638	402,905	0.07	2.93	9,637.93	402,902	1	3,384
in the	2,439	7,199	27,405	375,500	9,638	402,905	29,844	382,699	697.23	8,940.77	29,146.77	373,758	1,940	2,293

An initial bigram or already fabricated combination of bigrams (i.e. a 3-gram or longer) is merged with a subsequent bigram, while simultaneously association scores are continually averaged over all bigrams involved, until either a cut-off value for *n* is reached or average association strength for a longer sequence falls below that of an already established one (i.e. the longer sequence not constituting the more strongly collocated *n*-gram). This results in two very different types of sequences, the first one reflecting more traditional *n*-grams of a priori fixed lengths, while the latter has an appropriate value for *n* emerge from the data itself, thus forming 'true' *n*-grams (in the sense of a true variably *n*) best adapted to a particular corpus in an entirely data-driven manner.

Table 4.9: Association scores calculated for the first 26 bigrams in ICE-HK W1A-001 (plus overall token means of written ICE-HK). Bigrams surpassing the threshold values for a measure are highlighted in gray, and boxes indicate bigrams merged within the dynamic-length approch.

bigrams	*MI*	*t*	G^2	*g*	$\Delta P_{2|1}$
the greatness	3.20	1.26	9.86	-0.68	0.0001
greatness in	4.83	1.36	16.24	-1.09	0.6433
in the	1.81	35.27	4,476.62	16.49	0.1850
the number	2.23	7.38	224.25	8.85	0.0025
number of	4.49	13.65	1,595.20	10.03	0.7503
of vocabulary	1.94	1.48	7.42	0.16	0.0002
vocabulary in	2.10	1.33	6.22	0.34	0.0766
in the	1.81	35.27	4,476.62	16.49	0.1850
the english	1.36	4.37	54.92	7.72	0.0011
english language	8.50	5.46	439.69	6.72	0.1093
language is	1.34	1.35	4.77	1.84	0.0242
is amazing	3.81	1.31	10.49	-0.54	0.0003
amazing to	1.77	0.71	1.58	-2.44	0.0785
to most	-2.58	-8.60	-28.13	-1.51	-0.0011
most learners	5.23	0.97	7.73	-2.21	0.0018
learners which	4.03	0.94	5.41	-2.48	0.0469
which are	3.10	7.86	297.22	9.60	0.0554
are foreign	0.88	0.46	0.44	-2.54	0.0002
foreign to	-1.29	-1.45	-1.64	-2.60	-0.0193
to the	0.54	11.74	286.42	14.86	0.0339
the language	1.41	3.05	27.36	5.64	0.0005
language itself	5.64	0.98	8.49	-1.35	0.0078
in fact	4.53	8.82	614.83	6.38	0.0086
the greatness	3.20	1.26	9.86	-0.68	0.0001
greatness in	4.83	1.36	16.24	-1.09	0.6433
in the	1.81	35.27	4476.62	0.19	0.1850
mean $_{\text{HK wrt}}$	4.37	3.35	361.77	2.10	0.1076

The present analysis thus approaches *n*-grams from two very different perspectives but offers a methodologically consistent approach in which both types of sequences are the result of consecutive mergers of successive bigrams, while being different in the ways in which cut-offs are defined. Additionally, however, it makes sense to provide a different selection of basic bigrams for both approaches: Static-length *n*-grams

can be calculated across the entirety of the bigram data, since every bigram will be equally reflected in one (2-grams), two (3-grams) or more (4-grams, etc.) longer sequences. Even if one variety disprefers a bigram, resulting in negative association values, this bigram need not be discarded since its dispreference might contrast with preference in another variety, and furthermore several varieties display gradual differences in the levels of (dis)preference. Dynamic-length *n*-grams, on the other hand, define their length on the basis of rising or falling average association scores, continuing the merging procedure only if average association scores increase. Thus, the inclusion of negative association scores in the bigram data would almost universally lead to these sequences to begin in a less collocated bigram, up to and including mutually repellent items.[65] This is why, for dynamic-length sequences, bigram lists are first trimmed to contain only those surpassing a threshold value of collocability (which is taken to distinguish 'strong' collocations from 'weak' ones or even mutually repellent component items). Based on the discussion of association scores in Section 3.3.2, threshold values for each association measure were set as shown in Table 4.10 (with *MI* additionally restricted by a bigram frequency threshold of $O \geq 5$).

Table 4.10: Threshold values for the selection of bigrams for dynamic-length *n*-grams (association scores need to be greater than these values).

MI	***t***	***G²***	***g***	***ΔP***
3 ($O \geq 5$)	2.576	6.64	5.5	0

Bigrams surpassing these threshold values are highlighted by shaded cells in Table 4.9, and bigrams merged into longer sequences within the dynamic-length approach are furthermore indicated by boxes. Figure 4.2 provides a visual summary of the similarities and differences between the two approaches.

[65] Consider that a program always requires the selection of an initial bigram. If mutually repellent items were included, a previous sequence would terminate the merging procedure due to dropping association levels in case of the addition of the lowly associated bigram. In turn, merger of this bigram with a successive one would be highly likely to produce rising average association. In that process, the overall dataset would be strongly skewed towards repellent bigrams to introduce *n*-grams, with the exception of utterance-initial bigrams. Certainly, such *n*-grams do not constitute sensible choices for describing the linguistic patterning preferred by the speakers of a particular variety. That said, it should be noted that this decision is not necessarily taken by all authors following an iterative approach (cf. e.g. Gries & Mukherjee 2010).

With all sets of co-occurrence data thus generated, it is possible to apply these to the statistical analysis of patterns of (dis-)similarity in terms of association strengths across varieties. This will be the focus of the following section.

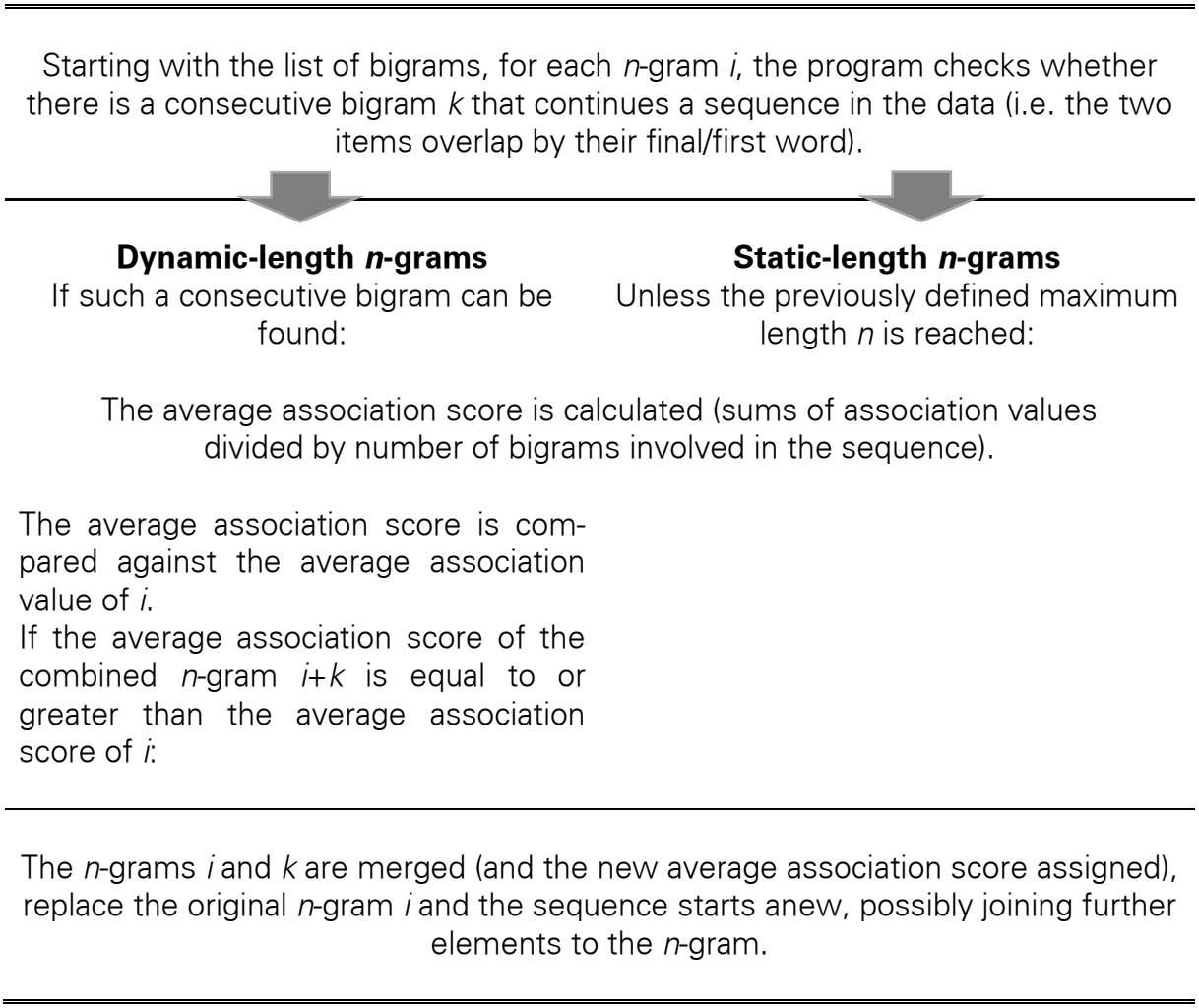


Figure 4.2: Methodological steps in the extraction of static- and dynamic-length *n*-grams

4.4 Evaluating the Data

With the *n*-gram data prepared, it is possible to move on to the statistical evaluation of the various lists of sequences and their association scores in order to discover whether different degrees of attraction (or also repulsion, in case of static-length *n*-grams) can be mapped onto language-external characteristics of the different national varieties (cf. Chapter 3). Section 4.4.1 presents cluster analysis as the family of methods chosen for the analysis of the current dataset and ways of interpreting and

evaluating cluster structures, while Section 4.4.2 lays out the research questions and hypotheses under which the data will be analyzed in Chapter 5.

4.4.1 Interpreting Co-occurrence Data with Clustering Methods

The methodological steps laid out in the previous sections consecutively led to the creation of a vast amount of co-occurrence data for each of the varieties under scrutiny: Even after reducing the dataset to exhibit only *n*-gram types (vs. tokens) and their collocational features, on average 193,310 lexical and 11,329 POS spoken as well as 170,367 lexical and 10,220 POS written bigram types for each 28 corpus parts remained to be handled by the analysis (in contrast to 569,767 lexical and 574,712 POS spoken tokens as well as 360,880 lexical and 363,136 POS tokens in writing, as discussed above). As laid out before, these bigrams only constituted one part of the analysis and were more importantly used to generate longer *n*-grams of either static or dynamic lengths. Actual discussion of the (numbers of) items thus generated will be reserved for the actual analysis in Chapter 5, but suffice it to say that token frequencies for static-length *n*-grams (cf. Sections 5.3 and 5.4) were found to be consecutively lower for the longer sequences than for bigrams (given the fact that successively more sequences of words fail to form a 3-gram, 4-gram, etc. within an utterance boundary), but types were found to diversify rapidly (up to 438,229 and 263,042 lexical well as 219,383 and 116,532 POS types in speech and writing, respectively), most drastically with either the switch from 2-grams to 3-grams (lexical sequences) or 3-grams to 4-grams (POS sequences).[66] While static-length sequences are identical across all measures (with only association values diverging), dynamic-length sequence frequencies instead depend on the impact of the threshold values particular to each measure as well as on the number of (consecutive) bigrams available for merging after this selection process. Thus, the discussion of these strongly varying frequencies is reserved until the analysis in Sections 5.1 and 5.2.

In terms of analytical methods, the present study's supposition that gradual differences of *some* extent between the varieties will provide a systematic way of

[66] Dynamic-length sequences are used to gauge the best lengths for static-length *n*-grams, but without anticipating the precise analysis in Chapter 5, it can already be said that only sequences up to four items were deemed sensible within the present dataset, with 5-grams consituting too steep of a drop in available tokens, in addition to miniscule types shared between varieties.

distinguishing these (without a particular interest in the association scores of an individual item) precludes a hypothesis-testing approach following a strict 'the more X the more/less Y' formulation of expected outcomes (cf. also Gries 2008b). Instead of trying to predict one likely outcome from a set of possible candidates (e.g. phase within Schneider's 2007 dynamic model) on the basis of a (small) number of dependent variables, each *n*-gram in fact presents its own variable. The analysis thus concerns the identification of maximally similar groups within the data instead of a clear correlation between two or more variables, and the overarching question becomes whether, overall, the degrees of similarity and difference between all *n*-grams display sufficiently strong patterns that groups of varieties can be identified which, in a second, interpretative step, may match the groups that emerge from the delineation of language-external properties of the individual linguistic settings.

For tasks such as the one required by the current data, cluster analysis presents a family of methods which aim at finding groups of similar objects within less clearly structured data (cf. e.g. Moisl 2015: 10). The unifying concern behind the diverse spectrum of clustering techniques lies in the task to be fulfilled rather than in the approach itself, which can be highly diverse. The objective of clustering approaches consists of grouping a number of objects (in this case corpus parts) in such a way as to form groups (or clusters) of objects which are more similar to one another than to those objects in other clusters. It thus helps in dealing with very large and complex datasets in an objective and replicable way (Moisl 2015: 301–302) – or as Gries & Hilpert (2008: 62) summarize, "cluster analyses allow us to perceive patterns at levels of granularity that human observers are incapable of noticing." Gries (2008b: 337) presents an sample application of the method of Hierarchical Cluster Analysis (HCA) particularly intuitive for linguists on (fictitious) data on some English consonants (Figure 4.3). Since cluster analysis can accommodate even variables of different types, the resultant 'dendrogram' visually represents (dis-)similarities in data which potentially composed of a diverse range articulatory characteristics or even accoustic measurements.

On the most general level, two groups can be distinguished, separated by comparatively long branches, separating the data into a group of only plosives and a second group of nasals, liquids and fricatives. (Note that actual 'height' values depend most on the choice of (dis-)similarity score (cf. below) and less on the data itself, which

they only reflect in a mediated fashion.) Beyond this general constellation, the first group also shows two subclusters, one containing all bilabial plosives, the other the alveolar plosives and a velar plosive (shown at relatively high distance to the alveolar sounds). In the second major cluster, a similar stepwise pattern is contained twice, with the set being distinguished broadly into nasals and liquids in one subcluster as well as into fricatives in the second. Within the first subcluster, two sounds with alveolar manner of articulation are represented as more similar to each other than each to the /m/ sound, while within the second subcluster labiodental fricatives are found most similar first, but without much internal difference within both subclusters.

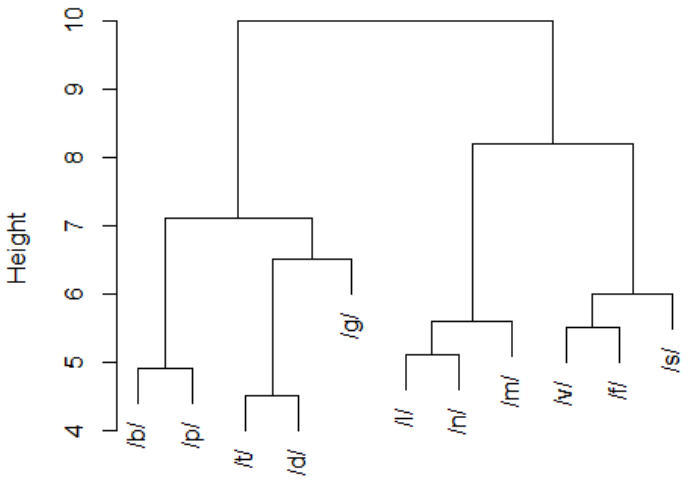


Figure 4.3: Dendrogram of a fictitious cluster analysis of English consonants (Gries 2008b: 337)

The situation with the *n*-gram data of the current study is both similar to and different from this introductory example: It is similar in that any single *n*-gram is not itself relevant for assigning a variety to a cluster (just as much as an unknown host of variables may have led to the above dendrogram), but overall similarity of many variables will lead to the formation of a group more similar internally than to other external clusters. It is different in the way that for the consonant data, a hypothesis-testing approach might follow, delving into individual contributing variables and specifically testing their impact, while with the *n*-gram data the intention cannot be to single out a few *n*-grams as independent/predictor variables for consecutive testing (contrast e.g. Biber 1993 or Biber 2004a for cluster analyses with more discrete dimensions). Instead of generating hypotheses from the cluster results (which is why cluster analysis is sometimes called hypothesis-generating instead of hypothesis-testing, cf. Gries 2008b: 337), the cluster structure itself becomes the object of testing for significant groups and the most sensible ways of segmenting the overall dendrogram, and evaluating

whether the clusters retrieved can be systematically brought in line with descriptive parameters (e.g. phases of the Dynamic Model). Issues in sensibly partitioning the data even arise in the mostly clear-cut example in Figure 4.3, since it is up to the analyst to realize that manner of articulation sensibly divides the data on the largest level, while on a smaller scale place emerges as the descriptive category over, e.g. voicing, or that even at relative distance, /g/ still belongs to the 'plosive' part of the structure. Relatedly, are there two, three, four, etc. cluster, i.e. at which height should the dendrogram be 'cut'? In everyday scenarios, the data moreover does not behave as nicely as the one above, so consider if only a single object were clustered differently or if the number of objects increased: Quickly, partitioning the data becomes a major challenge, which is why data-driven evaluation techniques are an aid over visual inspection.

Cluster validation is also a necessity due to the fact that clustering techniques will always detect some form of pattern even if these are not strictly warranted by the data: For most clustering methods, clusters may even emerge from entirely random data, since "all are based on the notion of cluster centres, and all are consequently predisposed to find convex linearly-separable clusters" (Moisl 2015: 226): "Clustering algorithms have the annoying habit of finding clusters even when the data are generated randomly" (Smith & Dubes 1980: 177; for an example cf. also Greenacre 2011: 5). In real-world scenarios, however, the "validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage." (Jain & Dubes 1988) A competing approach is to apply different clustering algorithms in turn and triangulate their findings in order to arrive at those groups of objects most frequently supported by the data at hand. The present analysis does both but focuses on triangulation of methods in order to facilitate a uniform approach to the 40 different datasets (two types of base data, five association measures and sequences of dynamic lengths as well as static-length 2-, 3- and 4-grams). Clustering and evaluation methods employed by the current study will be discussed in turn below.

Hierarchical Agglomerative Clustering

Hierarchical agglomerative clustering (here carried out with *R*'s `hclust()` function) is a bottom-up algorithm which initially treats every data point (variety) as a cluster of its own and then merges (agglomerates) maximally similar pairs of clusters successively until, at the final step, the two largest groups of clusters merge into the overall cluster containing all data points (Sarstedt & Mooi 2014: 281, Moisl 2015: 10–11; cf. also the `hclust` documentation, R Core Team 2019). Through the consecutive merging of pairs of clusters (either already merged pairs or initial unary data points), it produces a hierarchical structure that is also called a tree (with the root being the cluster that contains all data points, leaves being the individual data points, and branches in-between). The nature of this clustering approach makes it particularly worthwhile when a hierarchical system of some sort is suspected to operate on the data – whether these be clearly delineated groups or a stepwise pattern of (dis-)similarity (Seif 2018).

Performing a hierarchical cluster analysis requires two further methodological considerations: For one, data (i.e. *n*-gram association strengths for each corpus part) cannot immediately be entered into the analysis but needs to be transformed into a distance matrix through the choice of one among several different metrics. Additionally, the way that amalgamation is performed also depends on the selection of a specific procedure in the form of an amalgamation rule. Considering the first choice, i.e. that of a distance metric, it can fortunately be said that hierarchical approaches are not "sensitive to the choice of distance metric" (Seif 2018). Instead of trying to find the single 'perfect' metric, the goal should thus rather be not to find an entirely unfitting one. Apart from having to choose a metric that fits the type of data (ratio-scaled in the present case), the central question is whether the distance matrix is supposed to reflect magnitude (in the present case overall strength of association) or whether the general shape of the distribution (similar profile of association strengths, but not necessarily similarly strong on average) is given center stage.[67] Gries (2008b: 345; but cf. also Sarstedt & Mooi 2014: 281 for further illustration) provides yet another illustrative example (reproduced in Figure 4.4).

[67] Confer also this excellent discussion on stats.stackexchange.com: https://stats.stackexchange.com/questions/80377/which-distance-to-use-e-g-manhattan-euclidean-bray-curtis-etc

Depending on the choice of measure, either y2 or y3 might be returned as more similar to y1: The first two (y1,y2) are at great mutual distances in terms of their data values but display identical curvatures, i.e. their values vary by the same pattern. The latter pair (y1,y3), while very different in shape, are far more similar in their average values and mutually closest along the entire spectrum of data points.

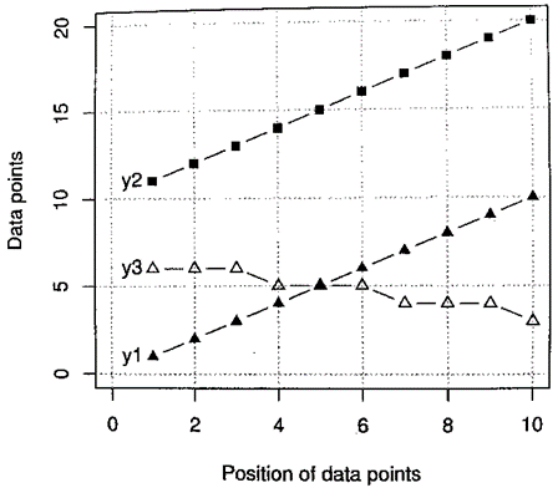


Figure 4.4: Three fictitious datasets and their (dis-)similarities

Thus, selection of a similarity measure entails a choice in terms of the 'kind' of similarity. For the present study, a curvature-based approach appears more reasonable, since similar relative preferences for the sequences under scrutiny should be emphasized over the (average) distance across all bigrams. Moreover, the contingency tables underlying the bigram association scores require that for any strongly attracting bigram another will be assigned a low score. As such, association values will be relatively similarly dispersed across components, which limits the use of a magnitude-based similarity assessment and instead further supports a focus on curvature. A measure available for both major HCA implementations followed in the present study is found in Pearson's non-centered distance measure r_U, which is related to the respective correlation coefficient (fitting the description of its characteristics above) and thus similar in form:[68]

$$r_U = 1 - \frac{\sum_{i=1}^{n} x_i + y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2 \ \sum_{i=1}^{n}(y_i)^2}}$$

[68] This choice also likens the HCA part of the clustering approach to that pursued in Gries & Mukherjee (2010).

Having thus decided on the method for the generation of the distance matrix, the second question to be answered concerns the best amalgamation method for the dataset. Again, a substantial number of methods exist for this purpose (though not the hundreds available as distance metrics), with some aiming at identifying the minimal or maximal difference between any item in two clusters (single-linkage or complete-linkage, respectively), while "other methods can be regarded as aiming for clusters with characteristics somewhere between the single and complete link methods." (`hclust` documentation, R Core Team 2019). One of the most wide-spread amalgamation rules, and one particularly favored in linguistics, is Ward's method, which instead of choosing maximally similar or distant individual object, aims at limiting the overall within-cluster variance (error sums-of-squares), thus "producing groups that minimize within-group dispersion at each binary fusion." (Murtagh & Legendre 2014: 275; cf. also Moisl 2015: 203ff. 212-213 for a discussion of cluster amalgamation) The method is particularly advisable in case clusters of about equal sizes can be expected and if the dataset does not contain outliers (Sarstedt & Mooi 2014: 291). While the application of Ward's method likens the current methodology to many other linguistic applications of cluster analysis, it needs to be said that the current study builds on a revised amalgamation rule (`ward.D2` instead of `ward.D`, or `ward` in *R* versions 3.0.3 and before), since the long-time standard implementation of Ward's method in *R* actually forgoes the squaring of dissimilarities required by Ward's criterion (Murtagh & Legendre 2014).

It has already been addressed that cluster analysis may require a substantial amount of inspection and intuitive analysis of the dataset. This is a fact often embraced by researchers, as e.g. Moisl (2015: 215) summarizes:

> Which is the best cut, that is, the once that best captures the cluster structure of the data? There have been attempts to formalize selection of a best cut […], but the results have been mixed, and the current position is that the best cut is the one that makes most sense to experts in the subject from which the data comes.

However, since this forgoes the data-driven approach of the present study and allows for potentially conflicting analyses only resolved through the application of further clustering techniques, hierarchical analysis in the present study is objectivized through the application of two methods of substantiation of the findings. The first technique to be discussed addresses the certainty for individual subclusters through the application of

bootstrapping, i.e. multiple analyses of the same data in order to assess the degree of stability (or random findings, cf. above), and will be discussed below. The second, to be addressed in the section thereafter, concerns the identification of an optimal cut-off point within a dendrogram, i.e. the question which overall segmentation (as opposed to substantiation of individual subclusters) can be most sensibly applied to describe the entirety of the data. Together, these two methods allow for a more systematic hierarchical approach to the many datasets within the present study.

Addressing uncertainty in hierarchical clustering

Significance testing, much like more regular statistical tests, can be applied to a cluster dendrogram in order to evaluate the degree of substantiation that the cluster finds in the data. As mentioned above, (hierarchical) clustering approaches will always assign a data point to a cluster, no matter whether the similarities within the cluster (and dissimilarities outside) are strong or weak (or indeed a chance result): There might, for instance, be very little actual distinction between clusters but due to the nature of the binary mergers, pairs of data points will still be presented as most mutually similar. In the present study, confidence levels can be added to the cluster dendrogram through the application of the `pvclust()` function (Suzuki & Shimodaira 2006, 2015), which discerns the stability of hierarchical clusters based on bootstrapping of the data, i.e. the random generation of thousands of subsets (here n=10,000) and their comparison against the larger dataset in order to "indicate the extent to which the cluster result captures the intrinsic cluster structure of the data" (Moisl 2015: 246). While `pvclust()` calculates two types of values, 'bootstrap probabilities' (BP) and 'approximately unbiased' (AU), the latter presents the less biased indicator (Suzuki & Shimodaira 2006: 1541) and is thus preferred within the present study. The authors argue that "[o]ne can consider that clusters (edges) with high AU values (e.g. 95%) are strongly supported by data." (Suzuki & Shimodaira 2015). The analysis will, however, also make note of those clusters that barely miss this cut-off point, since even AU=94 clusters show stability in 9,400 out of 10,000 cases and should be acknowledged.

With the help of AU bootstrap values, clusters strongly supported (AU≥95) by the data can be identified. Figure 4.5 repeats the (fictitious) dendrogram from Figure 4.3, adding (black dotted) boxes around clusters retrieved by `pvclust()`. Stable clusters

can be detected on different heights of the dendrogram, stating which objects are usually found within the same clusters even in the bootstrapped data. Thus, Figure 4.5 would indicate the bilabial and alveolar plosives as well as the fricatives group as strongly supported, additionally indicating stability of the plosives overall. This could be taken to indicate less reliability of the nasals+liquid cluster, as well as returning more evidence for the distinguishing plosives into three rather than two types.

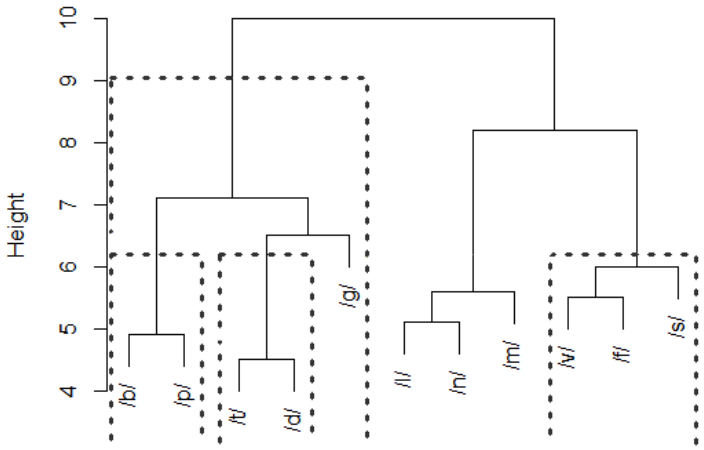


Figure 4.5: Dendrogram of the fictitious consonant data with dashed black lines indicating clusters achieving values of AU≥95

Deciding on the level of clusteredness

While the `pvclust`-based interpretation of the cluster solution helps in identifying relevant parts of the overall structure, it does not necessarily aid in the process of overall partitioning of the data, i.e. the question into how many clusters the data can most sensibly be segmented while also accounting for *all* objects at once. Particularly, it would not be advisable to segment the dendrogram along many different heights but rather aim at finding clusters that are on a similar overall height and set a cut-off point for interpretation correspondingly. Greenacre (2011: 2) summarizes that cluster segmentation most commonly follows a 'largest jump' approach:

> Faced with the dendrogram resulting from a hierarchical cluster analysis, the researcher has to make a decision where to cut the dendrogram to define the clusters. This is almost always done by the rule-of-thumb of looking for a large jump in the node heights, where there has been a large increase in the dissimilarity measure at that point to move to the next merging of the objects. (Greenacre 2011: 2)

There are two issues with this rule of thumb: First, in many cases, several jumps may be similar in size, and the question will present itself which additional, slightly smaller jumps should still be considered for segmentation of the overall cluster structure (i.e. the goal of drawing one horizontal line through the dendrogram which partitions the

entire data). In the dendrogram in Figure 4.6a, for instance, the largest jump height is found for two clusters (the uppermost horizontal line), but cutting at up to four clusters still shows substantially larger jumps than the levels below. Thus, a data-driven way of supporting a particular partition is required, particularly for even more extensive datasets such as in the present study, which may exhibit up to 28 objects (i.e. varieties). The second problem arises from the fact that even random data may lead to visually sensible cluster partitions, and so the actual dendrogram obtained has to be compared against what could have occurred if the data were a result of random variation. Figure 4.6b presents a dendrogram based on a random permutation of the data in Figure 4.6a, which means that values for each object (e.g. association strengths for an *n*-gram) have been randomly reshuffled. If reshuffling is repeated many times (10,000 times in the present study), the jump heights found in the original data (Figure 4.6a) can be contrasted against what is obtained in the randomized versions (Figure 4.6b). If a jump is found to be larger than what could be expected by chance (i.e. in the random permutations), a division at that height can be taken to be supported.

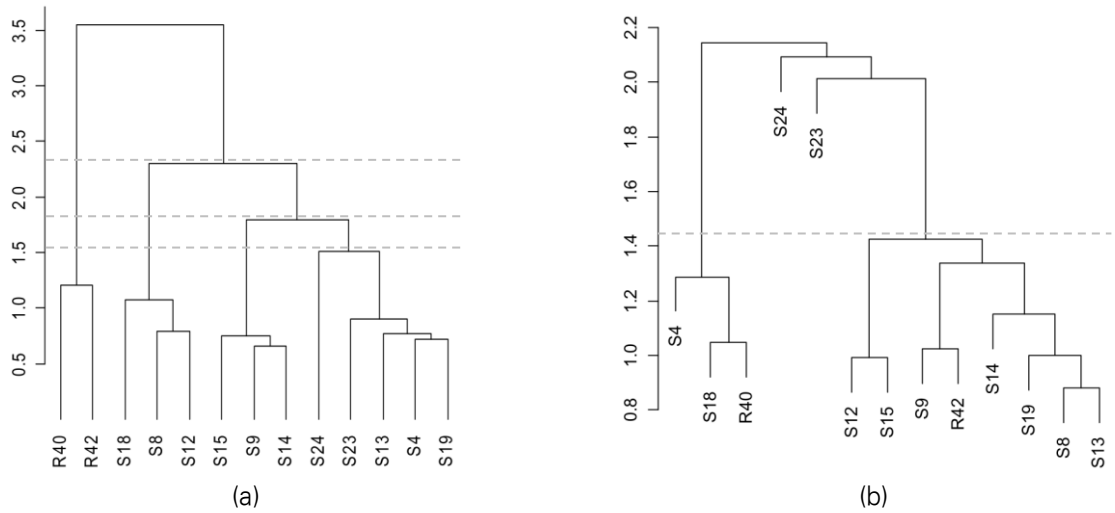


Figure 4.6: Cluster dendrogram of some base data (panel a) next to a random permutation (panel b). Visualizations taken from Greenacre (2011).

Greenacre's (2011) "simple permutation test for clusteredness" offers a data-driven solution for both of these problems. It provides an assessment of the best number of clusters *k* in a dendrogram by identifying the largest significant 'jumps' in node heights, i.e. the earliest points at which large amounts of variation can be explained (in contrast to the eventual fragmentation of the data into individual objects,

which automatically accounts for all variance). Similar to `pvclust`, it builds on random permutations of the data and comparing the original data against its chance alternatives:

> [T]he question is whether one can identify a level where the nodes below this level are significantly lower than one might expect by chance. At the same time, one might consider the topmost node and ask whether a random version of the data could have led to a dendrogram with a higher topmost node. (Greenacre 2011: 7)

The test returns a list of nodes, their respective heights in the dendrogram (their position on the vertical axis) as well as, most centrally, the sizes of the jumps to the next-coarser segmentation (column 'Nr. clusters') and their associated *p*-values (cf. Table 4.11, which provides further fictitious data). Here, a small amount of subjective inference is required in that comparatively large jumps with *p*-value≤0.05 need to be identified by hand. In the present example, segmentation into 2, 3 or 4 clusters finds the largest jumps, but these fail with regard to the respective *p*-values and thus need to be disregarded. More fine-grained solutions with smaller jumps are found at 5 and 6 clusters, of which only the former passes the significance level. A few later jumps are also found significant but only produce miniscule jumps. In this case, 5 clusters consequently provides the preferred solution.

Table 4.11: Sample results produced by Greenacre's permutation test for clusteredness

Node	Height	Jump	p-value	Nr. clusters	
1	0.0021	0.0002	0.192	10	Small and not significant.
2	0.0023	0.0006	0.055	9	
3	0.0029	0.0009	0.038	8	Significant but very small.
4	0.0038	0.0000	0.020	7	
5	0.0038	0.0062	0.121	6	► Relatively large but not significant.
6	0.0100	0.0062	0.001	5	► Relatively large and significant.
7	0.0162	0.0363	0.155	4	Large but not significant.
8	0.0525	0.0322	0.777	3	
9	0.0847	0.2684	0.267	2	
10	0.3531	*NA*	0.541	1	

For ease of inspection, a visual aid will be provided instead of the crowded tabular display. The respective version for the above sample data is shown in Figure 4.7a, indicating jump heights through lengths of the bars as well as *p*-values for each jump, the latter of which are plotted at the height of a dashed line indicating average jump heights across the entire (sample) dataset. However, since ever more fine-grained segmentations continually diminish the explanatory power of the clustering approach, only up to 11 clusters will ever be displayed in this graph (of which only 10 are available

for the consonant data).[69] The results of cutting the dendrogram of the consonant data into 5 clusters is in turn visualized in Figure 4.7b, indicated by solid gray lines. The ficticious clustering results would indicate that the plosives group should best be separated into three subgroups while liquids+nasals and fricatives provide sensible categories.

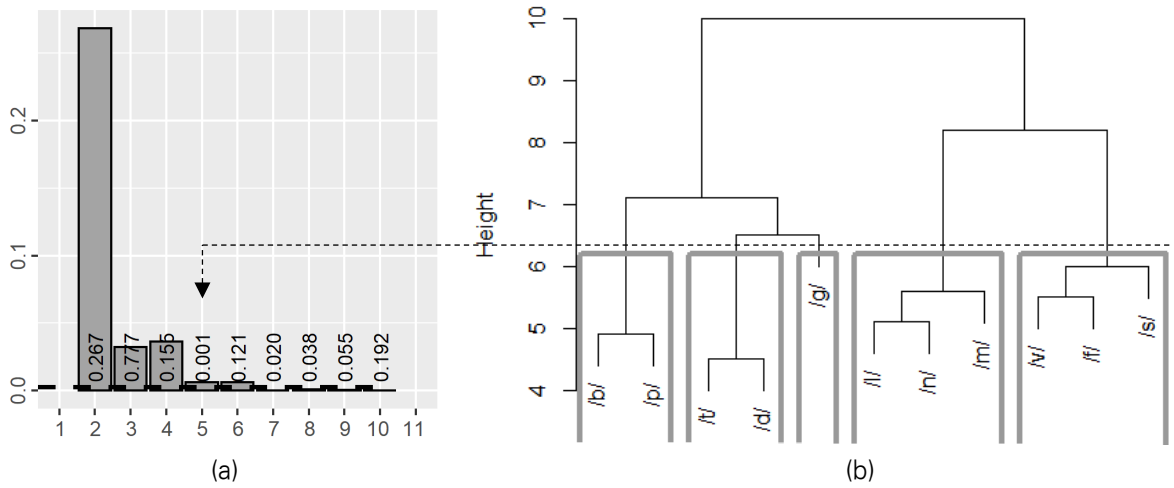


Figure 4.7: Visual representation of the results of Greenacre's permutation test (panel a) next to an application of the 5-cluster segmentation to the fictitious consonant data (panel b).

K-means clustering

Like hierarchical clustering, the k-means approach represents another staple technique of finding patterns in complex data. The name implies a variable number of mean values, and indeed this summarizes the method quite succinctly already: A number of *k* clusters is decided a priori and an equivalent number of data points is randomly selected from the entire data to represent the initial elements of each of the *k* clusters ('prototype objects', cf. Moisl 2015: 182). Based on sums of squares (thus similar in concept to Ward's method for HCA but different in implementation in that it does not employ distance measures), the algorithm assigns each element to the closest of the *k* clusters (cf. e.g. Moisl 2015: 181–182, Sarstedt & Mooi 2014: 294–295, Zelterman 2015: 295). After all elements are assigned to a cluster, the algorithm calculates the mean value of all elements assigned to each of the *k* clusters, and employs these as the new values against which all elements are compared within a repeated run of the process. As soon as newly assigned mean values bring about no further changes to

[69] Since there are 16 written corpus parts and 12 spoken ones, theoretically 15 clusters could be found in the written data, while 11 presents the maximum for speech. For the combined spoken-plus-written data, however, up to 27 clusters could theoretically emerge. These do not, however, become relevant in terms of jump heights and *p*-values throughout the study.

the elements involved in each cluster, the algorithm stops, outputting the cluster structure.

K-means clustering is an appealing alternative to a hierarchical approach since it clearly groups elements in contrast to the greater reliance on human inspection in most HCA approaches. Instead, any number of clusters can be decided on beforehand. However, this also presents two challenges: Firstly, the question arises to which extent the random allocation of the initial *k* clusters influences the final clustering results, since several applications of the method to the same data might lead to different results due to the random nature. Fortunately, this can be solved once again through the creation of random permutations of the data (*n*=10,000). Even with reliability thus controlled for, a second issue lies in the very nature of the definition of *k*. While in some applications, a most appropriate *k* is available beforehand, for the current study a different approach was devised in that each dataset was subjected to several runs of the k-means algorithm for any number of clusters *k* between 1 and 11. Figure 4.8 shows how sums of squares in between-cluster variation increase as those within clusters are diminished (overall variability explained vs. variability within each cluster, expressed as percentages of total sums of squares): Essentially, at *k*=1 clusters, all variability is found inside a single cluster, which thus provides no sensible explanatory categories for the data. Conversely, if all objects are allocated to their separate clusters, all within-cluster variability is accounted for (since there is no variability in clusters of only one object each), but all variability is instead found between clusters (which neither provides sensible categories, since no objects are grouped together). Either case does not explain the data sensibly, and instead an in-between approach is required which weighs informativity of each cluster against the effort of producing further groups (cf. Zelterman 2015: 297).

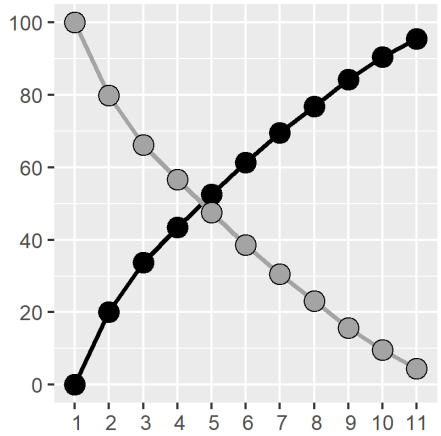


Figure 4.8: Percent variability explained (black) and within-cluster variation (gray) plotted against the number of k-means clusters

Favorable values for *k* can be inferred from the intersect of these two values (middle ground between too many and too few clusters) as well as from any greater rises/falls in either line (the 'elbow criterion', which reproduces looking for the greatest jumps in HCA). In this fictitious example, *k*=5 clusters would emerge as the segmentation most supported by the data.

Phylogenetic clustering using NeighborNet

The final clustering method to be employed on the *n*-gram dataset originally stems from bioinformatics but has found its way into comparative linguistics as well. The NeighborNet algorithm (Schliep 2011, Schliep et al. 2017) concerns the production of network structures of originally evolutionary relationships (phylogenetics) but can just as much be carried over to linguistic data (compare Lunkenheimer's 2013 application of the method to the eWAVE data). It takes the same distance matrix as hierarchical clustering as input and also agglomerates the data into ever larger structures. Unlike hierarchical approaches, however, the resulting clusters do not follow a hierarchical binary structure and can indeed overlap.

Interpretation of the network diagram succeeds by comparing the lengths of paths between data points (cf. Lunkenheimer 2013: 858): The length of the shortest path between two objects roughly translates (since no visualization can be a perfect representation of the underlying data) to the linguistic distance between them. The longer the path between two varieties, the greater their difference in terms of relative preference for the sequences under scrutiny, while a shorter path conversely indicates more mutual similarity. Since paths are plotted for each object to all others, parallel

lines capture mutual (dis-)similarity between several objects. Sets of parallel lines between two groups of varieties thus indicate that a certain distance is observed between all members of the first group and those in the second. If these parallel lines are long, this consequently informs of a split in the data. In Figure 4.9 (which provides a sample NeighborNet taken from Lunkenheimer 2013), a large set of parallel lines is found between clusters A and B, indicating a major split between these groups of varieties. Smaller sets of such parallel lines are in turn found for all members of cluster 1, 2 and 3. In the latter case, the two varieties separate from all others by long lines while being themselves connected by relatively short paths. This stands in contrast to the remains of cluster B: There, mostly boxy shapes are found, which indicate distance between all objects involved, and only very small numbers of parallel lines are found in mutual proximity. Long parallel lines and few boxy shapes thus represent stronger distinctions in the data, while a high frequency of boxes and shorter distances indicate more mutual similarity.

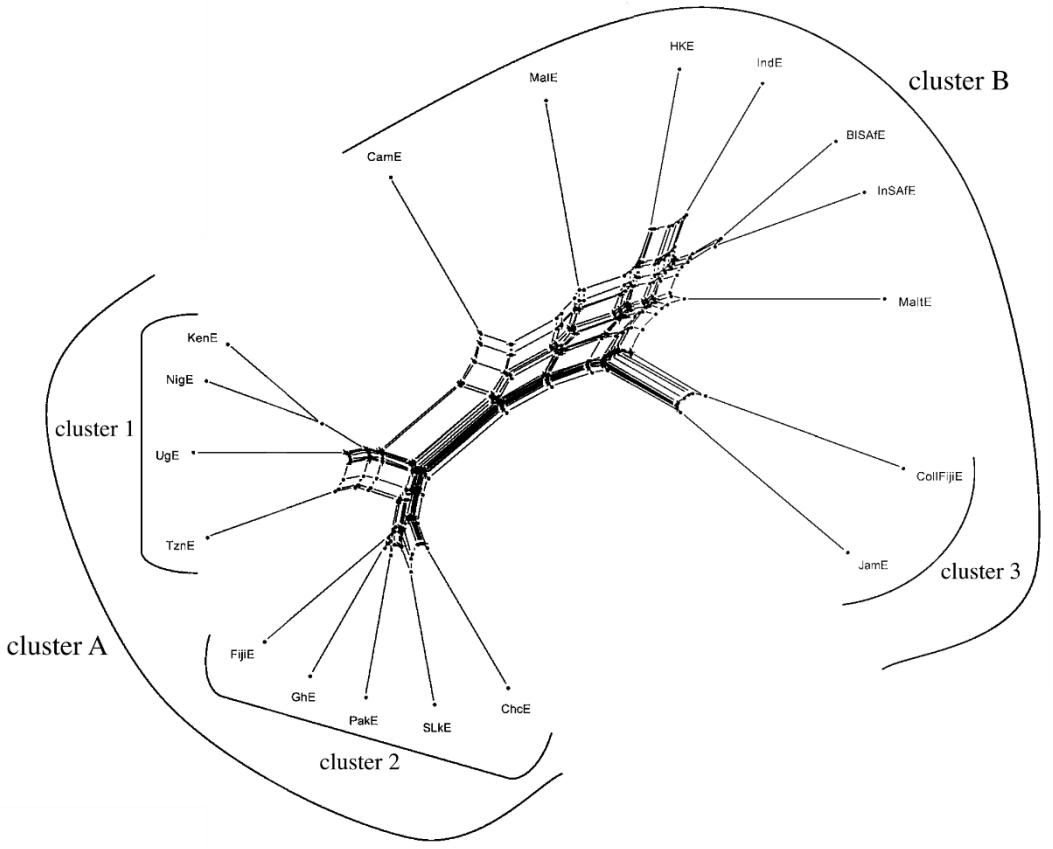


Figure 4.9: Sample NeighborNet visualization (taken from Lunkenheimer 2013: 859)

In contrast to the other two methods, this final method forgoes any enforcement of cluster structure, either in the form of a fixed number of *k* groups (k-means) or

through an underlying disposition to produce binary groups (HCA). However, the more intuitive analysis of the network graphs is also prone to subjectivity, particularly if more fine-grained structures are analyzed (cf. Figure 4.9). As such, results obtained on the basis of the NeighborNet method will be viewed more in contrast to those from the other approaches, in order to find out whether clusters established there can also be confirmed in this three-dimensional representation of pairwise (dis-)similarities. In case results of the NeighborNets strongly diverge from those of the other methods, only major splits in the data will be discussed since the reliability of smaller clusters cannot be guaranteed within the present framework without substantiation from other methods.

4.4.2 Questions, Assumptions and Hypotheses

The methods presented in the previous section form the backbone of the analytical approach for the current study. Methodologically, they will be applied in a triangulative fashion, assessing which clusters emerge from each of the individual methods while focusing on inter-method substantiated clusters. Thus, inter-method 'cluster stability' lies at the heart of the analysis:

> Stability means that the cluster membership of individuals does not change, or only changes little when different clustering methods are used to cluster the objects. Thus, when different methods produce similar results, we claim stability. (Sarstedt & Mooi 2014: 299)

Five methods are applied to the data in total: HCA plus stable subgroups (`pvclust`) and significant jumps (Greenacre's permutation test) as well as k-means clustering and the NeighborNet method. However, the HCA portion will usually focus little on the precise stepwise patterns found within the dendrogram, since these require substantial human inference. Instead, a more data-driven approach will be favored, and thus greater informative value is assigned to stable clusters identified through `pvclust` as well as significant jump heights. If available, clusters mutually agreed upon by these two methods will be preferred. Analyses will be carried out for spoken and written corpus parts separately (*n*-grams shared between all spoken or written corpus parts) as well as for the entirety of the ICE data (intersect of the merged spoken and written *n*-gram lists). Visual inspections of the NeighborNets will then form the second analytical step, and while the algorithm does not produce clear clusters, agreement between its results and those of the previous HCA and consecutive k-means approach

can be addressed. Since NeighborNets for the entirety of the ICE data (all 28 spoken and written parts) almost universally become too cluttered for presentation, only the spoken and written network graphs will be discussed but the combined spoken+written graph will be disregarded. Finally, k-means clusters will be produced (again for all three types of datasets). In case two solutions appear equally mandated by the data (e.g. the intersect occurs between two values of *k* or an earlier elbow point is clearly identifiable), competing clusterings will be discussed but one result will be selected as the preferred output for the subsequent evaluation. In such cases of ambiguity, the common practice is to choose the segmentation that "makes most sense to experts in the subject from which the data comes." (Moisl 2015: 215) Individual analyses (one for each association measure, lexical or POS base data, and dynamic-length sequences as well static-length 2-, 3-, and 4-grams will mostly restrict themselves to a description of groups of varieties identified, while the main analysis and interpretation of the findings is deferred until Chapter 6, which offers a discussion of results from a diverse set of perspectives.

The final methodological point to be considered after the successive steps of the approach outlined above are thus presented concerns the interpretative framework under which the methods will be applied. The main assumption underlying the current study certainly lies in the expectation that the patterning emerging from the *n*-gram data will be explicable on the basis of some (combination of) language-external descriptors of any kind. To this end, the present study takes on a bird's-eye perspective on lexicogrammatical variation by subjecting lexical as well as part-of-speech sequences of different types to a methodology which allows to discover patterns in complex data beyond concretely specified hypotheses of the 'the more/less X, the more/less Y' type. Instead, the current approach is devised in such a way that patterns of usage and preference of lexical and grammatical *n*-grams can be mapped across the entirety of World Englishes available through the ICE corpus. It is expected that these patterns relate to (largely) language-external factors as captured in major models for the description of World Englishes. As has been laid out in Chapter 2, these models, on the level of varietal granularity (i.e. relatively coarse-grained 'national varieties') reflect either the forms of the predominant speaker groups, phases in the Dynamic Model or otherwise degrees of relation on the bases of mutual cultural interaction,

regional proximity or shared substrate languages, the latter three most commonly co-occurring instead of forming clearly separate categories. Thus, the central concern of the present study can be encapsulated in the following research question:

> **Is there systematic collocational variation in the entirety of the ICE data (spoken and written modes as well as all combined data), and, if so, does that variation correlate systematically with categories informed by major models of World Englishes?**

The models and descriptive frameworks presented in Chapter 2, while very different in character and scope, all suggest certain expectations towards which patterning is most likely to emerge from the *n*-gram data: The very traditional ENL-ESL-EFL classification would imply a segmentation of the data into these respective groups (i.e. a binary segmentation, since EFL is not represented in the data). Kachru's (1992b: 356) Three Circles model, while different in spirit from this previous classification, still conforms to it in that membership of varieties in one of its circles overlaps with its classification in the previous model, thus similarly suggesting two major clusters (Inner vs. Outer Circle). Models focusing on proximity and regional standards, from Strevens's (1992: 33) world map of English over Görlach's (1988) Circle model of English to McArthur's (1987) Circle of World Englishes, on the other hand, imply a consecutive differentiation from standard(izing) varieties over regional standards to local forms. In this regard, it is conceptually compatible to the theory of epicenters in English (which itself remains untestable within the present approach), and would thus support a more regionally informed clustering of varieties, i.e. South and East Asian as well as East and West African clusters. For the Inner Circle varieties, these models would posit American, British as well as Antipodean/Australasian branches, which likely will be too fine-grained a segmentation for the ICE data (five varieties, one of which only available in writing).

The most prominent consistent approach to the description of English today, however, is certainly found in Schneider's (2007) Dynamic Model. Like many other studies following the publication of the Dynamic Model, the current analysis also centrally concerns the empirical testing of the model's phases on actual linguistic data. Specifically, the phases of the dynamic model are described in terms of several language-external characteristics, and only concern linguistic effects in a very general form. While structural nativization takes center stage and is the main driving force of the process of variety formation, and while lexicogrammatical patterning is in turn central

to the nativization processes (cf. Chapters 2 and 3), it is unclear whether a consistent pattern of lexicogrammatical (dis-)similarity can be discerned between varieties on different stages in the Dynamic Model. In terms of the cluster structure, this model would suggest some form of clustering of varieties within similar phases. This might take the form of a broad distinction into more or less 'advanced' varieties such as a binary segmentation into phase 4 and above vs. all others, or an isolation of less advanced varieties of e.g. phases 2-3. In case of a perfect reflection of the Dynamic Model within the data, a stepwise pattern of (dis-)similarity might also manifest itself, with varieties on major stages of the model merged into coherent groups while consecutively found more dissimilar to previous stages. Table 4.12 summarizes estimates of phases for all varieties captured within ICE, particularly drawing on Schneider's (2007) initial overview but also incorporating those in Hundt (2018) and Werner (2013) as well as sources on specific varieties: Huber (2014) and Buschfeld et al. 2018 on Ghanaian English, Meierkord 2012 on Ugandan English, Mukherjee 2007 on Indian English and Bernaisch (2015) on Sri Lankan English, as well as Hickey (2016) on Irish English.

Table 4.12: ICE varieties and respective phases in the Dynamic Model

Phase 1:	**Phase 2:**	**Phase 3:**	**Phase 4:**	**Phase 5:**
Foundation	**Exonormative Stabilization**	**Nativization**	**Endonormative Stabilization**	**Differentiation**
	TZ	GH		CAN
	HK	NIG	SIN	NZ
	(PHI[70])	IND		GB
	SL	JA		IRL
	UG			USA
	KY			
	EA_{SPK}			

Beyond the largely language-external descriptors presented above, however, some language-internal considerations are also applicable to the patterns discovered within the *n*-gram data: Firstly, the data will show whether the distinction in modes (spoken vs. written) as a design feature of the ICE corpus will re-emerge from the patterns of

[70] In contrast to other varieties progressing along the developmental cline, English in the Philippines is a likely case of regression (or "restriction" in Moag's (1992) cycle. While some authors see "[s]igns foreshadowing codification in phase 4" (Schneider 2007: 143), others place it squarely within the phase 2 category (Hundt 2018: 239). Thus, the variety's assignment can be regarded as less secured than that of others.

use of the *n*-grams shared between varieties. While strong differences between written and spoken language should not be a surprise, it needs to be recalled that at least within the overall combined datasets, only *n*-grams shared between the written and spoken parts are present and only the association scores resulting from the actual use of these patterns inform the clustering results. The separate sets of spoken and written data, on the other hand, will make it possible to determine whether (dis-)similarities between varieties will follow similar explanatory lines for both modes or whether these will differ, as well as whether overall similarity of the varietal patternings will be greater in one mode than in the other. If conventional expectation is to be trusted, the written language should follow more established patterns in this regard, with stronger differences between varieties emerging in the spoken forms, i.e. "[c]onvergence in writing–divergence in speech" (Mair 2007).

Finally, the association scores of *n*-grams will themselves be a worthwhile object of study, in that different measures highlight distinct types of co-occurrence and thus different ways of using language. Since most of the commonly-applied association measures are strongly correlated to one another (Levshina 2015: 238), differences can be expected to arise mostly on the level of less stable groups and more fine-grained structures within the data, which may be influenced by differences between measures. The present study cannot, however, provide a systematic evaluation of the performance of each measure, since some form of benchmark (a list of target *n*-grams) would be required for this purpose. Thus, it can only contrast the stability of findings for each measure across sequence lengths, data types (lexical or grammatical) or modes, or alternatively identify for any such combination of variables those measures producing results incompatible with the general trend.

5 Significant Sequences in World Englishes

5.1 Dynamic-length Lexical *N*-grams

Extraction of lexical *n*-grams resulted in 28 bigram lists and token and type frequencies as presented in Table 5.1.[71] Token frequencies therein are a direct consequence of the varying component sizes[72] but may also reflect utterance structures contained in the data: For instance, single-word utterances would not make bigrams, which may be of particular effect in speech, while some punctuation (e.g. quotation marks) was taken to separate text units within the written data (cf. Section 4.3.1).

Table 5.1: Frequencies of lexical bigram types and tokens, plus TTRs, extracted from the ICE data

Component	Spoken Mode			Written Mode		
	Tokens	Types	TTR	Tokens	Types	TTR
CAN	548,879	181,612	3.02	355,056	177,337	2.00
EA	541,571	186,215	2.91	–	–	–
KY	–	–	–	341,475	156,710	2.18
TZ	–	–	–	350,829	156,038	2.25
GB	556,918	199,702	2.79	368,038	177,610	2.07
GH	–	–	–	354,652	160,058	2.22
HK	638,104	197,534	3.23	412,543	186,709	2.21
IND	609,391	211,166	2.89	357,909	172,423	2.08
IRL	540,610	186,127	2.90	372,046	175,853	2.12
JA	573,282	189,878	3.02	355,050	165,414	2.15
NIG	532,977	174,980	3.05	349,507	160,174	2.18
NZ	607,220	204,377	2.97	373,149	182,866	2.04
PHI	596,552	206,875	2.88	378,432	184,217	2.05
SIN	542,955	191,651	2.83	348,151	165,300	2.11
SL	548,741	189,603	2.89	342,723	162,577	2.11
UG	–	–	–	349,953	160,610	2.18
USA	–	–	–	364,562	181,974	2.00
mean ($\bar{x}$)	**569,767**	**193,310**	**2.95**	**360,880**	**170,367**	**2.12**
sd (s)	33,083	10,356	0.12	17,049	10,287	0.07

For ease of comparison, Table 5.1 also provides type-token ratios, which indicate the average number of tokens per type. This shows a larger repetition of the same bigram types in the spoken corpus parts (TTR≈2.95) than in the written data

[71] Please recall that three ICE components are only finished in terms of their written parts at the time of writing (GH, UG, USA), while ICE-EA contains separate written material for Kenya and Tanzania, which has been compiled into separate datasets in the present study (cf. Section 4.2.1).
[72] Some exceed the common 600,000+400,000 token marks, in particular ICE-HK.

(TTR≈2.12), but relatively stable mean TTRs across components.[73] These bigrams will be discussed further during the analysis of static-length 2-grams in Section 5.2, but for the present purposes only provide the basis for the generation of longer dynamic-length sequences. Before the merging procedure, however, the variety-specific lists of bigrams for each of the five association measures were first trimmed to only those surpassing the respective association thresholds (cf. Section 4.3.2), which are employed to retrieve only those bigrams representing 'true' collocations for the merging procedure. The threshold values are repeated here in Table 5.2.

Table 5.2: Threshold values for the selection of bigrams for dynamic-length *n*-grams (association scores need to be greater than these values).

MI	***t***	***G²***	***g***	***ΔP***
3 ($O \geq 5$)	2.576	6.64	5.5	0

Token and type frequencies in the trimmed lists are shown in Tables 5.3 and 5.4, which furthermore indicate the impact of the association thresholds ('threshold effect'). While the low threshold used for *ΔP* has only a minor impact on bigram counts, more pronounced drops in token and, particularly, type frequencies are effected for *G²* bigrams and even more for *MI*, *t* and *g*. Note, however, that reductions in types and tokens are not directly proportional: *MI*'s focus on rare items awards high association values to bigrams with relatively low TTRs while measures with a high-frequency bias like the t-score favor higher-TTR items. Thus, different numbers of tokens are affected by the removal of similar numbers of types.[74]

The trimmed bigram lists in turn formed the basis for the formation of variable-length *n*-grams: Consecutive merging of adjacent items was carried out until average association scores would drop with the addition of the subsequent bigram (cf. Section 4.3.2). Analyses for each of the five association measures will be carried out below.[75]

[73] Mean TTRs are means of the TTRs from the corpus parts (thus the SD), and not calculated from the mean token and type frequencies.

[74] If the threshold were set at a value of 0, *MI*, *t* and *G²* would all retain the same tokens, since they produce negative values only for O<E.

[75] Type frequencies can be higher in the analyses to follow than in Tables 5.3 and 5.4. This is due to the way *n*-grams are generated: While a shorter *n*-gram token is deleted from the data if it can be fused with a consecutive bigram, this procedure is repeated for all instances of this *n*-gram. A later token may not be fused with its consecutive bigram and thus remain in the data as its shorter form, increasing type diversity.

Table 5.3: Spoken lexical bigram token and type frequencies, plus TTRs, after the application of threshold values

Spoken	MI			t			G^2			g			*ΔP*		
component	**tokens**	**types**	**TTR**	**tokens**	**types**	**TTR**	**tokens**	**types**	**TTR**	**tokens**	**types**	**TTR**	**tokens**	**types**	**TTR**
CAN	125,084	5,840	21.4	225,430	5,489	41.1	382,214	97,216	3.9	184,364	1,804	102.2	489,987	164,453	3.0
EA	128,922	6,703	19.2	219,779	5,926	37.1	390,502	102,936	3.8	170,854	1,844	92.7	497,423	171,820	2.9
GB	116,298	5,831	19.9	215,994	5,427	39.8	386,785	109,530	3.5	178,640	1,859	96.1	499,189	183,078	2.7
HK	148,317	6,818	21.8	270,048	6,307	42.8	436,836	99,069	4.4	226,932	1,995	113.8	559,510	176,178	3.2
IND	137,502	6,986	19.7	243,491	6,215	39.2	429,020	113,883	3.8	197,916	2,051	96.5	549,654	192,227	2.9
IRL	113,781	5,594	20.3	212,846	5,228	40.7	375,155	100,495	3.7	178,297	1,838	97.0	484,361	169,737	2.9
JA	123,673	6,162	20.1	233,894	5,771	40.5	398,488	100,183	4.0	193,239	1,940	99.6	511,762	172,081	3.0
NIG	119,117	5,802	20.5	217,128	5,428	40.0	365,402	89,531	4.1	177,619	1,710	103.9	471,986	157,003	3.0
NZ	131,774	6,202	21.2	245,270	5,922	41.4	419,125	108,819	3.9	203,809	2,006	101.6	541,341	185,232	2.9
PHI	138,129	6,907	20.0	236,814	6,125	38.7	422,388	114,614	3.7	194,933	2,003	97.3	536,950	189,373	2.8
SIN	118,114	5,796	20.4	210,829	5,341	39.5	374,965	101,742	3.7	173,155	1,832	94.5	486,316	174,508	2.8
SL	119,040	5,945	20.0	214,516	5,318	40.3	377,653	100,664	3.8	179,167	1,766	101.5	487,029	172,014	2.8
mean ($\bar{x}$)	**126,646**	**6,216**	**20.4**	**228,837**	**5,708**	**40.1**	**396,544**	**103,224**	**3.8**	**188,244**	**1,887**	**99.7**	**509,626**	**175,642**	**2.9**
sd (s)	10,091	481	0.7	16,964	366	1.4	23,178	6,957	0.2	15,367	104	5.3	28,285	9,833	0.1
Threshold effect	-78%	-97%		-60%	-97%		-30%	-47%		-67%	-99%		-11%	-9%	

Table 5.4: Written lexical bigram token and type frequencies, plus TTRs, after the application of threshold values

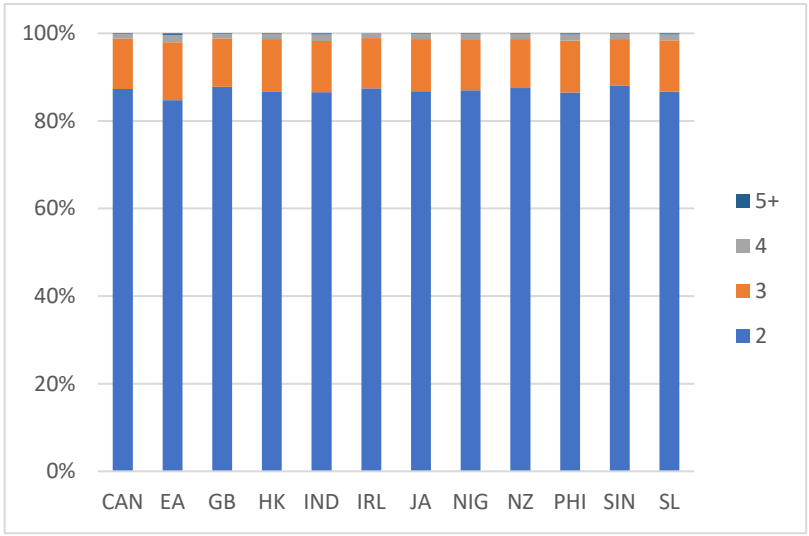
Written component	MI tokens	MI types	MI TTR	t tokens	t types	t TTR	G^2 tokens	G^2 types	G^2 TTR	g tokens	g types	g TTR	*ΔP* tokens	*ΔP* types	*ΔP* TTR
CAN	55,255	4,160	13.3	91,345	3,236	28.2	254,367	111,516	2.3	68,763	1,086	63.3	334,913	169,323	2.0
KY	60,723	4,274	14.2	100,778	3,467	29.1	244,083	94,714	2.6	75,789	1,161	65.3	320,343	148,658	2.2
TZ	66,439	4,630	14.3	106,939	3,700	28.9	252,508	94,356	2.7	78,618	1,166	67.4	329,592	147,630	2.2
GB	61,682	4,252	14.5	104,004	3,457	30.1	265,047	110,100	2.4	80,433	1,245	64.6	346,431	168,974	2.1
GH	62,928	4,566	13.8	105,451	3,659	28.8	256,401	97,630	2.6	77,201	1,132	68.2	333,489	151,779	2.2
HK	83,706	5,741	14.6	126,878	4,424	28.7	302,116	115,381	2.6	89,772	1,287	69.8	389,699	177,103	2.2
IND	62,043	4,625	13.4	98,742	3,611	27.3	260,637	109,006	2.4	71,266	1,110	64.2	337,854	164,339	2.1
IRL	63,911	4,571	14.0	104,987	3,553	29.5	266,909	108,630	2.5	79,576	1,217	65.4	350,562	167,326	2.1
JA	61,738	4,448	13.9	100,956	3,476	29.0	253,878	101,480	2.5	75,791	1,131	67.0	333,811	156,890	2.1
NIG	60,797	4,305	14.1	102,731	3,517	29.2	252,113	98,139	2.6	76,489	1,136	67.3	328,435	151,815	2.2
NZ	60,978	4,290	14.2	101,058	3,415	29.6	265,455	112,346	2.4	77,777	1,204	64.6	351,474	174,006	2.0
PHI	61,433	4,620	13.3	100,889	3,511	28.7	272,378	115,854	2.4	75,080	1,119	67.1	356,896	175,632	2.0
SIN	61,672	4,564	13.5	95,750	3,451	27.7	251,517	102,743	2.4	69,812	1,067	65.4	328,584	157,428	2.1
SL	61,157	4,470	13.7	96,178	3,498	27.5	248,341	101,711	2.4	68,921	1,058	65.1	323,487	154,465	2.1
UG	65,380	4,722	13.8	99,666	3,677	27.1	252,446	99,138	2.5	71,154	1,127	63.1	330,167	152,401	2.2
USA	59,321	4,405	13.5	93,486	3,443	27.2	262,147	115,102	2.3	69,342	1,129	61.4	343,831	173,597	2.0
mean ($\bar{x}$)	**63,073**	**4,540**	**13.9**	**101,865**	**3,568**	**28.5**	**260,021**	**105,490**	**2.5**	**75,362**	**1,148**	**65.6**	**339,973**	**161,960**	**2.1**
sd (s)	5,844	349	0.4	7,698	247	0.9	13,119	7,325	0.1	5,357	61	2.1	16,382	10,054	0.1
Threshold effect	-83%	-97%		-72%	-98%		-28%	-38%		-79%	-99%		-6%	-5%	

5.1.1 MI-score

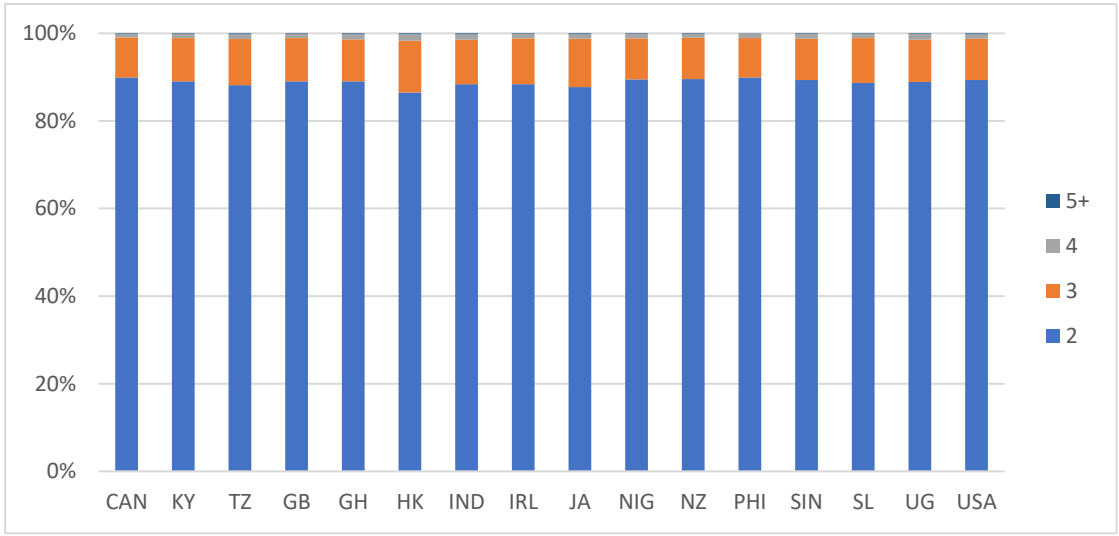
Fusing bigrams into *n*-grams (cf. Chapter 4, particularly Figure 4.2) on the basis of the *MI* measure yielded, on average, 110,368 (*s*=8,417) spoken and 55,998 (*s*=4,688) written *n*-grams of 10,070 (*s*=934) and 6,711 (*s*=567) types, entailing mean TTRs of 11.0 (*s*=0.5) and 8.3 (*s*=0.1). While token frequencies dropped slightly from the bigram numbers after their (relatively strict) thresholds (-13% and -11%), moderately many new types were generated (+62% and +48%). Average lengths of the resulting tokens were almost identical between modes at 2.15 (*s*=0.01) units for tokens in speech and 2.13 (*s*=0.01) in writing, but types were found to be slightly longer in speech (2.51 units, *s*=0.02) than writing (2.43, *s*=0.02). This should result in relatively similar distributions of *n*-gram lengths between speech and writing (with slightly longer types in writing), and indeed Figure 5.1 supports this. Additionally, *MI* is found to generate mostly very short sequences, likely a result of the significant thresholds imposed on *MI* drastically reducing the numbers of consecutive bigrams available for merging. Concerning longer sequences, only 3- and 4-grams can be said to become somewhat numerous, with the latter only noticeable within the type data. Longer sequences were only generated in fractional numbers and visualization is capped correspondingly.

Merging (technically: creating the intersect) of the varietal *n*-gram lists only retained items shared between all datasets. Figure 5.2 visualizes the distribution of overlap between any two sets: It confirms that no single intersect drastically lowers the number of shared items, and only a positive (i.e. unproblematic) outlier is detected (written SIN+HK). Even though many sequences are shared between any two varieties, the intersect of all varieties still shows very large reductions in types frequencies to 647 (-94%) and 490 (-91%) items for speech and writing, respectively.

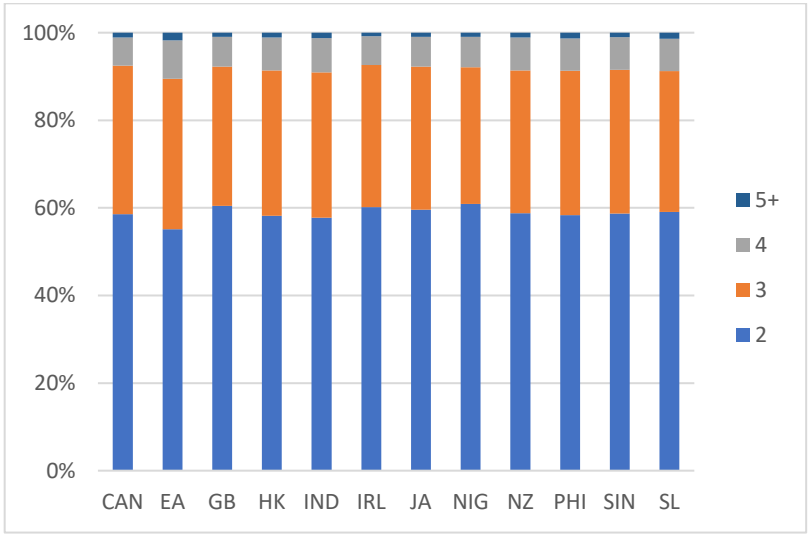
Given the low frequencies of types of 4 units and longer, none are found in the spoken and written datasets (henceforth referred to as the SPK and WRT) or in the overall combined data (henceforth ALL), which is the intersect of SPK and WRT (Table 5.5). Table 5.6 presents the top and bottom items in terms of association scores, illustrating how *MI* frequently assigns higher values to fixed expressions or such incorporating rarer items (e.g. *et al*, *per*). This becomes particularly visible with the lowest-ranking *n*-grams, which always contain at least one high-frequency item.



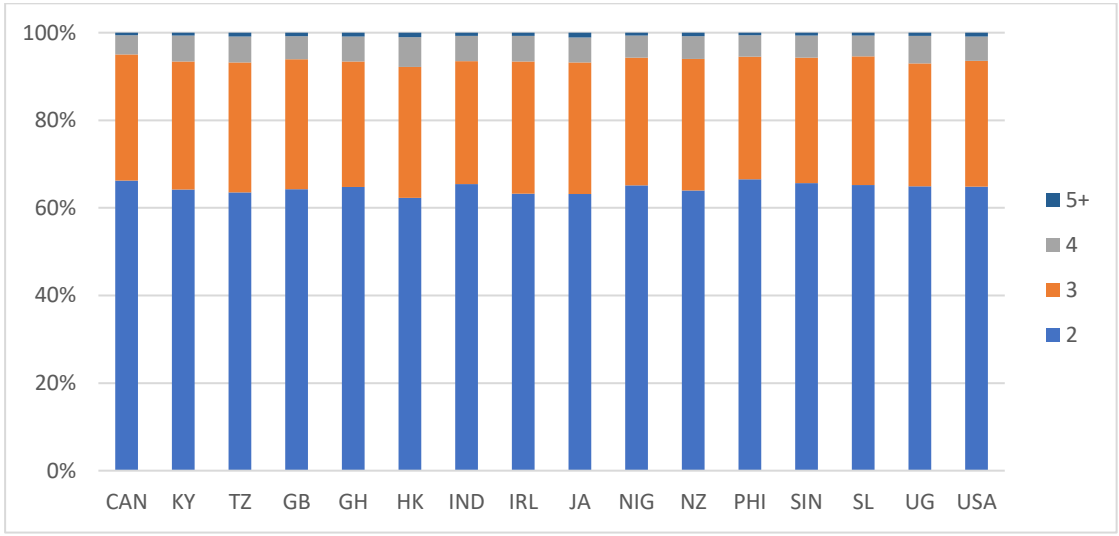
Token frequencies: Spoken data



Token frequencies: Written data



Type frequencies: Spoken data



Type frequencies: Written data

Figure 5.1: Distribution of lexical *MI n*-gram lengths across the varietal datasets

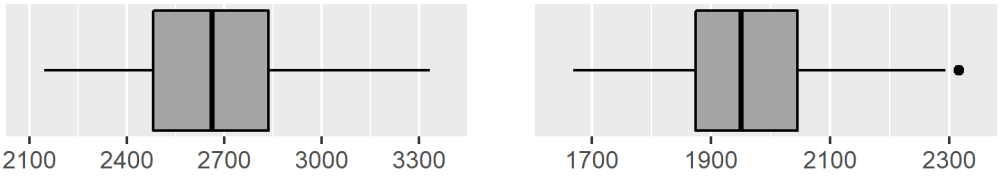


Figure 5.2: Number of shared lexical *MI n*-grams between any two datasets

Table 5.5: Lexical *MI n*-gram type frequencies by length in the intersects of the variety-specific datasets

***N*-gram length**	**ALL**	**SPK**	**WRT**
2	258 (96%)	603 (93%)	455 (93%)
3	11 (4%)	44 (7%)	35 (7%)
total	269	647	490

Table 5.6: Lexical *MI n*-grams with highest and lowest association scores

SPK		**WRT**	
***n*-gram**	***MI***	***n*-gram**	***MI***
prime minister	11.31	et al	12.42
united states	11.11	per cent	10.67
little bit	9.07	united states	10.56
nineteen ninety	9.07	six months	9.02
years ago	8.92	years ago	8.88
difference between	8.73	pointed out	8.54
might not	3.40	they should	3.42
when we	3.34	to find	3.41
to take	3.30	they could	3.41
a sense	3.29	it would	3.40
to get	3.29	the highest	3.33
the world	3.26	the entire	3.31

Hierarchical clustering (Figure 5.3) reveals a clear separation of spoken and written modes as the only stable and significant cluster in ALL but no further stable groups after bootstrapping using `pvclust` (black, dotted rectangles). In particular, neither SPK nor WRT provide further stable clusters.

Testing for significant jumps in node heights (cf. Figure 5.4; clusters indicated as gray solid rectangles in Figure 5.3) confirms the binary separation inside ALL. It also finds a large and significant jump at k=3, splitting the spoken branch into JA+Inner Circle (henceforth IC) and the remaining Outer Circle varieties (henceforth OC). Testing of SPK marks the large jumps for three clusters as significant, separating the IC varieties from EA+IND and all other varieties. Segmentation at significant above-average jump height k=5 additionally identifies a Southeast Asian SIN+HK+PHI group, JA+SL, and isolates NIG. Within WRT, the first significant jump is found at k=6, resulting in a combined IC cluster and several smaller and mostly proximity-based groups (except for UG merging with West African varieties).

The NeighborNets in Figure 5.5 mostly support the groups identified within the hierarchical analysis, but many boxy shapes also indicate ambiguous structures mirroring the low substantiation levels found by `pvclust`. IC (+JA) is established in SPK, while WRT indicates two related IC groups but also similarity of GB and HK (which itself is relatively similar to SIN) and furthermore shows some proximity within the African group. The other OC groups are less clearly supported.

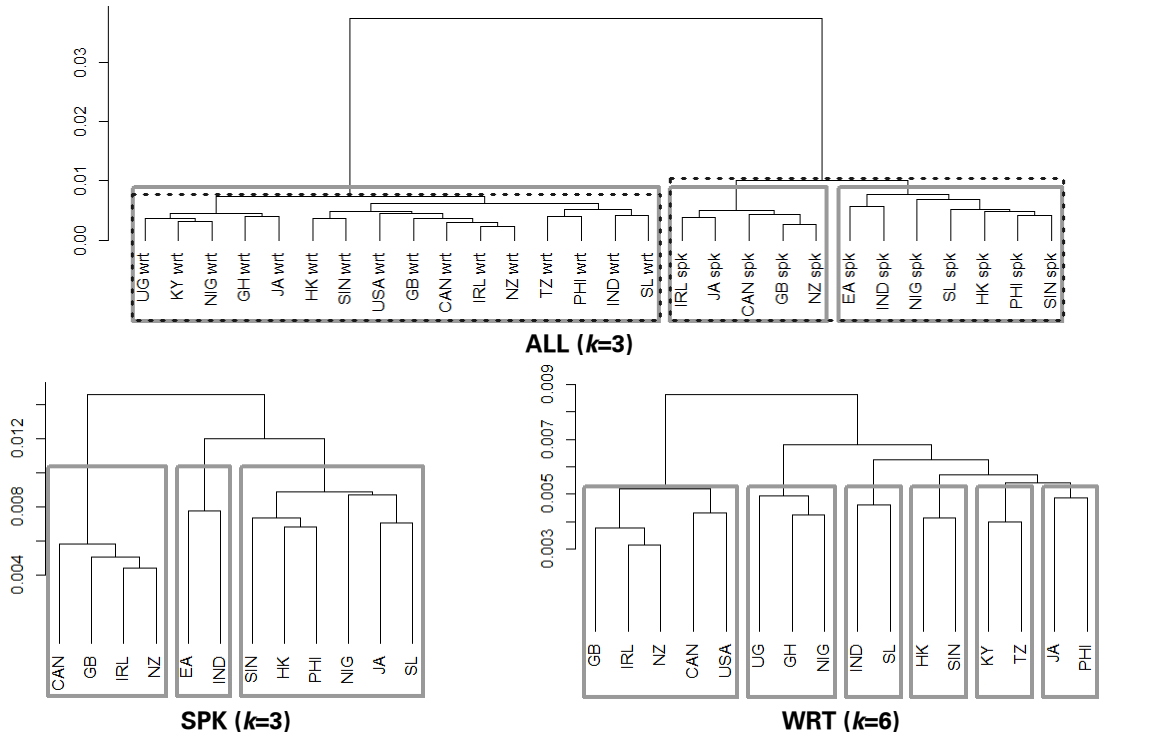


Figure 5.3: Hierarchical clustering results for lexical *MI n*-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

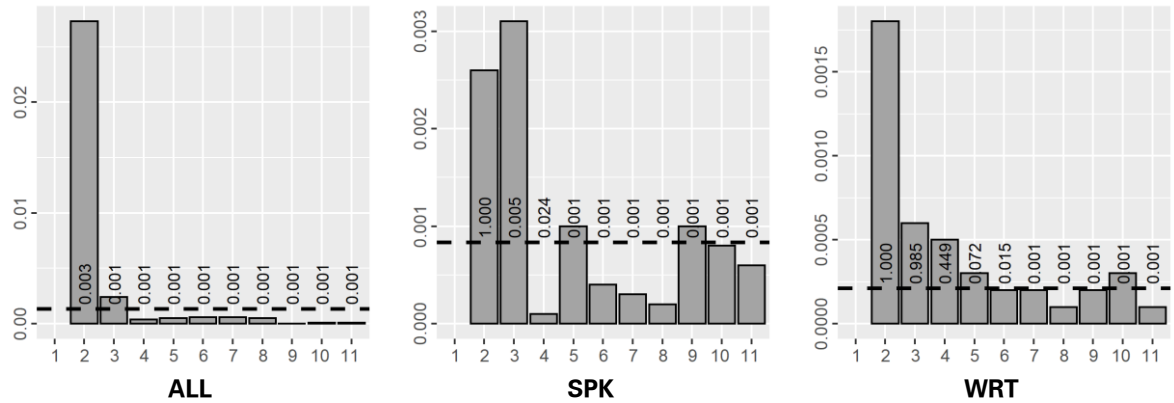


Figure 5.4: Jumps in node heights and respective *p*-values for lexical *MI n*-grams

Employing k-means to enforce a separation into *k* clusters at the intersection between explained and within-cluster variation (Figure 5.6; for clustering results cf. Table 5.7) indicates somewhat finer structures than the previous methods: While a large 'elbow' can be detected at *k*=2 for ALL, the intersect is found for a 4-cluster solution,

splitting the spoken and written branches largely into the Inner and Outer Circle (but cf. HK_{WRT} and JA_{SPK}). For SPK, overlap occurs between *k*=4 and *k*=5, reinstating the IC (+JA) group (i.e. IC, sometimes plus JA) in addition to smaller regional groups and further unary nodes. WRT favors even finer clusters than before (*k*=7), resulting in an IC cluster plus fine regional groups (except for JA+PHI).

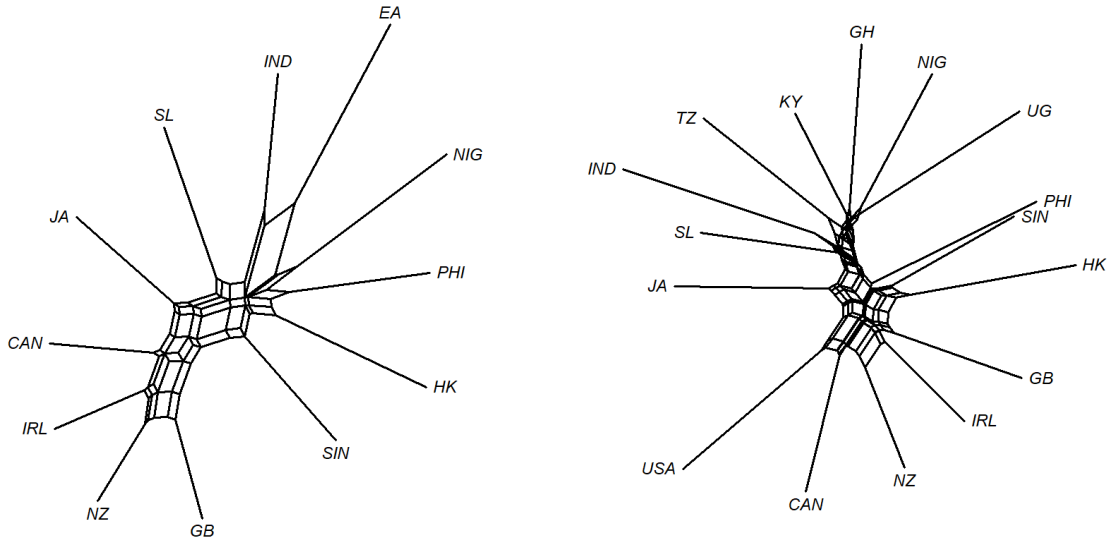


Figure 5.5: NeighborNets of the spoken and written data for lexical *MI n*-grams

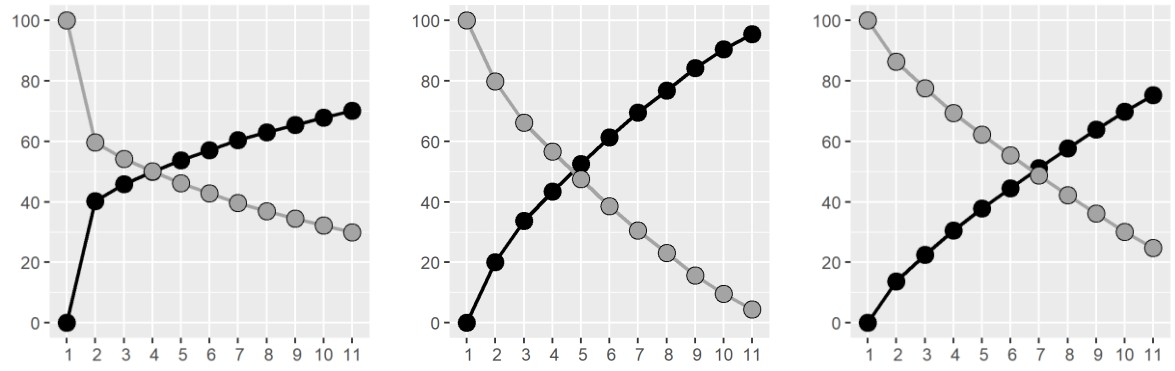


Figure 5.6: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *MI n*-grams

Table 5.7: K-means clustering results for specific values of *k* for lexical *MI n*-grams

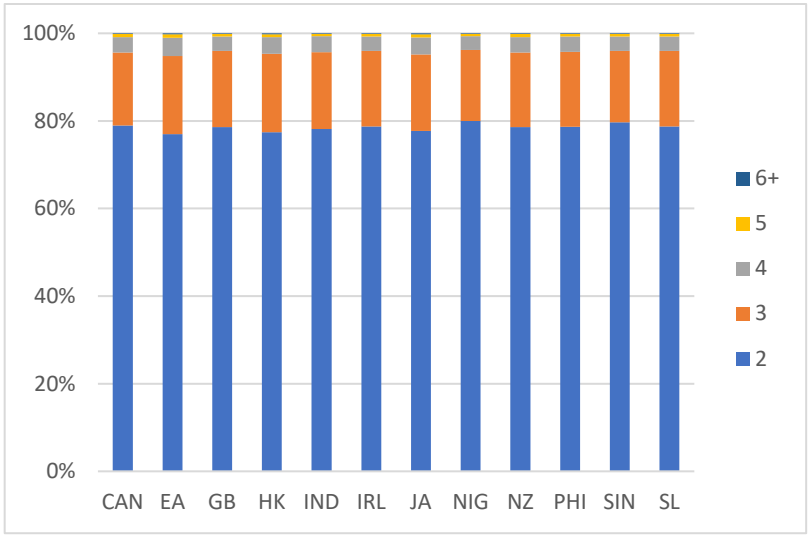
	ALL (*k*=4)		**SPK (*k*=5)**		**WRT (*k*=7)**
1	CAN, GB, HK, IRL, NZ, SIN, USA_{WRT}	1	CAN, GB, IRL, JA, NZ	1	CAN, GB, IRL, NZ
2	KY, TZ, GH, IND, JA, NIG, PHI, SL, UG_{WRT}	2	HK, PHI, SIN	2	HK, SIN
3	CAN, GB, IRL, JA, NZ_{SPK}	3	EA	3	KY, TZ
4	EA, HK, IND, NIG, PHI, SIN, SL_{SPK}	4	NIG	4	GH, NIG, UG
		5	IND, SL	5	IND, SL
				6	USA
				7	JA, PHI

5.1.2 T-score

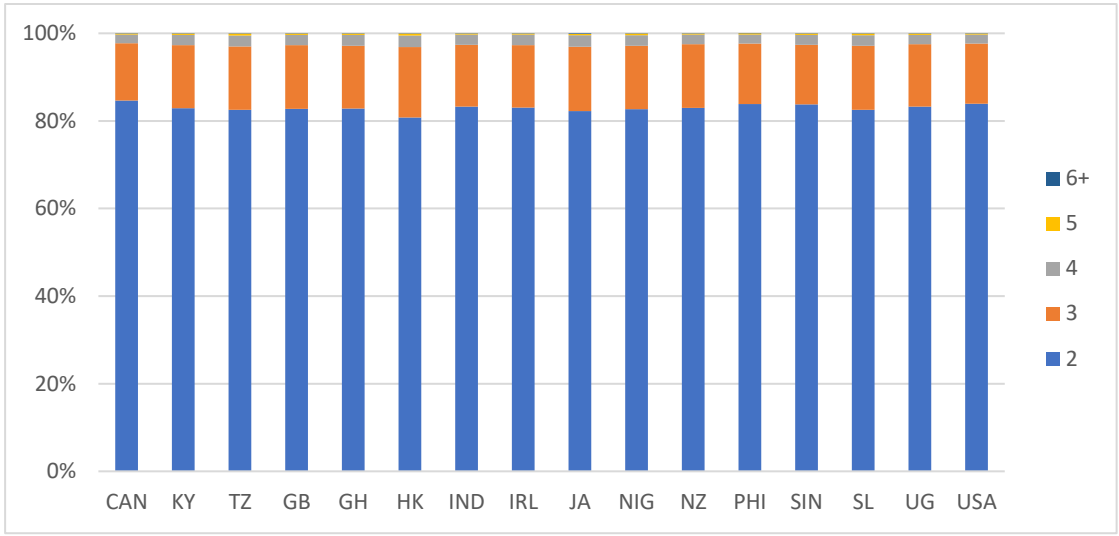
Merging of bigrams on the basis of the t-score produced an average of 180,464 (*s*=12,490) spoken and 84,779 (*s*=5,697) *n*-grams of 19,968 (*s*=1,642) and 8,955 (*s*=714) types. These frequencies constitute slight reductions in token counts by (-21% and -17%) but strong increases in token frequencies (+250% and +151%), greatly normalizing the large TTRs found for post-threshold bigrams to more normal *n*-gram TTRs of 9.1 (*s*=0.5) and 9.5 (*s*=0.2). Mean token lengths lie at 2.27 (*s*=0.01) and 2.20 (*s*=0.01) but type lengths are considerably higher at 3.06 (*s*=0.03) and 2.81 (*s*=0.03), indicating relatively fewer new tokens generated through bigram merging, but a more sizeable proportion of longer types. The lower threshold value for *t* than *MI* thus appears to have led to a larger amount of consecutive bigrams to remain in the data for consecutive merging. The distributions within Figure 5.7 correspondingly lay open the drastic divergences between token and type frequencies of various lengths: While token lengths are only barely constituted of sequences of $n \geq 4$, lengths shift upwards by one item for types, making 3- and 4-grams figure prominently. Noticeable differences between the two modes concern the higher frequencies of longer tokens and particularly types in the spoken data, which may correspond to *t*'s preference for frequent fixed sequences.

Combining the varietal *n*-gram data reveals a relatively large frequency of mutually shared items, resulting from the larger number of items in each varietal dataset. Some outliers can be observed towards the positive ends of the scales (NZ+CAN, NZ+IRL and NZ+GB in the spoken and GB/IRL and GB/NZ in the written data), but no negative outliers are detected which might cause excessive deletion of shared *n*-grams (cf. Figure 5.8). Reductions through merging of the data still accrue -92% and -90% losses in types (retaining 1,521 and 919 items) compared to variety averages.

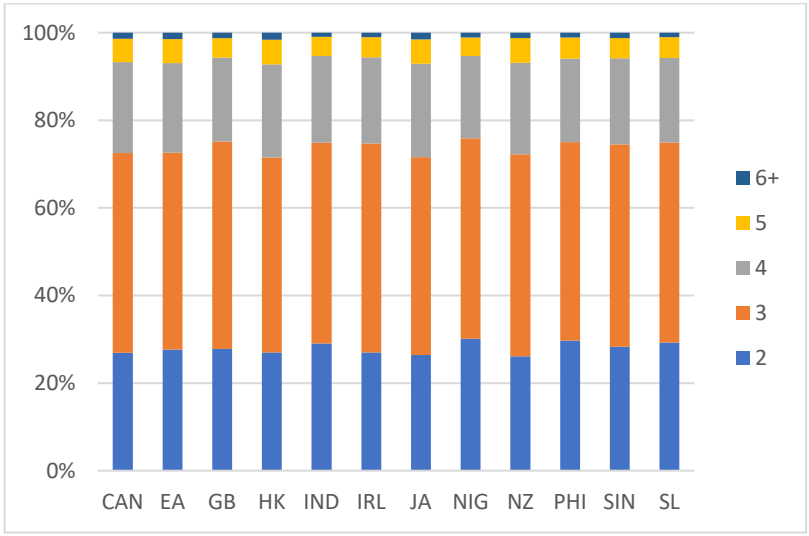
The relatively high frequencies of longer types in the separate varietal datasets leads to a retention of at least some items of lengths greater than 3 units, particularly within the spoken data (Table 5.8). Table 5.9 shows that, as may be expected, preferred t-score-based *n*-grams largely result from high-frequency items. This is underscored by the fact that even some longer sequences of high-frequency items can be found among those with the highest association scores.



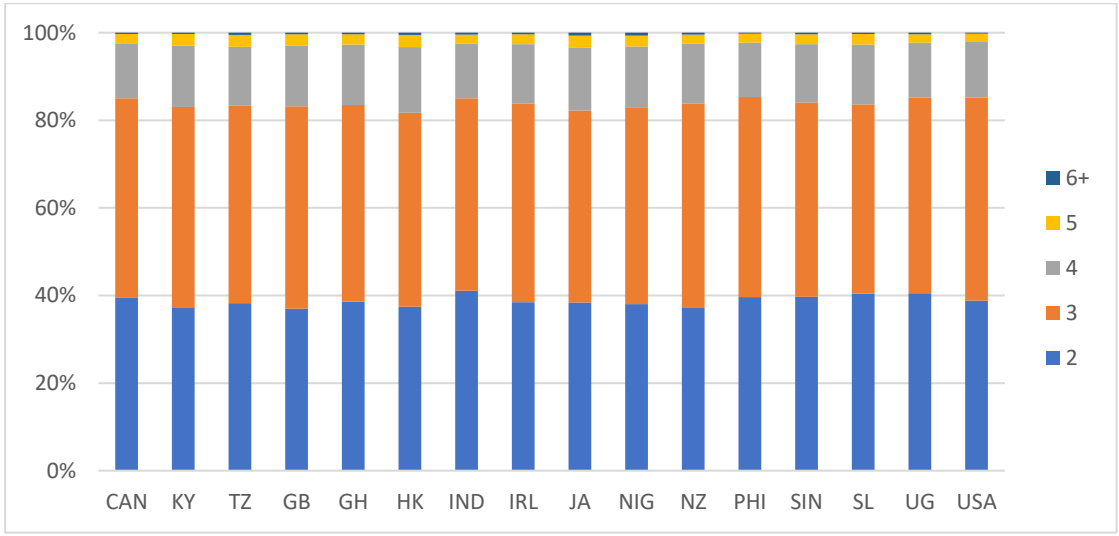
Token frequencies: Spoken data



Token frequencies: Written data



Type frequencies: Spoken data



Type frequencies: Written data

Figure 5.7: Distribution of for lexical *t n*-gram lengths across the varietal datasets

Figure 5.8: Number of shared lexical *t n*-grams between any two datasets

Table 5.8: Lexical *t n*-gram type frequencies by length in the intersects of the variety-specific datasets

***N*-gram length**	**ALL**	**SPK**	**WRT**
2	518 (85%)	1,104 (73%)	710 (77%)
3	88 (14%)	398 (26%)	205 (22%)
4	3 (1%)	19 (1%)	4 (1%)
total	609	1,521	919

Table 5.9: Lexical *t n*-grams with highest and lowest association scores

SPK		**WRT**	
***n*-gram**	***t***	***n*-gram**	***t***
of the	44.00	of the	41.61
you know	43.13	in the	34.50
in the	40.50	one of the	26.95
i think	35.55	part of the	26.42
if you know	33.18	out of the	26.11
do you know	32.62	use of the	25.69
said it	3.49	a letter	3.48
depend on	3.46	not necessarily	3.45
the results	3.44	derived from	3.44
made up	3.39	they did	3.43
doing this	3.39	into account	3.36
gone through	3.33	told him	3.26

Within the context of the hierarchical cluster analysis (Figure 5.9), *t*-based *n*-grams demonstrate a relative propensity for generating stable clusters: In all three datasets, numerous smaller clusters hold true after bootstrapping using `pvclust` (black dotted lines), highlighting small regional similarities in addition to the IC groups (but contrast the IND+EA spoken group). Moreover, larger clusters are only found in SPK, while WRT only finds stable binary groups (ALL additionally returning the written branch overall as stable).

Jumps in node heights (Figure 5.10) produce less fine-grained distinctions. ALL supports up to *k*=4, which splits off the IC varieties in speech first and then EA+IND. SPK prefers *k*=3, resulting in an IC cluster, EA+IND and the remaining varieties, partitioning the latter into HK+SIN, NIG and JA+PHI+SL at *k*=5. The first significant jump at *k*=6 in WRT resembles a regional distinction (two African clusters and a northern American group). It also somewhat corresponds to phases, with several stage 4 varieties within one cluster and the most institutionalized SIN clustered with the IC varieties.

NeighborNet clustering (Figure 5.11) particularly supports the spoken IC and HK+SIN clusters and finds NIG sitting uneasily between IND+EA and the rest of the OC group. In WRT, the two African clusters can be identified, which jointly emerge at some distance from all other varieties, but IND is placed between this larger cluster and the remaining OC varieties. The IC varieties are found at large distances from all other varieties, but HK+SIN sits in an intermediate position and a North-American branch splits off from the larger IC group.

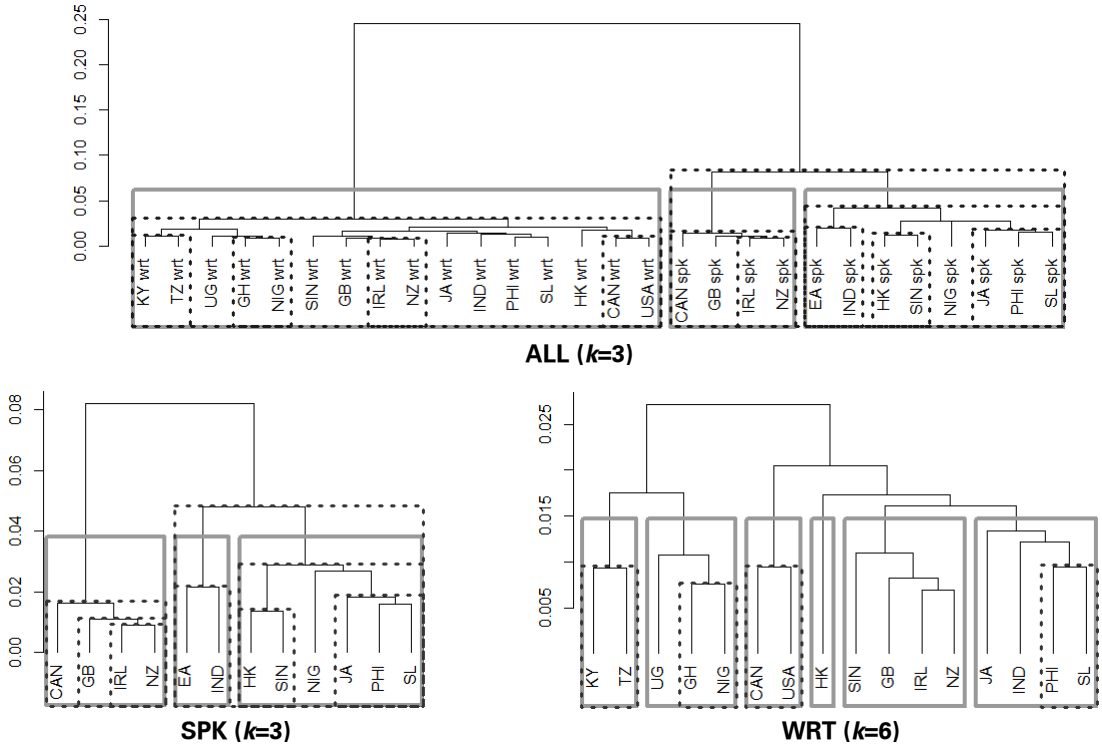


Figure 5.9: Hierarchical clustering results for lexical *t n*-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

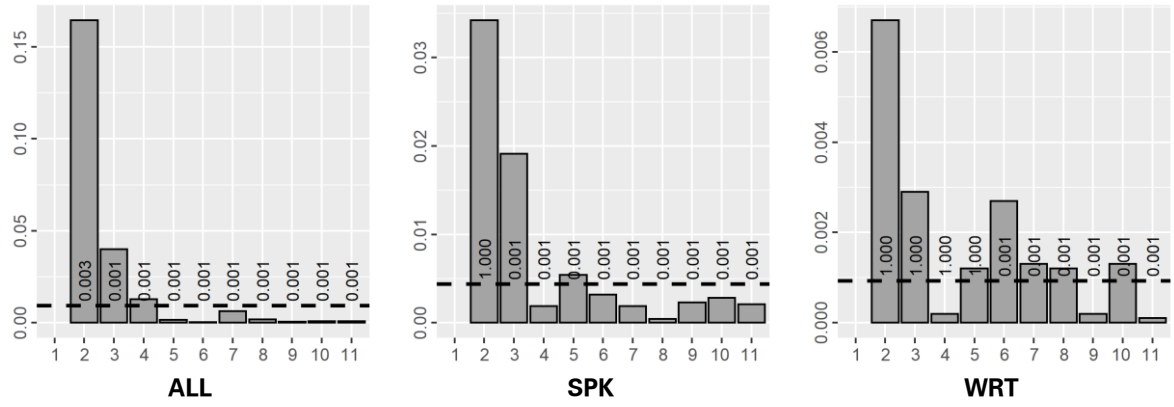
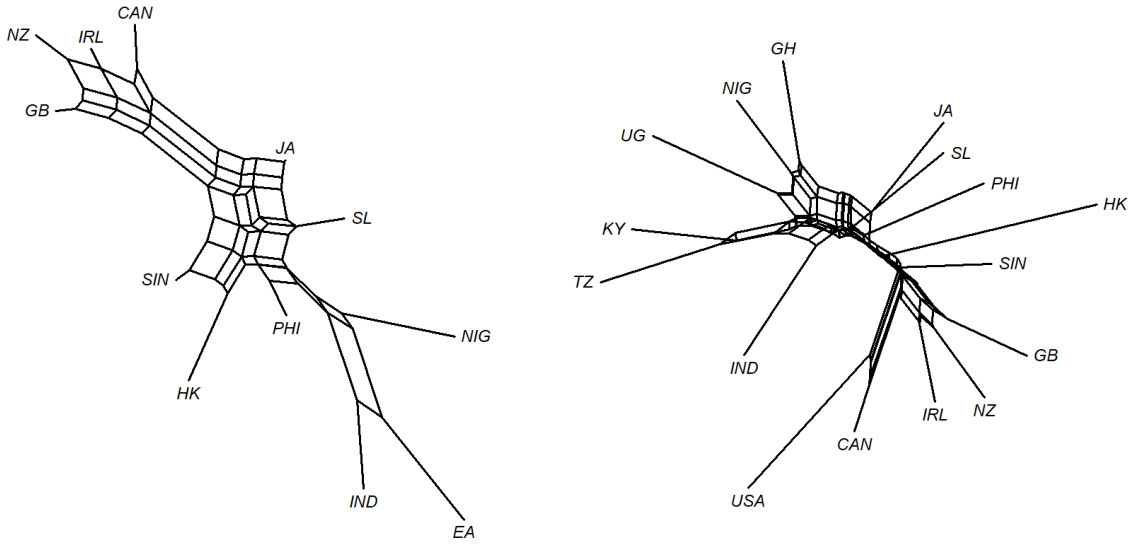


Figure 5.10: Jumps in node heights and respective *p*-values lexical *t n*-grams

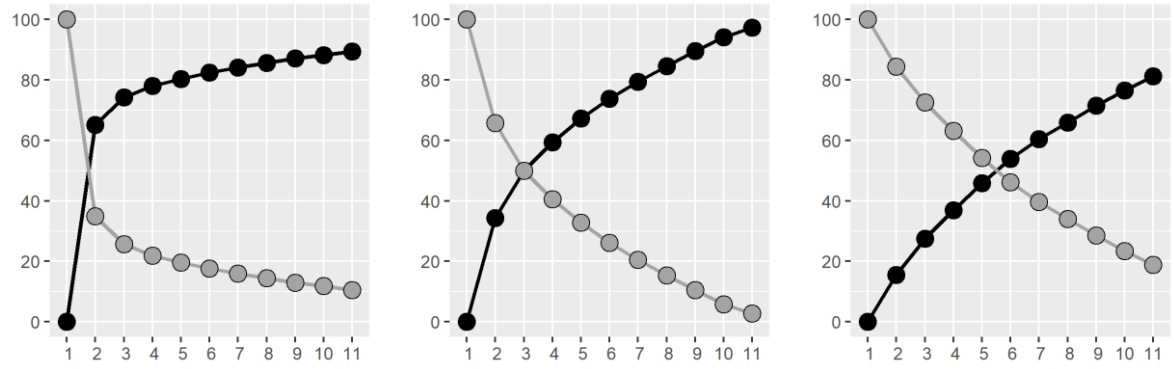
K-means variances (Figure 5.12) show a clear preference for *k*=2 or maximally *k*=3 (judging from the 'elbow') in ALL, as well as confirming three spoken and five or six written groups, leading to segmentations as shown in Table 5.10. These perfectly

reproduce the HCA results except for allocating SIN to a mostly Asian cluster for *k*=6 in WRT (*k*=5 would conflate this group with GH+NIG+UG).



SPK **WRT**

Figure 5.11: NeighborNets of the spoken and written data lexical *t n*-grams



ALL **SPK** **WRT**

Figure 5.12: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters lexical *t n*-grams

Table 5.10: K-means clustering results for specific values of *k* lexical *t n*-grams

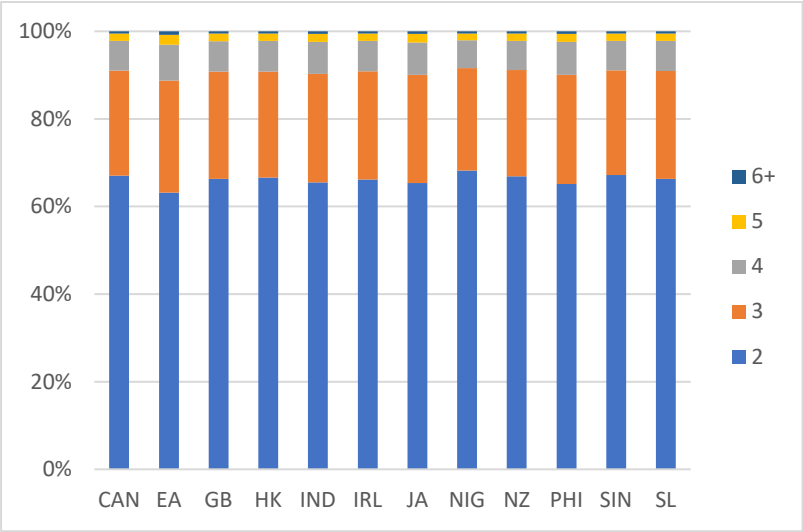
	ALL (*k*=3)		**SPK (*k*=3)**		**WRT (*k*=6)**
1	EA, HK, IND, JA, NIG, PHI, SIN, SL $_{SPK}$	1	HK, JA, NIG, PHI, SIN, SL	1	GH, NIG, UG
2	CAN, GB, IRL, NZ $_{SPK}$	2	CAN, GB, IRL, NZ	2	IND, JA, PHI, SIN, SL
3	All written corpus parts	3	EA, IND	3	CAN, USA
				4	GB, IRL, NZ
				5	KY, TZ
				6	HK

5.1.3 Log Likelihood

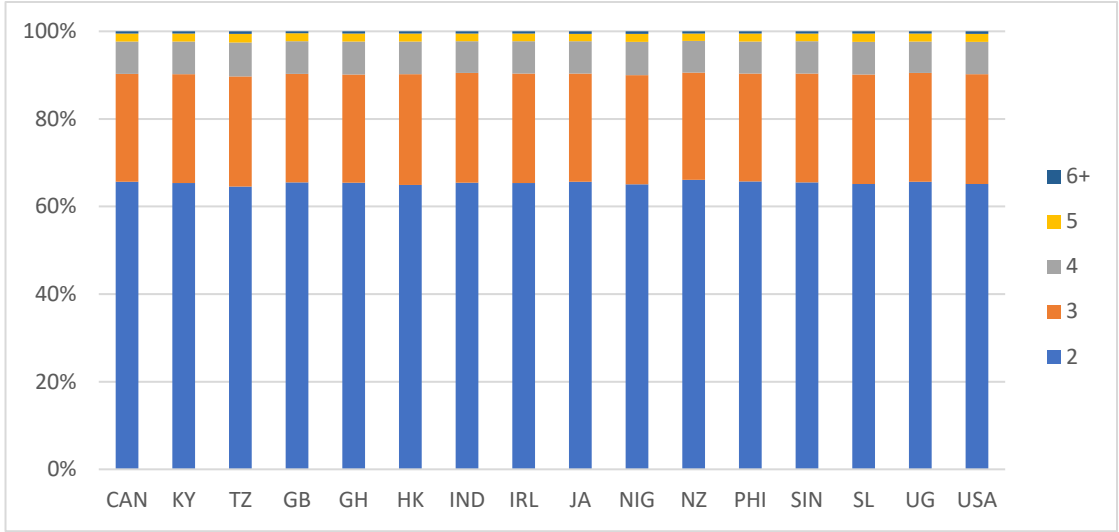
On the basis of the second-largest number of bigrams retained after the application of threshold values, G^2 also produced the largest number of *n*-gram tokens and types. On average, 271,159 (*s*=15,263) and 176,468 (*s*=8,960) tokens were generated in speech and writing respectively, which were constituted from 115,796 (*s*=6,688) and 101,309 (*s*=5,800) types. This represents a reduction of token counts by -32% in both modes, a slight increase in spoken type frequencies (+12%) and actually a decrease in written types (-4%). Average TTRs were found at 2.3 (*s*=0.1) and 1.7 (*s*=0.1), which again represent the second-lowest of all lexical *n*-grams. Mean token lengths are virtually identical between modes (2.46, *s*=0.02 and 2.47, *s*=0.01), but the average type is marginally longer in speech (2.81, *s*=0.03) than in writing (2.69, *s*=0.02). Figure 5.13 reveals some preference for longer types (not quite as pronounced as in case of *t*), which also come out in noticeable frequencies in the token data. G^2 thus presents itself as the traditional measure with the highest propensity for generating longer sequences. It should be noted, however, that this may be more a result of the less impactful threshold values imposed on bigram association scores than a feature of the measure itself.

Figure 5.14 reveals average shared item numbers far above the other measures but not proportional to the size of the G^2 data (only c. 11% average overlap in contrast to c. 30% of other measures). While most types are lost during the merging of individual varietal datasets (-98% and -99% to absolute numbers of 2,534 and 1,431 items), Figure 5.14 demonstrates that again none of the mergers unduly affect overall item frequencies.

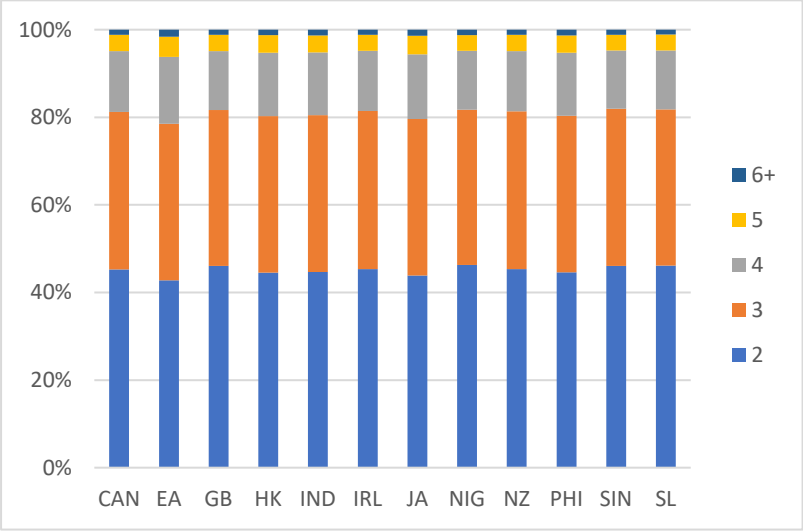
The relatively large numbers of both variety-specific as well as mutually shared items do not, however, translate into a retention of particularly many longer sequences in the shared datasets (Table 5.11), at least in relative comparison to the other association measures. This may be a result of very heterogenous *n*-grams produced by G^2 on the basis of too lax association thresholds. Strangely, Table 5.12 shows that even some lower-scoring items appear to represent relatively collocated forms, *what are* or *are getting*. However, it also highlights that the top association scores for G^2 may be highly inflated for some high-frequency items.



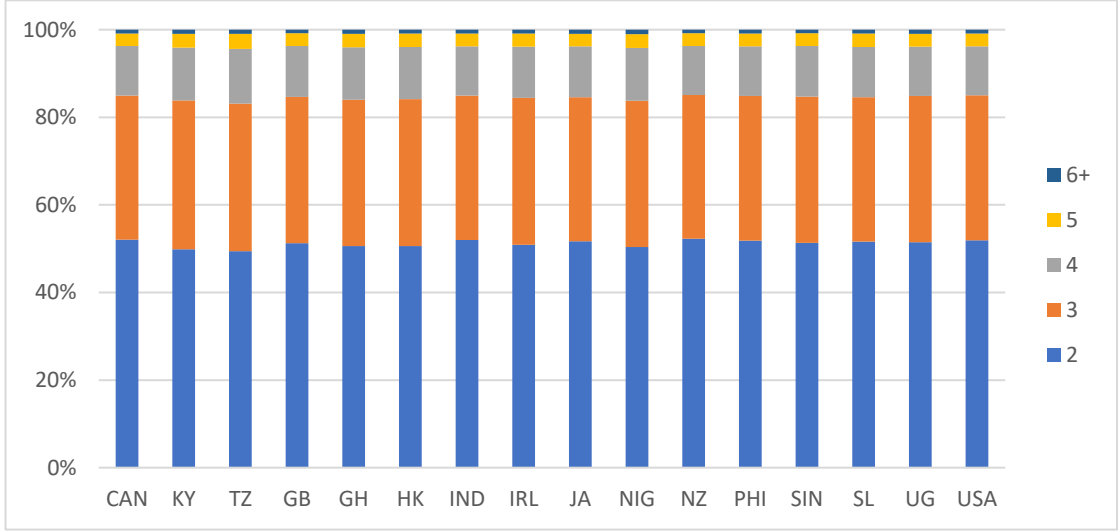
Token frequencies: Spoken data



Token frequencies: Written data



Type frequencies: Spoken data



Type frequencies: Written data

Figure 5.13: Distribution of lexical *G*² *n*-gram lengths across the varietal datasets

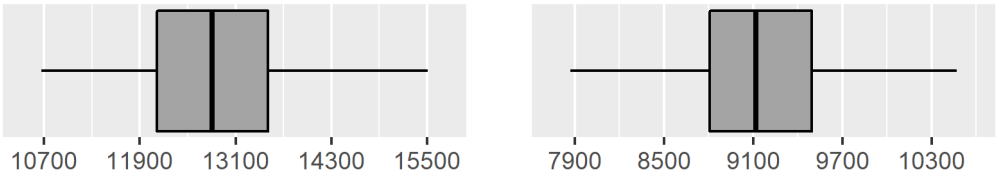


Figure 5.14: Number of shared lexical G^2 *n*-grams between any two datasets

Table 5.11: Lexical G^2 *n*-gram type frequencies by length in the intersects of the variety-specific datasets

N-gram length	ALL	SPK	WRT
2	894 (90%)	2,030 (80%)	1,242 (87%)
3	103 (10%)	486 (19%)	184 (13%)
4	1 (0%)	18 (1%)	5 (0%)
total	998	2,534	1,431

Table 5.12: Lexical G^2 *n*-grams with highest and lowest association scores

SPK		WRT	
n-gram	G^2	*n*-gram	G^2
you know	16998.33	of the	6298.86
i think	11949.03	it is	4396.09
if you know	10222.06	in the	4371.90
do you know	10008.67	part of the	3624.14
did you know	9105.95	one of the	3473.03
then you know	8650.69	use of the	3413.81
in both	22.63	on this	23.78
all our	22.34	what are	23.18
the event	22.01	their work	23.17
different countries	21.88	are getting	22.94
take away	21.79	of people	20.92
say this	19.52	others are	19.04

Analysis using hierarchical clustering (Figure 5.15) shows relatively few stable clusters given the large number of *n*-grams within the present data. Stability of the spoken branch fails to materialize in ALL and only EA+IND$_{SPK}$ is found stable in both types of spoken data. Writing only detects stable clusters in JA GH+NIG in both datasets, and additionally finds KY+TZ in WRT.

Significant jumps in node heights (Figure 5.16) favor the usual binary separation for ALL, but splits up to *k*=4 display jumps above average, entailing further segmentation of the spoken branch into IC, EA+IND and the remaining varieties. The same groups would be retrieved in SPK at *k*=3, but only *k*=5 achieves significance, resulting in a segmentation not easily analyzable under an areal or evolutionary perspective. WRT present similar issues, and the only complete cluster analyzable under any given hypothesis can be found in the USA+CAN+PHI group, i.e. a case of potential AmE epicentral influences.

NeighborNet clusters (Figure 5.17) more strongly identify a joint spoken IC cluster and additionally retrieve HK+SIN as well as isolation of EA+IND. In WRT, IC varieties separate from the remaining ones together with SIN and PHI but also form discrete groups. In the latter case, PHI tends more strongly towards USA+CAN than SIN to GB+IRL+NZ. The two African written clusters found in the HCA (with some similarity to IND and JA) are also supported in WRT.

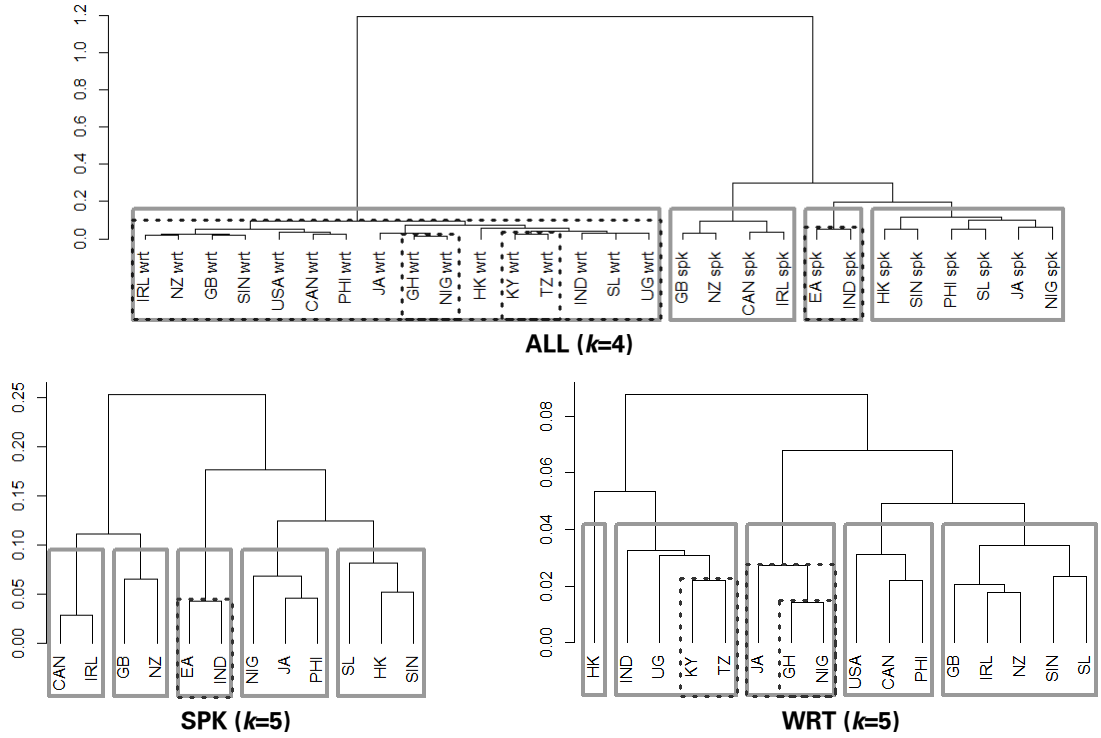


Figure 5.15: Hierarchical clustering results for lexical G^2 *n*-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

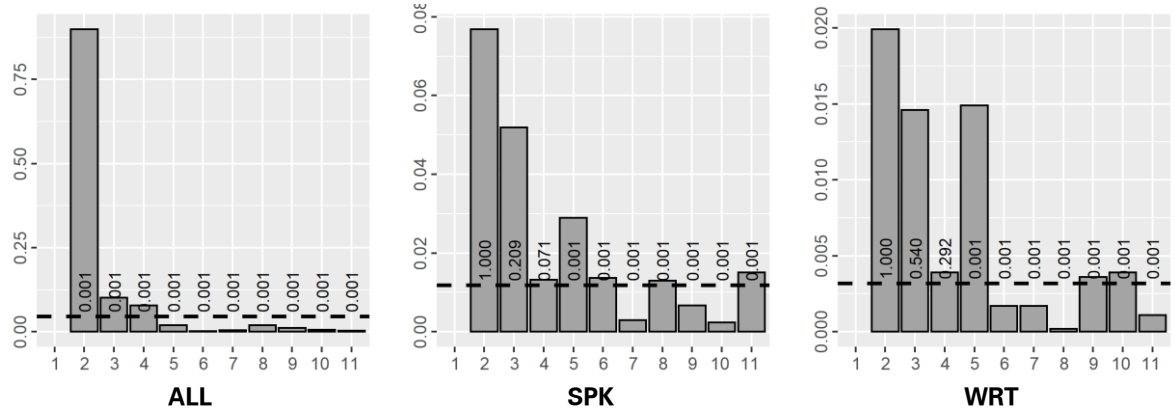
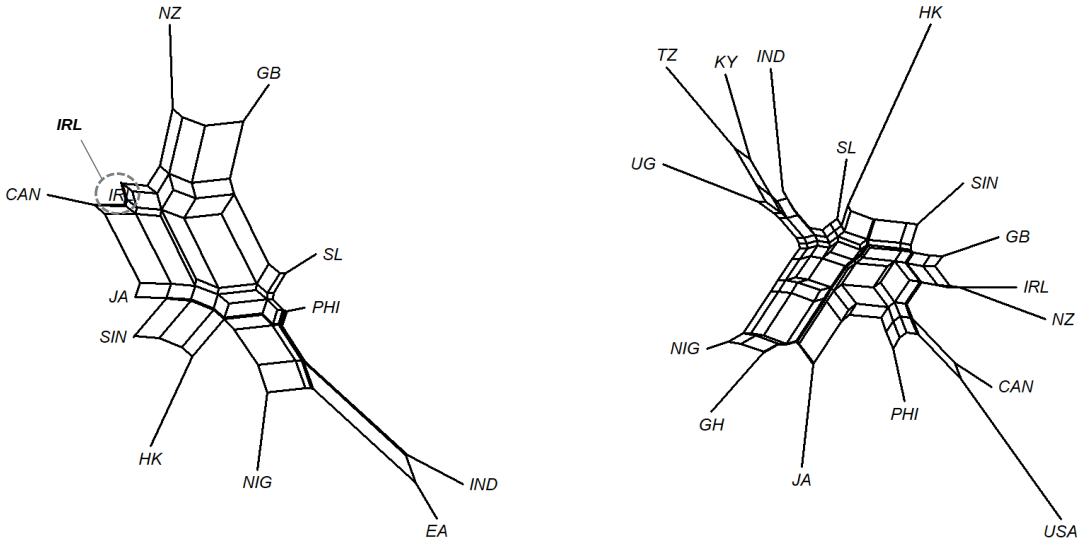


Figure 5.16: Jumps in node heights and respective *p*-values for lexical G^2 *n*-grams

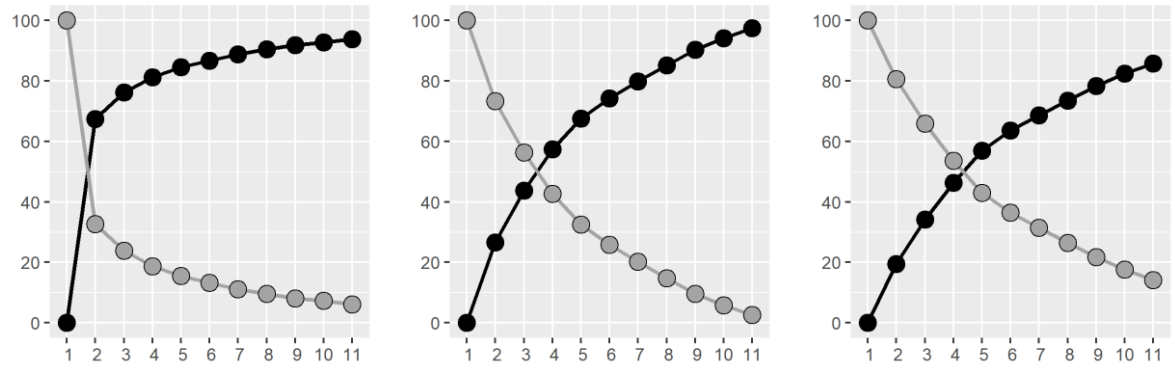
Turning to k-means variances (Figure 5.18), the preferred *k*=2 for ALL leads to the spoken-written distinction, which at *k*=3 is subdivided into Inner and Outer Circle spoken varieties (Table 5.13). SPK indicates good values for *k* at 3 and 4. While *k*=3 retrieves the IC group and distinguishes HK+SIN from the remaining varieties, *k*=4

produces less meaningful clusters by removing SIN from HK and merging it CAN+IRL (+JA) while isolating GB+NZ. Writing indicates 4 or 5 clusters, which only diverge in allocating HK either to the GH+JA+NIG cluster or to a separate unary node. Again, the results present challenges for a coherent explanation, casting G^2 as a more problematic measure under the present methodology. It will have to be explored how the present results relate to static-length *n*-grams in later sections.



SPK **WRT**

Figure 5.17: NeighborNets of the spoken and written data for lexical G^2 *n*-grams



ALL **SPK** **WRT**

Figure 5.18: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical G^2 *n*-grams

Table 5.13: K-means clustering results for specific values of *k* for lexical G^2 *n*-grams

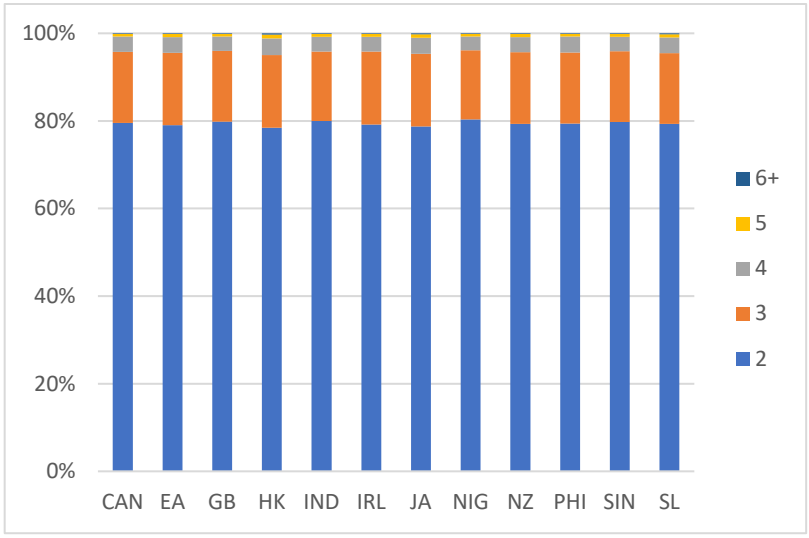
	ALL (*k*=3)		**SPK (*k*=4)**		**WRT (*k*=5)**
1	All written corpus parts	1	EA, IND, JA, NIG, PHI, SL	1	GH, JA, NIG
2	CAN, GB, IRL, NZ $_{SPK}$	2	HK, SIN	2	HK
3	EA, HK, IND, JA, NIG, PHI, SIN, SL $_{SPK}$	3	CAN, GB, IRL, NZ	3	GB, IRL, NZ, PHI, SIN
				4	CAN, USA
				5	KY, TZ, IND, SL, UG

5.1.4 Lexical Gravity

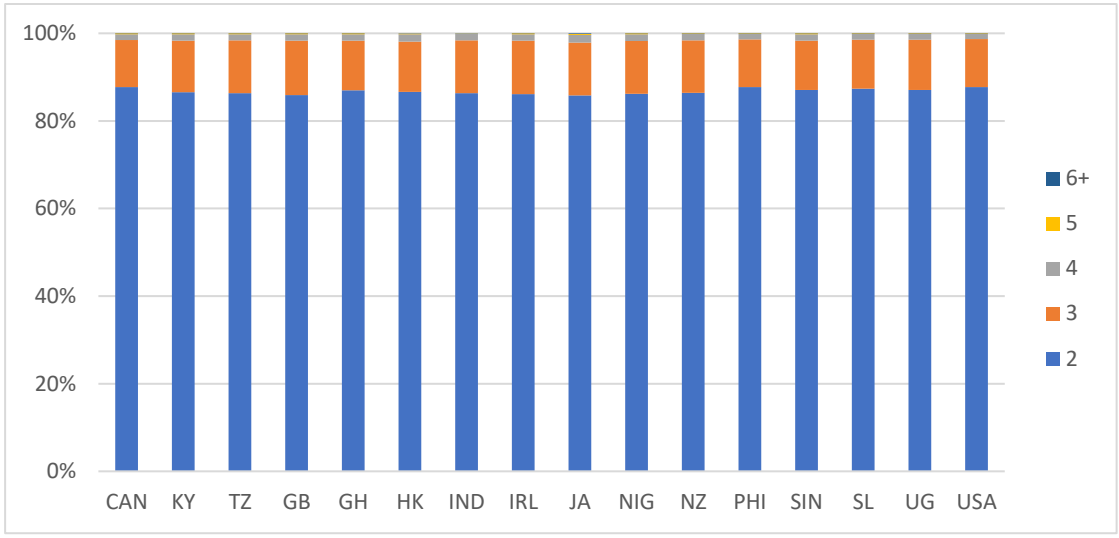
N-grams derived on the basis of the lexical-gravity measure yielded mean token frequencies of 149,000 (*s*=11,496) and 65,457 (*s*=4,407) for speech and writing, respectively, with type frequencies of 12,043 (*s*=1,076) and 3,851 (*s*=287), representing a loss of overall tokens by -21% and -13% but a drastic increase over bigram types by 538% and 235%, i.e. particularly in speech. While TTR thus greatly dropped in contrast to the post-threshold bigram data, average values are still the highest of all measures at 12.4 (*s*=0.5) and 17.0 (*s*=0.6). Average token lengths were almost identical for speech (2.27, *s*=0.01) and writing (2.20, *s*=0.01) while types are somewhat longer in speech (given the larger extent of generation of longer sequences observed above), with values of 3.01 (*s*=0.03) and 2.81 (*s*=0.03). In terms of the distribution of *n*-gram lengths (Figure 5.19), *g* presents similar results to the previously-discussed *t* but with even more longer sequences for the type data, particularly in speech. Tokens are almost entirely restricted to sequences up to length 3, which corroborates the findings from all measures but G^2. While 4-grams are an exception in the token data, 5-grams still constitute a sizeable proportion of the type data, particularly in speech.

Again, while merging the data, only positive outliers are discovered (Figure 5.20). These concern the mergers of the written datasets for IRL+NZ, IRL+GB and NZ+GB. Even though relative reductions in type frequencies are (if not by much) the lowest in contrast to other measures (-92% and -86%), absolute sizes of the merged datasets are at the second-lowest rank overall (976 and 521 types remaining).

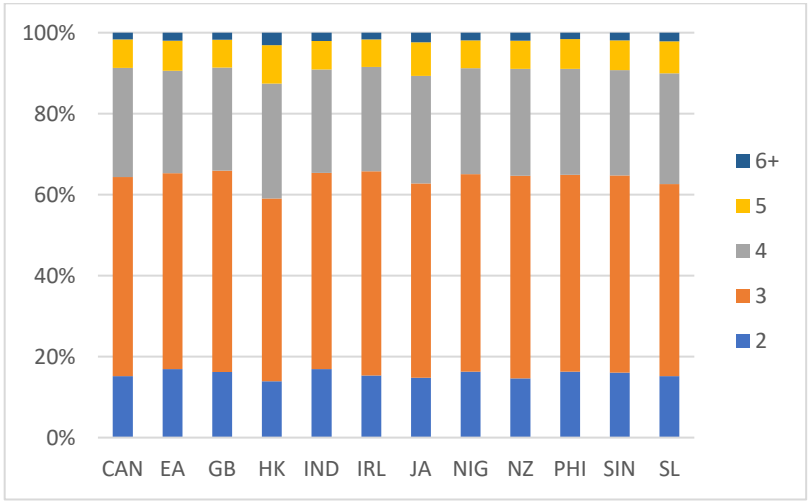
Lexical gravity retains a surprisingly large number of longer sequences among the retained *n*-grams (Table 5.14). While this is more of a relative effect for ALL and WRT, it also extends to the second-largest absolute numbers of all measures for SPK, making the small set of bigrams comparatively successful in generating longer sequences. This is corroborated by the second-lowest relative frequencies of retained 2-grams (yielding to Delta P). Table 5.15 shows that *g* appears to assign quite consistently high scores to the same types even across different modes. Top-scoring sequences also appear to rest on high-frequency items. This, however, can even be said to some extent for the lowest-ranking collocates, indicating that surpassing the high association threshold is largely restricted combinations of at least moderately high frequencies.



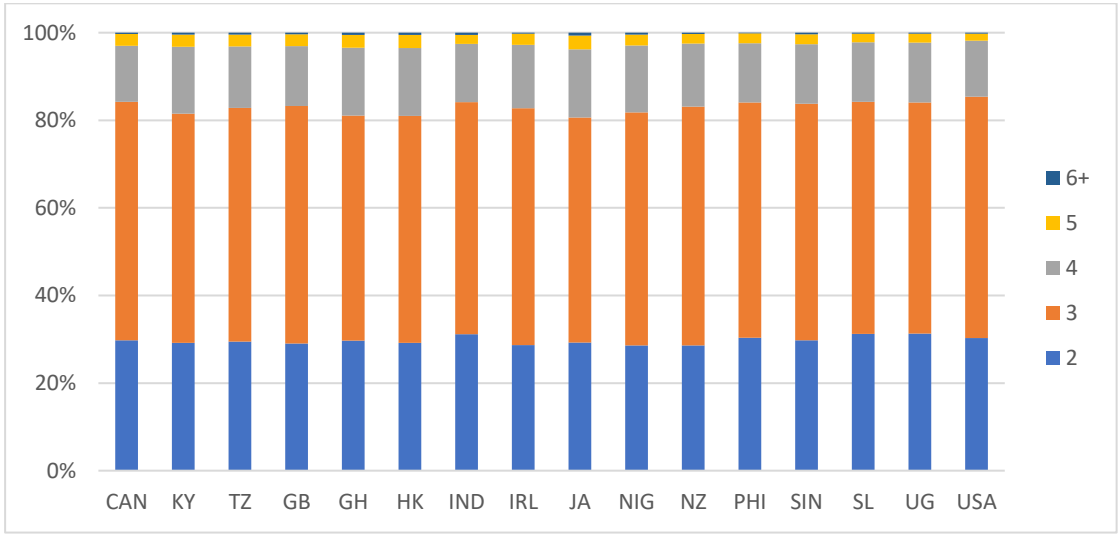
Token frequencies: Spoken data



Token frequencies: Written data



Type frequencies: Spoken data



Type frequencies: Written data

Figure 5.19: Distribution of lexical *g n*-gram lengths across the varietal datasets

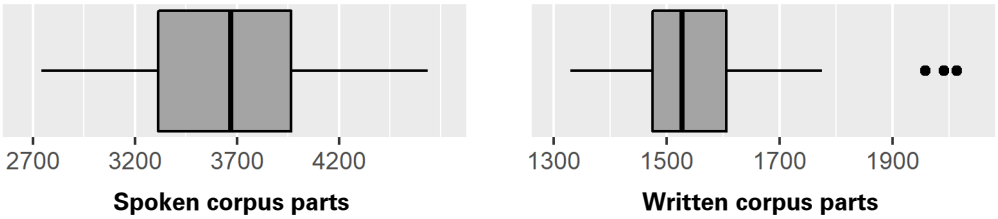


Figure 5.20: Number of shared lexical *g* *n*-grams between any two datasets

Table 5.14: Lexical *g* *n*-gram type frequencies by length in the intersects of the variety-specific datasets

***N*-gram length**	**ALL**	**SPK**	**WRT**
2	279 (79%)	591 (61%)	379 (73%)
3	74 (21%)	350 (36%)	137 (26%)
4	2 (1%)	35 (4%)	5 (1%)
total	355	976	521

Table 5.15: Lexical *g* *n*-grams with highest and lowest association scores

SPK		**WRT**	
***n*-gram**	***g***	***n*-gram**	***g***
of the	17.54	of the	18.05
in the	16.40	in the	16.51
one of the	15.07	one of the	15.04
some of the	14.90	to the	14.99
to the	14.64	it is	14.63
this is	14.52	some of the	14.54
way to	6.51	at home	6.55
three years	6.49	is being	6.54
to show	6.46	for his	6.53
now that	6.38	in any	6.51
should have	6.38	to look	6.48
for all	6.37	responsible for	6.45

Given the small size of the *g*-based data, HCA finds surprising many substantiated clusters (Figure 5.21). These not only support ALL's spoken-written distinction but also several clusters therein as well as in the separate sets. The data indicate one IC cluster in speech but two clusters distinguished by region in writing (North America vs. a rather British-epicentral one, henceforth IC_{NA} and IC_{GB}). Both types of spoken data support the HK+SIN and JA+PHI+SL(+NIG) groups, while ALL's spoken branch furthermore retrieves EA+IND and WRT supports the NIG+GH+UG cluster. The second African written group KY+TZ only barely misses strong support at AU=94.[76]

Significant jumps (Figure 5.22) most prominently indicate *k*=3 for ALL, subdividing the spoken branch into IC and OC. Larger values first identify a written IC cluster before separating HK+SIN in speech and then Africa in writing at *k*=6. For SPK, *k*=3 separates IC from two (barely found insignificant) OC clusters, while further significant

[76] Recall that AU scores provide the strictest form of stability assessment in `pvclust`. As such, AU=94 is still informative, indicating support within ≥9,400 out of 10,000 permutations.

jumps first isolate HK+SIN and then EA, IND and NIG. For WRT, clusters at *k*=6 present the first significant jumps, overlapping closely with the stable clusters above and dividing the data into two IC groups, two African clusters (but with UG being assigned to West Africa, as otherwise common) and a remaining OC group from which HK (the variety with the 'lowest' evolutionary stage) is split off.

NeighborNet analysis (Figure 5.23) retrieves spoken IC, isolates EA+IND (and NIG to a lesser extent) and somewhat supports HK+SIN. WRT identifies the IC and African clusters but also shows similarity within each larger group, and also indicates HK+SIN.

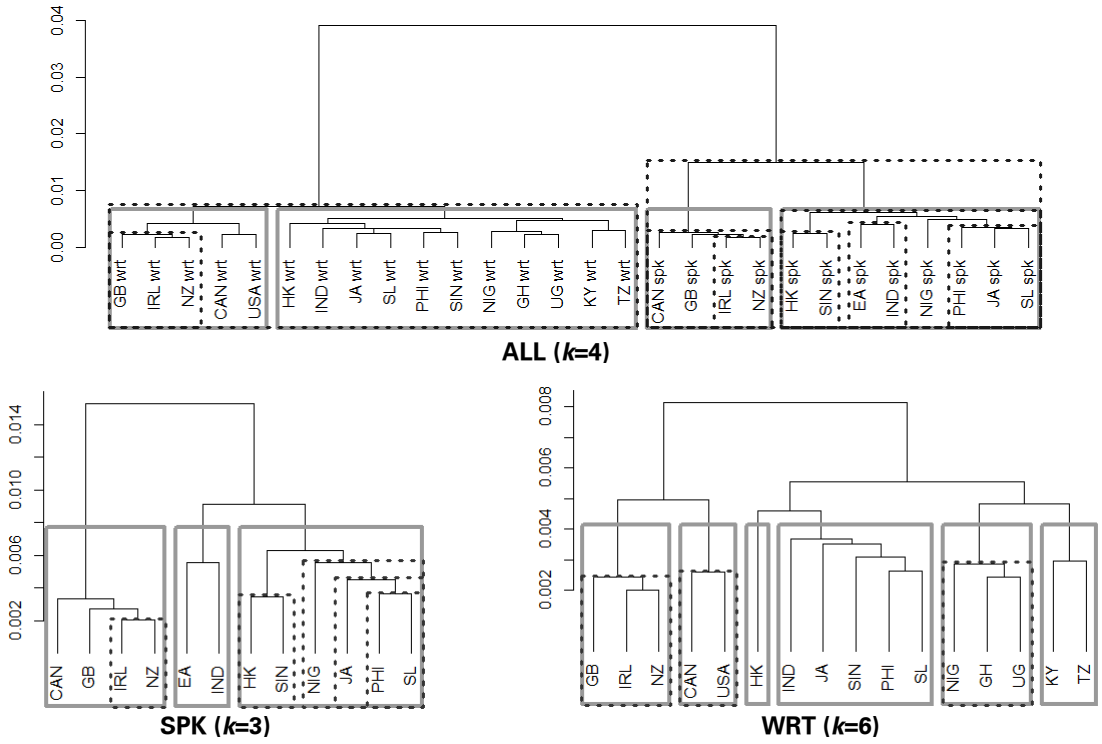


Figure 5.21: Hierarchical clustering results for lexical *g* *n*-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

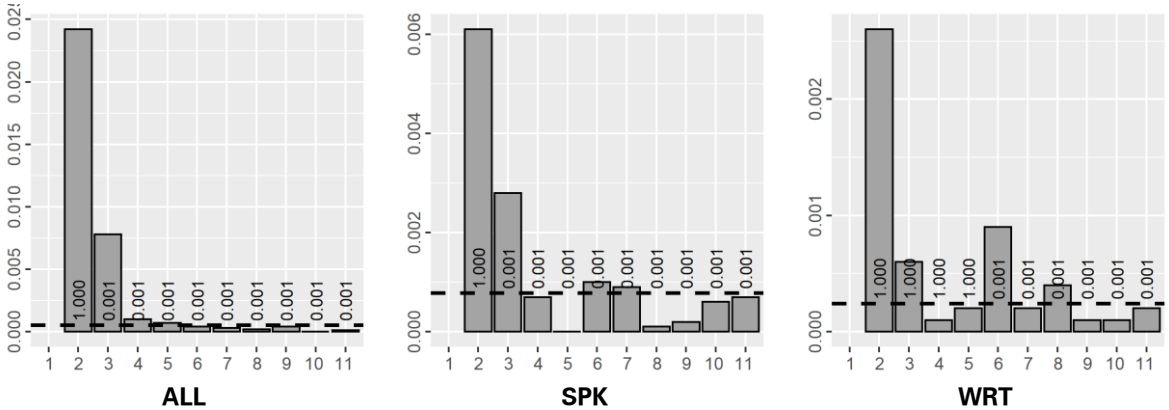
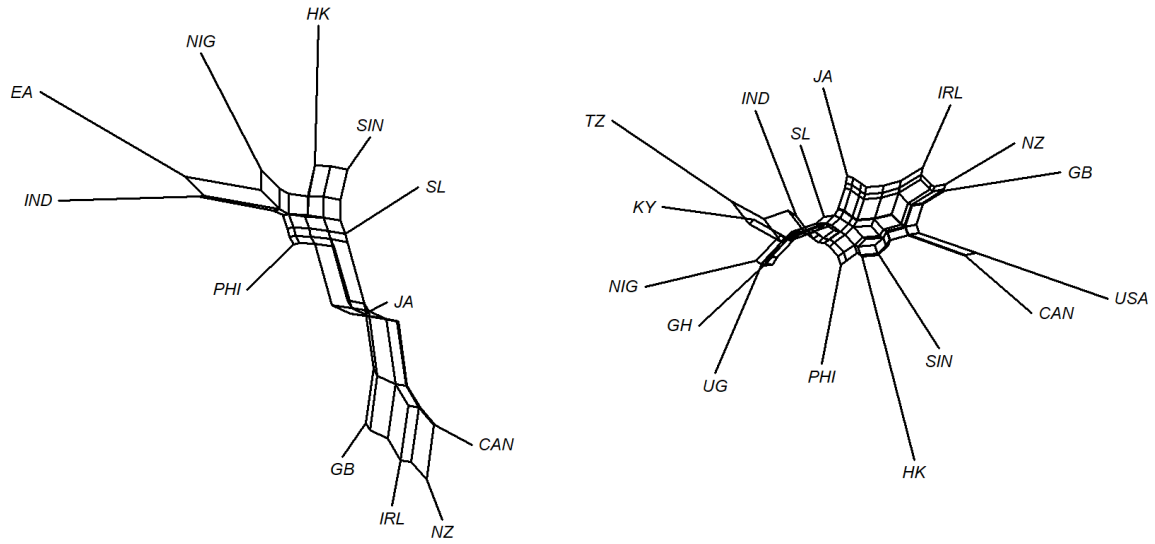


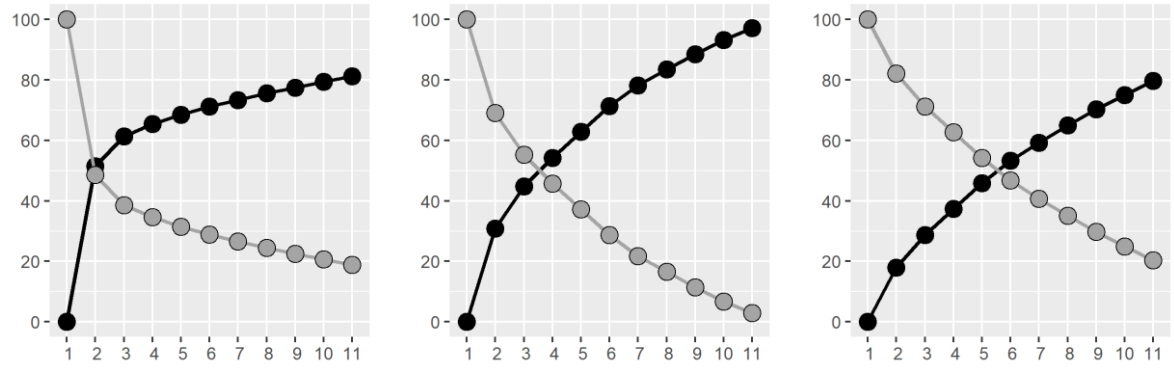
Figure 5.22: Jumps in node heights and respective *p*-values for lexical *g* *n*-grams

Turning to k-means as the final method, results are again mostly identical to previous findings: For ALL, *k*=2 is most strongly indicated and retrieves speech vs. writing, while *k*=3 identifies spoken IC. SPK suggests *k*=3 and *k*=4, which only differ in splitting HK+SIN from the non-EA+IND Outer Circle cluster. WRT at *k*=6 results in identical clusters as in the HCA, but a less fragmented *k*=5 solution diverges considerably by assigning IND to EA (for the first time in writing) and producing a Southeast Asian HK+SIN+PHI cluster in addition to all remaining varieties.



SPK **WRT**

Figure 5.23: NeighborNets of the spoken and written data for lexical *g n*-grams



ALL **SPK** **WRT**

Figure 5.24: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *g n*-grams

Table 5.16: K-means clustering results for specific values of *k* for lexical *g n*-grams

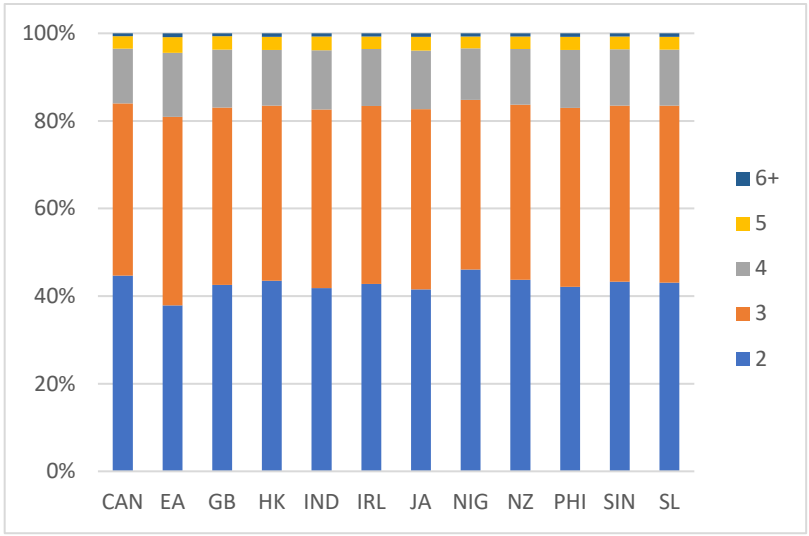
	ALL (*k*=3)		**SPK (*k*=4)**		**WRT (*k*=5)**
1	All written corpus parts	1	HK, SIN	1	GB, IRL, NZ
2	CAN, GB, IRL, NZ $_{SPK}$	2	CAN, GB, IRL, NZ	2	CAN, USA
3	EA, HK, IND, JA, NIG, PHI, SIN, SL $_{SPK}$	3	JA, NIG, PHI, SL	3	KY, TZ, IND
		4	EA, IND	4	HK, PHI, SIN
				5	GH, JA, NIG, SL, UG

5.1.5 Delta $P_{2|1}$

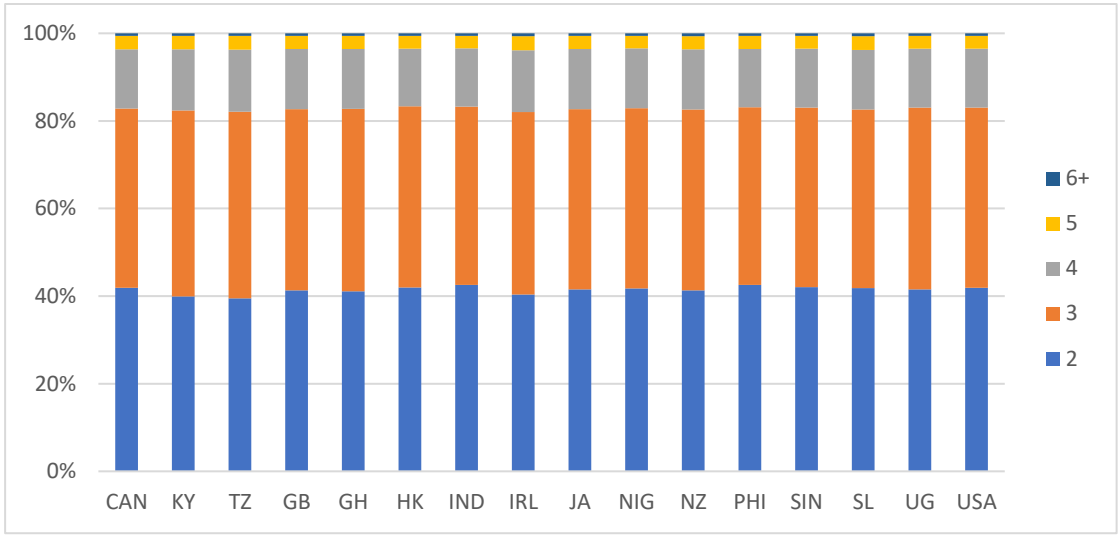
Generation of *n*-grams on the basis of *ΔP* bigrams produced an even larger set of sequences than G^2, with an average of 285,224 (*s*=15,629) and 188,828 (*s*=9,600) spoken and written tokens, respectively, constituted from 156,830 (*s*=8,011) and 127,961 (*s*=6,922) types. Even given the large size of the data, token numbers were reduced most strongly of all measures (-44% in both modes), indicating a very lax threshold for *ΔP*. Relatively fewer types were deleted (-11% and -21% respectively), but still *ΔP* stands out as the only measure which effected significant drops in type frequencies (G^2 removing 4% of types in writing). Mean TTRs resulted in the lowest overall, with 1.8 (*s*=0.1) and 1.5 (*s*=0.0). The average token was found to be basically identical in length between speech (2.79 units, *s*=0.03) and writing (2.80, *s*=0.01) but types are longer in speech (3.15, *s*=0.03 vs. 3.02, *s*=0.02). The distributions in Figure 5.25 show that this is mainly the result of a higher number of retained 2-grams in the written data. Overall, the distributions are much more similar between tokens and types than for all other measures, with 2-grams only providing about twice the number of tokens than types.

While a large disparity between initial variety-specific *n*-grams and the set of shared sequences (reductions by 98-99% to 2,667 and 1,543 items) needs to be noted, no single merging of datasets is responsible for this drop, as Figure 5.26 shows (the positive outlier in the spoken data being constituted by GB/NZ). The dramatic decrease in overall frequencies can thus rather be seen as a consequence of the relatively unfiltered bigram data leading to a retention (and creation) of relatively weakly collocated sequences specific to each individual dataset.

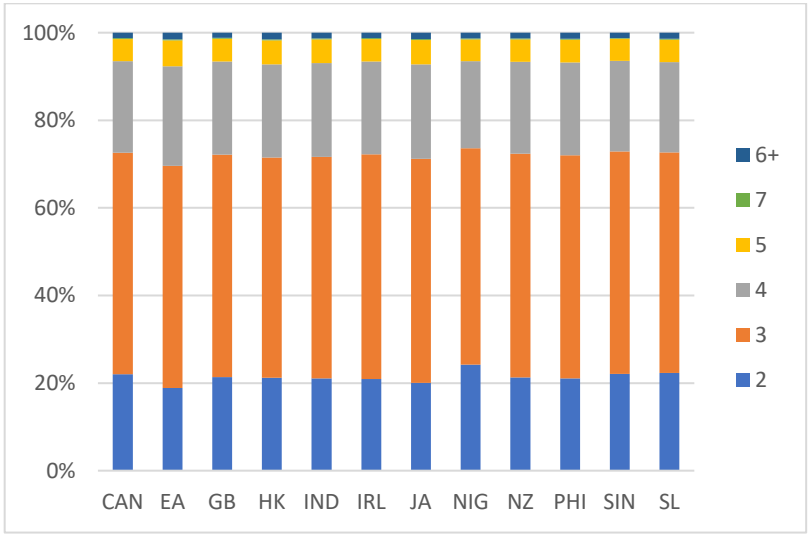
Even given the large relative reductions, merging of the individual varietal datasets produced the higher absolute number of shared sequences, including several longer ones (Table 5.17). Also, while relatively not as successful in producing longer shared sequences as the lexical-gravity measure, the *ΔP* data still display a fair share of 4-grams. As shown in Table 5.18, the retained sequences are largely comprised of sequences in which a consecutive element is strongly determined by the preceding item, in particular prepositional choices. Strong determination of a preceding element, conversely, merits the lowest association scores for this measure, as is its design.



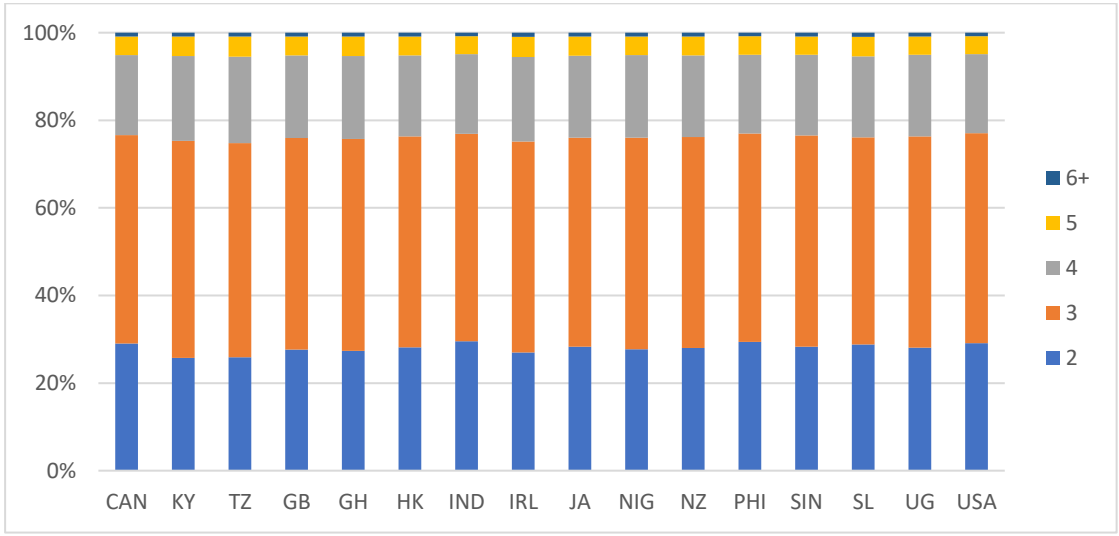
Token frequencies: Spoken data



Token frequencies: Written data



Type frequencies: Spoken data



Type frequencies: Written data

Figure 5.25: Distribution of lexical *ΔP n*-gram lengths across the varietal datasets

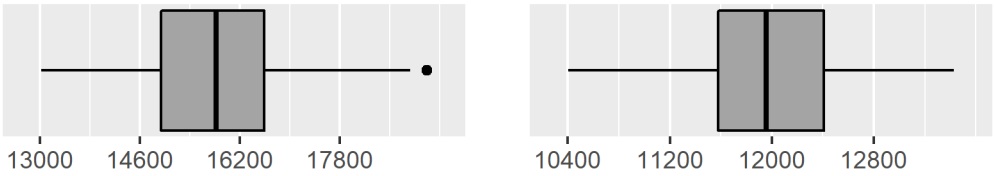


Figure 5.26: Number of shared lexical *ΔP* *n*-grams between any two datasets

Table 5.17: Lexical *ΔP* *n*-gram type frequencies by length in the intersects of the variety-specific datasets

***N*-gram length**	**ALL**	**SPK**	**WRT**
2	703 (69%)	1,636 (61%)	1,001 (65%)
3	308 (30%)	986 (37%)	532 (34%)
4	5 (0%)	45 (2%)	10 (1%)
total	1,016	2,667	1,543

Table 5.18: Lexical *ΔP* *n*-grams with highest and lowest association scores

SPK		**WRT**	
***n*-gram**	***ΔP***	***n*-gram**	***ΔP***
according to	0.9589	according to	0.9631
able to	0.9292	regardless of	0.9410
supposed to	0.9233	depending on	0.9162
kind of	0.8970	et al	0.9125
sort of	0.8807	trying to	0.8220
trying to	0.8700	lack of	0.8194
the evening	0.0004	the answer	0.0003
of problems	0.0004	the dark	0.0003
the background	0.0004	the meantime	0.0002
the girl	0.0003	the background	0.0002
the exam	0.0003	the bed	0.0002
the picture	0.0003	the contrary	0.0002

As already observed for the other low-threshold measure G^2, the *ΔP*-based hierarchical analysis (Figure 5.27) finds relatively few stable clusters in the relatively large number of retained *n*-grams. In addition to the spoken/written distinction in ALL, only the IC_{GB} subcluster is identified in the spoken branch as also in SPK, where an additionally HK+SIN group is detected. Furthermore, the $EA+IND_{SPK}$ cluster barely misses significance in ALL at AU=94.

Significant jumps (Figure 5.28) favor k=5 both in SPK as well as in WRT. This retrieves one IC cluster in SPK in addition to unary EA and IND and two further groups of which one is the infrequent combination of HK+SIN with PHI. In WRT, two IC clusters are produced, with HK associated with IC_{GB}) and a coherent African group emerges while IND is separated from the remaining varieties. ALL returns values of k up to 5 before producing below-average jumps and creating unary clusters. This results in an IC clusters in both modes (joined by SL in writing) the EA+IND combination, all other varieties merged into combined clusters.

The NeighborNets in Figure 5.29 support a separate IC spoken cluster and also HK+SIN, but otherwise show many boxy shapes. EA is found to separate from the data together with IND as well as NIG, the latter of which are however mutually very distinct. WRT indicates either one or two IC clusters, and HK is found somewhat similar to NZ. JA is found in greater proximity to IC than OC but still shows strong difference to the former.

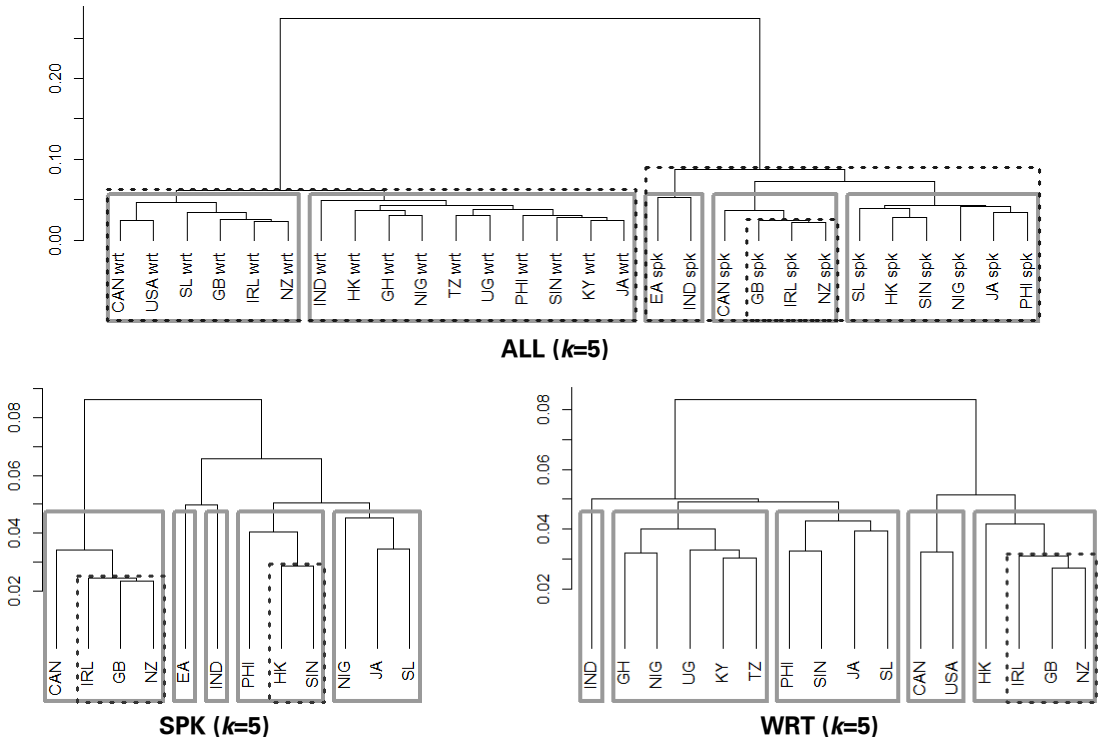


Figure 5.27: Hierarchical clustering results for lexical *ΔP n*-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

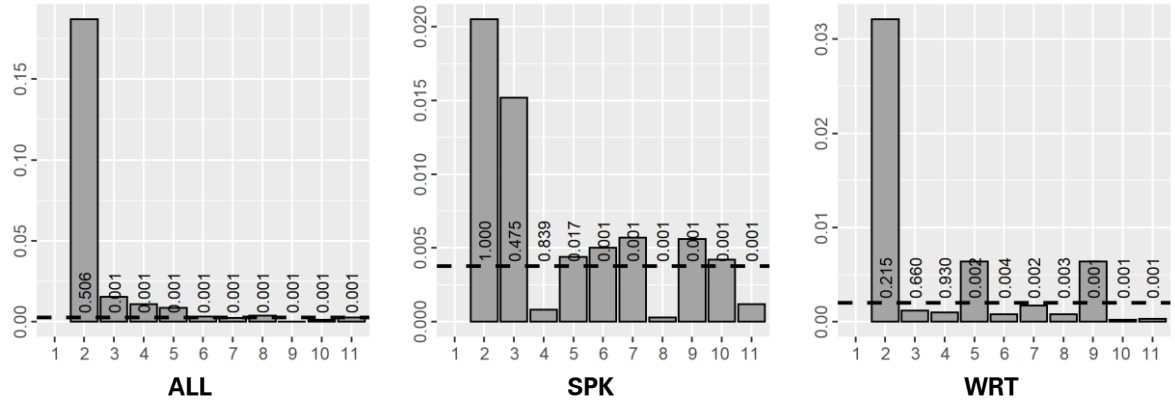


Figure 5.28: Jumps in node heights and respective *p*-values for lexical *ΔP n*-grams

The k-means method indicates best clustering results at values for *k* relatively similar to the hierarchical approach (Figure 5.30 and Table 5.19). For ALL, *k*=4 merits the same separation of the spoken branch as within the HCA but a homogenous written branch, which at *k*=5 further identifies an African (+IND) subcluster. For SPK at *k*=5,

identical results to the hierarchical analysis are produced while *k*=4 merges the non-EA+IND Outer Circle subclusters. For WRT, hierarchical clusterings are also confirmed except for HK isolated within a separate cluster and JA associated to the IC_{NA} group.

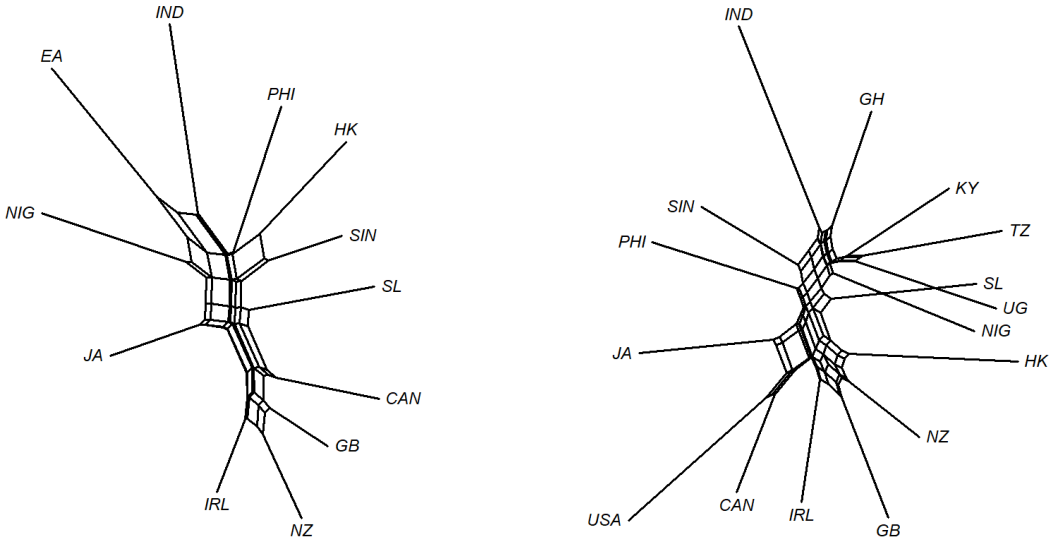


Figure 5.29: NeighborNets of the spoken and written data for lexical *ΔP n*-grams

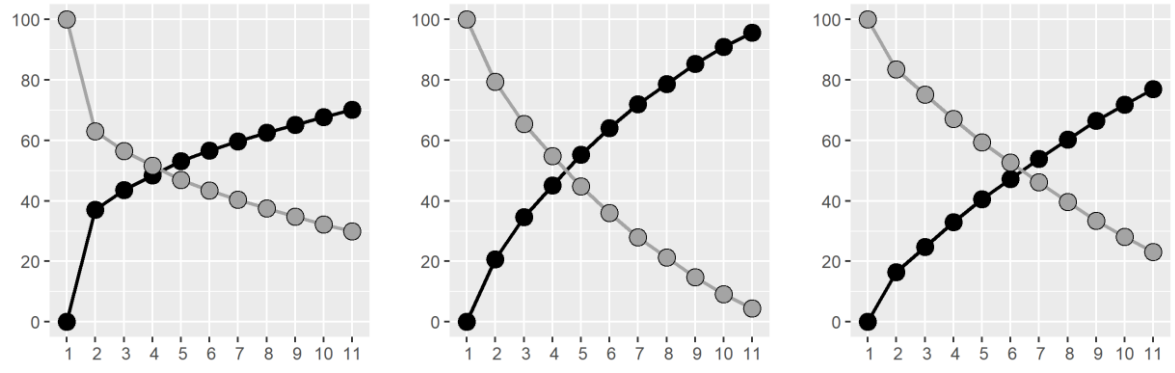


Figure 5.30: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *ΔP n*-grams

Table 5.19: K-means clustering results for specific values of *k* for lexical *ΔP n*-grams

	ALL (*k*=4)		**SPK (*k*=4)**		**WRT (*k*=6)**
1	HK, JA, NIG, PHI, SIN, SL $_{SPK}$	1	EA	1	PHI, SIN, SL
2	EA, IND $_{SPK}$	2	IND	2	KY, TZ, GH, NIG, UG
3	CAN, GB, IRL, NZ $_{SPK}$	3	HK, PHI, SIN, JA, NIG, SL	3	CAN, JA, USA
4	All written corpus parts	4	CAN, GB, IRL, NZ	4	HK
				5	GB, IRL, NZ
				6	IND

5.2 Dynamic-length POS-grams

Extraction of bigrams from the POS-annotated version of the cleaned and homogenized ICE components again yielded 28 bigram lists (modes treated separately), for which token and type frequencies are indicated in Table 5.20. These numbers are somewhat different from those observed for lexical bigrams, which in particular is due to specifics of the POS annotation (e.g. genitives marked by separate tags). Otherwise, the same disclaimers apply as before (dependence on utterance structure, potential splitting of utterances at certain types of punctuation). Given the much more restricted set of possible types (i.e. particularly the tagset, but note the retention of some lexical items discussed in Section 4.2.1), TTRs are far higher than found before, but still speech displays a larger repetition (by again c. 40%) of the same types (TTR≈50.83) than writing (TTR≈35.56).[77] As before, bigram frequencies only serve as the backdrop against the generation of longer sequences, but will be addressed separately for static-length 2-grams.

Table 5.20: Frequencies of POS bigram types and tokens, plus TTRs, extracted from the ICE data

Component	Spoken Mode			Written Mode		
	Tokens	Types	TTR	Tokens	Types	TTR
CAN	556,148	11,453	48.56	357,307	9,784	36.52
EA	544,096	11,117	48.94	–	–	–
KY	–	–	–	342,832	9,819	34.92
TZ	–	–	–	352,248	10,000	35.22
GB	562,647	11,614	48.45	370,013	10,223	36.19
GH	–	–	–	356,731	10,283	34.69
HK	642,663	11,467	56.04	414,961	10,048	41.30
IND	612,547	11,141	54.98	362,290	10,227	35.42
IRL	546,008	11,430	47.77	373,741	10,068	37.12
JA	578,638	11,847	48.84	358,373	10,843	33.05
NIG	537,040	10,596	50.68	351,513	10,437	33.68
NZ	613,789	11,496	53.39	374,898	10,382	36.11
PHI	601,509	11,363	52.94	380,619	10,424	36.51
SIN	547,714	11,297	48.48	350,137	10,140	34.53
SL	553,745	11,122	49.79	345,491	10,606	32.58
UG	–	–	–	352,097	10,153	34.68
USA	–	–	–	366,932	10,087	36.38
mean ($\bar{x}$)	**574,712**	**11,329**	**50.83**	**363,136**	**10,220**	**35.56**
sd (s)	33,101	301	2.84	17,085	267	1.94

Application of association thresholds results in frequencies as shown in Tables 5.21 and 22. As observed for lexical bigrams, TTRs dramatically increase, in particular for

[77] Mean TTRs again are means of the TTRs from the corpus parts (thus the SD), and not calculated from the mean token and type frequencies.

those measures with more restrictive threshold values, indicating relatively stronger retention of more frequent types (reinforced for *MI* by the additional frequency threshold). The relative order of the impact of these thresholds is almost identical to the lexical data except for *t*'s threshold effect being reduced so that it is no longer on par with that of *MI*. Generally, effects on type frequencies are largely identical to the lexical data for those measures already previously strongly affected by the thresholds (*MI, t, g*; least so for *t*). For the relatively low-threshold measures (G^2, ΔP), they do however increase significantly, indicating that more types do not achieve the threshold values. Threshold effects on token frequencies differ more strongly from what was previously observed. They are only barely higher in case of G^2 but more so for ΔP and also, curiously, for *MI*, but surface much less strongly for *t* and *g* in comparison to the lexical data counterpart. The greatly reduced variety of types through the POS annotation thus appears to 'favor' those measures which have already displayed a tendency towards high-frequency items after the application of thresholds in the previous section. Conversely, the measures overall not as strongly affected by their thresholds now are more drastically impacted in type frequencies through the reduction in diversity brought about by POS annotation.

Apart from the effects of the POS annotation on the data itself as well as somewhat different effects of the association thresholds on the POS-annotated data, all merging of bigrams to longer sequences was carried out in the same manner as for the lexical data.[78]

[78] Again, note that type frequencies can be higher in the analyses to follow given the particulars of the generation procedure.

Table 5.21: Spoken POS bigram token and type frequencies, plus TTRs, after the application of threshold values

Spoken	MI			t			G^2			g			*ΔP*		
component	**tokens**	**types**	**TTR**	**tokens**	**types**	**TTR**	**tokens**	**types**	**TTR**	**tokens**	**types**	**TTR**	**tokens**	**types**	**TTR**
CAN	78,797	430	183.2	364,447	1,373	265.4	374,728	2,448	153.1	236,271	172	1,373.7	402,135	5,193	77.4
EA	82,056	497	165.1	382,361	1,469	260.3	395,665	2,674	148.0	268,704	178	1,509.6	422,105	5,604	75.3
GB	78,936	462	170.9	371,782	1,395	266.5	384,228	2,482	154.8	247,985	165	1,502.9	411,526	5,328	77.2
HK	95,119	448	212.3	414,587	1,387	298.9	427,389	2,451	174.4	300,609	169	1,778.8	455,976	5,192	87.8
IND	88,678	461	192.4	408,966	1,378	296.8	420,104	2,366	177.6	296,289	177	1,673.9	447,080	5,071	88.2
IRL	73,848	419	176.2	360,063	1,357	265.3	372,191	2,381	156.3	230,958	167	1,383.0	404,123	5,268	76.7
JA	82,029	488	168.1	382,006	1,440	265.3	394,371	2,614	150.9	259,949	178	1,460.4	426,347	5,596	76.2
NIG	77,637	404	192.2	344,477	1,213	284.0	356,398	2,201	161.9	234,256	149	1,572.2	382,807	4,691	81.6
NZ	75,177	387	194.3	398,343	1,421	280.3	411,326	2,416	170.3	264,066	165	1,600.4	441,999	5,178	85.4
PHI	91,224	478	190.8	391,829	1,461	268.2	406,254	2,491	163.1	273,180	168	1,626.1	439,559	5,367	81.9
SIN	81,921	451	181.6	353,733	1,354	261.3	364,437	2,390	152.5	245,514	167	1,470.1	396,977	5,347	74.2
SL	74,524	418	178.3	349,508	1,321	264.6	359,164	2,268	158.4	239,656	163	1,470.3	392,804	5,041	77.9
mean ($\bar{x}$)	**81,662**	**445**	**183.8**	**376,842**	**1,381**	**273.1**	**388,855**	**2,432**	**160.1**	**258,120**	**168**	**1,535.1**	**418,620**	**5,240**	**80.0**
sd (s)	6,506	33	12.8	22,342	66	13.0	23,050	125	9.2	22,411	8	114.5	22,752	237	4.7
Threshold effect	-86%	-96%		-34%	-88%		-32%	-79%		-55%	-99%		-27%	-54%	

Table 5.22: Written POS bigram token and type frequencies, plus TTRs, after the application of threshold values

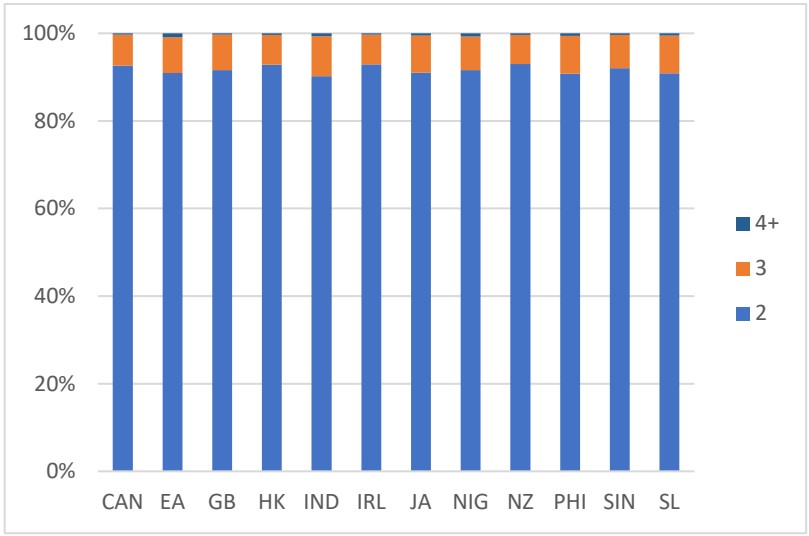
Written		MI			t			G^2			g			*ΔP*	
component	tokens	types	TTR	tokens	types	TTR	tokens	types	TTR	tokens	types	TTR	tokens	types	TTR
CAN	40,240	413	97.4	242,750	1,124	216.0	253,831	2,592	97.9	187,578	137	1,369.2	270,905	5,402	50.1
KY	44,391	437	101.6	238,745	1,119	213.4	249,560	2,709	92.1	179,010	142	1,260.6	266,953	5,488	48.6
TZ	42,839	449	95.4	243,953	1,064	229.3	254,050	2,864	88.7	195,516	147	1,330.0	271,459	5,850	46.4
GB	45,882	454	101.1	256,982	1,208	212.7	265,158	2,758	96.1	194,542	148	1,314.5	285,077	5,602	50.9
GH	43,953	442	99.4	250,563	1,107	226.3	260,720	3,078	84.7	196,708	147	1,338.1	275,527	6,038	45.6
HK	52,040	462	112.6	285,176	1,182	241.3	294,895	2,796	105.5	228,640	143	1,598.9	310,562	5,525	56.2
IND	42,869	501	85.6	245,564	1,124	218.5	255,527	2,998	85.2	198,921	140	1,420.9	271,877	5,832	46.6
IRL	44,441	459	96.8	255,805	1,160	220.5	264,510	2,790	94.8	202,116	145	1,393.9	286,505	5,718	50.1
JA	43,334	465	93.2	243,130	1,177	206.6	253,369	3,172	79.9	195,334	151	1,293.6	273,736	6,363	43.0
NIG	40,917	437	93.6	243,300	1,071	227.2	252,457	2,996	84.3	194,300	147	1,321.8	268,101	6,122	43.8
NZ	45,318	434	104.4	253,166	1,150	220.1	261,653	2,703	96.8	199,052	151	1,318.2	286,460	5,769	49.7
PHI	43,432	445	97.6	257,689	1,144	225.3	266,552	3,026	88.1	208,165	141	1,476.3	284,383	6,011	47.3
SIN	42,427	451	94.1	238,618	1,117	213.6	249,014	3,031	82.2	188,292	148	1,272.2	265,616	5,864	45.3
SL	40,070	480	83.5	229,458	1,091	210.3	239,641	3,089	77.6	192,110	148	1,298.0	258,227	6,231	41.4
UG	42,475	439	96.8	236,805	1,106	214.1	246,455	3,011	81.9	193,825	149	1,300.8	262,892	5,947	44.2
USA	42,125	428	98.4	244,569	1,103	221.7	255,971	2,849	89.8	195,910	148	1,323.7	276,928	5,651	49.0
mean ($\bar{x}$)	**43,547**	**450**	**97.0**	**247,892**	**1,128**	**219.8**	**257,710**	**2,904**	**89.1**	**196,876**	**146**	**1,351.9**	**275,951**	**5,838**	**47.4**
sd (s)	2,713	20	6.6	12,250	39	8.4	11,898	162	7.4	10,312	4	83.6	12,187	264	3.5
Threshold effect	-88%	-96%		-32%	-89%		-29%	-72%		-46%	-99%		-24%	-43%	

5.2.1 MI-score

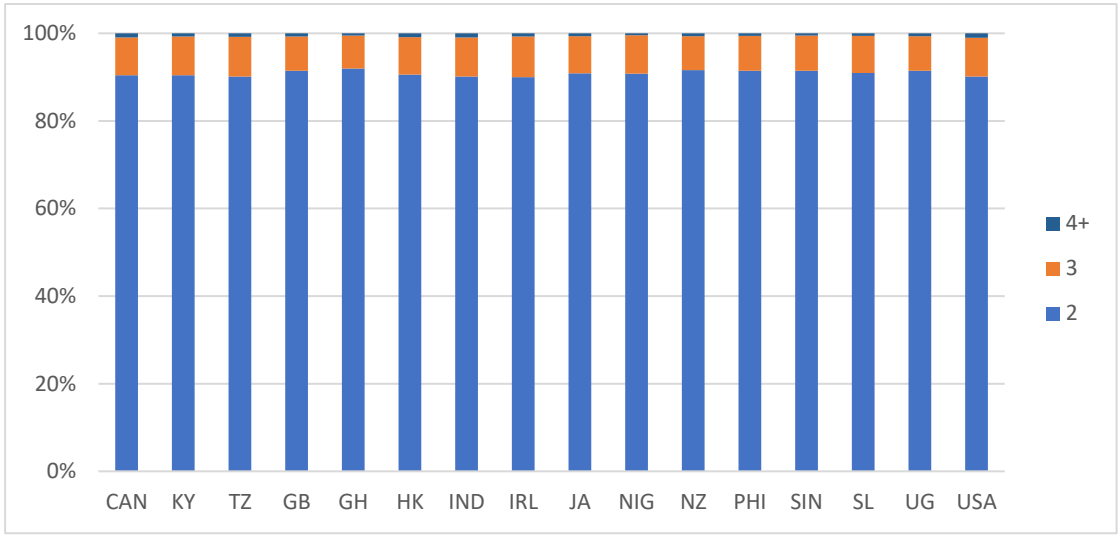
MI POS *n*-grams totaled 75,036 (*s*=5,864) and 39,615 (*s*=788) tokens of 736 (*s*=86) and 788 (*s*=45) types for speech and writing, respectively, resulting in TTRs of 102.9 (*s*=9.3) and 50.4 (*s*=3.8). For tokens, this represents frequency reductions from the bigram data by -8% and -9%, while for types, an additional +65% and +75% are generated. The average token was found to be of identical lengths between speech (2.09 items, *s*=0.01) and writing (2.10, *s*=0.01), and types were also barely different in lengths (2.57, *s*=0.07 vs. 2.62, *s*=0.04). The distributions of item lengths in Figure 5.31 also indicate relative homogeneity between speech and writing, with some exception in case of the spoken IC varieties retaining more shorter sequences than all other varieties, at the particular expense of 4-grams. Figure 5.31 further reveals a shift in items lengths between tokens and types, in that 4-grams consistently show higher relative frequencies for types than any length above 2 for tokens.

Losses accrued during the merging procedure are lower than for the lexical data and lowest among POS-grams at -79% and -77%, but only very low absolute frequencies are observed (153 and 185 types in speech and writing). For the first time, a negative outlier (Figure 5.32) can be found in the number of shared written sequences between CAN and GH, but since the difference to the next-lowest frequency was found to lie at two items and both of these varietal datasets also displayed very high overlaps with other data, no intervention was deemed necessary.

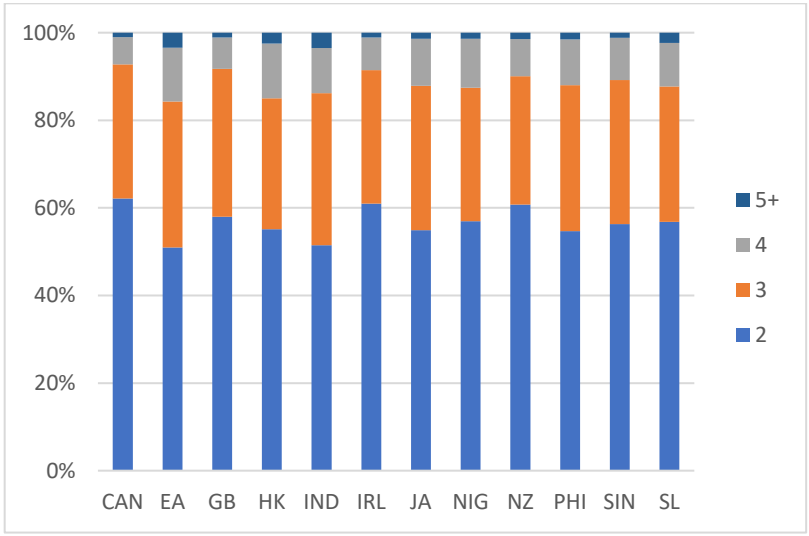
Low absolute numbers of items and limited relative importance of sequences at n≥4 results in none of these sequences being shared between varieties with the exception of a single 5-gram in writing (Table 5.23). In contrast to the numerical dominance of (shared) spoken sequences in the lexical data, a larger overlap of written types can be observed for the POS data. Just as for lexical sequences, however, Table 5.24 again testifies to *MI*'s preference for rarer sequences, on the one hand by showing how the measure awards high scores to either lexical (i.e. less frequent) items or to rare combinations of POS tags, e.g. two consecutive adverbs for appositional structures (REX) or plural after singular pronouns (PN1/2, PPX1/2), which may both be cases of false starts/corrections. Frequent structures such as determiner+nouns or article+verb combinations, comparative constructions, etc. score low with the *MI* measure.



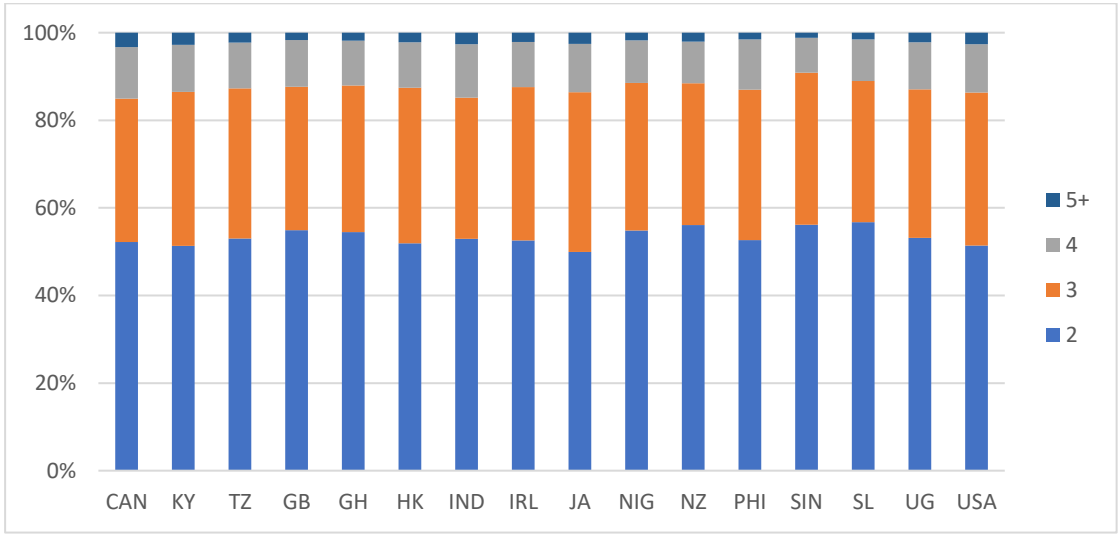
Token frequencies: Spoken data



Token frequencies: Written data



Type frequencies: Spoken data



Type frequencies: Written data

Figure 5.31: Distribution of POS *MI n*-gram lengths across the varietal datasets

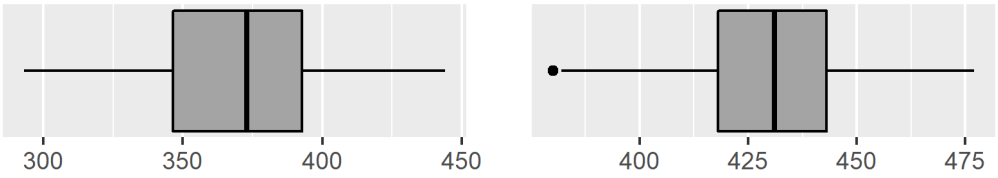


Figure 5.32: Number of shared POS *MI* *n*-grams between any two datasets

Table 5.23: POS *MI* *n*-gram type frequencies by length in the intersects of the variety-specific datasets

***N*-gram length**	**ALL**	**SPK**	**WRT**
2	101 (96%)	139 (91%)	152 (82%)
3	4 (4%)	14 (9%)	32 (17%)
5	0 (0%)	0 (0%)	1 (1%)
total	105	153	185

Table 5.24: POS *MI* *n*-grams with highest and lowest association scores

SPK		**WRT**	
***n*-gram**	***MI***	***n*-gram**	***MI***
even though	10.53	even though	10.09
rather than	10.10	rather than	9.45
other than	10.09	other than	9.40
REX REX	9.34	PN1 PN2	8.65
PPX1 PPX2	8.53	PPX1 PPX2	8.47
DAR than	8.19	REX REX	8.08
DD2 NN2	3.34	DD1 NNT1	3.45
PPY VV0	3.34	PPY VM	3.40
when PPIS2	3.32	PPIS1 VVD	3.35
PPH1 VBDZ	3.29	RGT JJ	3.30
PNQS VVD	3.28	RGR JJ	3.22
VM XX	3.27	AT JJT	3.13

Within the hierarchical analysis (Figure 5.33), issues with the relatively small number of items begin to surface in barely any clusters registering as stable over bootstrapping iterations using `pvclust`, and only a partial African KY+TZ+NIG cluster is found in writing and the IC group in speech. Interestingly, it is the ALL dataset substantiating these clusters based on an overall smaller number of *n*-grams (but contrasting speech and writing), while SPK and WRT retrieve no stable clusters.

Significant jumps (Figure 5.34) in ALL indicate at most the separateness of EA+IND_{SPK} at k=3 beyond the spoken/written distinction. SPK only retrieves the first significant jump at k=7, isolating HK in addition to the more commonly unary EA, IND and NIG and identifying PHI+SL in addition to a IC cluster without GB, which is allocated to a regionally as well as evolutionarily diverse cluster with JA and SIN. WRT indicates the best solution at k=4, identifying an IC cluster (+PHI), an African group as well as the two Asian clusters HK+SIN and IND+SL(+JA).

The NeighborNets in Figure 5.35, while agreeing with most of these results for writing, diverge more strongly for speech. SPK finds the IC cluster, some difference of HK+SIN from other OC varieties and slightly higher similarity of EA to NIG than IND. NIG and JA are also found to be similar. WRT, meanwhile, distinguishes an IC cluster in which NZ is removed from the usual IC_{GB} group, identifies HK+SIN, IND+SL and indicates NIG+GH+KY and overall distance of Africa + IND+SL from the remaining varieties.

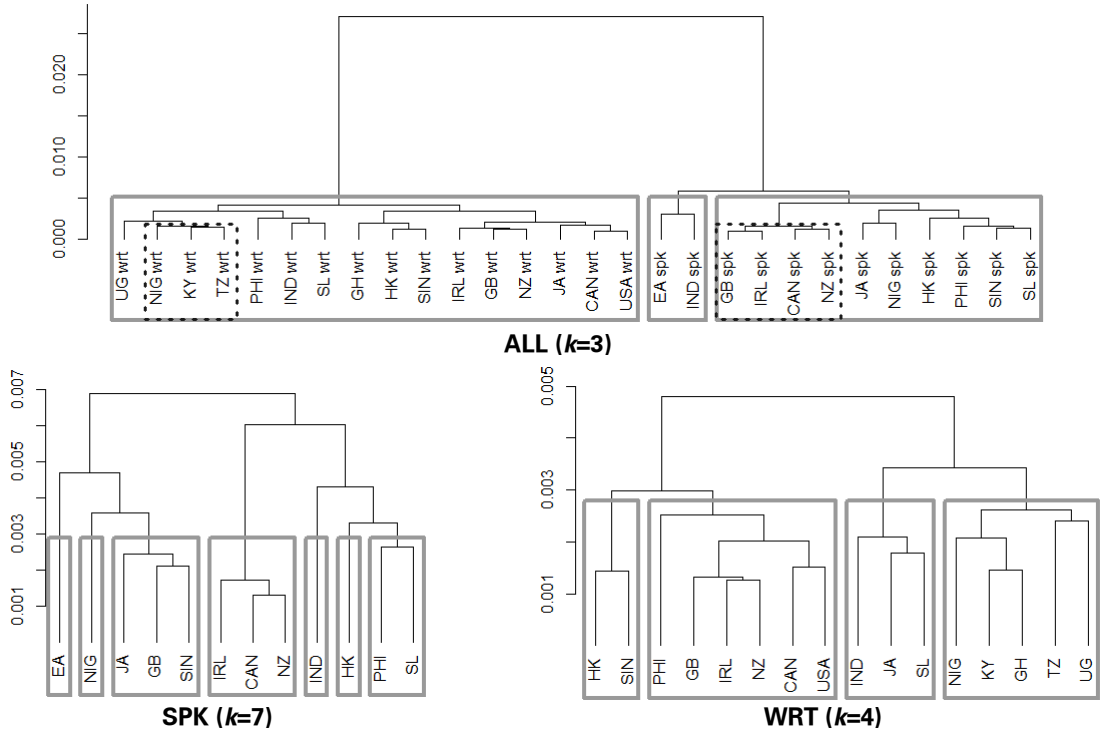


Figure 5.33: Hierarchical clustering results for POS *MI n*-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

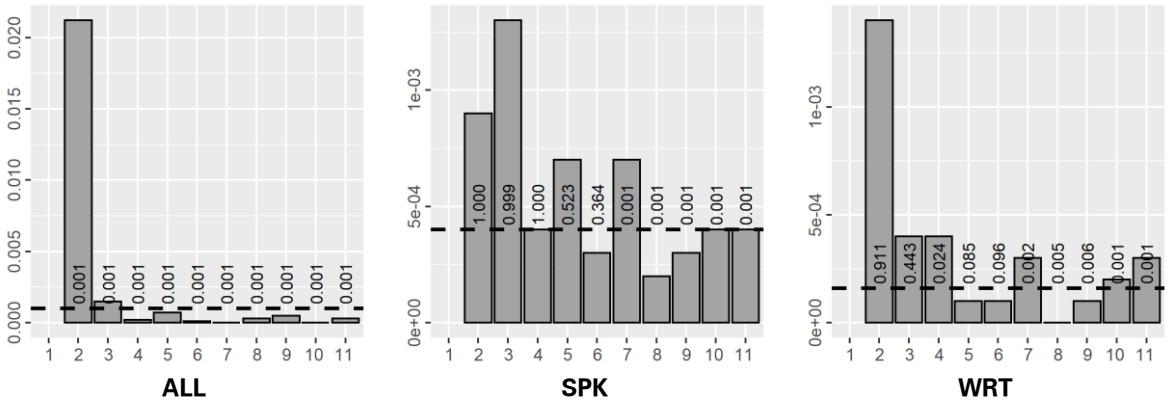


Figure 5.34: Jumps in node heights and respective *p*-values for POS *MI n*-grams

K-means clustering (Figure 5.36 and Table 5.25) is found in strong support of a simple binary separation of ALL, while finding fewer clusters in SPK (*k*=4) but more in WRT (*k*=6). The spoken results can, for the first time, be taken to represent

evolutionary phases, as long as the dividing line for the OC group is drawn between 'phase 2 to early 3', 'phase 3 to early 4' and 'fully phase 4', with only NIG not quite fitting the latter group. Writing, however, more clearly exhibits tendencies of clustering varieties from proximal regions/epicenters, but also shows some results incongruous to this (the above 'intermediate' varieties JA+PHI as cluster #1, and KY removed from the East African varieties and included in a West African cluster). The results for writing, however, can also be said to agree with the NeighborNets above.

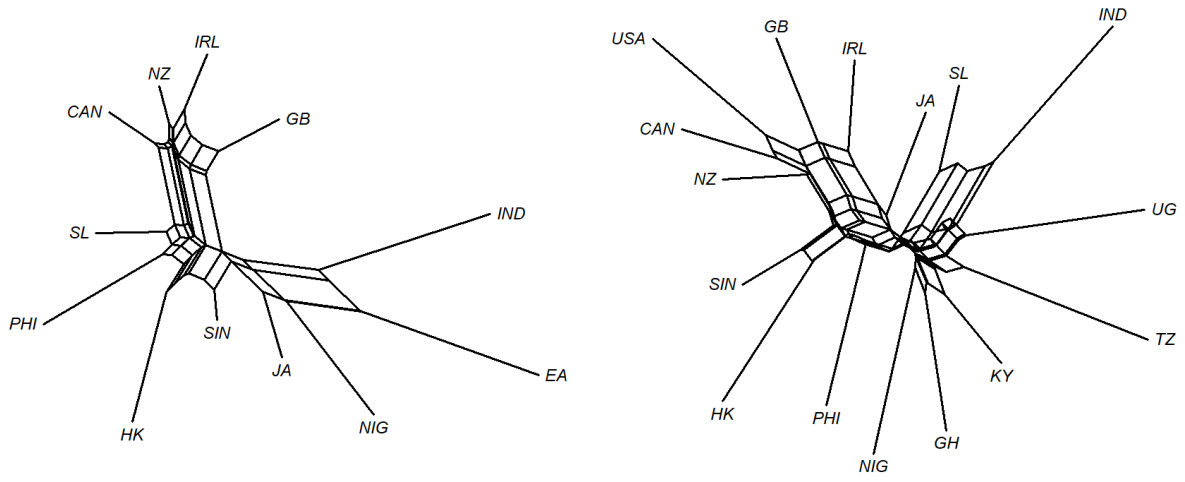


Figure 5.35: NeighborNets of the spoken and written data for POS *MI n*-grams

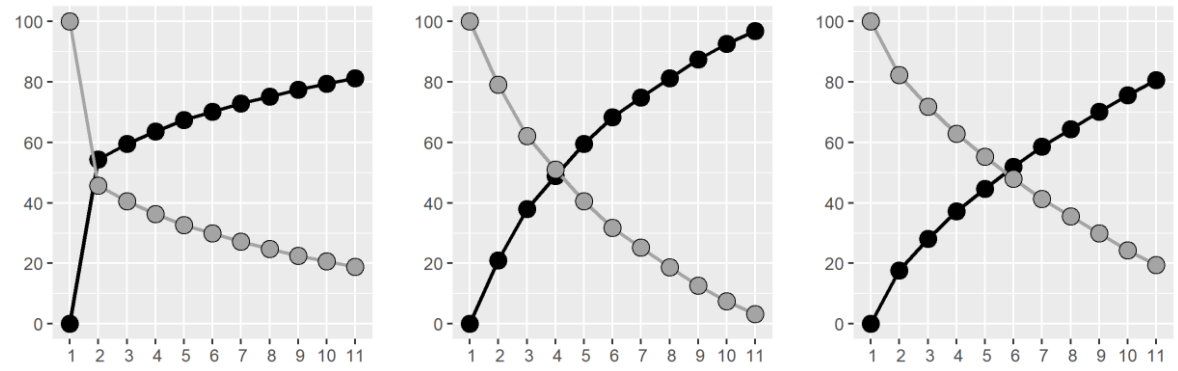


Figure 5.36: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *MI n*-grams

Table 5.25: K-means clustering results for specific values of *k* for POS *MI n*-grams

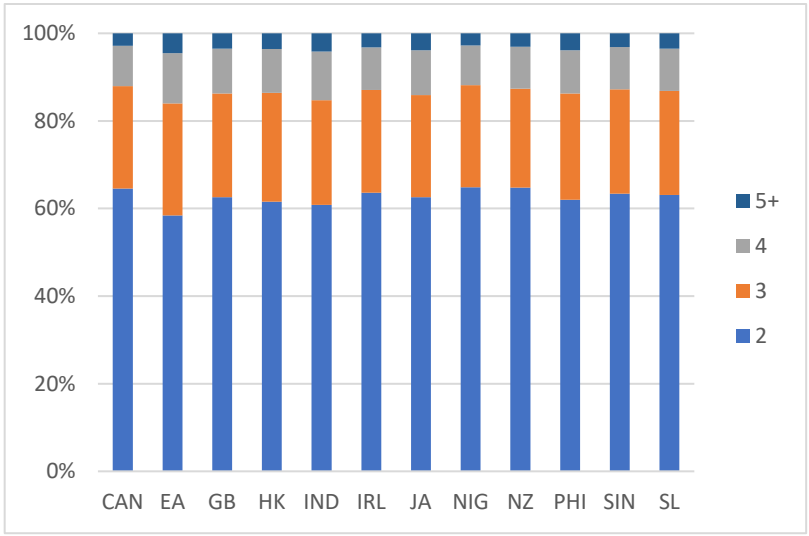
	ALL (*k*=2)		SPK (*k*=4)		WRT (*k*=6)
1	All spoken corpus parts	1	EA	1	JA, PHI
2	All written corpus parts	2	HK, IND, PHI, SL	2	CAN, GB, IRL, NZ, USA
		3	CAN, GB, IRL, NZ	3	HK, SIN
		4	JA, NIG, SIN	4	TZ, UG
				5	KY, GH, NIG
				6	IND, SL

5.2.2 T-score

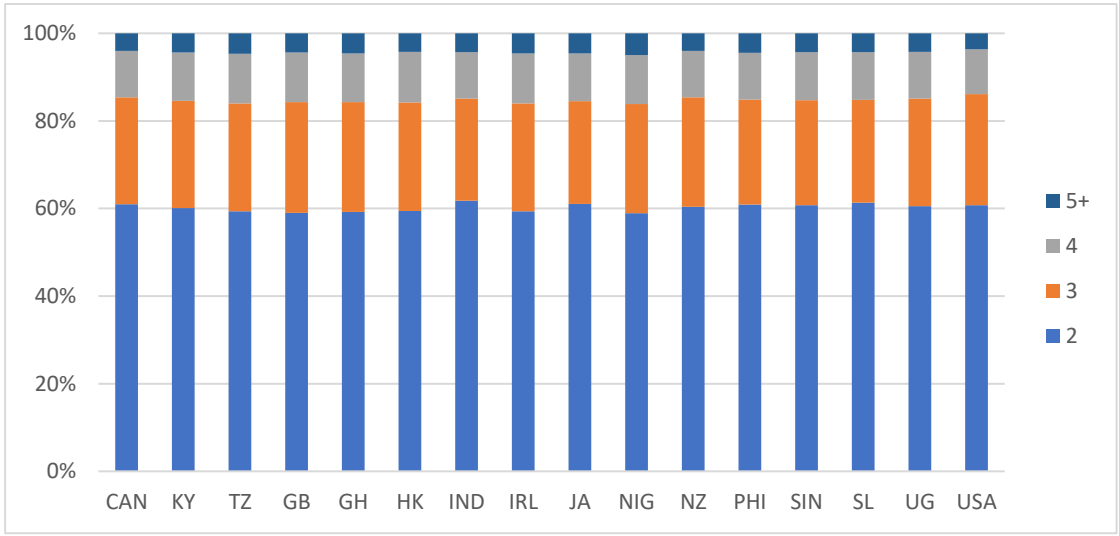
Generation of *n*-grams based on the t-score produced 242,451 (*s*=12,469) and 154,018 (*s*=7,264) tokens of 14,978 (*s*=1,428) and 10,182 (*s*=535) types, resulting in TTRs of 16.3 (*s*=1.4) and 15.1 (*s*=0.7). While this equates to regular reductions in token frequencies against bigram counts by -36% and -38%, an immense increase in type frequencies can be detected at +985% and +803%, indicating large numbers of newly generated sequences. Average sequences were found almost identical in length both for tokens (2.55, *s*=0.04 vs. 2.61, *s*=0.02) as well as for types (4.02, *s*=0.07 vs. 4.06, *s*=0.05). The large increases in average lengths from tokens to types harmonizes with the increases in frequency in painting *t*-based sequences as very successful in finding consecutive items with high mutual attraction. Figure 5.37 confirms the dominance of longer sequences for types: While sequences of $n \geq 5$ elements are still a noticeable portion of the token data, for types sequences of lengths 4 and above constitute the majority of items, whereas in particular 2-grams become a minor category in the type data.

Merging of the varietal datasets incurs losses of types by -87%and -83% (slightly less than G^2 and ΔP and slightly more than *g*), still resulting in relatively high absolute shared item frequencies of 1,927 and 1,686 (similar to G^2 and ΔP but much higher than *MI* and *g*). No single merger of two varieties displays too strong an effect towards decreasing the combined sets (Figure 5.38), while the positive outliers represent the mergers of IRL/NZ, IRL/GB as well as GB/NZ.

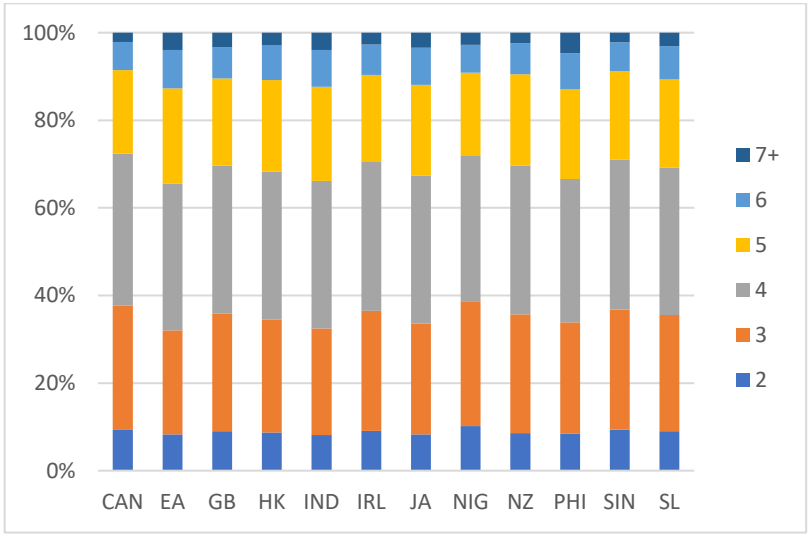
With the preponderance of longer types, these also factor significantly within the list of shared sequences presented in Table 5.26, and sequences up to *n*=6 can be observed, even if only 5-grams still surface in relevant proportions. Table 5.27 reveals that *t* again pronouncedly displays its preference for high-frequency items, assigning top association scores in particular to determinative, pre- and postmodifying structures surrounding the most frequent word class of nouns, but also listing *to*-infinitives highly.



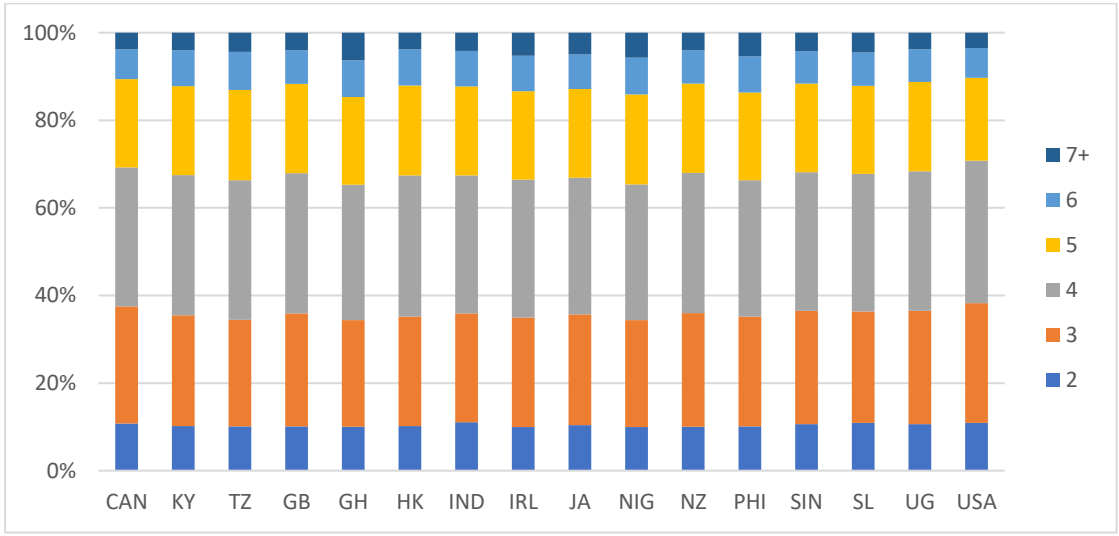
Token frequencies: Spoken data



Token frequencies: Written data



Type frequencies: Spoken data



Type frequencies: Written data

Figure 5.37: Distribution of POS *t n*-gram lengths across the varietal datasets

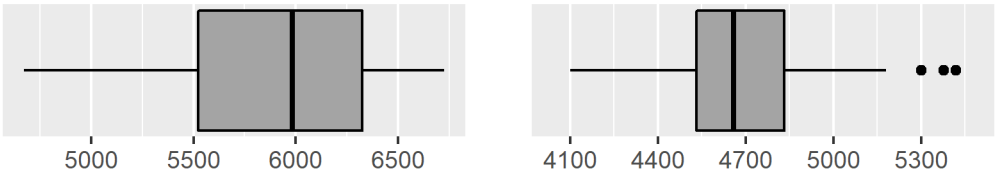


Figure 5.38: Number of shared POS *t n*-grams between any two datasets

Table 5.26: POS *t n*-gram type frequencies by length in the intersects of the variety-specific datasets

***N*-gram length**	**ALL**	**SPK**	**WRT**
2	450 (41%)	668 (35%)	542 (32%)
3	383 (34%)	725 (38%)	637 (38%)
4	202 (18%)	399 (21%)	382 (23%)
5	72 (6%)	130 (7%)	113 (7%)
6	4 (0%)	5 (0%)	12 (1%)
total	1,111	1,927	1,686

Table 5.27: POS *t n*-grams with highest and lowest association scores

SPK		**WRT**	
***n*-gram**	***t***	***n*-gram**	***t***
AT NN1	94.14	AT NN1	76.53
to VVI	82.30	to VVI	71.51
JJ NN1	76.84	JJ NN1	67.45
of AT NN1	69.30	NN1 of	64.94
NN1 of	68.15	of AT NN1 of	61.02
in AT NN1	67.60	AT JJ NN1 of	59.26
NNT2 before	3.97	as long	3.72
VHZ VDN	3.95	long as	3.72
or RRR	3.73	PPHO1 RP	3.68
VVN before	3.73	now that	3.66
VVZ into	3.67	from RL	3.64
DAR NN	3.30	VDI PN1	3.50

HCA finds more clusters in case of *t* than *MI* (Figure 5.39), but except for the spoken/written branches, only written subclusters emerge within ALL. The two IC clusters retrieved in WRT are also confirmed in ALL, but are furthermore merged with the stable HK+SIN and IND+SL+PHI+JA groups, respectively, indicating relative similarities but also a degree of separateness of HK and SIN from the other Asian varieties (the written African cluster barely misses significance).

Significant jump heights (Figure 5.40) latch onto the greater diversity within the spoken branch of ALL, segmenting it into IC and two OC clusters (separating EA and IND) at the last above-average jump at k=4. SPK obtains almost identical results, only assigning JA to the IC group, and WRT offers a segmentation largely overlapping with the smaller stable clusters in ALL, i.e. removing the two IC groups from the OC varieties, isolating HK and distinguishing between an African and an Asian (+JA) cluster.

NeighborNet analysis (Figure 5.41) supports these segmentations, furthermore indicating an intermediary position of JA_{SPK} which may account for its divergent clustering in the HCA. It also finds NIG to share some features with EA+IND and supports HK+SIN. The written data indicate two very distinct IC groups. HK+SIN does, however, share some features with IC_{NA}, and IC_{GB} with parts of the African data. Yet, separation of the African cluster (with UG in between NIG and GH) is more strongly indicated. IND and SL are also found in close proximity.

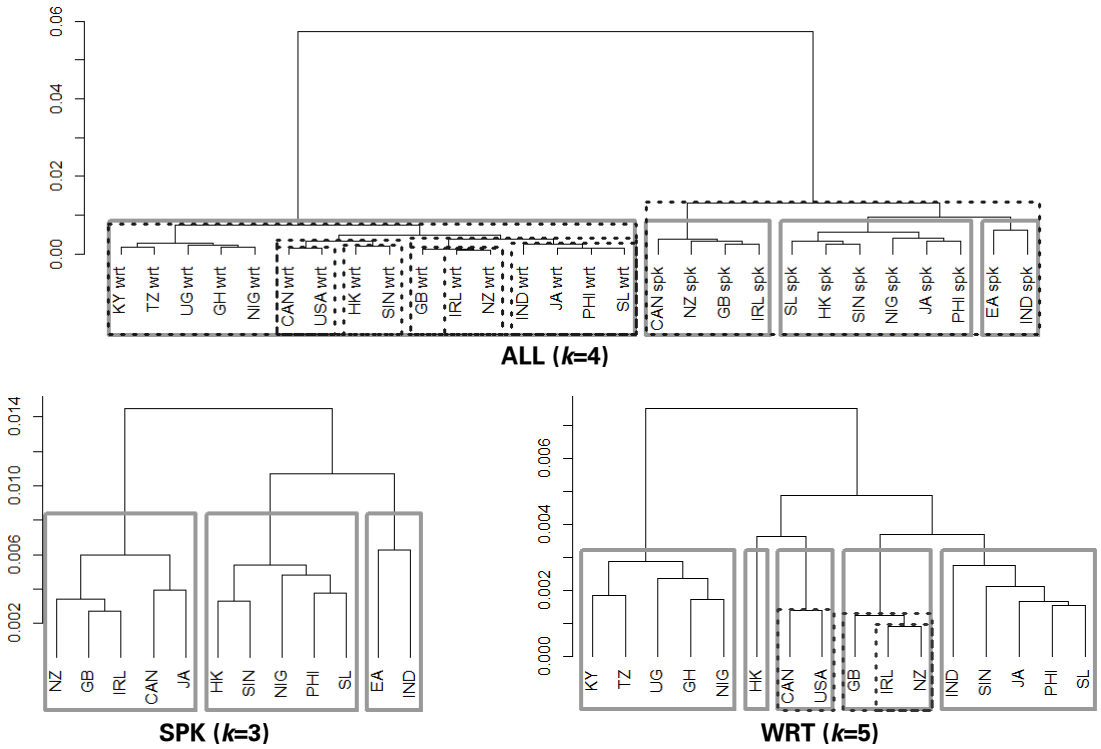


Figure 5.39: Hierarchical clustering results for POS *t n*-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

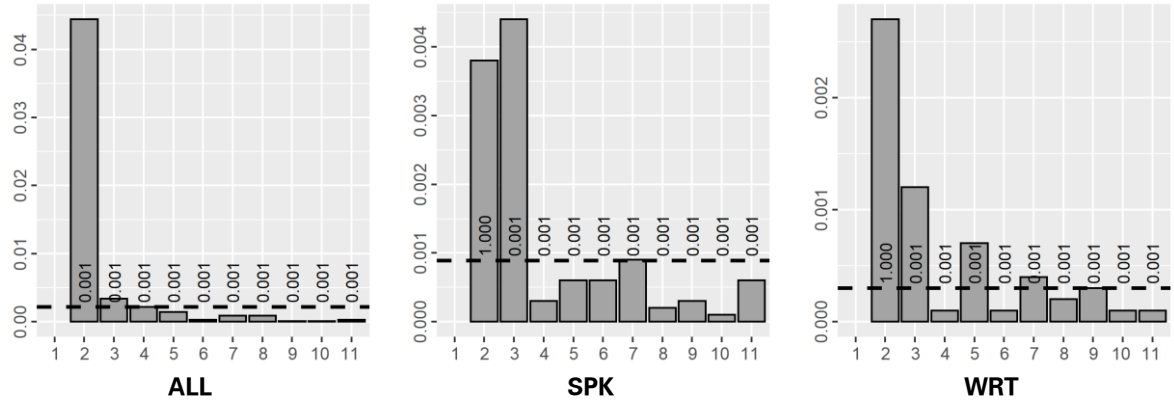
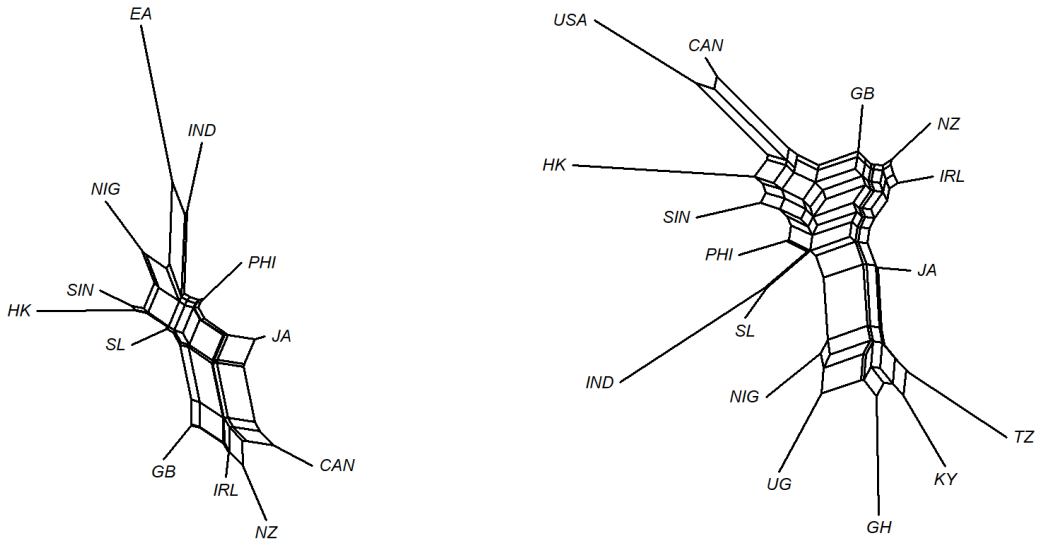


Figure 5.40: Jumps in node heights and respective *p*-values for POS *t n*-grams

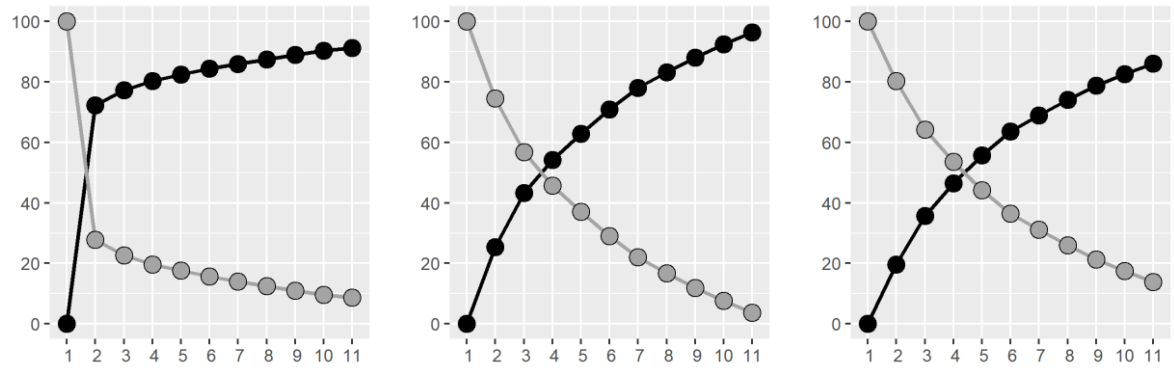
K-means analysis (Figure 5.42 and Table 5.28) indicate the usual clear segmentation of ALL into the two major branches. SPK shows the best intersect at *k*=4, separating HK, IND and EA from the remaining OC varieties and the IC cluster. This presents

some agreement to both a regional as well as an evolutionary perspective, in that some of the least 'advanced' varieties are isolated, but this already appears less adequate in case of IND, SL, and PHI. The written data, except for most of the African varieties grouped together, shows no discernible structure at all. What is worse, the somewhat less indicated values of 3 and 5 for *k*, respectively, lead to very different results but perfectly replicate the hierarchical analysis (except for JA_{SPK} moving to the non-EA+IND cluster), demonstrating the potential effects of the low degree of stability found in the separate datasets above.



SPK **WRT**

Figure 5.41: NeighborNets of the spoken and written data for POS *t n*-grams



ALL **SPK** **WRT**

Figure 5.42: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *t n*-grams

Table 5.28: K-means clustering results for specific values of *k* for POS *t n*-grams

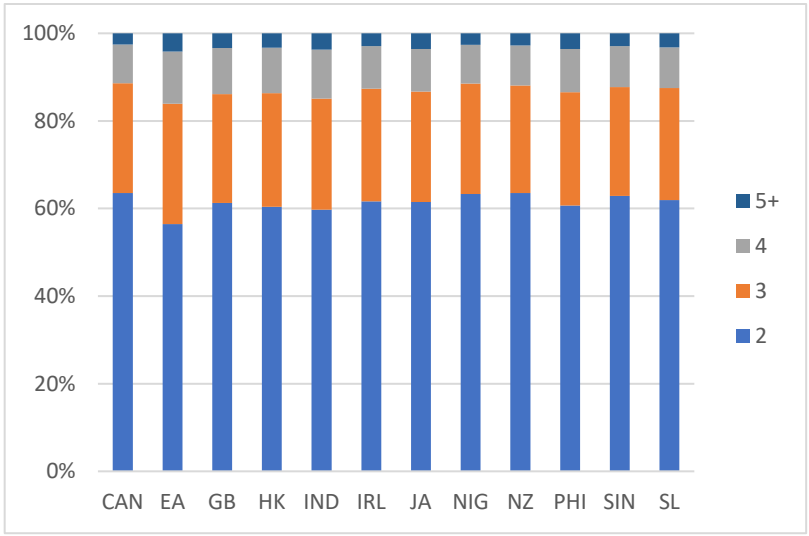
	ALL (*k*=2)		SPK (*k*=4)		WRT (*k*=4)
1	All spoken corpus parts	1	CAN, GB, IRL, NZ	1	HK
2	All written corpus parts	2	JA, NIG, PHI, SIN, SL	2	CAN, SIN, SL, UG, USA
		3	EA	3	KY, TZ, GH, NIG
		4	HK, IND	4	GB, IND, IRL, JA, NZ, PHI

5.2.3 Log Likelihood

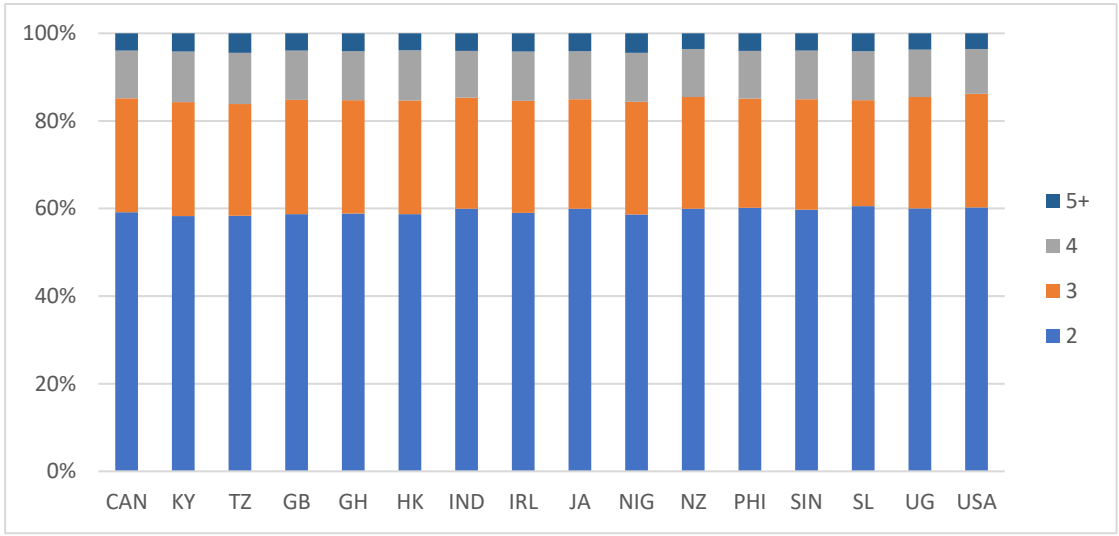
An average of 249,736 (*s*=12,989) and 160,448 (*s*=7,321) tokens (spoken and written, respectively) were produced on the basis of the log-likelihood measure, instantiated by 17,322 (*s*=1,482) and 12,897 (*s*=380) types. This results in mean TTRs of 14.5 (*s*=1.1) and 12.4 (*s*=0.5) and represents decreases in token frequencies from the bi-gram data of -36% and -38% but increases in type counts by +612% and +344%, G^2 thus again presenting itself as an intermediate measure between *MI* and *t*, but leaning more towards numbers found for the latter measure. Mean lengths are found to be almost identical for tokens (2.56, *s*=0.04 vs. 2.61, *s*=0.01) and only slightly diverging for types (3.80, *s*=0.06 vs. 3.67, *s*=0.04), also tending more towards the findings for *t*. Distributions of lengths (Figure 5.43) also show largely similar relative frequencies in case of tokens, but more 2-gram types, particularly in writing, at the expense of the longest sequences (6-grams being the longest of relevant frequency). Also, the same shift of relative proportion of sequences of $n \geq 4$ from a small minority (10-15%) to the majority of types can be observed, but is slightly less strong than for *t*.

During merging of the varietal datasets, no combination of any two triggers a drastic decrease in the frequency of shared sequences, and only positive outliers are detected in the written data effected by the mergers of GB with IRL and NZ (Figure 5.44). The resultant data constitute the second-largest *n*-grams set of any of the POS-annotated data, surpassing the *t*-based set but remaining smaller than that for ΔP. Reductions in item counts amounted to a relative -88% and -85%, resulting in 2,142 spoken and 1,977 written sequences.

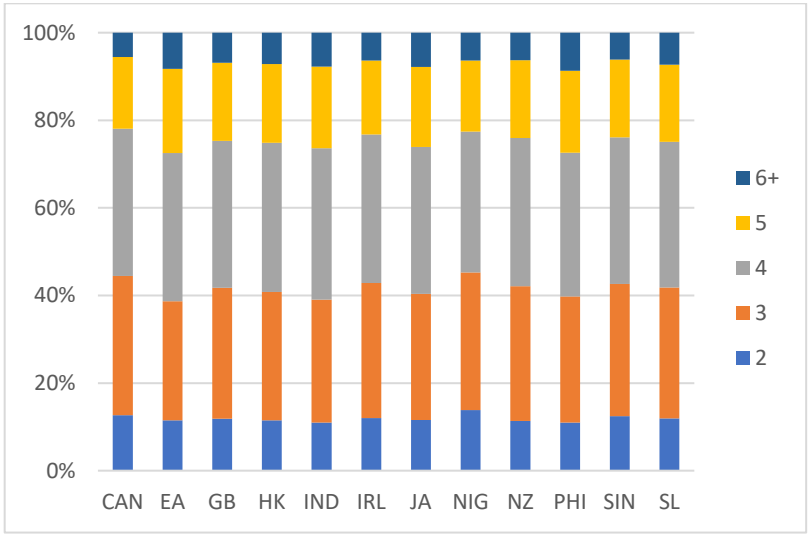
Again, substantial numbers of longer sequences can be found in the shared datasets, particularly at length 4 but also above (Table 5.29). The most strongly collocated items (Table 5.30) again display the inflated association scores previously observed for lexical association under the present measure. Highly collocated sequences appear to more commonly contain some form of verb of more frequent tag types, while rarer POS annotation or such more restricted in scope (e.g. VBDZ representing *was*) is found in higher frequency for lower-scoring items. Moreover, the strongest collocational sequences display a predominance of 3-grams over the otherwise more regular 2-grams.



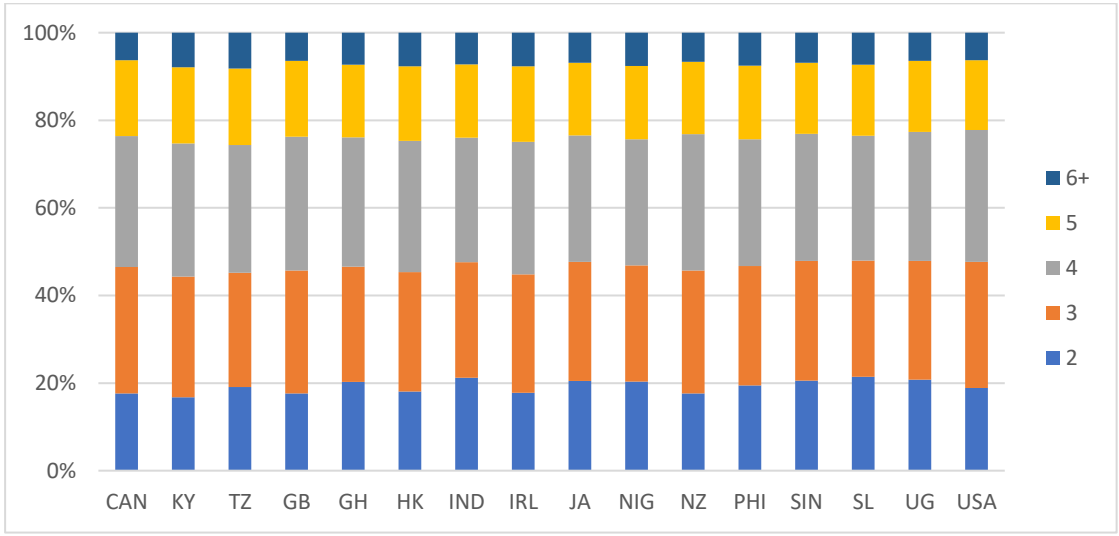
Token frequencies: Spoken data



Token frequencies: Written data



Type frequencies: Spoken data



Type frequencies: Written data

Figure 5.43: Distribution of POS G^2 *n*-gram lengths across the varietal datasets

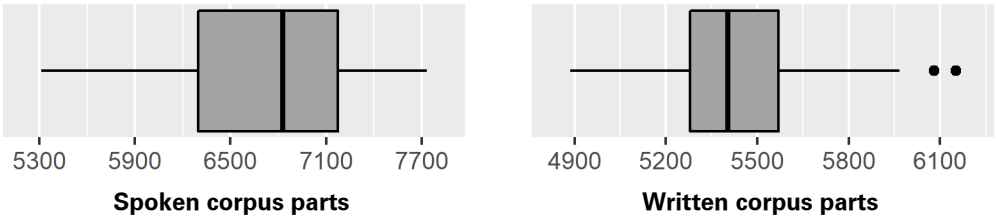


Figure 5.44: Number of shared POS G^2 *n*-grams between any two datasets

Table 5.29: POS G^2 *n*-gram type frequencies by length in the intersects of the variety-specific datasets

N-gram length	ALL	SPK	WRT
2	544 (42%)	780 (36%)	684 (35%)
3	454 (35%)	830 (39%)	734 (37%)
4	200 (15%)	399 (19%)	412 (21%)
5	80 (6%)	130 (6%)	137 (7%)
6	3 (0%)	3 (0%)	10 (1%)
total	1,291	2,142	1,977

Table 5.30: POS G^2 *n*-grams with highest and lowest association scores

SPK		WRT	
***n*-gram**	**G^2**	***n*-gram**	**G^2**
to VVI	50856.20	to VVI	42029.78
AT NN1	40103.78	AT NN1	24742.37
VVGK to VVI	28043.32	VVN to VVI	22385.52
JK to VVI	26804.34	JK to VVI	21828.00
VVN to VVI	26524.08	order to VVI	21381.95
VMK to VVI	26210.45	VVGK to VVI	21341.98
VBDZ MC1	30.40	if DD1	26.23
NN2 at	28.57	of PNQO	26.21
VVI RRR	26.64	NN2 concerning	24.89
DA1 for	24.27	VV0 from	24.82
XX VDN	22.56	VV0 into	22.69
NN1 among	20.18	VVI RRR	22.58

Hierarchical clustering (Figure 5.45) finds a relatively large number of stable groups but restricted overlap between datasets and clusters not conforming to any apparent pattern. Moreover, the spoken/written distinction fails to materialize, potentially due to the (unsubstantiated) allocation of EA_{SPK} to written KY+TZ+GH. This partial African cluster finds substantiation in ALL as well as WRT, as do IC_{GB}(+)JA+NIG, HK+SIN and PHI+SL. The latter two are always clustered with CAN and USA which form a separate subcluster in WRT. The spoken data are more heterogenous, jointly supporting only a strange CAN+JA+NIG+SIN group (with one or two subclusters) while ALL merges these with JA to an equally nonsensical group, and also merges IC_{GB}+SL.

Significant jumps (Figure 5.46) only support up to k=3 for ALL, separating the spoken+written (but incomplete) African group from speech and writing. For SPK, the single stable CAN+JA+NIG+SIN cluster is separated from IC_{GB}+SL, IND+PHI and unary EA and HK at the first significant k=5. WRT finds significant above-average

jumps between *k*=3 and *k*=6, consecutively splitting off the only non-substantiated varieties (IND, UG) before arriving at either the coarser or finer stable clusters.

The strange clusters obtained above find support in the NeighborNets (Figure 5.47), which also obtain only little meaningful structure except for HK+SIN (found close to USA) and some separateness of the African varieties in both modes. Furthermore, many varieties are found at relative mutual equidistance, resulting in mostly boxy shapes.

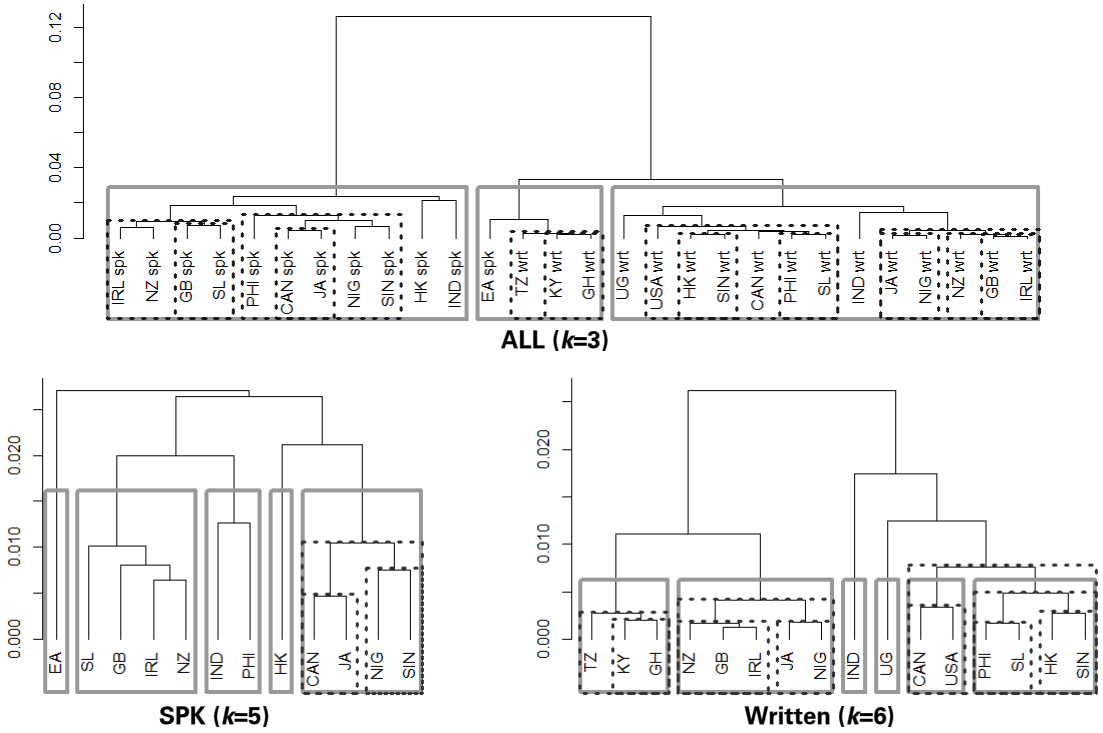


Figure 5.45: Hierarchical clustering results for POS G^2 *n*-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

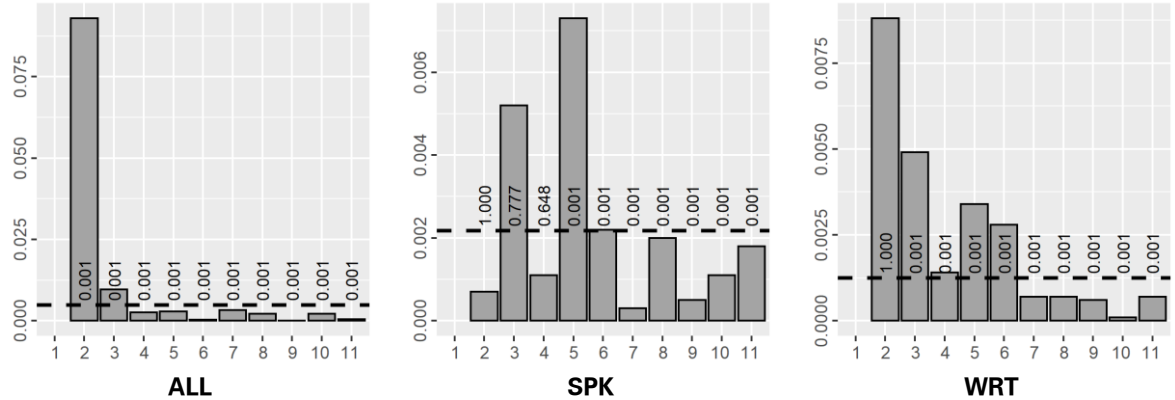


Figure 5.46: Jumps in node heights and respective *p*-values for POS G^2 *n*-grams

K-means (Figure 5.48 and Table 5.21) arrives at yet another segmentation, indicating a clear preference for *k*=2 for ALL (at *k*=3 splitting of EA+HK+IND from the spoken

branch), and returning equally implausible segmentations for SPK at *k*=4 and WRT at *k*=3. At the slightly less indicated values of 3 and 4, respectively, SPK merges GB with HK+IND and the remains of cluster #4 with #3, while writing clusters UG with USA. Overall, G^2 POS *n*-grams present a confounding case of relative stability within methods (`pvclust` and results at different values for *k*) but divergence between methods, and furthermore producing linguistically nonsensical results. It may that this is an effect of the relative equidistance observed above leading to varying results over methods.

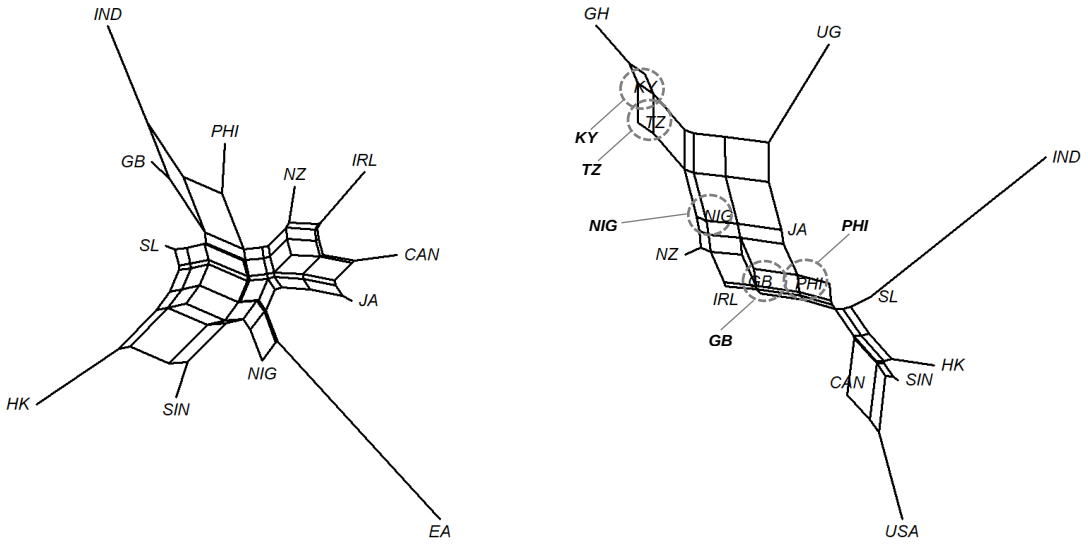


Figure 5.47: NeighborNets of the spoken and written data for POS G^2 *n*-grams

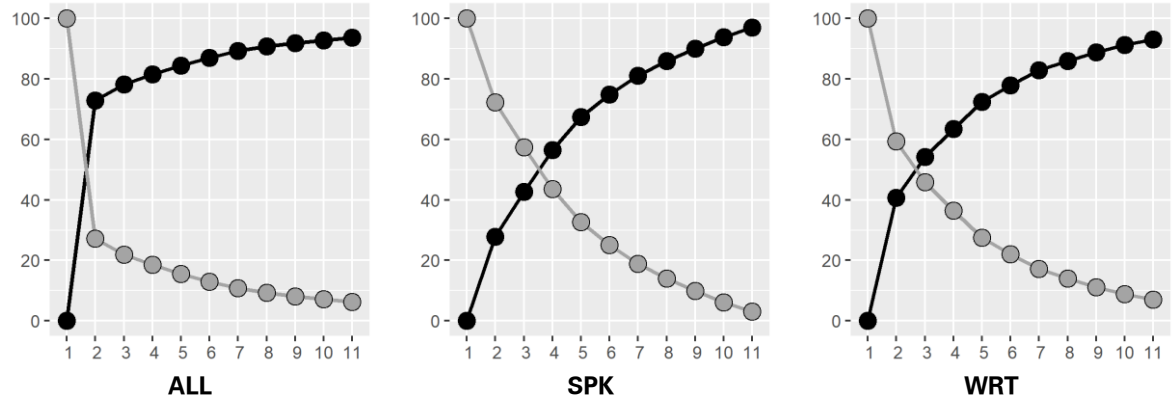


Figure 5.48: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS G^2 *n*-grams

Table 5.31: K-means clustering results for specific values of *k* for POS G^2 *n*-grams

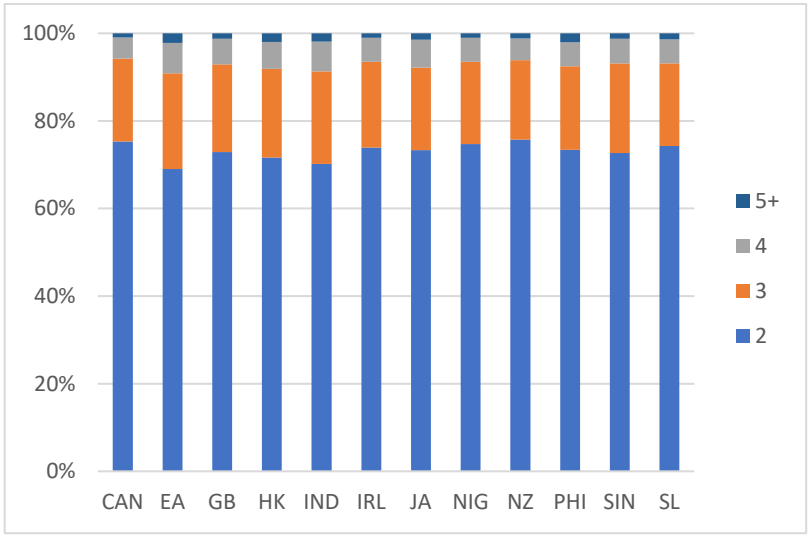
	ALL (*k*=2)		SPK (*k*=4)		WRT (*k*=3)
1	All spoken corpus parts	1	EA	1	GB, HK, IRL
2	All written corpus parts	2	HK, IND	2	CAN, IND, PHI, SIN, SL, UG, USA
		3	CAN, NIG, PHI, SIN, SL		
		4	GB, IRL, JA, NZ	3	KY, TZ, GH, JA, NIG, NZ

5.2.4 Lexical Gravity

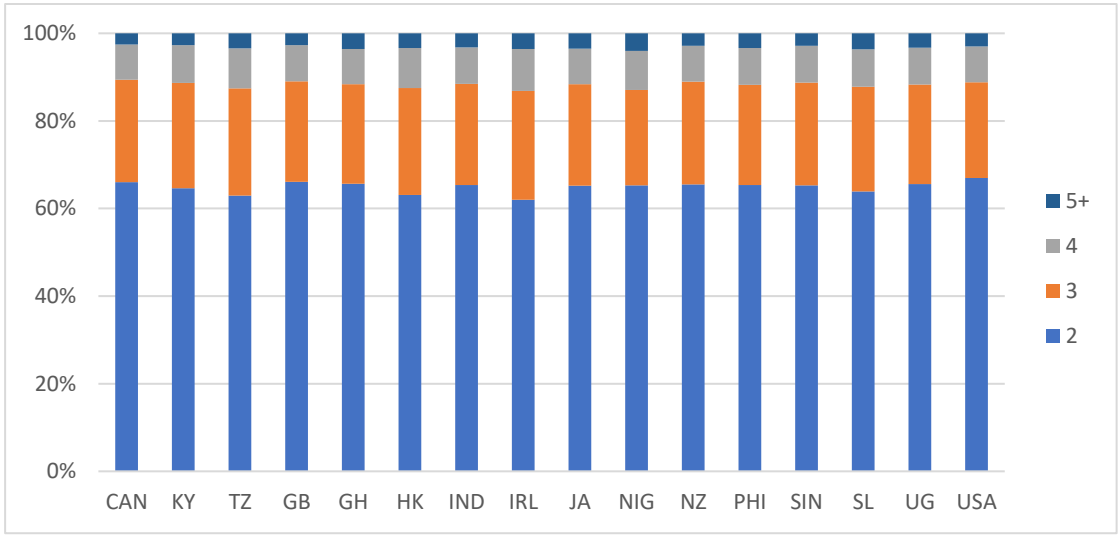
Application of the merging procedure to the lexical-gravity data led to an average of 189,466 (*s*=13,434) spoken and 129,971 (*s*=6,220) written tokens being generated, which were of 1,737 (*s*=421) and 2.372 (*s*=264) types. This represents decreases of token frequencies slightly below those of other measures (except *MI*) by -27% and -34%, but enormous increases in type frequencies by +934% and +1525%. This resulted in the largest average TTRs of 114.1 (*s*=21.4) and 55.4 (*s*=6.3), surpassing those of *MI* (the second measure strongly affected by its threshold). Unusually high standard deviations across varieties point to very different reactions of the datasets to the lexical gravity measure. Mean lengths were found noticeably lower for speech than for writing, with average tokens at 2.36 (*s*=0.03) and 2.51 (*s*=0.02) units and types diverging even more strongly (4.41, *s*=0.23 vs. 4.84, *s*=0.13). Distributional analysis (Figure 5.49) reveals that the relatively pronounced standard deviations above manifest themselves in irregular proportions of *n*-gram type lengths over varieties: While the token data is relatively uniform in its preference for 2-grams (with slightly higher frequencies for n≥3 in writing), the type data even includes sizeable proportions of 7-grams, which in several cases constitute the more numerous group than 2-grams.

Drops in type frequencies caused by merging of the dataset are found at normal values of -83% and -81%, but absolute frequencies are among the lowest (288 and 455 types). Figure 5.50 confirms that no combination of two varieties causes losses of exceptionally many shared sequences (the positive outlier being spoken HK/PHI).

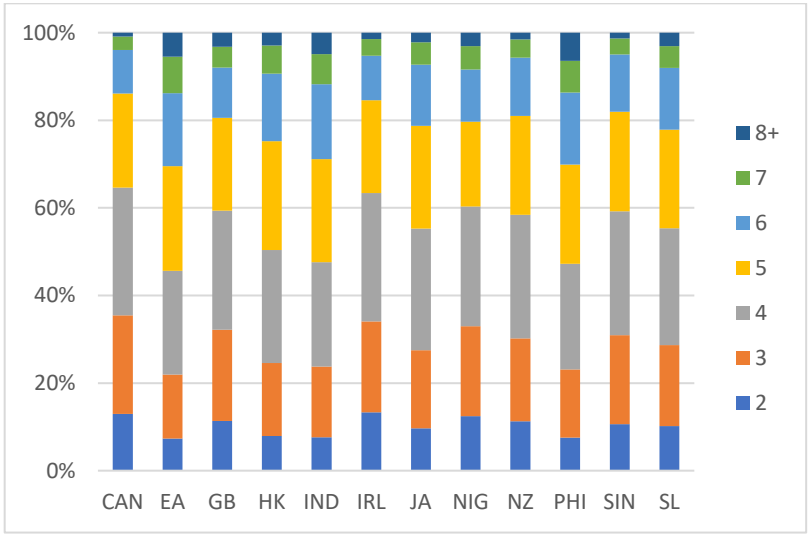
Even given the small size of the data, many longer items are shared between varieties (Table 5.32), with 5-grams and longer sequences more frequent in relative terms than with observed for all other measures. Like in the lexical analysis, *g* appears very successful in producing these longer sequences, and several 4-grams can be found among the top collocates, while bottom items are only derived from 2-grams (Table 5.33). Top collocates mostly revolve around nouns (pre- and post-determination) and contain more high-frequency items (nouns, but also *of* or *in*), demonstrating that the association threshold trims the data to more frequent items. However, differences in association scores between top and bottom *n*-grams are not as pronounced as for other measures.



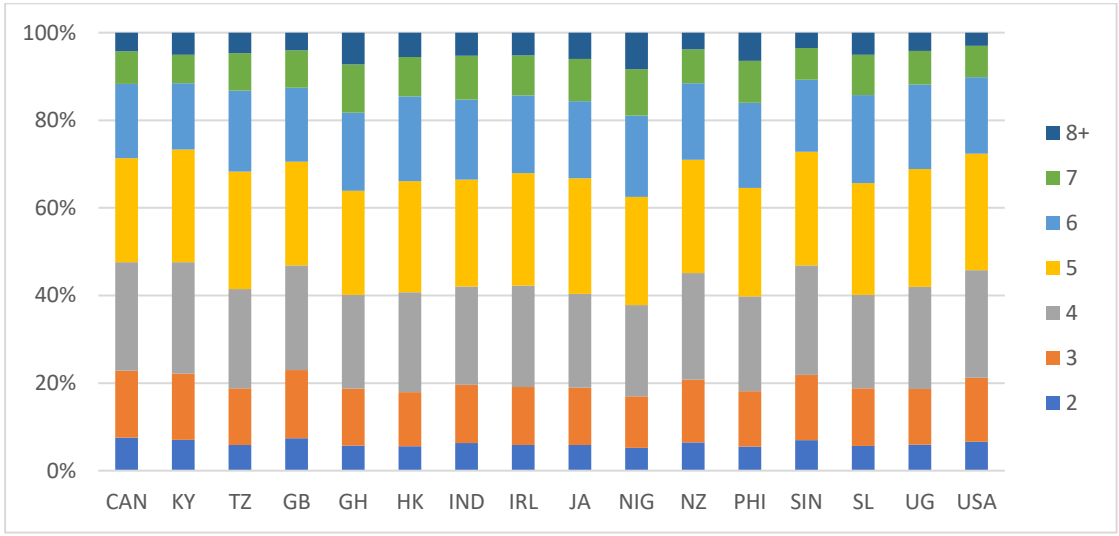
Token frequencies: Spoken data



Token frequencies: Written data



Type frequencies: Spoken data



Type frequencies: Written data

Figure 5.49: Distribution of POS *g* *n*-gram lengths across the varietal datasets

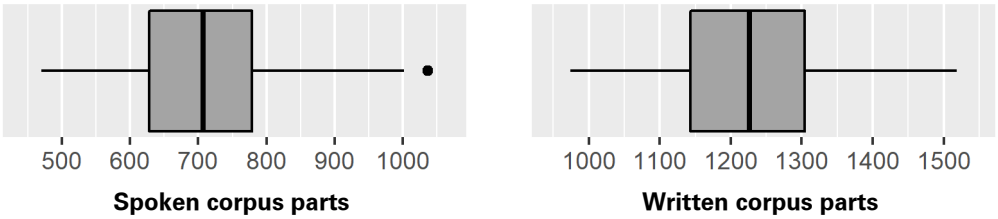


Figure 5.50: Number of shared POS *g* *n*-grams between any two datasets

Table 5.32: POS *g* *n*-gram type frequencies by length in the intersects of the variety-specific datasets

***N*-gram length**	**ALL**	**SPK**	**WRT**
2	64 (31%)	91 (32%)	99 (22%)
3	62 (15%)	89 (31%)	132 (29%)
4	57 (27%)	73 (25%)	156 (34%)
5	22 (11%)	31 (11%)	59 (13%)
6	3 (1%)	4 (1%)	8 (2%)
7	0 (0%)	0 (0%)	1 (0%)
total	208	288	455

Table 5.33: POS *g* *n*-grams with highest and lowest association scores

SPK		**WRT**	
***n*-gram**	***g***	***n*-gram**	***g***
AT NN1	11.17	JJ NN1	12.22
JJ NN1	11.12	NN1 of	11.70
NN1 of	10.59	AT NN1	11.64
to VVI	10.39	AT JJ NN1 of	11.35
in AT NN1 of	10.25	JJ NN2	11.28
of AT NN1 of	10.24	AT JJ NN1	11.18
VV0 AT	6.19	in AT1	6.06
and AT	6.14	VVD RP	6.05
with AT	6.14	VVI to	6.01
RR VV0	6.13	PPX1 PPX2	6.01
VDZ XX	6.12	VVI RP	5.98
VBZ VVG	5.99	VVZ to	5.91

Hierarchical analysis (Figure 5.51) only finds limited support for stable clusters, potentially due to the small size of the dataset (cf. *MI*), even causing the spoken/written distinction to miss significance. IC clusters emerge in both types of written data (IC_{GB} lacking IRL in ALL), as does a KY+TZ cluster. Cluster in speech are, however, relatively different, with only ALL identifying an IC cluster and a further IND+PHI group found in ALL extended considerably in SPK to either IND+PHI+JA+NIG or merged with all varieties except EA.

Segmentation by jump heights (Figure 5.52) only supports up to *k*=4 in ALL, indicating separateness of EA, IND and PHI first after the spoken/written distinction and then partitioning off the IC varieties. SPK vastly prefers *k*=4, also identifying an IC cluster, separateness of (only) EA and partitioning the remaining varieties by no apparent structure. Writing only strongly supports *k*=2, separating IC+HK from OC varieties, and at *k*=3 further indicating a separate status of KY+TZ.

The NeighborNets in Figure 5.53 retrieve more typical structures than the HCA. SPK strongly supports an IC cluster, similar distance of EA to IND as NIG and a HK+SIN group. WRT finds the usual regional separation within the IC cluster as well as similarity within the African subgroups (KY+TZ and GH+NIG in particular), but IND and SL are also found in intermediary positions between a larger African part of the cluster and the remaining varieties.

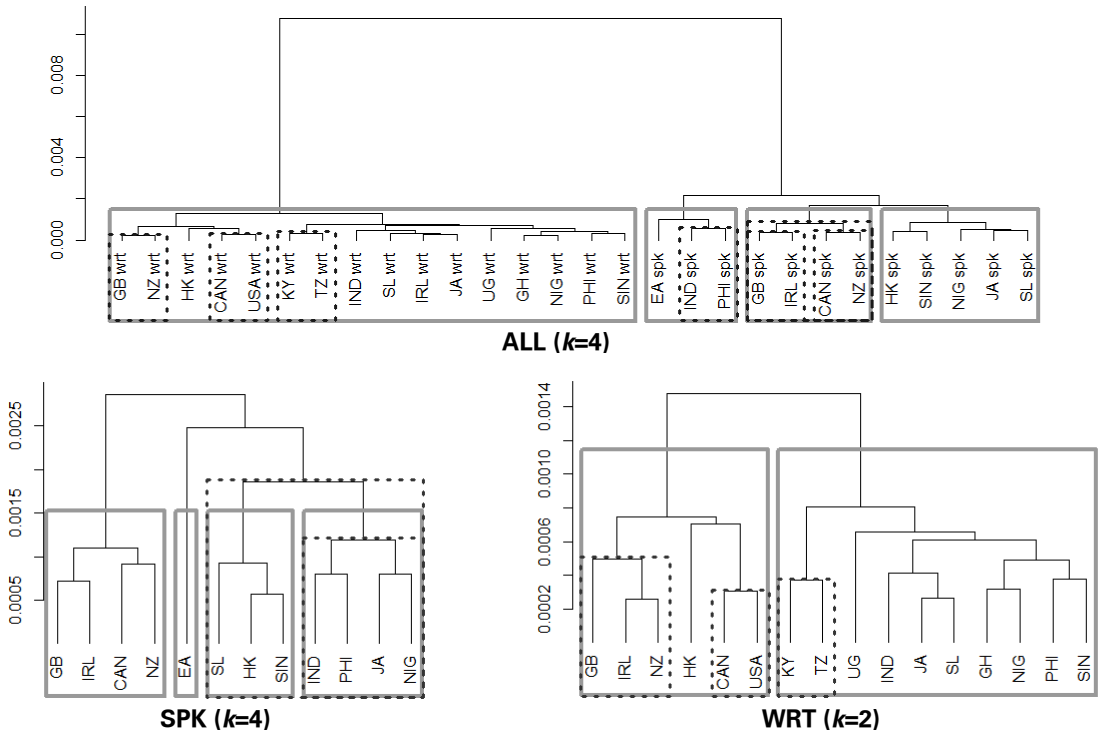


Figure 5.51: Hierarchical clustering results for POS *g n*-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

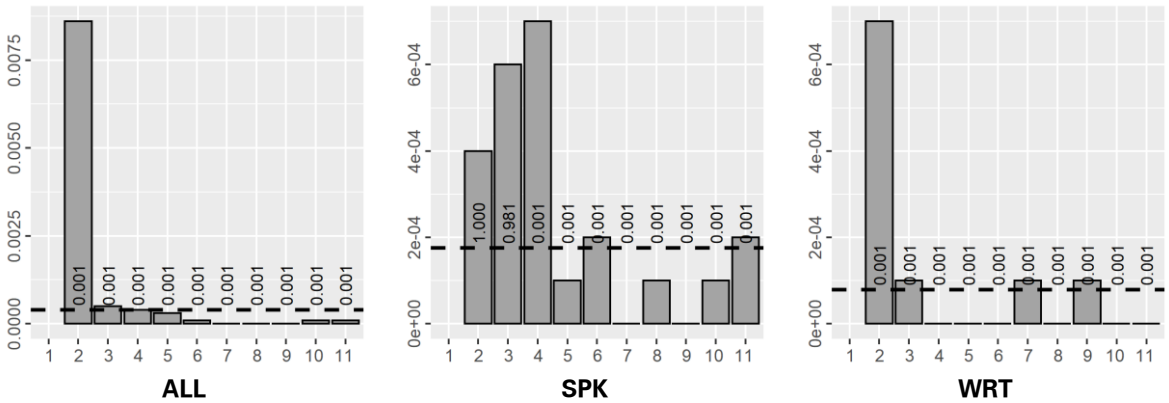


Figure 5.52: Jumps in node heights and respective *p*-values for POS *g n*-grams

Clustering with the k-means algorithm (Figure 5.54 and Table 5.34) leads to the usual *k*=2 for ALL, which, however, introduces a very particular exception from the typical spoken/written distinction in that the EA dataset merges with the written varieties. This is continued for *k*=3 by combining EA_{SPK} with the written forms of relatively

exonormative HK, KY as well as IC-without-IRL, furthering a perspective modeled on exonormativity. For SPK, the more strongly indicated *k*=4 (given the elbow point) is identical to the hierarchical solution except for NIG, and at *k*=3 merges all OC data except EA into a single group. WRT prefers *k*=4, again indicating mutual similarities of KY+TZ as well as of HK to the (incomplete) IC group.

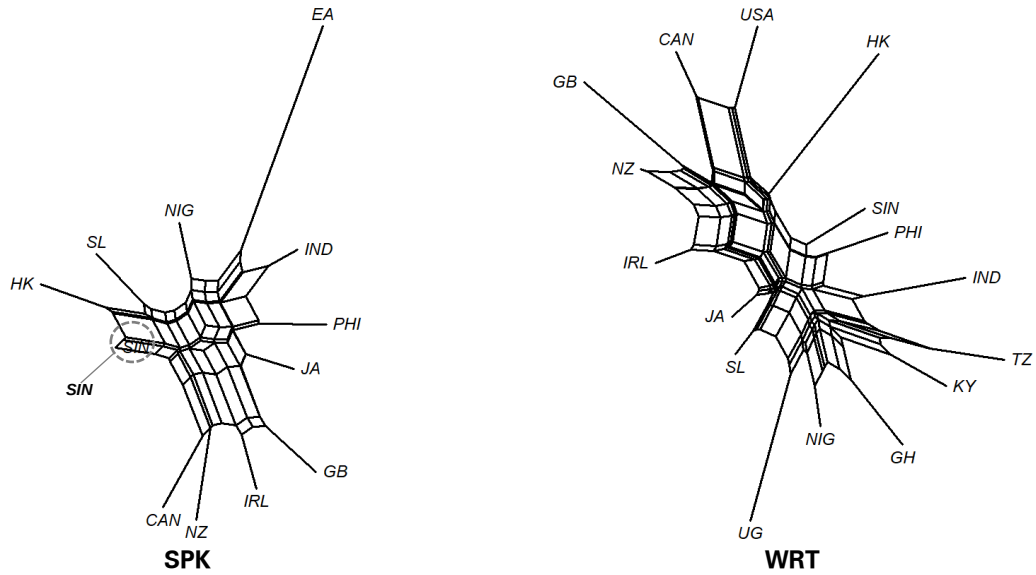


Figure 5.53: NeighborNets of the spoken and written data for POS *g n*-grams

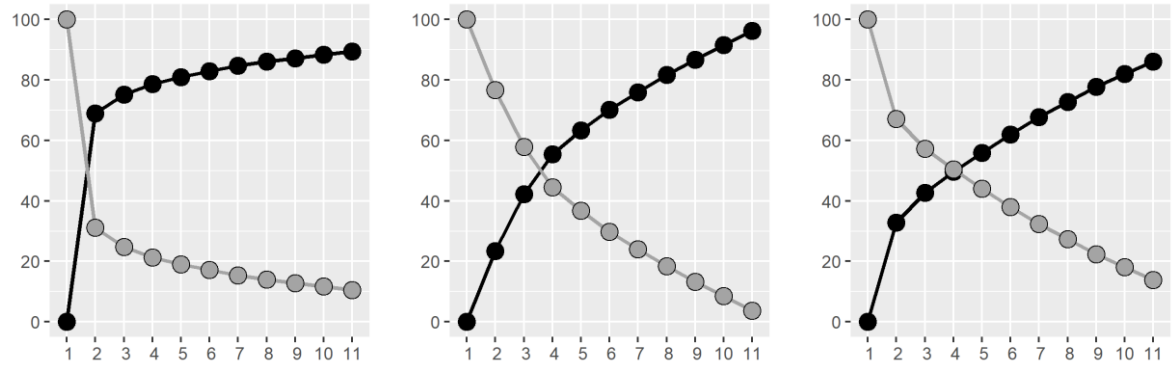


Figure 5.54: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *g n*-grams

Table 5.34: K-means clustering results for specific values of *k* for POS *g n*-grams

	ALL (*k*=2)		SPK (*k*=4)		WRT (*k*=4)
1	All written corpus parts + EA_{SPK}	1	IND, JA, PHI	1	KY, TZ
		2	CAN, GB, IRL, NZ	2	GH, JA, NIG, PHI, SL, UG
2	All spoken corpus parts - EA_{SPK}[79]	3	EA	3	IND, IRL, SIN
		4	HK, NIG, SIN, SL	4	CAN, GB, HK, NZ, USA

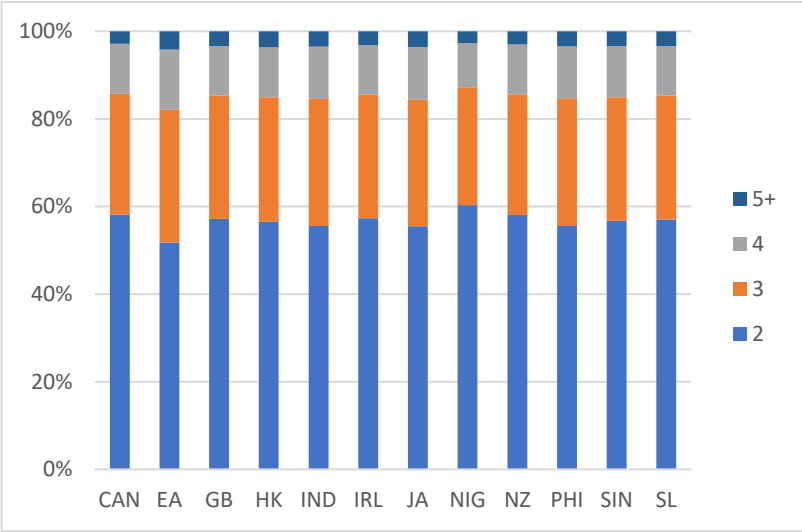
[79] The shorthand form A-B is occasionally employed to provide a label for a mostly coherent cluster A (e.g. "All spoken corpus parts"), from which one variety B is missing.

5.2.5 Delta $P_{2|1}$

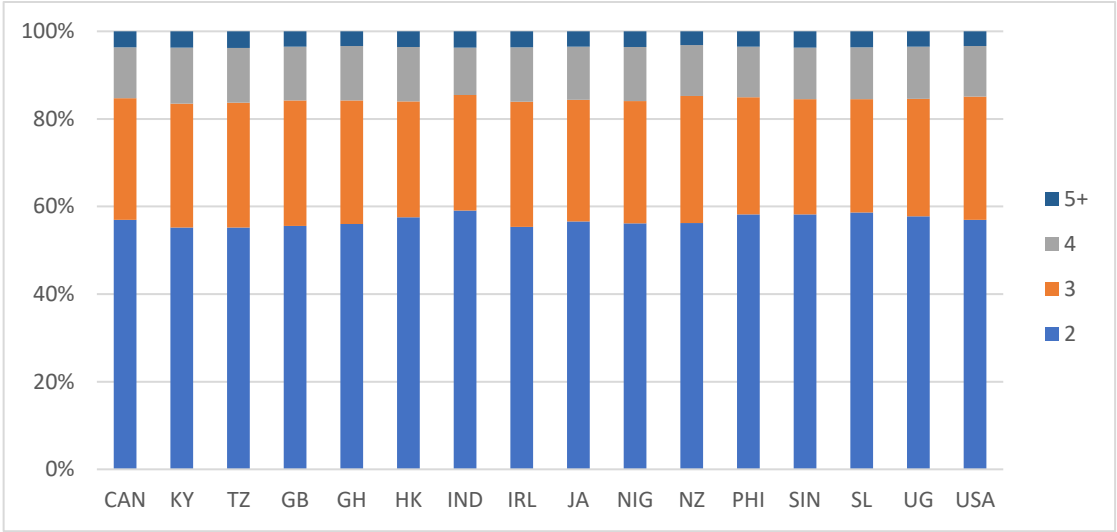
Delta P-based *n*-grams, the data with the least impactful threshold, resulted in the retention and creation of the largest number of tokens, at 257,517 (s=12,601) and 169,193 (s=7,556) items, but in particular produced by far the most types with 24,650 (s=1,618) spoken as well as 17,232 (s=624) written sequences. This represents changes from the bigram data by -38% and -39% for tokens and +370% and +195% for types, resulting in the lowest overall TTRs of 10.5 (s=0.7) and 9.8 (s=0.5). Average tokens were found at identical lengths 2.63 (s=0.04 and s=0.02) but types are longer in speech at 3.75 units (s=0.04) over 3.52 (s=0,04) in writing. While token length distributions consequently are very similar across modes (Figure 5.55), the difference in type lengths shows itself within the distributions in particular through more frequent 2-gram types in writing as opposed to speech. Even given this difference, 2-gram types are among relatively less prominent in both modes, being surpassed in frequency by up to 5-grams in speech and 4-grams in writing.

Losses incurred during dataset merging are on the higher end of the spectrum with -89% and -87% but still retain the largest absolute number of items (2,598 and 2,178 in speech and writing, respectively). Figure 5.56 attests to no single merger exerting a significantly negative effect on retained items; the single positive outlier in the written data is constituted by IRL/NZ.

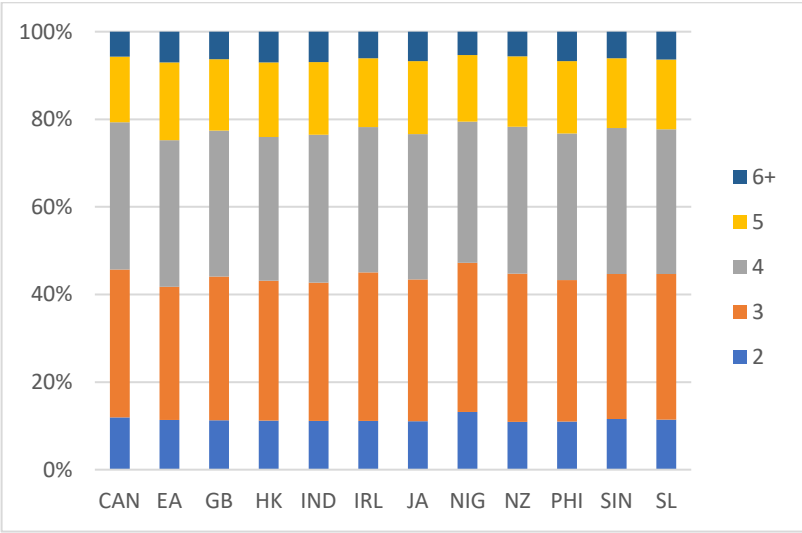
As with previous measures, longer sequences can also be found in the shared *n*-gram data (Table 5.35), but frequencies are negligible for the two longest forms. Again, it shows that larger numbers of items surpassing threshold values do not necessarily lead to substantial relative amounts of longer shared sequences, and thresholds fulfill their task. Top collocates (Table 5.36) are strongly shaped by lexical sequences that of a highly fixed nature in contrast to the more open placeholder categories of the POS annotation and also exhibit the rightward directionality in the measure's design. The only POS tag found among the strongest collocates (JK) represents highly collocated words like *able* or *willing* in 'be able/willing to'.



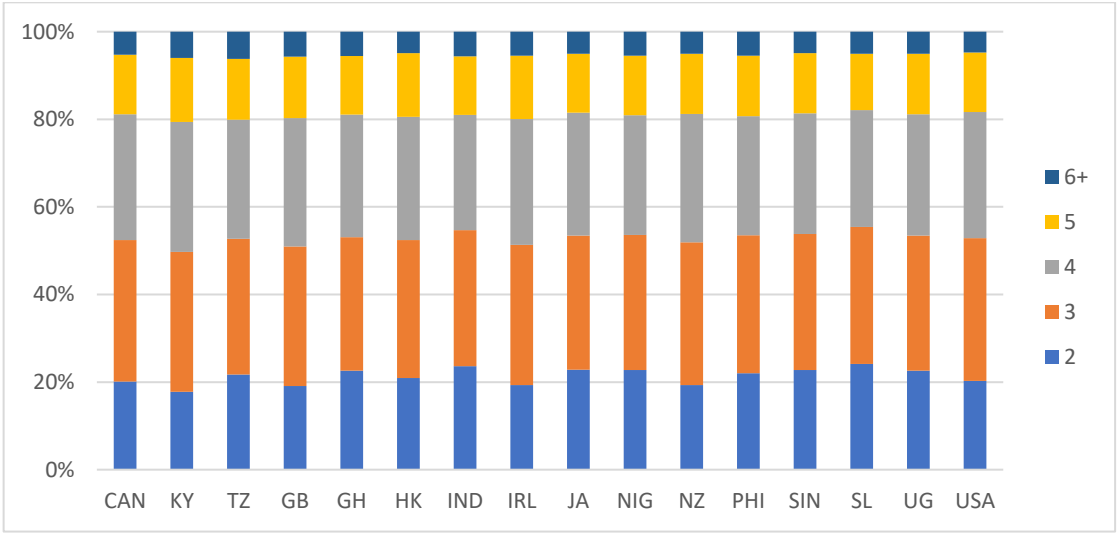
Token frequencies: Spoken data



Token frequencies: Written data



Type frequencies: Spoken data



Type frequencies: Written data

Figure 5.55: Distribution of POS *ΔP* *n*-gram lengths across the varietal datasets

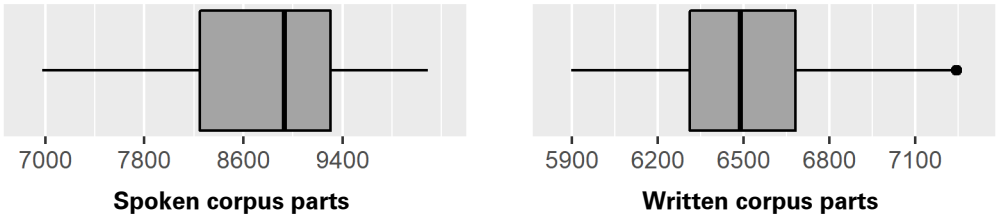


Figure 5.56: Number of shared POS *ΔP n*-grams between any two datasets

Table 5.35: POS *ΔP n*-gram type frequencies by length in the intersects of the variety-specific datasets

N-gram length	ALL	SPK	WRT
2	617 (44%)	901 (35%)	790 (36%)
3	496 (36%)	1,043 (40%)	840 (39%)
4	247 (18%)	568 (22%)	452 (21%)
5	34 (2%)	84 (3%)	95 (4%)
6	0 (0%)	2 (0%)	1 (0%)
total	1,394	2,598	2,178

Table 5.36: POS *ΔP n*-grams with highest and lowest association scores

SPK		WRT	
***n*-gram**	***ΔP***	***n*-gram**	***ΔP***
rather than	0.9992	rather than	0.9987
depending on	0.9944	apart from	0.9949
away from	0.9935	depending on	0.9928
instead of	0.9762	JK to	0.9684
JK to	0.9721	prior to	0.9680
according to	0.9659	according to	0.9659
VV0 DAR	0.0006	NN1 against	0.0004
NN2 over	0.0004	RR since	0.0004
NN1 over	0.0004	of 0	0.0003
JJ MC2	0.0004	NN1 towards	0.0003
NN1 without	0.0003	NN1 within	0.0002
NN1 under	0.0002	NN1 until	0.0001

Hierarchical analysis of the *ΔP*-based data (Figure 5.57) indicate a similar clustering of EA_{SPK} to the written branch as observed for *g*. While the spoken branch is similarly not substantiated, writing achieves significance overall and is thus clearly distinguished from EA_{SPK}. Within SPK, EA+IND is stably contrasted against all other varieties, while ALL only retrieves a single spoken (and incomplete) cluster in IC_{GB}. Both types of written data identify two IC clusters, and WRT additionally reconstitutes the EA_{WRT} (KY+TZ) data.

Significant jumps (Figure 5.58) within ALL vastly prefer a binary partition (EA_{SPK} merged with writing). Successive above-average jumps are only achieved when EA_{SPK} is separated from the written branch at *k*=4 and speech segments into IC and OC before splitting off IND at and *k*=5. and OC in speech. SPK similarly indicates unary nodes for EA and IND and otherwise a separation into Inner and Outer Circles. WRT produces the first significant jump at *k*=5, partitioning the set into an East African

group, IC_{GB} and a more epicentral IC_{NA}+PHI in addition to isolating IND and grouping all remaining (regionally and evolutionarily mixed) varieties together.

NeighborNet analysis (Figure 5.59) indicates isolation of EA, IND and NIG within SPK and also finds CAN and NZ slightly separate from GB+IRL inside their mutual IC cluster (which indicates some shared features of CAN and PHI). HK and SIN are also found to be relatively proximal. In WRT, two distinct IC clusters emerge and similarity between IC_{NA} and PHI is observed. The African subclusters show some mutual distance to the remaining varieties, and IND and SL are found to share some features.

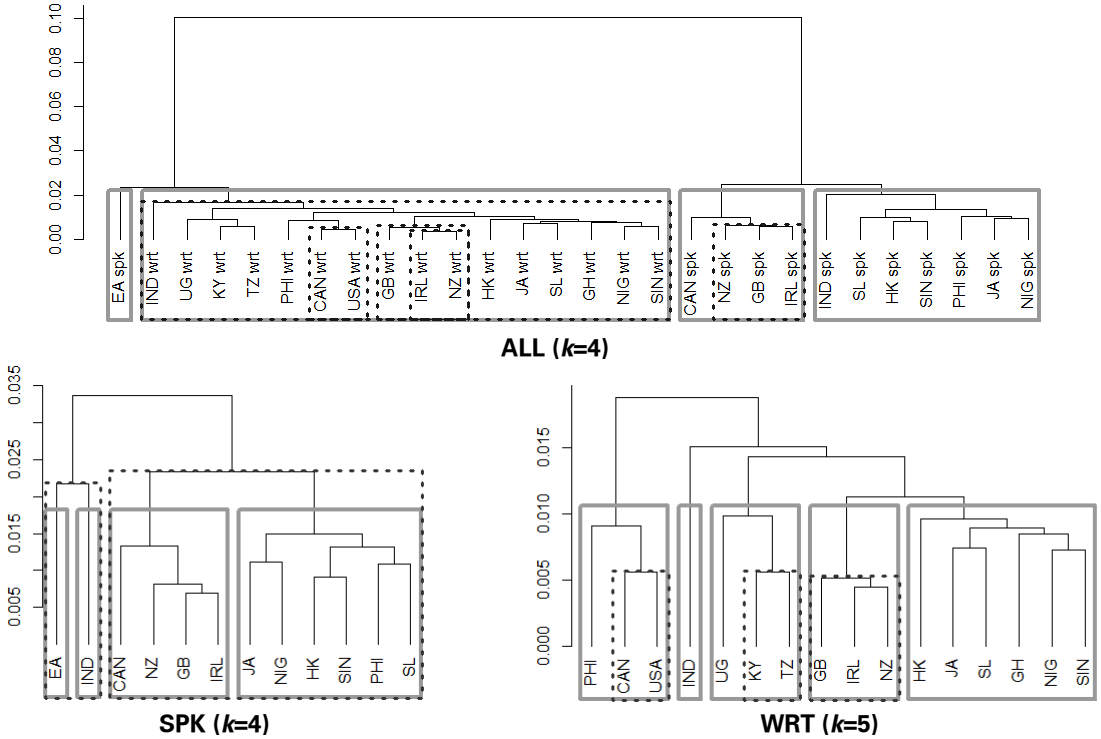


Figure 5.57: Hierarchical clustering results for POS *ΔP n*-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

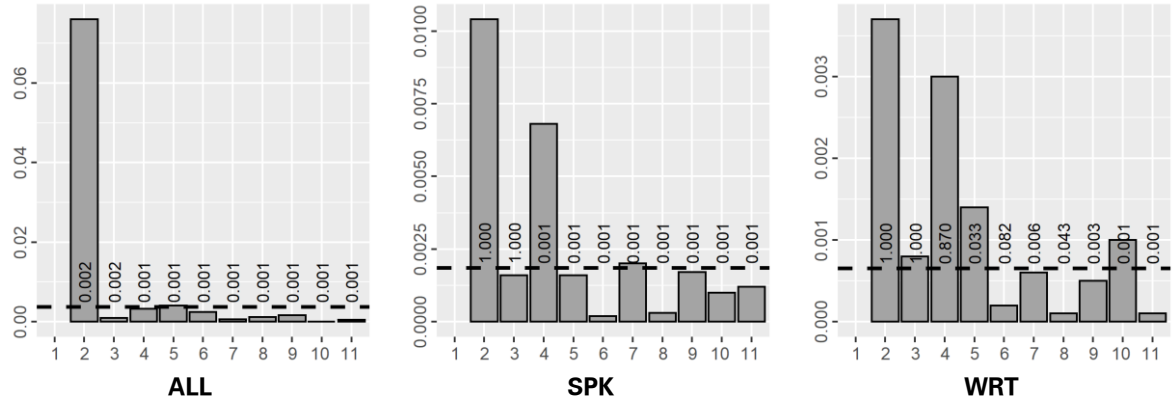


Figure 5.58: Jumps in node heights and respective *p*-values for POS *ΔP n*-grams

K-means clusters (Figure 5.60 and Table 5.37) indicate up to *k*=3 for ALL, separating speech from writing before splitting off EA+IND$_{SPK}$ from the spoken varieties. For SPK, the clearly indicated *k*=4 similarly leads to a separation of EA and IND from the IC and OC data. For WRT, relatively large values are preferred, mirroring the hierarchical analysis at *k*=5 but furthermore creating a second cluster of more exonormative varieties (HK+GH) at the actual intersect at *k*=6.

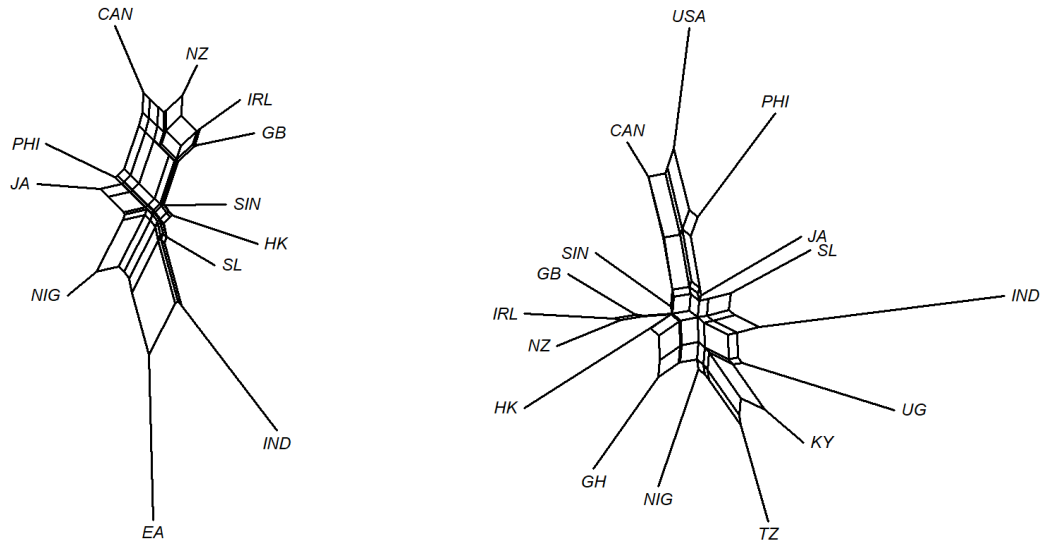


Figure 5.59: NeighborNets of the spoken and written data for POS *ΔP n*-grams

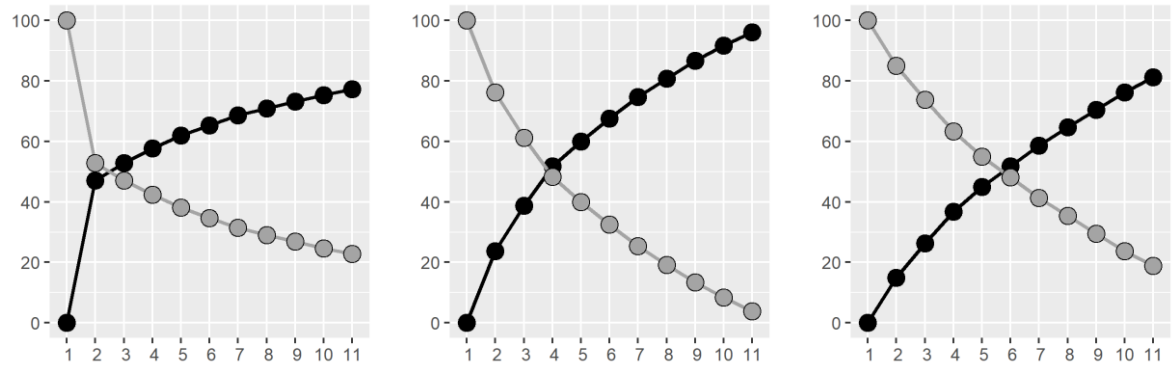


Figure 5.60: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *ΔP n*-grams

Table 5.37: K-means clustering results for specific values of *k* for POS *ΔP n*-grams

	ALL (*k*=3)		**SPK (*k*=4)**		**WRT (*k*=6)**
1	CAN, GB, HK, IRL, JA, NIG, NZ, PHI, SIN, SL $_{SPK}$	1	IND	1	KY, TZ, UG
2	EA, IND $_{SPK}$	2	CAN, GB, IRL, NZ	2	JA, NIG, SIN, SL
3	All written corpus parts	3	EA	3	IND
		4	HK, JA, NIG, PHI, SIN, SL	4	GB, IRL, NZ
				5	GH, HK
				6	CAN, PHI, USA

5.3 Static-length Lexical *N*-grams

Length distributions of dynamic-length *n*-grams identified within the previous analyses can be used as an indicator of sequence lengths most preferred by the varieties under scrutiny within the present study. In the lexical data, the largest proportion of tokens is provided by 2-grams, with relative frequencies above 60% except in case of *ΔP* (where 2- and 3-grams constitute c. 40% each). For types, 3-grams need to be additionally considered in order to account for similar shares of the datasets, at which point >90% of tokens are explained in all but the ΔP data. Further including 4-grams accounts for ≥90% of all types and >95% of tokens in all datasets, while 5-grams form a numerically insignificant category in all cases. Certainly, not all datasets behave identically, and in particular *MI* sticks out as the exception, favoring shorter sequences much more strongly (up to length 3 already accounting for c. 90% of types). It thus appears that items of lengths between 2 and 4 are preferred strongly by the data.

Extracting static-length sequences between 2, 3, and 4 units in length and calculating their frequencies and respective association values results in the average scores represented in Table 5.38. Several aspects of these distributions are interesting and deserve discussion. Firstly, a steady decrease in token frequencies can be identified. Reasons for which have already been discussed in the introductory pages to the previous analyses and concern the smaller share of full 4-word sequences within utterances, in that not all of these even contain four words. In contrast, type frequencies increase across lengths but show a major jump between 2- and 3-/4-grams. As such, 2-grams concentrate on far fewer types, particularly in speech, while the creation of longer sequences appears to enforce variability in sequence types. Average association scores appear to follow these patterns of change across lengths: In the case of tokens, they remain relatively similar overall at different lengths but still show the unidirectional and largely linear pattern observed above. For types, the division between the frequencies of 2-grams and longer sequences also reflects in stronger changes in average scores between lengths 2 and 3 rather than 3 and 4.

Table 5.38: Static-length lexical *n*-gram average frequencies and association values

	Tokens						Types					
n	**Freq.**	***MI***	***t***	***G²***	***g***	***ΔP***	**Freq.**	***MI***	***t***	***G²***	***g***	***ΔP***
Spoken												
2	569,767	3.39	4.16	509.87	3.22	0.0793	193,310	4.64	0.64	14.19	-1.66	0.0924
3	520,468	3.34	4.08	491.33	3.18	0.0802	381,317	3.48	2.46	240.42	1.74	0.0760
4	475,661	3.34	4.00	480.08	3.15	0.0809	438,229	3.34	3.56	407.93	2.81	0.0784
Written												
2	360,880	4.30	3.04	227.19	2.03	0.1086	170,367	5.52	0.86	14.08	-1.28	0.1189
3	316,141	4.14	3.10	233.40	2.09	0.1085	261,820	4.27	2.33	146.67	1.27	0.1052
4	276,973	4.07	3.11	234.42	2.12	0.1082	263,042	4.08	2.93	212.30	1.93	0.1066

Merging (/intersecting) the varietal datasets necessarily removes a great amount of types from the data, resulting in frequencies as shown in Table 5.39. Again, this data provides valuable insights into relevant types of *n*-grams within the present analysis: Firstly, it becomes apparent that consistently more shared sequences are found between spoken components than the written ones. At length 2, the relative differences in number reflect the 60-40 distribution of spoken and written modes in the ICE corpora. At larger lengths, however, these differences are neither proportional to the relative sizes of the spoken and written ICE parts nor do they conform to the numbers of 2-, 3- or 4-grams found above. Instead, writing appears to diversify much more strongly within the national components at greater sequence lengths than spoken language. Secondly, striking drops in shared type frequencies can be observed across lengths. While spoken 2-grams still account for 6.4% of the average spoken corpus part, the number of written 4-grams only represents 0.07% of the types in the written parts. Thus, 4-grams should be taken with the necessary precautions since they only represent a minor aspect of the 'common core' of routinely-used sequences. This also makes a potential analysis of 5-grams a moot point and instead puts emphasis on 2- and 3-grams for the lexical analysis of static-length sequences.

Table 5.39: Lexical *n*-gram type frequencies by length in the intersects of the variety-specific datasets

***N*-gram length**	**ALL**	**SPK**	**WRT**
2	5,863	12,362	8,074
3	1,376	5,575	2,105
4	93	787	174

A final check of the data concerns the number of items in the intersects of any two varietal datasets, since negative outliers can lead to increased losses of types during the merging of datasets. Since only varying association scores of otherwise identical *n*-gram type data will be scrutinized in the analyses to follow (unlike in the variable-length approach), only one such evaluation of pairwise overlap needs to be conducted

at each length. Figure 5.61 indicates the number of mutual overlap of types between any two varietal datasets in speech as well as in writing. It can be seen that 4-grams show the lowest number of shared sequences between any two varieties, which in turn leads to the very low frequencies observed above. This is true for both speech and writing, but writing is again shown to lose types more quickly than speech, with stronger relative losses at each stage. Outliers only emerge for written 3- and 4-grams and mostly concern positive outliers: GB/IRL at length 3 and KY/TZ at both lengths. While negative outliers are regarded as a warning in the present framework, the single one obtained for 3-grams (IND/USA) will be ignored for the present analysis on the grounds that it only barely forms an outlier, is not found as a systematic feature at different lengths and was not returned in any of the analyses before.

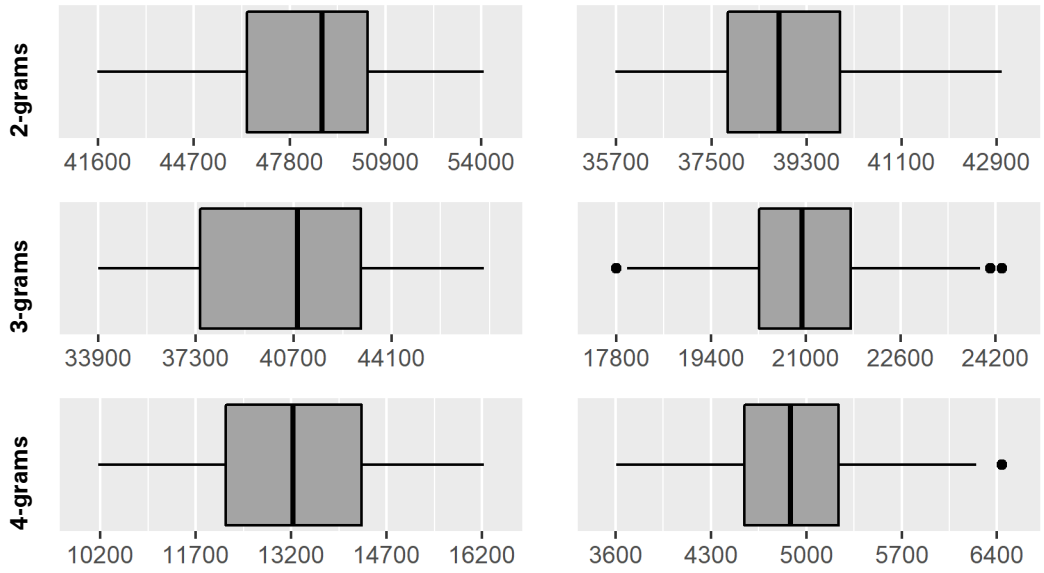


Figure 5.61: Number of shared lexical 2-, 3- and 4-grams between any two datasets

5.3.1 MI-score

N-gram association scores on the basis of *MI* (Table 5.40) show the measure's strong preference for fixed, specialized terms and rare items. This is particularly pronounced for shorter sequences, and consecutively lessened for longer ones. While many top 3-grams constitute extended forms of the shorter sequences, this is far less the case for combinations of four items. None of the top spoken 4-grams and only two written types include types from the lists of shorter sequences. On the bottom of the spoken list, the measure places potential cases of false starts, repetitions, and other production errors. The lowest scores in writing, in contrast, are awarded to a relatively large number of subjectively good collocates. Their low scores can, however, be explained as a consequence of the low-frequency bias of *MI*, which disfavors components words of higher frequencies. As such, the lowest-ranking sequences are almost completely constructed from high-frequency items.

Table 5.40: Lexical *MI n*-grams with highest and lowest association scores

2-grams type	*MI*	3-grams type	*MI*	4-grams type	*MI*
Spoken					
civil servants	12.84	ten years ago	7.95	a little bit more	6.34
nineteenth century	12.58	the twentieth century	7.69	i beg your pardon	6.12
human beings	12.14	the united nations	7.41	a little bit of	5.72
twentieth century	11.99	the united states	7.36	in the united states	5.63
armed forces	11.76	the united kingdom	7.20	should be able to	5.62
prime minister	11.31	we're talking about	7.18	will be able to	5.54
the and	-6.19	to you the	-2.11	for the for the	-0.48
the to	-6.21	of of of	-2.21	of the of the	-0.54
a the	-6.25	would that be	-2.39	those of you who	-0.81
the is	-6.49	to the to	-2.77	and the the the	-1.14
i the	-6.84	and the and	-3.00	the the the the	-1.80
i to	-6.88	the and the	-3.00	and the and the	-1.94
Written					
et al	12.42	90 per cent	8.97	a wide variety of	5.82
twentieth century	12.23	two years ago	7.22	should be able to	5.43
19th century	11.21	the united kingdom	7.10	will be able to	5.29
raw materials	11.04	per cent of	6.97	per cent of the	5.23
united kingdom	10.95	the united states	6.90	have been able to	5.23
20th century	10.87	i am sure	6.85	in the united states	5.21
the will	-4.86	and to the	-0.72	the start of the	1.08
in to	-4.91	to which the	-0.76	the time of the	1.08
it the	-5.01	and that of	-1.10	at the back of	0.80
to and	-5.23	as is the	-1.34	the back of the	0.69
to in	-6.25	to that of	-2.00	from time to time	0.48
of and	-6.63	in and out	-2.75	with the help of	0.28

2-gram preferences calculated on the basis of *MI* amalgamate into several smaller clusters mostly according to a regional analysis in the hierarchical approach (Figure 5.62): In addition to returning stable differences between speech and writing in the ALL dataset, IC_{GB} is picked out in both branches, as is the African cluster in writing and Southeast-Asian HK+SIN in speech. Most of the results for ALL's written branch are supported in WRT (except for UG). For speech, a more comprehensive segmentation is found stable through a formation of complete IC and OC groups and differentiation of EA+IND from the latter (NIG occupying a somewhat separate position).

Analysis of jump heights (Figure 5.63) in ALL returns a four way split as the earliest significant segmentation, separating spoken and written IC (+JA) from OC varieties. A moderately large significant jump is found at k=7, segmenting the OC into Africa and Asia in writing and separating EA+IND and NIG from the remaining OC group in speech. For SPK, k=5 results in the first significant jumps, again separating EA, IND and NIG from the OC varieties. Note, however, that the two subsequent values show even larger jumps and may thus be preferred. This would separate the remaining OC cluster into a HK+SIN cluster and a counterintuitive JA+PHI cluster by splitting off SL, all the while retaining the IC cluster. WRT only finds jumps significant at k=6, which subdivides both the IC and African clusters and splits off IND from the OC varieties.

The NeighborNets (Figure 5.64) show only very small distances between most varieties. Still, an IC group can be confirmed in SPK, as is difference of EA, IND and NIG to the remaining data. WRT supports two proximal IC groups and separates Africa from the data, together with IND and SL.

K-means (Figure 5.65 and Table 5.41) prefers a fine segmentation. This coincides with the HCA, in that SPK is identical to a HCA at k=6 (significant but smaller jumps at k=5 merge clusters #2 and #5 to an Asia-without-IND cluster), and WRT only produces an additional HK node. The ALL data, however, differs from the HCA, and only varieties within stable clusters are confirmed. The spoken part identifies IC+JA, and supports the difference between unary EA, IND and NIG and other OC varieties. For writing, the African cluster and the smaller IND+SL confirm a regional perspective, and grouping PHI with CAN+USA may also be regarded as support for this. However, the remaining varieties form a less interpretable group not previously encountered.

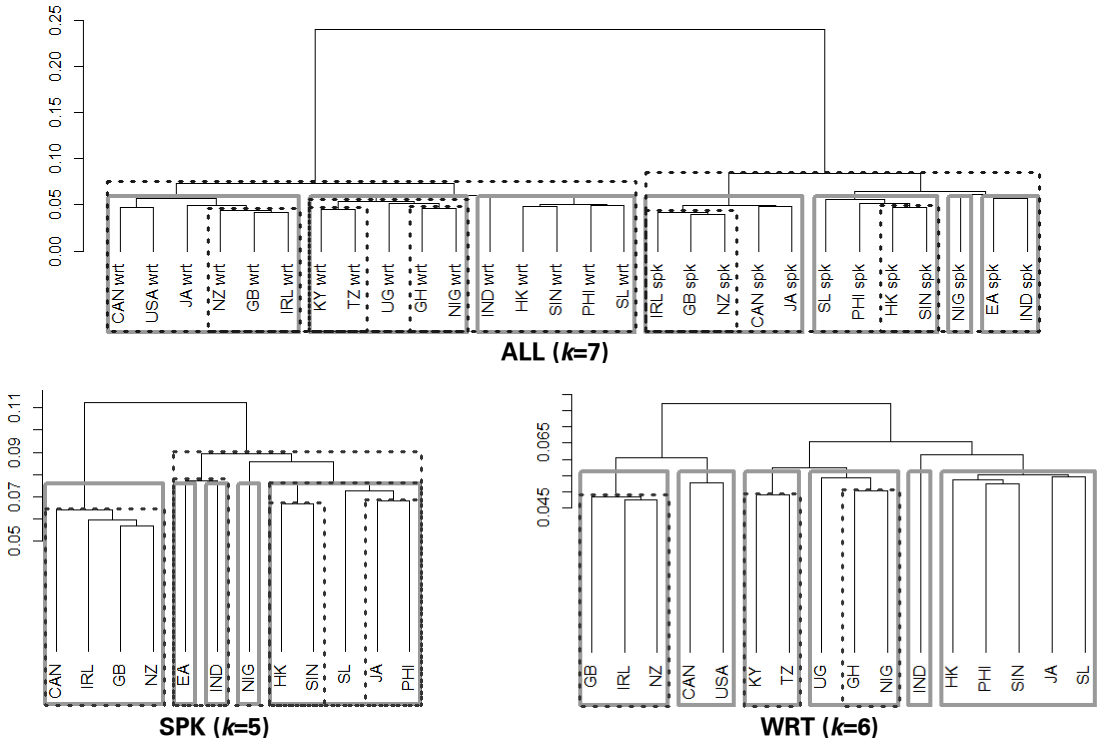


Figure 5.62: Hierarchical clustering results for lexical *MI* 2-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

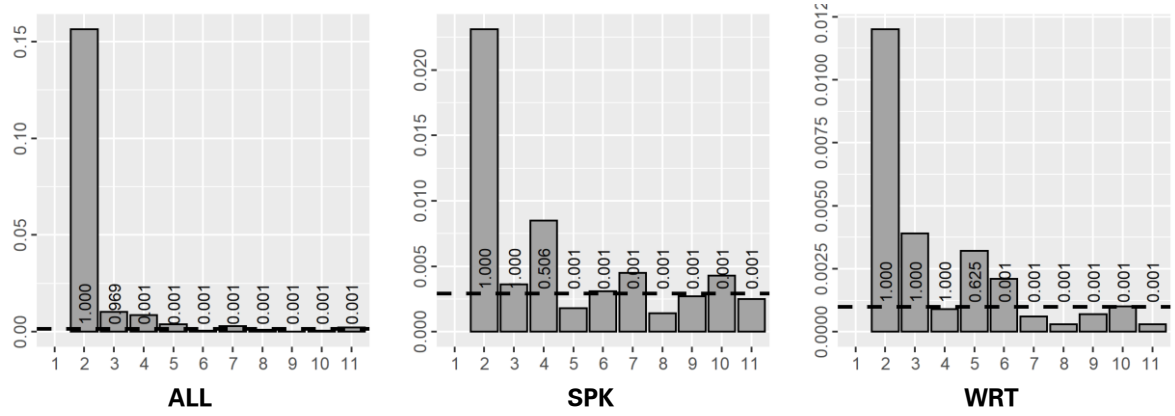


Figure 5.63: Jumps in node heights and respective *p*-values for lexical *MI* 2-grams

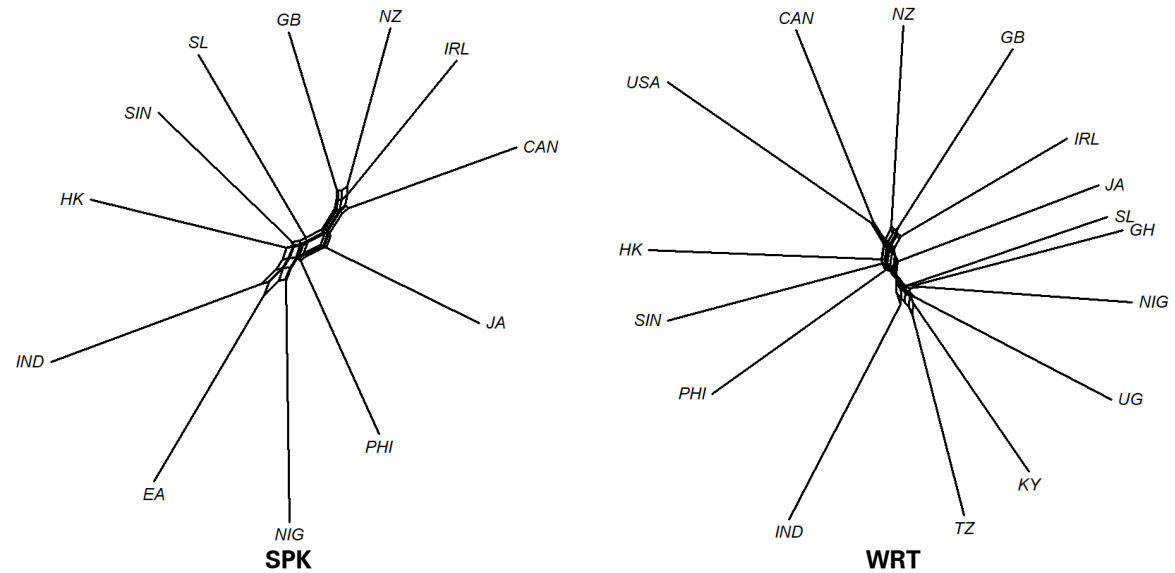


Figure 5.64: NeighborNets of the spoken and written data for lexical *MI* 2-grams

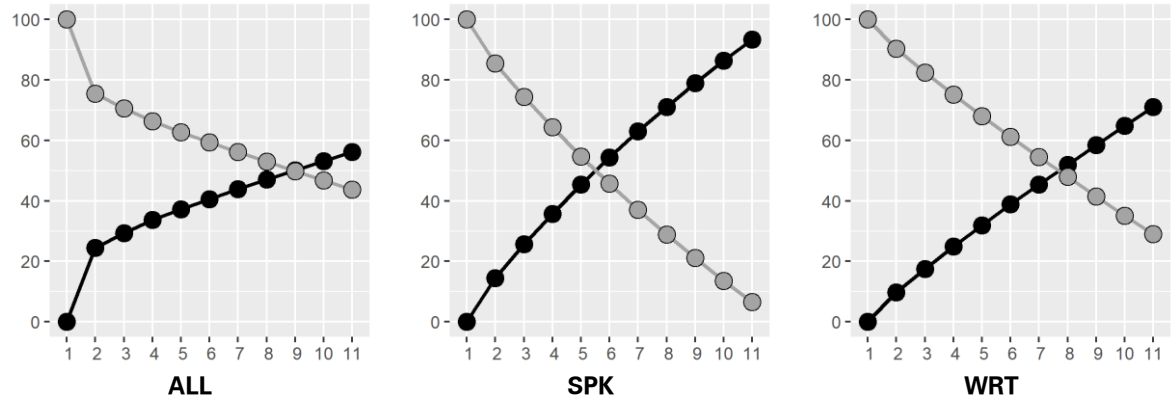


Figure 5.65: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *MI* 2-grams

Table 5.41: K-means clustering results for specific values of *k* for lexical *MI* 2-grams

	ALL (*k*=9)		SPK (*k*=6)		WRT (*k*=8)
1	NIG $_{SPK}$	1	CAN, GB, IRL, NZ	1	KY, TZ
2	EA $_{SPK}$	2	HK, SIN	2	JA, PHI, SIN, SL
3	IND $_{SPK}$	3	IND	3	CAN, USA
4	HK, PHI, SIN, SL $_{SPK}$	4	EA	4	IND
5	CAN, GB, IRL, JA, NZ $_{SPK}$	5	JA, PHI, SL	5	UG
6	GB, HK, IRL, JA, NZ, SIN $_{WRT}$	6	NIG	6	GH, NIG
				7	HK
7	IND, SL $_{WRT}$			8	GB, IRL, NZ
8	KY, TZ, GH, NIG, UG $_{WRT}$				
9	CAN, USA, PHI $_{WRT}$				

Analysis of MI-score **3-grams** continues to single out EA, IND and NIG as varieties separate from the remaining data. HCA (Figure 5.66) finds stable clusters in the spoken/written branches of ALL as well as in an OC (vs. IC) spoken group. This distinction is also found in SPK, and EA, IND and NIG are all removed from the remaining OC varieties. SPK further substantiates the IC group not found stable in ALL and provides further indication towards subdividing the remaining OC varieties into HK+SIN and JA+PHI+SL. Writing separates an African group in both datasets, with an additional Asian cluster in ALL and two IC clusters in WRT, respectively. IND is separated from other Asian OC in both sets of data.

Substantiated cuts (Figure 5.67) can be performed in ALL at *k*=3, separating spoken IC and OC, while larger values immediately section off EA and IND and also NIG after reaching the height of the stable African written cluster. While later *k*=8 presents an unusually fine-grained segmentation of ALL, its results for the spoken branch are replicated within SPK at the first significant jump at *k*=5. For WRT, *k*=5 is similarly indicated and results in groups largely in line with those found in ALL's written branch above, segmenting two OC (+JA) and an African cluster and separating IND from remaining the OC.

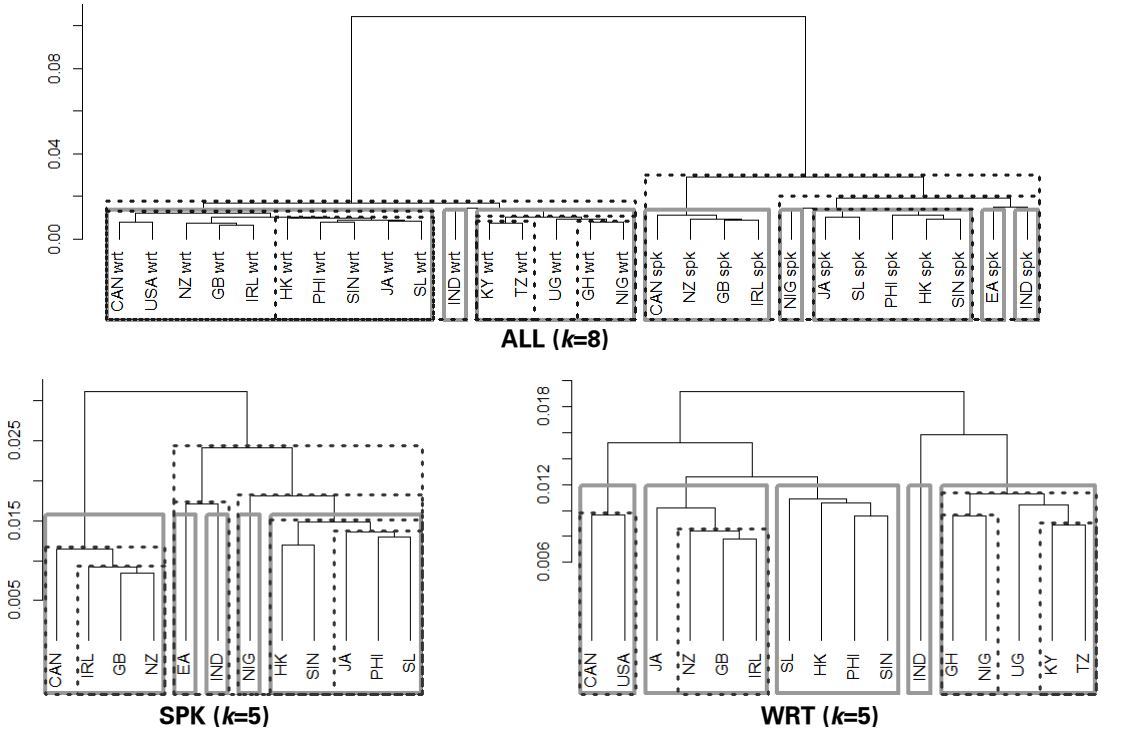


Figure 5.66: Hierarchical clustering results for lexical *MI* 3-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

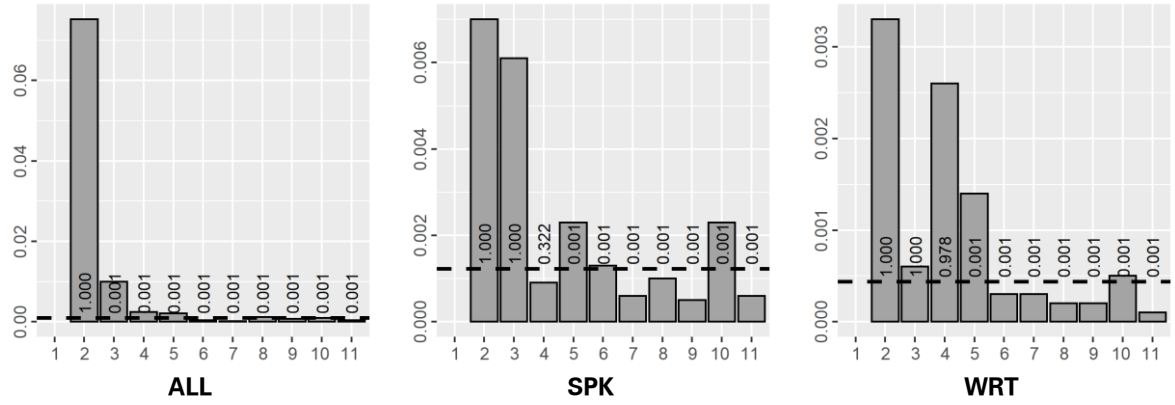


Figure 5.67: Jumps in node heights and respective *p*-values for lexical *MI* 3-grams

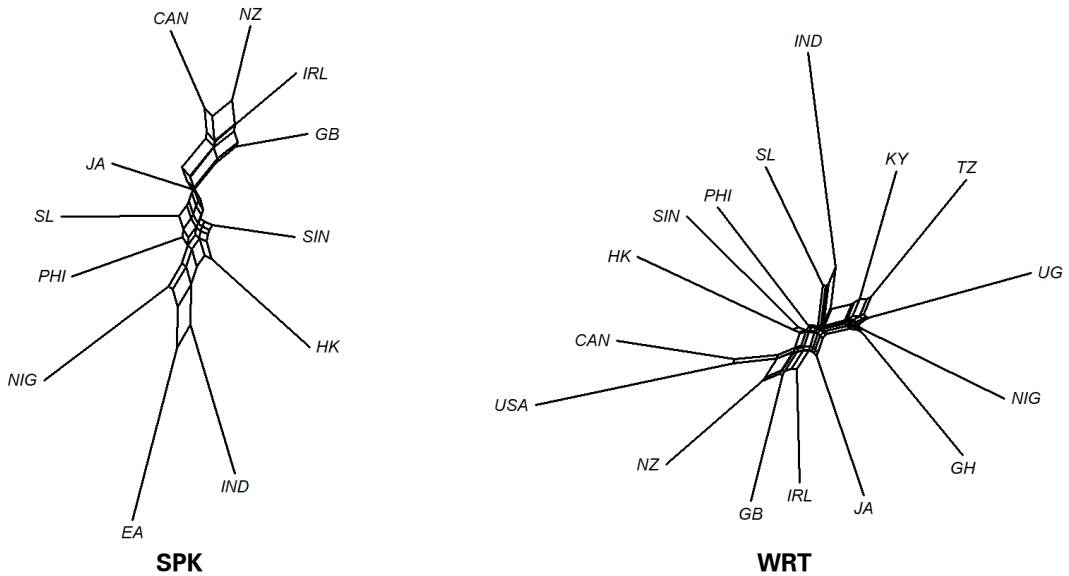


Figure 5.68: NeighborNets of the spoken and written data for lexical *MI* 3-grams

NeighborNet analysis (Figure 5.68) also indicates a separate status of IC (subdivided in writing) and a relatively coherent African cluster while Asian varieties are found in more intermediate positions. The separateness of EA, IND and NIG from other spoken varieties is also captured. In writing, IND is however grouped with SL, which appears more plausible than the hierarchical analysis.

K-means clusters, on the other hand, also find IND separated from other OC varieties in both SPK and WRT (Figure 5.69 and Table 5.42), as well as EA_{SPK} and NIG_{SPK} (the latter merging with the other OC at *k*=4). Results are generally almost identical to previous analyses (except for the broader *k*=3 approach to ALL only isolating spoken IC), only tending towards a finer regional association in writing than above.

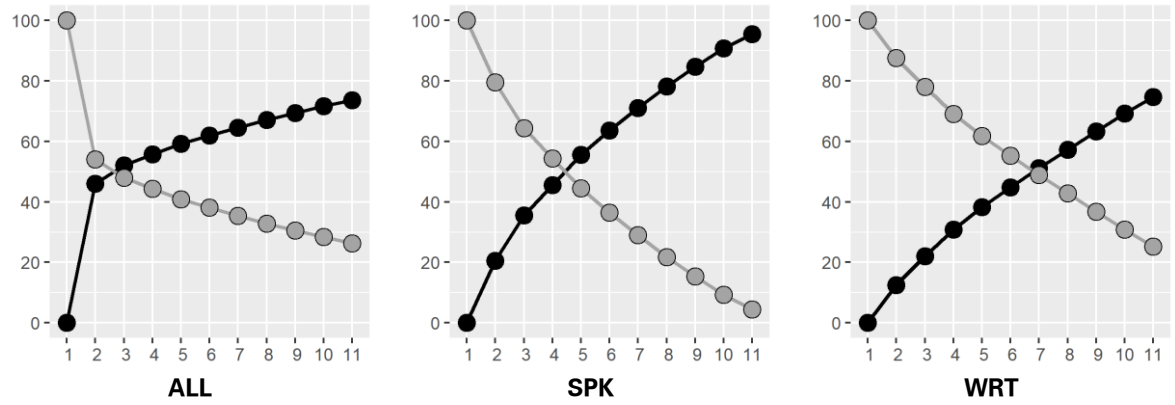


Figure 5.69: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *MI* 3-grams

Table 5.42: K-means clustering results for specific values of *k* for lexical *MI* 3-grams

	ALL (*k*=3)		**SPK (*k*=5)**		**WRT (*k*=7)**
1	EA, HK, IND, JA, NIG, PHI, SIN, SL_{SPK}	1	CAN, GB, IRL, NZ	1	GH, NIG
2	CAN, GB, IRL, NZ_{SPK}	2	EA	2	PHI, SL
3	All written corpus parts	3	NIG	3	KY, TZ, UG
		4	IND	4	HK, SIN
		5	HK, JA, PHI, SIN, SL	5	IND
				6	GB, IRL, JA, NZ
				7	CAN, USA

In contrast to shorter lengths, analysis results for MI-based **4-grams** produce a sudden relative absence of stable clusters (including the spoken/written distinction) and a respective likelihood of unreliability, even if some emerging groups mirror those of other datasets (Figure 5.70). For ALL, only a spoken IC cluster is found reliable. A similar group barely misses stability in SPK (AU=94, potentially due to the merger with JA), and separation of EA+IND from OC is detected. WRT shows only a single binary GB+NZ cluster, and exemplifies the inherent instability by reporting almost completely different segmentations than in ALL.

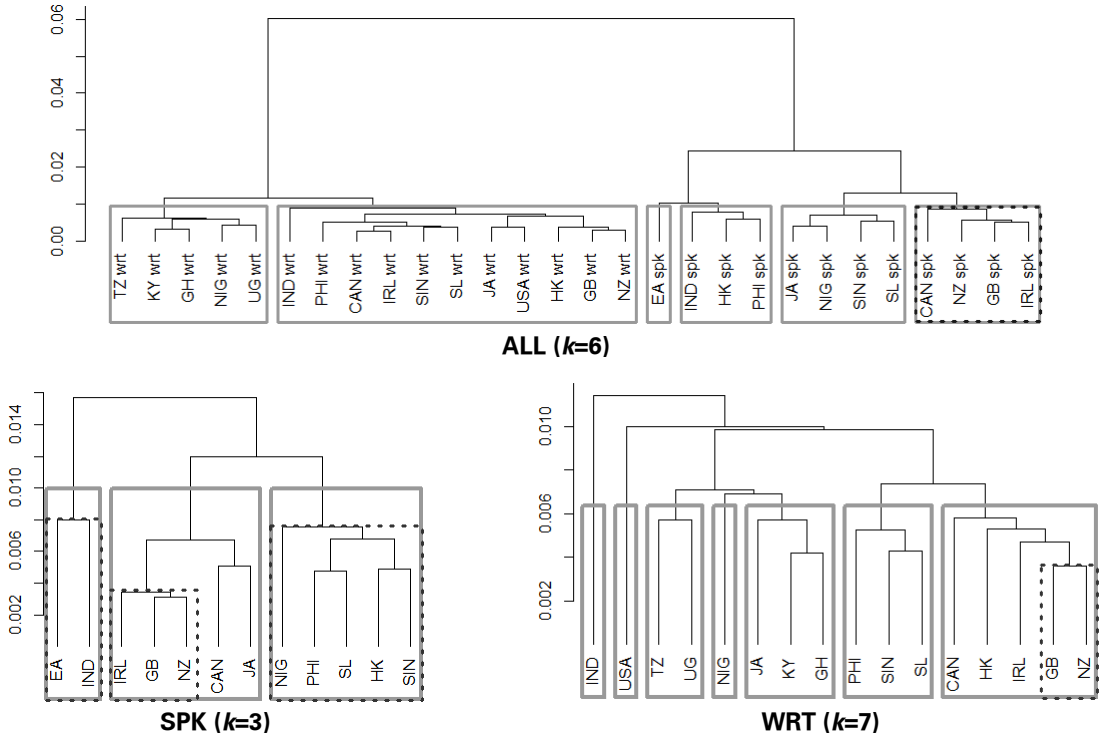


Figure 5.70: Hierarchical clustering results for lexical *MI* 4-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

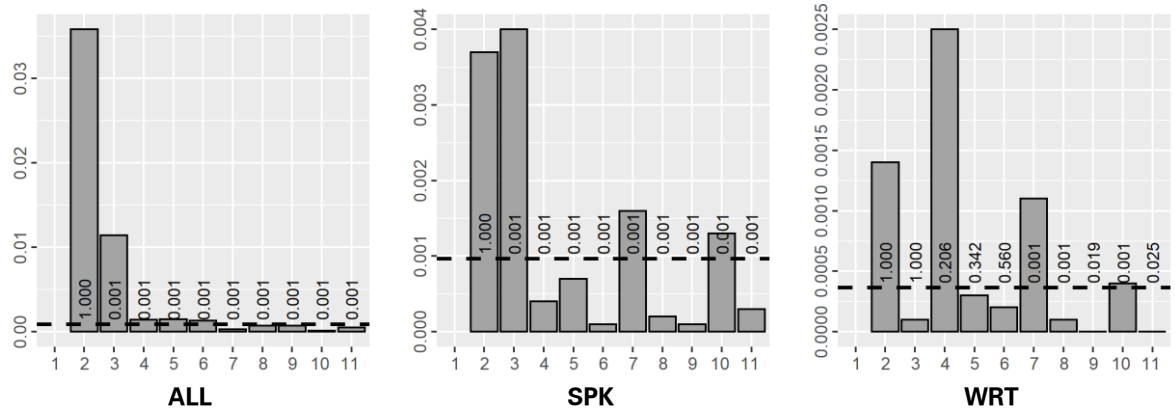


Figure 5.71: Jumps in node heights and respective *p*-values for lexical *MI* 4-grams

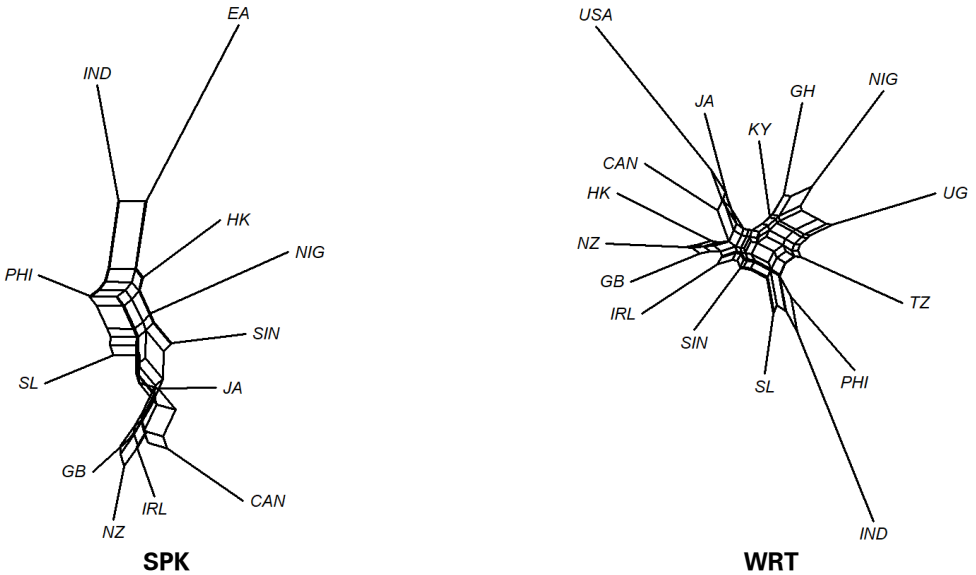


Figure 5.72: NeighborNets of the spoken and written data for lexical *MI* 4-grams

Jump heights (Figure 5.71) indicate significant differentiation in ALL after *k*=3, broadly isolating a group of less advanced spoken varieties in the dynamic model (EA+IND+HK+PHI) but contrasting against a highly heterogenous second group. The latter is only separated at *k*=4, when an IC group emerges. Jump heights remain similar up to *k*=6, where the written branch starts to distinguish the African group and the spoken branch segments off EA. If the stable IC cluster is to be taken as an indication of appropriate cutting height, this result may be slightly preferred over others. Still, most of the OC varieties cluster into counterintuitive groups (e.g. separating IND and SL or HK and SIN). For SPK, *k*=3 achieves significance and maps well onto the stable clusters (recall the AU= 94 for the overall IC group), thus underlining the established difference of EA+IND from other OC. In writing, *k*=7 is returned as the first significant jump, but only a mostly IC group appears meaningful. In both separate modes, the early split between individual nodes and the remaining varieties presents a very different picture from analyses based on other datasets.

The NeighborNets in Figure 5.72 also reflect a relatively unclear overall situation with many boxy structures. While an IC group can at least somewhat be established in speech (with JA quite similar to IC) and EA and IND are removed from the remaining data (but not overly similar to each other), there is only little indication of familiar or meaningful structures in writing. WRT finds some separation of IND+SL+PHI, some indication for an African group (but internally heterogenous), and while IC varieties form two loose groups (IC_{NA}+JA), these are interspersed by other varieties.

Analysis using k-means (Figure 5.73 and Table 5.43) does little to improve overall interpretability except affirming the discrete status of EA_{SPK} to other spoken varieties by clustering it with the written ones at *k*=2 for ALL, reinstating EA+HK+IND+PHI_{SPK} from the HCA results. Also, EA is isolated within SPK, where also the IC cluster reemerges. Results for WRT are similar to the hierarchical solution, picking out the African group but also presenting strong heterogeneity.

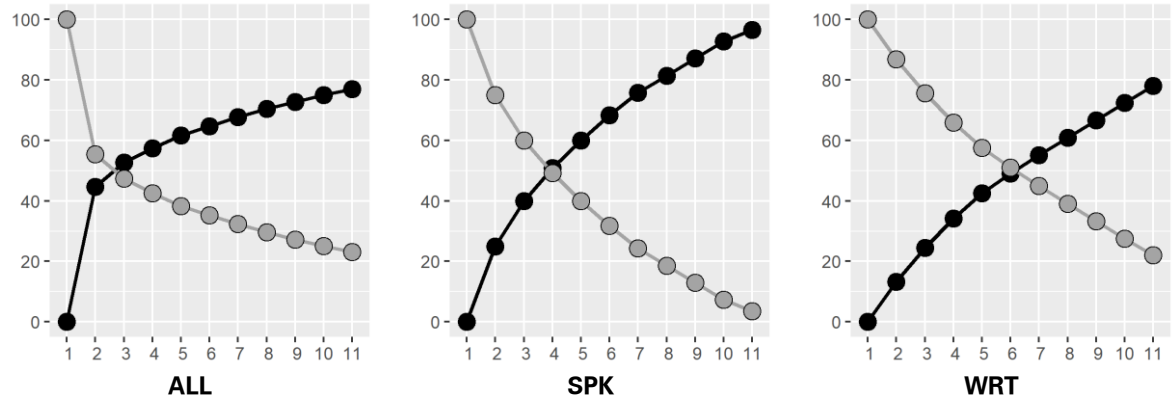


Figure 5.73: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *MI* 4-grams

Table 5.43: K-means clustering results for specific values of *k* for lexical *MI* 4-grams

	ALL (*k*=3)		SPK (*k*=4)		WRT (*k*=6)
1	CAN, GB, IRL, JA, NIG, NZ, SIN, SL $_{SPK}$	1	JA, NIG, SIN, SL	1	CAN, PHI, SIN, SL
2	EA, HK, IND, PHI $_{SPK}$	2	HK, IND, PHI	2	USA
3	All written corpus parts	3	EA	3	GH, NIG
		4	CAN, GB, IRL, NZ	4	IND
				5	GB, HK, IRL, JA, NZ
				6	KY, TZ, UG

5.3.2 T-score

T-score based *n*-grams demonstrate the frequency-dependent nature of the measure in the top and bottom types at each length (Table 5.44): Particularly in case of 2-grams, collocates differentiate by item frequency, placing combinations including function words at the top while very rare combinations, stemming from false starts, repetitions, missing punctuation or potentially errors in the corpus data are placed in the bottom rows of the table. Formation of longer sequences counteracts the effects of these outliers (the values should be taken as such, given the relatively stable mean association scores in Table 5.38 in this section's introduction) by averaging them with less extreme results. The result is a reduction in the overall range of association scores in particular effected by pronounced increases for the lowest collocates (opposed to relatively small decreases in the top types). Within the spoken data, bottom collocates still build upon verbal repetition and show much lower scores than found in the written data, where subjectively 'good' collocates can be found even with the lowest scores (recall that *t*=2.576 is commonly taken to define a true collocation).

Table 5.44: Lexical *t n*-grams with highest and lowest association scores

2-grams type	*t*	3-grams type	*t*	4-grams type	*t*
Spoken					
of the	44.00	if you know	33.18	you know you know	29.11
you know	43.13	do you know	32.62	a lot of the	27.80
in the	40.50	lot of the	31.36	i think you know	27.10
i think	35.55	of the same	31.19	you know in the	26.13
to be	29.71	of the first	30.75	do you know what	25.88
i don't	28.66	one of the	30.21	and then you know	25.71
i to	-210.10	the the the	-59.17	for the for the	-15.06
the is	-221.54	of the of	-81.51	the you know the	-34.87
a the	-230.82	the of the	-81.51	and the the the	-38.19
the to	-265.05	to the to	-125.38	of the of the	-39.67
the and	-271.70	and the and	-133.96	the the the the	-59.17
i the	-279.01	the and the	-133.96	and the and the	-88.05
Written					
of the	41.61	of the same	27.86	one of the most	21.06
in the	34.50	of the first	27.13	part of the world	20.86
it is	25.34	one of the	26.95	is one of the	19.49
to be	23.95	part of the	26.42	the end of the	19.49
on the	23.46	out of the	26.11	per cent of the	19.29
at the	18.78	of the government	25.75	of the fact that	18.93
it the	-83.81	referred to as	-11.05	a great deal of	3.44
in to	-90.12	and to provide	-11.60	in an effort to	3.41
and of	-94.89	of at least	-13.48	is the result of	2.87
to and	-117.29	to that of	-18.72	is made up of	1.43
to in	-152.21	in and out	-33.90	from time to time	0.80
of and	-210.73	and of course	-43.76	the extent to which	0.11

Hierarchical analysis of t-score-based **2-grams** show a relative absence of stable cluster structures. While hierarchical analysis (Figure 5.74) of ALL finds sufficient evidence to distinguish speech from writing, only an African and a binary JA+SIN cluster are identified in the written branch and also further substantiated by WRT. For speech, the separate data (SPK) show no stable clusters at all, while ALL distinguishes a GB+NZ cluster as well as a fragmentary Asian cluster joined with JA. Within the latter, $HK+SIN_{SPK}$ barely misses significance at AU=94, but so does $EA+IND_{SPK}$. Within WRT, IND+JA+SIN is similarly barely non-significant, providing a counterpoint to the merging of IND_{WRT} with EA observed above.

Deciding on a particular *k* is challenging given the low number of stable clusters. For ALL, apart from a binary split, significant jumps (Figure 5.75) only emerge for *k*=3 separating EA+NIG+EA from the remaining spoken branch. Later, *k*=6 is found only barely below average jump heights, additionally distinguishing the IC group in both modes and an African and mostly Asian cluster in writing.

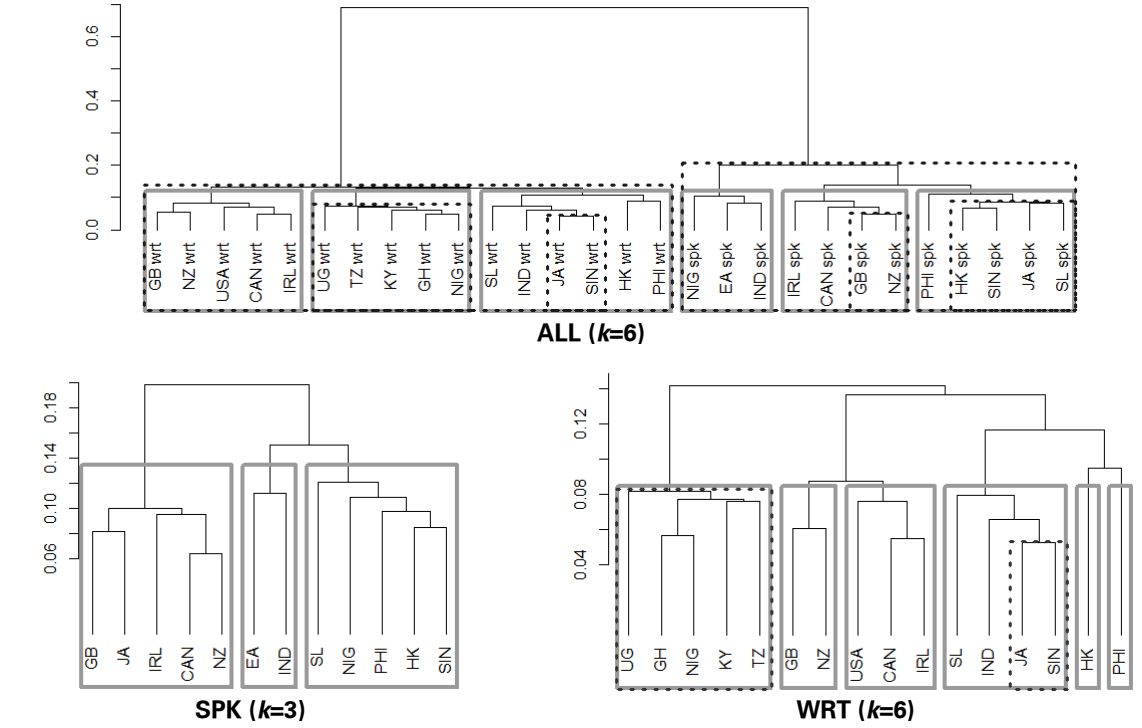


Figure 5.74: Hierarchical clustering results for lexical *t* 2-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

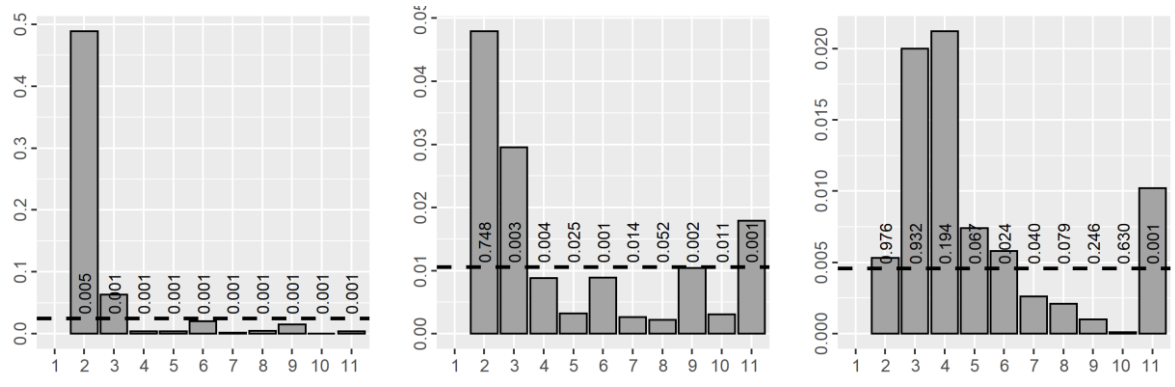


Figure 5.75: Jumps in node heights and respective *p*-values for lexical *t* 2-grams

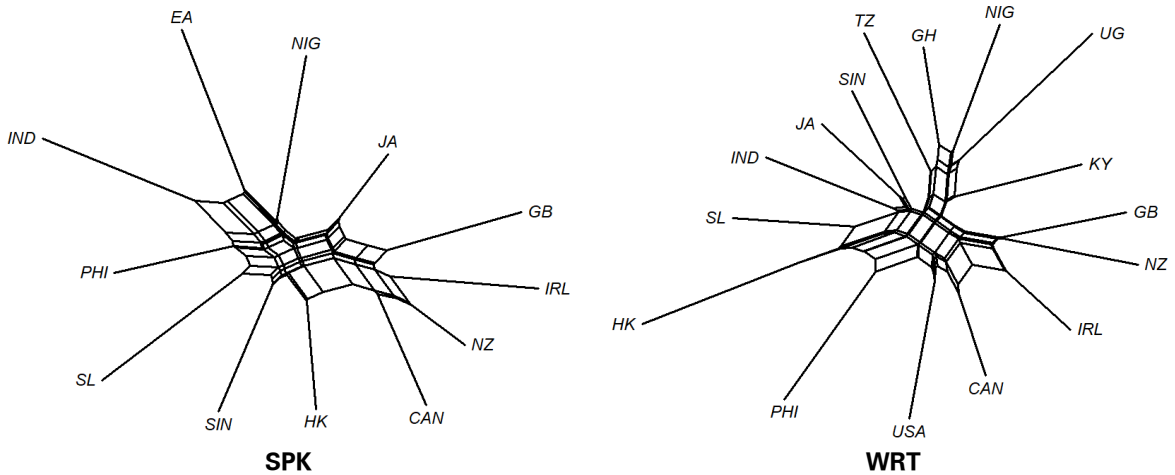


Figure 5.76: NeighborNets of the spoken and written data for lexical *t* 2-grams

For SPK, jumps at *k*=3 are the first to become significant but are only moderately larger than at *k*=11, which would result in total fragmentation of the dataset. The segmentation the lower value agrees with other analyses, however, by differentiation IC vs. OC, with EA+IND separated from the latter. For WRT, cutting at the first significant jump at *k*=6 also produces the large stable African cluster but separates IC varieties in an unfamiliar fashion and also splits off HK and PHI from the remaining varieties. What is more, the larger jump at *k*=11 equally indicates that much of the evidence for the remaining structure is shaky at best.

NeighborNet analysis (Figure 5.76) retrieves great heterogeneity within the data. Yet, in contrast to the HCA above, it shows that there is still good reason to divide the cluster at the heights of either the African group or either one or two IC clusters. SPK also supports separateness of EA+IND, while WRT indicates a stranger SL+HK+PHI.

K-means clustering (Figure 5.77 and Table 5.45) also supports distinguishing IC from OC in ALL's spoken branch at *k*=4 (but cf. PHI_{SPK}), partially confirmed by SPK at *k*=5, where the separate status of EA and NIG is further indicated. Writing isolates HK not only in WRT but even in ALL, further distinguishing the familiar two IC groups and the African cluster at *k*=6 (an equally possible *k*=5 merging the separate IC groups).

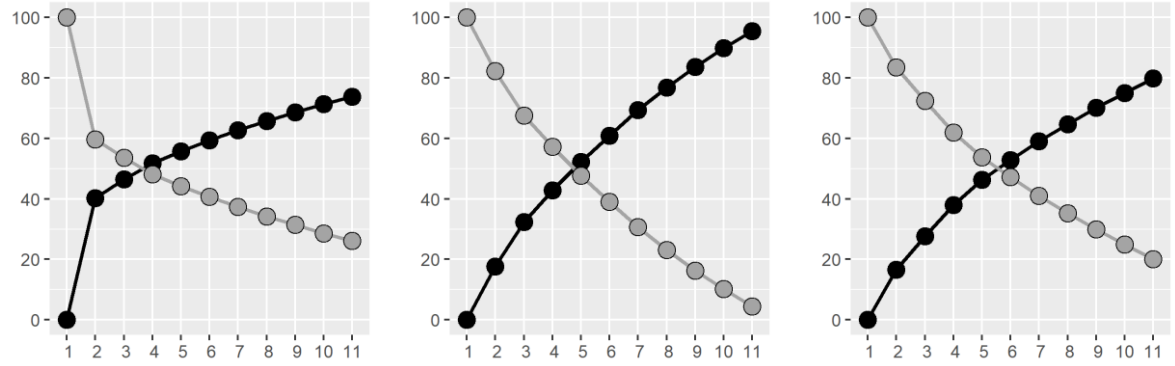


Figure 5.77: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *t* 2-grams

Table 5.45: K-means clustering results for specific values of *k* for lexical *t* 2-grams

	ALL (*k*=4)		**SPK (*k*=5)**		**WRT (*k*=6)**
1	All written corpus parts -HK	1	GB, IRL	1	IND, JA, SIN, SL
		2	EA	2	CAN, USA
2	HK_{WRT}	3	NIG	3	GB, IRL, NZ
3	CAN, GB, IRL, NZ, PHI_{SPK}	4	IND, PHI, SIN, SL	4	KY, TZ, GH, NIG, UG
4	EA, HK, IND, JA, NIG. SIN, SL_{SPK}	5	CAN, HK, JA, NZ	5	PHI
				6	HK

Hierarchical analysis of the **3-gram** data (Figure 5.78) reveals more stable clusters than for the shorter sequences, particularly within ALL: The spoken/written split is confirmed, as is a separation of IC vs. OC in speech (but not in writing), the special nature of EA+IND in speech, and other typical (HK+SIN) or less frequent (PHI, JA, SL) clusters. In the written branch, the overall cluster barely misses significance, potentially due to the inclusion of HK. SPK only adds an IC+JA cluster, while WRT, shows two stable African clusters (without UG).

Segmentation of the dendrograms using significant jumps (Figure 5.79) returns the spoken/written split within ALL as well as good indications for *k*=3 splitting separating spoken IC and OC. A similar match to stable clusters is found at (slightly below-average) *k*=5, identifying the African written cluster and the special status of EA and IND. Even finer distinction at *k*=7 begins to split off unary nodes, which may be taken to indicate a less stable allocation of PHI and NIG with their respective clusters. Segmenting SPK turns out to be challenging given the very different sizes of stable clusters and the fact that cutting at *k*=3 to reconcile the segmentation with the stable large IC cluster returns insignificant jumps. This makes *k*=5 appear like the most sensible alternative, indicating a separate status of GB from the remaining IC (+JA) and retrieving Asia (with IND separated) and Africa (IND, EA and NIG merging at *k*=4). For writing, *k*=7 or *k*=8 are the first segmentations at significant jump heights, and of those the latter fits better onto the stable clusters previously identified, cutting the data along mostly regional lines. Some correlation to the degree of endonormativity can be seen in the clustering of SIN with IC_{GB} as also in the isolation of more exonormatively oriented varieties HK, UG, KY+TZ.

Separation along the lines of Inner vs. Outer Circles is also retrieved by the NeighborNets (Figure 5.80), particularly in SPK. WRT more strongly separates IC_{NA} and supports an African group with internal regional differentiation. SPK also isolates EA, IND and (much less so) NIG from the OC varieties, and HK+SIN finds support in both data.

K-means (Figure 5.81 and Table 5.46) for ALL only really indicates the spoken/written distinction, picking out the IC spoken group (at *k*=3) also found in SPK and WRT. SPK further isolates EA and IND (but also GB) but WRT does not establish the African cluster(s), instead producing OC clusters of a very heterogenous kind.

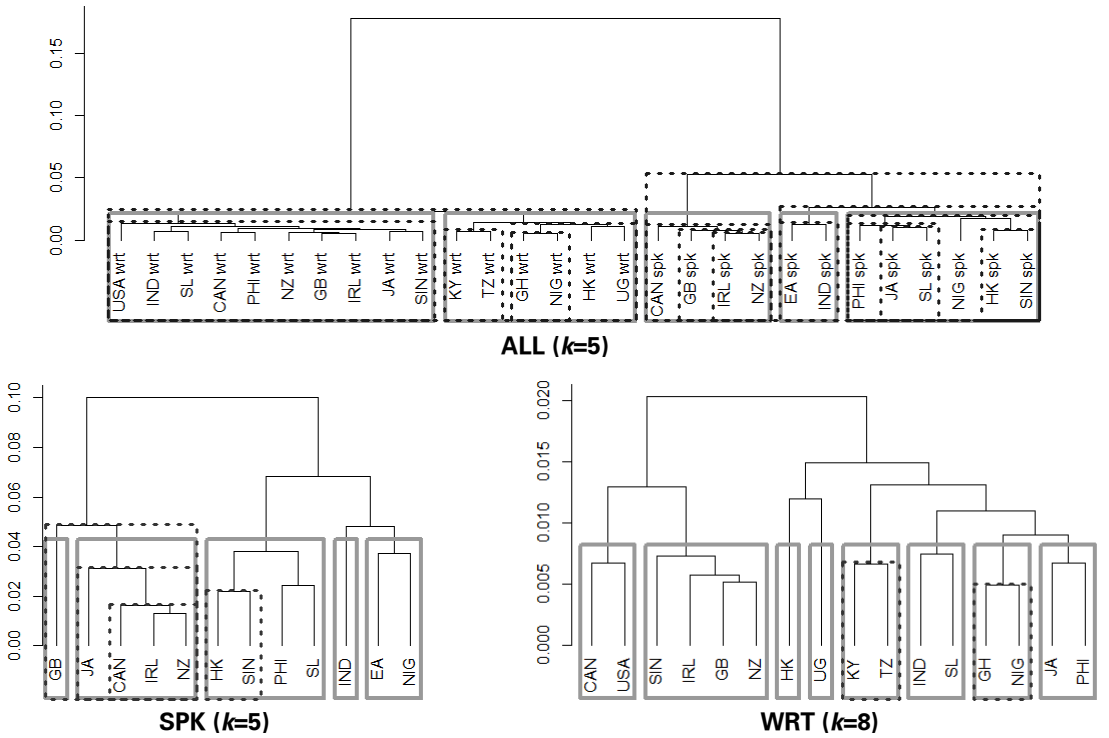


Figure 5.78: Hierarchical clustering results for lexical *t* 3-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

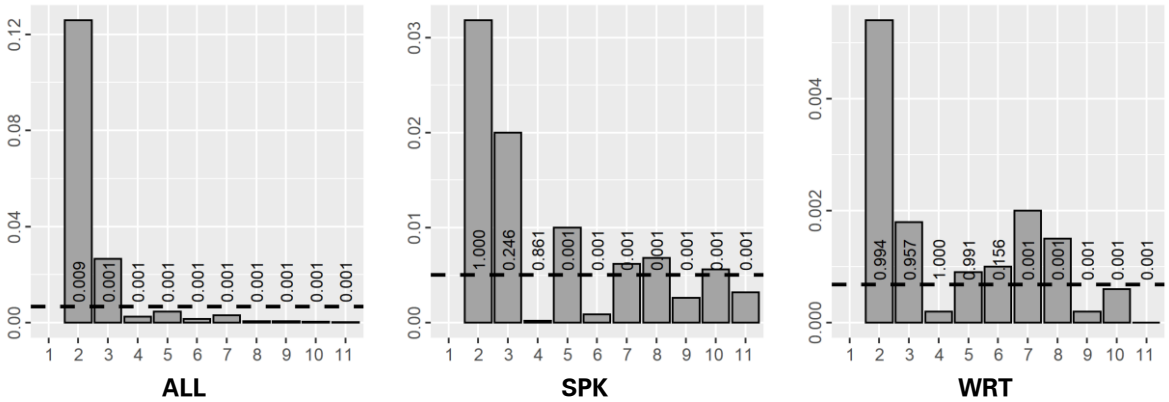


Figure 5.79: Jumps in node heights and respective *p*-values for lexical *t* 3-grams

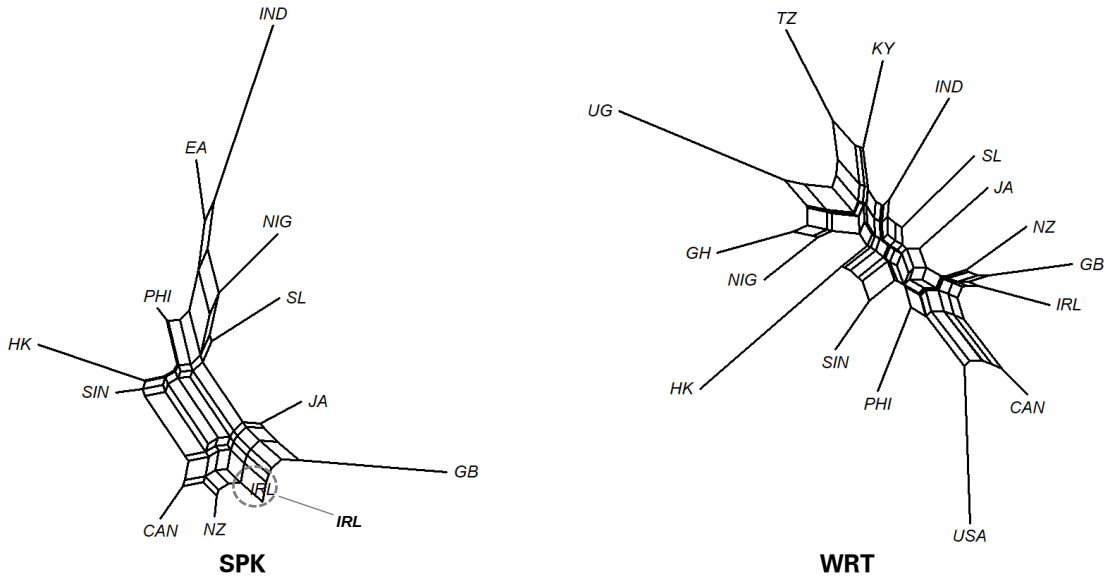


Figure 5.80: NeighborNets of the spoken and written data for lexical *t* 3-grams

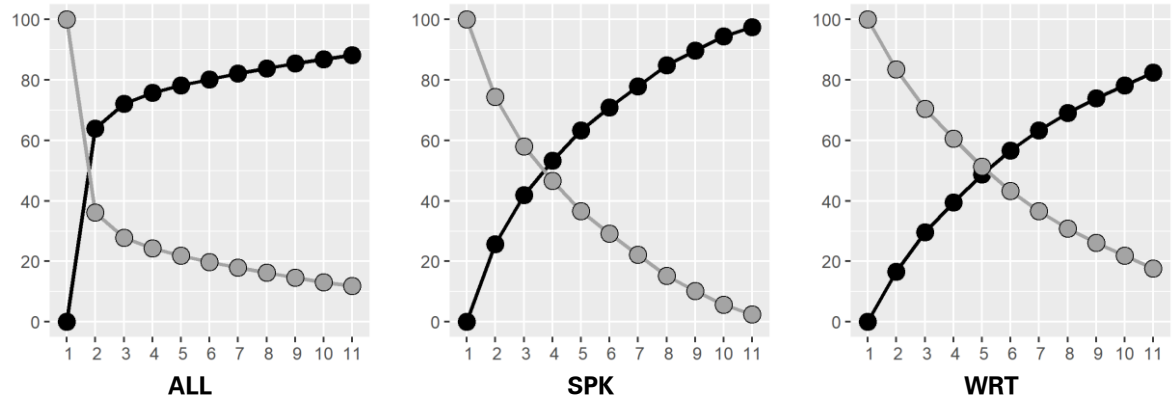


Figure 5.81: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *t* 3-grams

Table 5.46: K-means clustering results for specific values of *k* for lexical *t* 3-grams

	ALL (*k*=2)		SPK (*k*=4)		WRT (*k*=5)
1	All spoken corpus parts	1	HK, NIG, PHI, SIN, SL	1	KY, TZ, IND, SIN, SL, UG
2	All written corpus parts	2	EA, IND	2	HK
		3	CAN, IRL, JA, NZ	3	GB, IRL, NZ
		4	GB	4	GH, JA, NIG, PHI
				5	CAN, USA

Increasing sequence length to **4-grams** fails to continue the development to more stable clusters within the hierarchical analysis (Figure 5.82). For ALL, the spoken/written distinction is supplemented by a HK+SIN and EA+IND cluster in speech while the combined OC cluster barely misses significance. In SPK, GB is again separated from the stable remains of the IC group joined with JA, or merges within an OC group separated from EA, IND and GB. WRT, by contrast, shows no stable clusters, further indicating homogeneity.

Turning to significant jump heights (Figure 5.83) in order to less selectively cut the data, the spoken/written split in ALL is not returned as significant, and rather *k*=3 to *k*=6 is supported. At the lower values, this separates IC from OC in speech (coinciding with the barely-insignificant cluster). Cuts at larger values retain the spoken IC group and further separate EA+IND and HK+SIN(+NIG) before identifying a mostly African cluster in writing (but including IND and JA). WRT strongly favors a binary partition, distinguishing a mostly African cluster from the rest. Above-average jumps at *k*=3 further highlight a West African binary subcluster, and *k*=5 results in the separation of the remaining African data as well as creating a unary USA node.

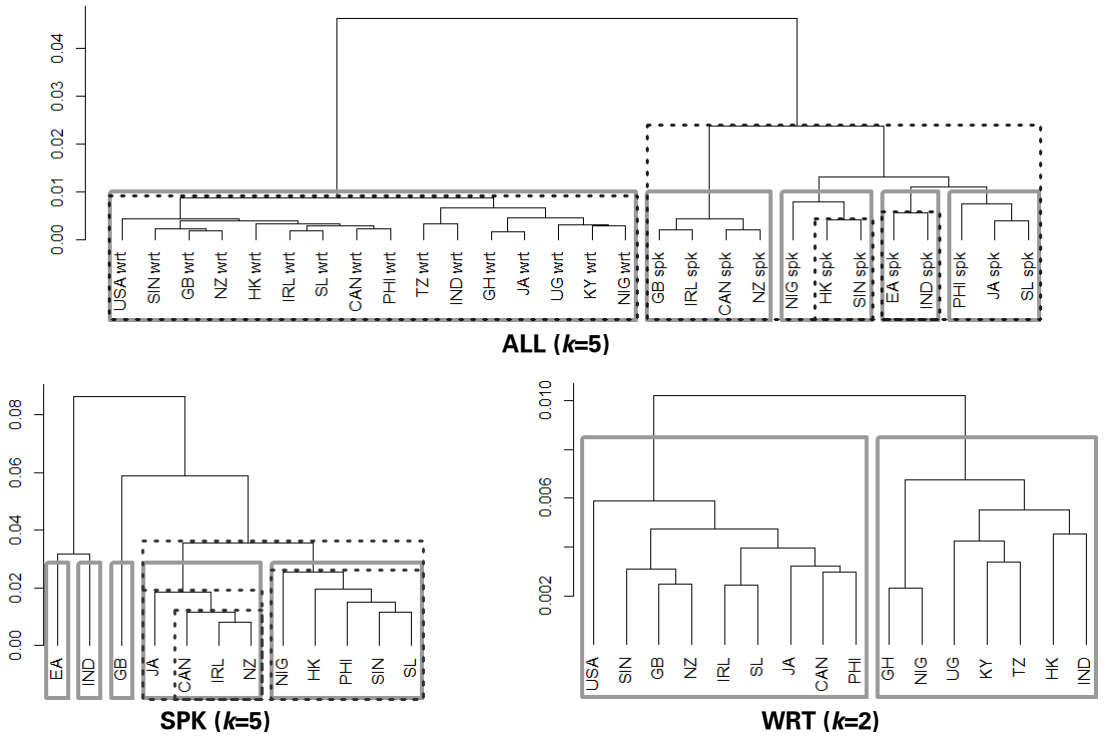


Figure 5.82: Hierarchical clustering results for lexical *t* 4-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

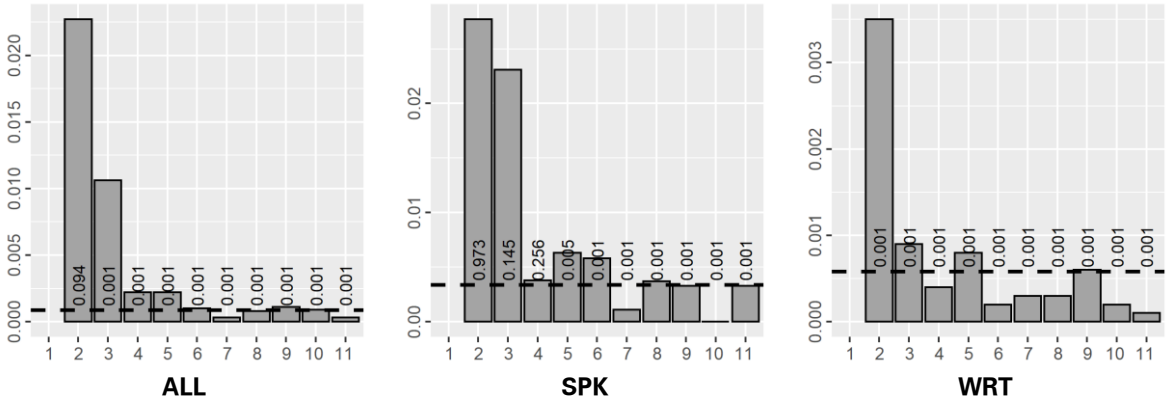


Figure 5.83: Jumps in node heights and respective *p*-values for lexical *t* 4-grams

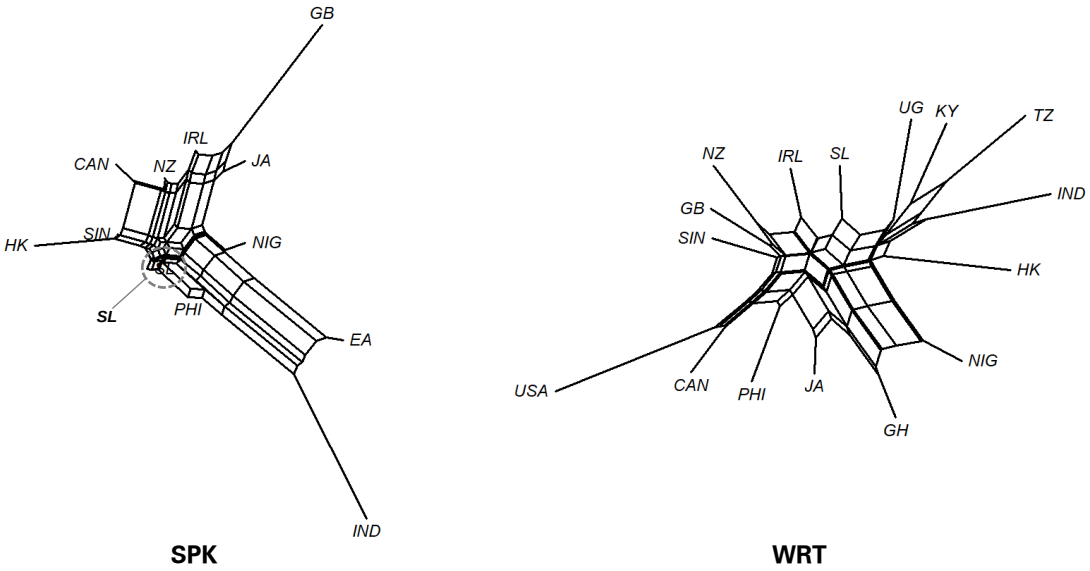


Figure 5.84: NeighborNets of the spoken and written data for lexical *t* 4-grams

The NeighborNets in Figure 5.84 indicate that the strange segmentation for SPK above may stem from proximity of JA to GB and the IC cluster as well as relatively small distances overall – with the exception of EA+IND, which are clearly different. In WRT, separate East and West African clusters emerge (the former in partial proximity with IND). The IC varieties are much less clearly separated from the remaining network than usual, while SIN and PHI associate closely with IC_{GB} and IC_{NA}, respectively.

K-means variance intersects (Figure 5.85) also point to relatively few clusters being identifiable (Table 5.47): ALL converges at the spoken/written distinction, while speech separates EA+IND and GB from the rest at *k*=3, subdividing the OC group at *k*=4. WRT presents four heterogenous clusters with some indication towards the usual IC groups, but the 'elbow' at a very different *k*=2 may be taken to indicate a competing binary solution taking the form of a merger of clusters 1+3 and 2+4, not conforming to any given hypothesis.

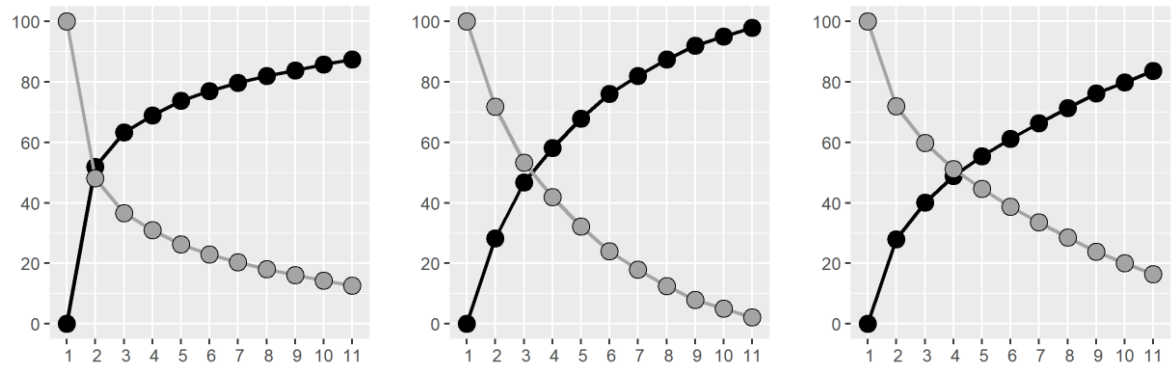


Figure 5.85: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *t* 4-grams

Table 5.47: K-means clustering results for specific values of *k* for lexical *t* 4-grams

	ALL (*k*=2)		SPK (*k*=4)		WRT (*k*=4)
1	All spoken corpus parts	1	EA, IND	1	CAN, SIN, USA
		2	GB	2	GB, HK, IRL, NZ
2	All written corpus parts	3	CAN, IRL, JA, NZ	3	GH, JA, NIG, PHI
		4	HK, NIG, PHI, SIN, SL	4	KY, TZ, IND, SL, UG

5.3.3 Log likelihood

Log likelihood *n*-gram types (Table 5.48) show a high degree of similarity to those found on the basis of the t-score, with most of the top and several bottom types being identical between the two measures, particularly for the shorter lengths. However, while types are often the same, differences in association results are somewhat more pronounced. Log likelihood reproduces the inflated association scores observed within

previous analyses but tends much more towards excessive positive values, in contrast to some exceptionally low scores observed for *t*. In further contrast, while negative values become less extreme for longer *t*-based sequences, changes are slightly more pronounced for positive values in case of G^2. Relatively speaking, G^2 finds even stronger support for the rather grammatical *n*-gram types commonly associated with *t*. On the basis of the (admittedly limited) data below, log likelihood should thus be seen as a relatively small quantitative variation on the dataset in contrast to some of the greater differences introduced through other measures.

Table 5.48: Lexical G^2 *n*-grams with highest and lowest association scores

2-grams type	G^2	3-grams type	G^2	4-grams type	G^2
Spoken					
you know	16998.33	if you know	10222.06	you know you know	11333.62
i think	11949.03	do you know	10008.67	i think you know	9656.63
of the	7577.89	did you know	9105.95	you know i mean	8118.80
i mean	7298.82	you know what	8907.86	i mean you know	8118.54
i don't	6789.98	are you know	8713.30	i think i think	7992.61
in the	6646.51	you know how	8703.62	you know i don't	7949.19
a the	-1869.56	know the the	-1040.47	from time to time	7.47
i the	-1949.17	to the to	-1063.71	for those of you	-30.14
the the	-2065.22	so the the	-1086.92	those of you who	-70.31
the of	-2239.00	and the and	-1304.63	and the and the	-859.42
the to	-2596.47	the and the	-1304.63	and the the the	-1366.48
the and	-2640.25	the the the	-2065.22	the the the the	-2065.22
Written					
of the	6298.86	of the same	3956.61	per cent of the	3086.17
it is	4396.09	part of the	3624.14	part of the world	2585.84
in the	4371.90	of the first	3572.06	one of the most	2429.88
has been	3209.92	one of the	3473.03	a part of the	2424.79
will be	2988.62	of the government	3420.86	the part of the	2415.61
have been	2950.00	use of the	3413.81	the end of the	2392.88
in and	-621.36	than that of	-72.82	the establishment of a	94.28
in to	-710.17	and to provide	-77.45	over a period of	67.49
to in	-770.28	and that of	-92.02	in an effort to	57.01
to and	-970.29	to that of	-179.24	is made up of	41.92
and of	-1072.57	in and out	-319.32	from time to time	19.30
of and	-1262.22	and of course	-371.59	the extent to which	17.15

Log likelihood **2-grams** produce consistently more stable clusters in writing than in speech (Figure 5.86), not even substantiation ALL's spoken branch, i.e. some spoken varieties are found too similar to written ones, and only EA+IND is found to be a stable cluster in both types of spoken data. The written data finds some support for African subclusters in both sets and distinguishes IC (+HK) from OC, within which a partial Asian PHI+SIN+SL set is found.

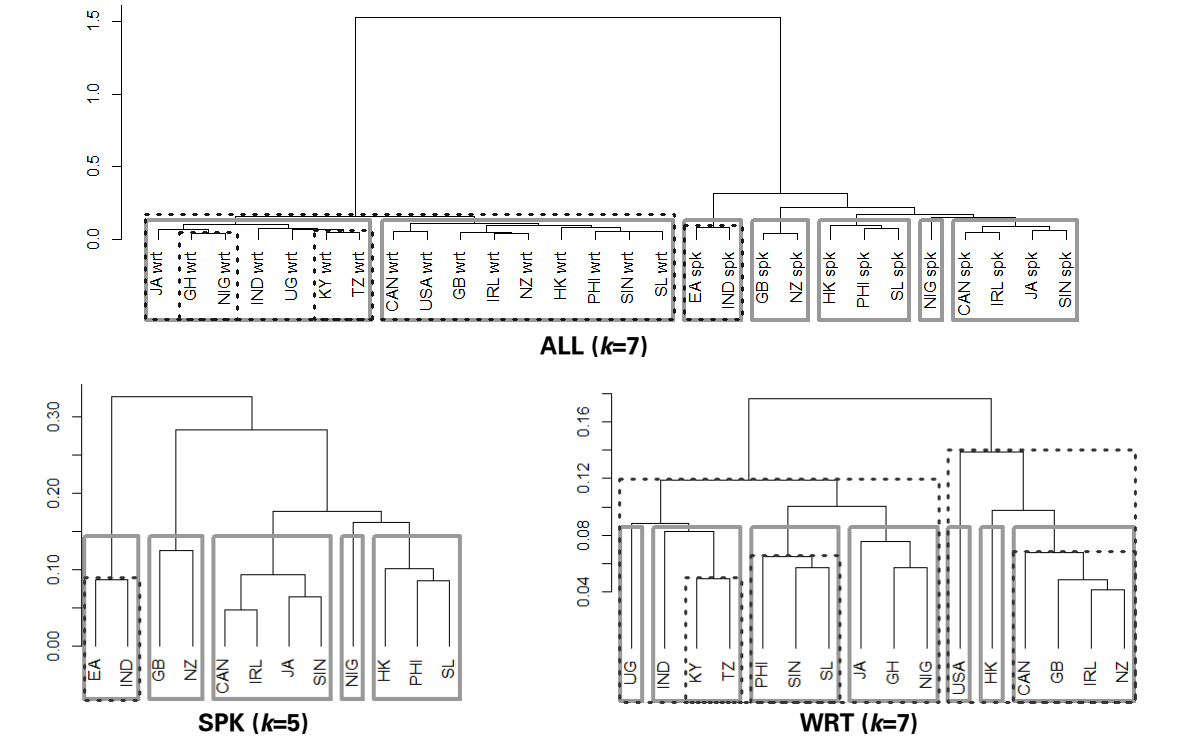


Figure 5.86: Hierarchical clustering results for lexical G^2 2-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

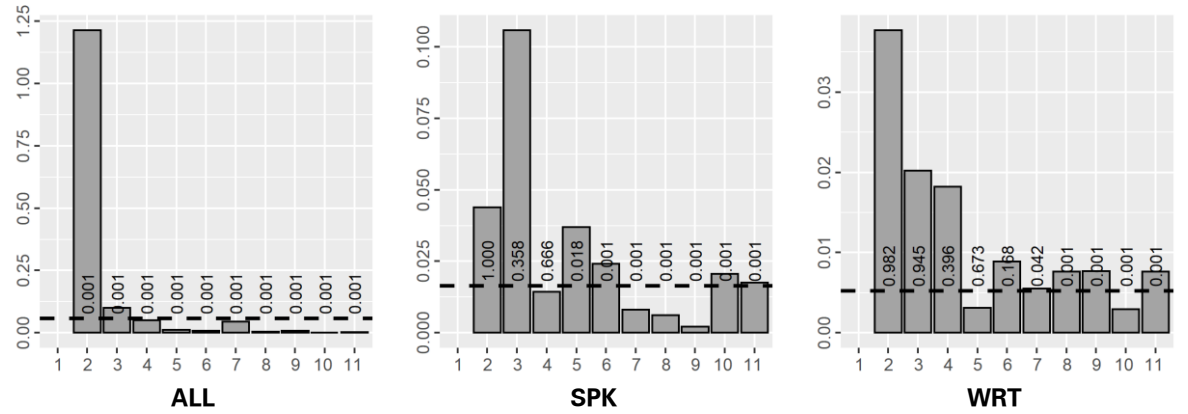


Figure 5.87: Jumps in node heights and respective *p*-values for lexical G^2 2-grams

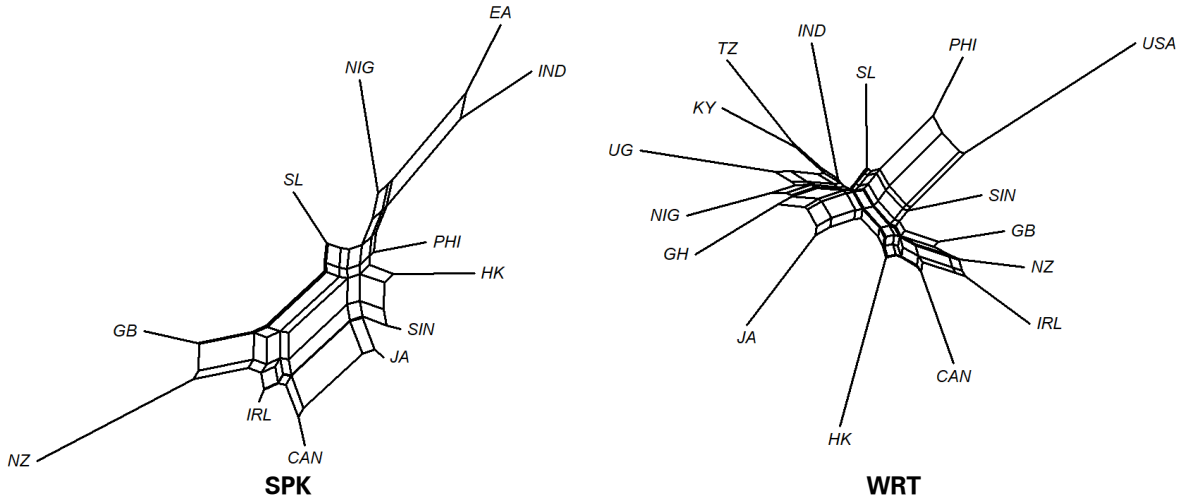


Figure 5.88: NeighborNets of the spoken and written data for lexical G^2 2-grams

ALL vastly prefers a binary split, but up to *k*=4 as well as *k*=7 show average jumps (Figure 5.87), identifying spoken EA+IND and GB+NZ before NIG, HK+PHI+SL and a stranger CAN+IRL+JA+SIN. This fine partition conforms to SPK at *k*=5, which roughly retrieves an exo-/endonormative separation. In writing, ALL identifies a mostly African group (+IND+JA) but WRT at the earliest significant *k*=7 fragments into unary nodes (UG, USA, HK) and vaguely regional clusters: KY+TZ(+IND), PHI+SIN+SL, GH+NIG+JA, GB+IRL+NZ+CAN.

NeighborNet analysis (Figure 5.88) of SPK identifies EA+IND and (less so) NIG as well as the IC group but also indicates about equal distance of CAN to GB and JA. In WRT, USA is found more similar to PHI than other IC varieties, but distances between the larger IC group (plus HK and SIN) and the remaining graph are still large. An African group (+IND+JA) and internal regional differentiation is again supported.

K-means (Figure 5.89 and Table 5.49) separates speech and writing in ALL, bringing out spoken IC at *k*=3. In SPK, this group is only found at the slightly less-indicated *k*=3 while isolating NZ at *k*=4. Otherwise, EA+IND is separated from the remaining OC varieties. WRT at *k*=6 splits off CAN+USA from a partial Asian group of various phases (#4), of which however IND and HK are allocated elsewhere. The data otherwise conform loosely to African subgroups each merged with a further single variety.

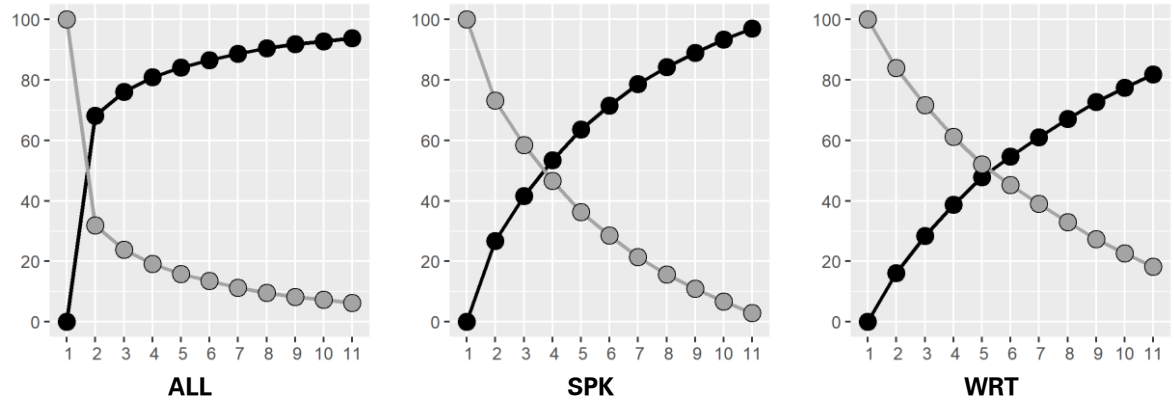


Figure 5.89: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical G^2 2-grams

Table 5.49: K-means clustering results for specific values of *k* for lexical G^2 2-grams

	ALL (*k*=3)		**SPK (*k*=4)**		**WRT (*k*=5)**
1	All spoken corpus parts	1	EA, IND	1	KY, TZ, IND, UG
2	All written corpus parts	2	HK, JA, NIG, PHI, SIN, SL	2	GH, JA, NIG
		3	CAN, GB, IRL	3	HK
		4	NZ	4	CAN, PHI, SIN, SL, USA
				5	GB, IRL, NZ

3-grams on the basis of log likelihood repeat and indeed intensify the effect of finding little substantiation in speech but more in writing (the spoken branch of ALL again not found stable). While a partial stable IC cluster is detected in ALL (Figure 5.90), this gives way to two binary groups somewhat separating by degree of endonormativity (PHI and SL being less endonormative than JA and NIG). The EA+IND cluster combines varieties of very different phases, but their separation from other OC is not surprising given previous analyses. Stable clusters of the SPK data are also detected (but not substantiated) in ALL, but the partial IC cluster in ALL is not present in SPK.

Significant jumps (Figure 5.91) for SPK are found after *k*=3, splitting off EA+IND and separating somewhat more endonormative varieties (plus CAN+IRL) from exonormative ones (plus GB+NZ). A finer segmentation is also indicated at *k*=6, overlapping with the stable clusters but only producing very small groups not presenting themselves as overly meaningful. WRT indicates good cuts already at *k*=2, repeating a rough segmentation by norm orientation also present in ALL's written branch at *k*=8 (which, at that height, also harmonizes with the spoken cuts). Further jumps above average are found at *k*=5, which overlap with the smaller stable clusters identified above and somewhat more closely conform to a fine-grained regional perspective.

The NeighborNets in Figure 5.92 visually highlight relative equidistance between the spoken varieties (except EA+IND). For WRT, some distinction can be confirmed for the EA data (i.e. KY+TZ without UG), the two West African varieties (+JA) and a North American epicentral cluster including PHI. Very little structure can be otherwise discerned, which fits the difficult segmentation above.

K-means analysis (Figure 5.93 and Table 5.50) similarly finds little support for larger clusters, indicating only *k*=3 for SPK and supporting *k*=2 not only for ALL (picking out EA+IND+PHI_{SPK} at *k*=3) but even somewhat for WRT. SPK splits off EA+IND (+SL) at *k*=3, and merges relatively more institutionalized varieties with CAN and IRL in cluster #3 (identical to the HCA). WRT produces no clusters of any sense, mixing regions and phases at *k*=3 (and conflating clusters #2 and #3 to an equally nonsensical group for the slightly less-indicated *k*=2).

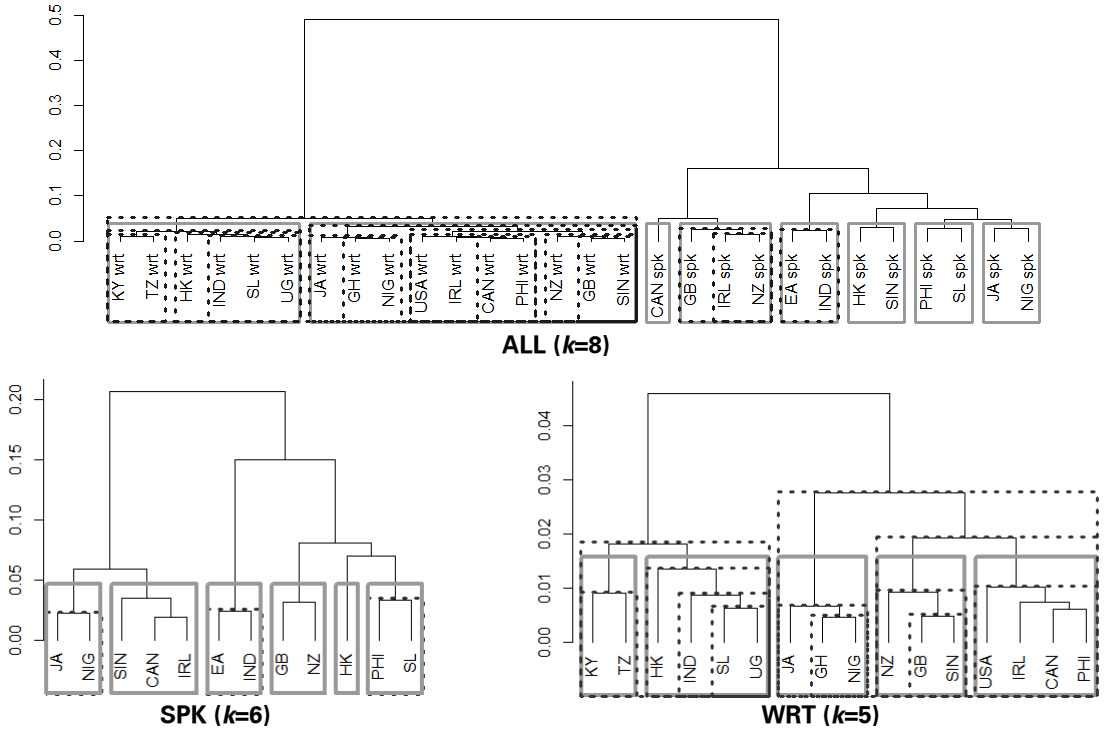


Figure 5.90: Hierarchical clustering results for lexical G^2 3-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

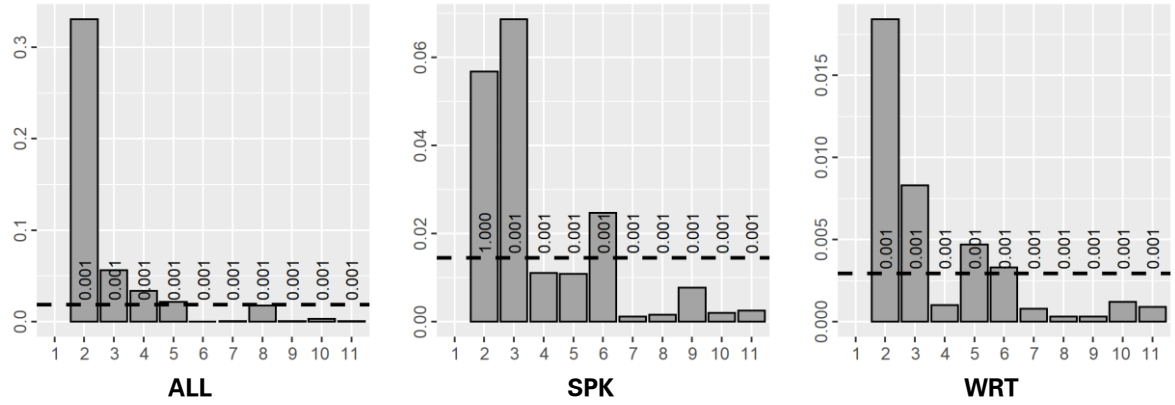


Figure 5.91: Jumps in node heights and respective *p*-values for lexical G^2 3-grams

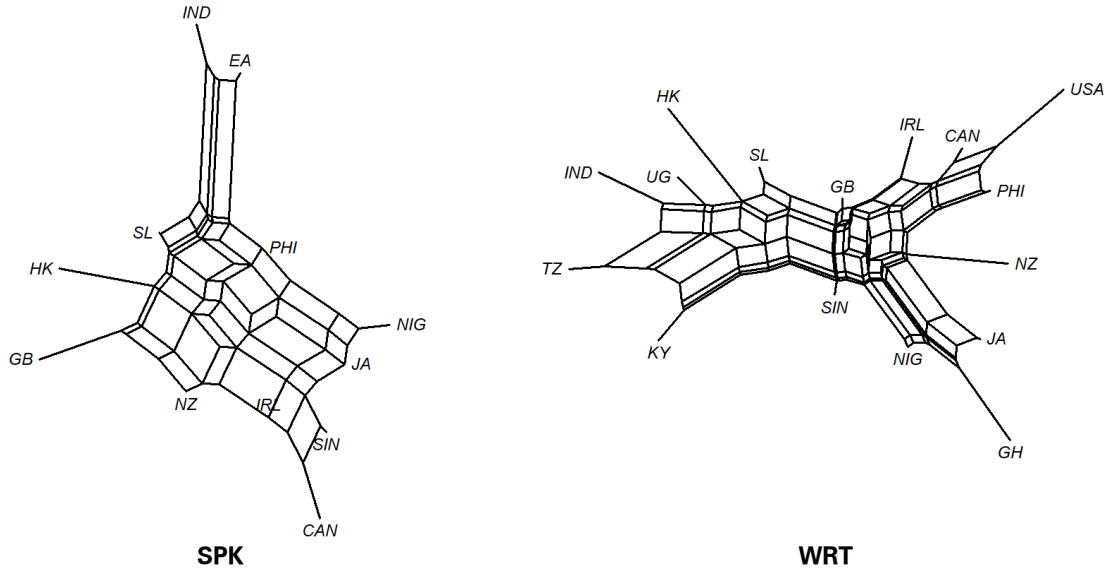


Figure 5.92: NeighborNets of the spoken and written data for lexical G^2 3-grams

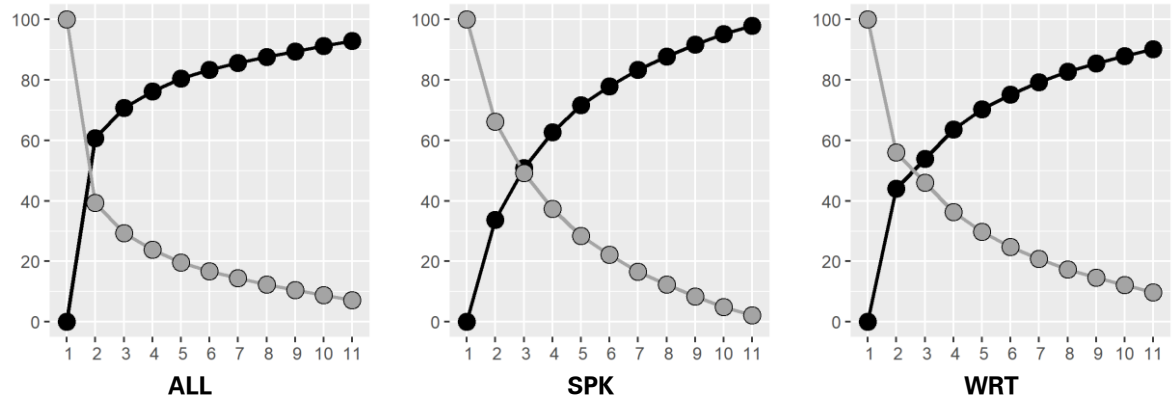


Figure 5.93: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical G^2 3-grams

Table 5.50: K-means clustering results for specific values of *k* for lexical G^2 3-grams

	ALL (*k*=2)		SPK (*k*=3)		WRT (*k*=3)
1	All spoken corpus parts	1	EA, IND, SL	1	GB, GH, IRL, JA, NIG, NZ, PHI
2	All written corpus parts	2	GB, HK, NZ, PHI	2	KY, TZ, HK, IND
		3	CAN, IRL, JA, NIG, SIN	3	CAN, SIN, SL, UG, USA

With the analysis of **4-grams**, the lessening difference between speech and writing is further intensified in the ALL data (Figure 5.94). Unlike in other analyses, this for once concerns not only EA and IND, but further the spoken forms of also PHI, JA and SL, which all show great similarity to USA_{WRT}. Only the EA+IND group is stable (in addition to two spoken IC groups as well as HK+SIN, so all of the results need to be considered under the lens of almost entirely lacking substantiation, making generalizations much less reliable. In the written branch, the next-closest North American data (CAN, but +IRL) is merged with PHI, and a GB+SIN cluster emerges. The latter is confirmed in WRT (but nothing else is found), but SPK provides only different stable clusters (CAN+JA, PHI+SL, IRL+NZ) with the exception of EA+IND.

In terms of significant jump heights (Figure 5.95), some results from previous log-likelihood analyses are carried over to 4-grams, e.g. fragmented IC clusters in all datasets and SIN more closely associated with IC than OC in writing. Furthermore, the closeness of particularly PHI to USA is also observed in WRT at *k*=4, which otherwise separates the data on the basis of institutionalization. ALL prefers a binary distinction between the spoken IC+NIG+HK+SIN cluster to all others, splitting off the remaining spoken varieties (+USA_{WRT}) only at *k*=4 and further supporting an even finer *k*=8 segmentation mostly clustering less institutionalized written varieties and isolating the USA_{WRT}-like binary spoken groups in addition to NIG.

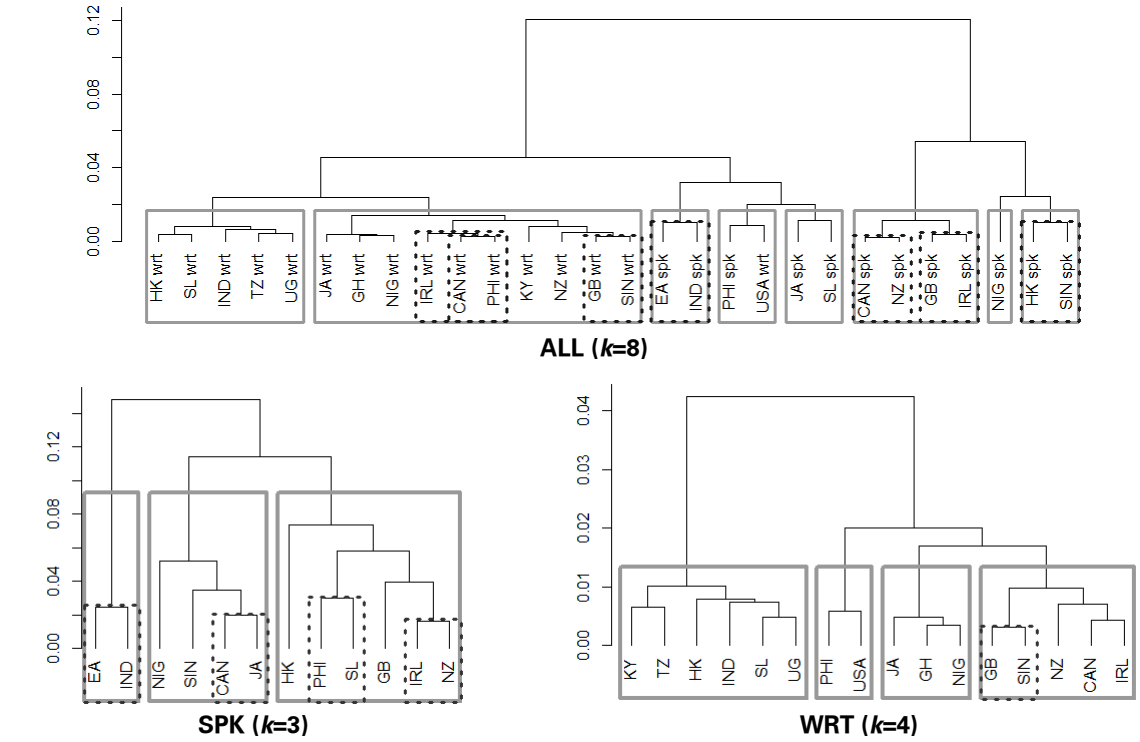


Figure 5.94: Hierarchical clustering results for lexical G^2 4-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

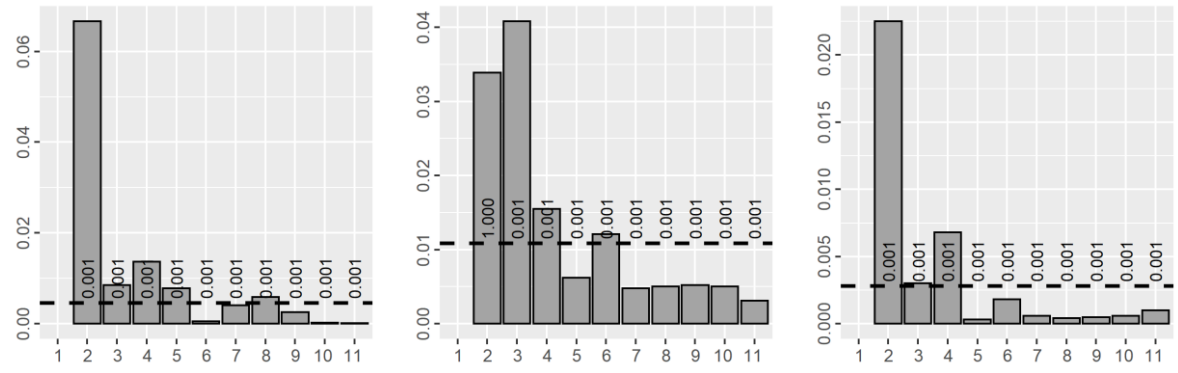


Figure 5.95: Jumps in node heights and respective *p*-values for lexical G^2 4-grams

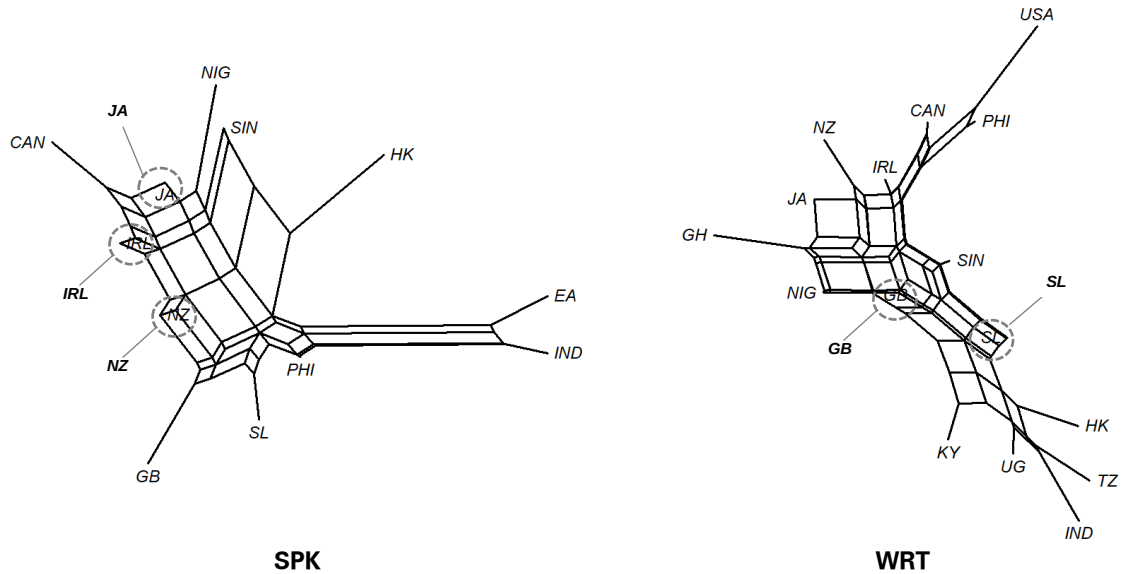


Figure 5.96: NeighborNets of the spoken and written data for lexical G^2 4-grams

NeighborNet analysis (Figure 5.96) only underscores the vague structures produced on the basis of the G^2 measure, only truly distinguishing EA+IND in speech. There is some indication towards similarity of GB, SL and PHI and very limited support for IC, but mostly heterogeneity is retrieved. WRT identifies most of the African varieties (+IND) and the USA+CAN+PHI cluster also found at shorter sequence lengths.

K-means clustering (Figure 5.97 and Table 5.51) brings to light the low substantiation in the present data by producing yet another result even at often the same cluster numbers. This even extends to the dissolved spoken/written separation, which is cleanly reinstated at *k*=2, while *k*=3 separates IND+PHI$_{SPK}$. SPK at *k*=3 merges SL and GB with the common EA+IND and further isolates HK, while WRT creates two entirely heterogenous clusters, not even allocating the above stable GB and SIN or CAN/USA and PHI to the same groups at *k*=2. Yet a different CAN+USA+PHI cluster is retrieved at (non-indicated) *k*=3.

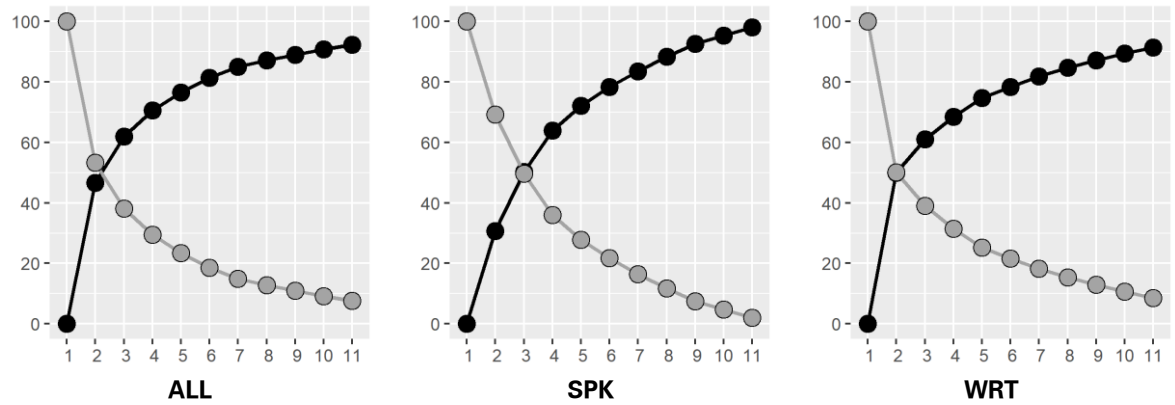


Figure 5.97: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical G^2 4-grams

Table 5.51: K-means clustering results for specific values of *k* for lexical G^2 4-grams

	ALL (*k*=3)		SPK (*k*=3)		WRT (*k*=2)
1	All written corpus parts	1	HK	1	GB, GH, IRL, JA, NIG, NZ, PHI
2	IND, PHI $_{SPK}$	2	CAN, IRL, JA, NIG, NZ, PHI, SIN	2	CAN, KY, TZ, HK, IND, SIN, SL, UG, USA
3	CAN, EA, GB, HK, IRL, JA, NIG, NZ, SIN, SL $_{SPK}$				

5.3.4 Lexical Gravity

Sequence association calculated on the basis of lexical gravity displays a similar tendency towards high-frequency items as also found for all other measures except *MI* (Table 5.52). Unlike other measures, however, bottom *n*-grams contain relatively few items that might be considered errors of some form. Also, longer sequences only rarely contain ones found at the respective ends of the collocability spectrum at

shorter lengths, which has been a repeated observation for all previous measures. Changes in association between shorter and longer sequences also affect top and bottom association values to similar degrees. Interestingly, though, negative association scores do not drop far below zero, and for 4-grams always surpass this value, which indicates that this would not represent a sufficient criterion for collocability (even if it implies O>E). However, the threshold of $g \geq 5.5$ recommended in the respective publications on the measure conversely fails to do justice to what intuitively should be relatively good collocates (e.g. *thank you very much*, *I beg your pardon*, *is there any*, *as simple as*). The previous *n*-gram analyses may thus have been too restrictive.

Table 5.52: Lexical *g* *n*-grams with highest and lowest association scores

2-grams		3-grams		4-grams	
type	***g***	**type**	***g***	**type**	***g***
Spoken					
of the	17.54	one of the	15.07	of the one of	13.36
in the	16.40	of the first	14.99	the one of the	13.36
to the	14.64	some of the	14.90	it is in the	13.26
this is	14.52	of the the	14.80	this is the first	13.25
and then	14.33	of the other	14.51	one of the most	13.23
on the	14.25	in the first	14.42	this is a very	13.15
this mean	-4.69	as simple as	-0.08	does not mean that	3.74
lot a	-4.90	by virtue of	-0.24	get in touch with	3.61
they mean	-4.96	the wake of	-0.42	is made up of	3.07
going i	-5.00	to thank the	-0.94	united states of america	3.03
know be	-5.17	the eve of	-1.31	from time to time	3.03
i been	-5.83	to amend the	-1.45	i beg your pardon	0.38
Written					
of the	18.05	one of the	15.04	one of the most	13.39
in the	16.51	of the first	14.61	part of the world	12.48
to the	14.99	some of the	14.54	most of the time	12.40
it is	14.63	out of the	14.13	is one of the	12.36
and the	14.50	of the most	14.07	that it is a	12.28
on the	14.21	most of the	14.04	the use of the	12.07
developed the	-2.97	the passage of	0.70	thank you very much	4.70
well the	-3.16	all sorts of	0.64	is made up of	4.38
you be	-3.18	is there any	0.60	a great deal of	4.11
done the	-3.40	the advent of	0.10	from time to time	3.91
you i	-3.54	the occurrence of	0.08	a wide variety of	3.76
there a	-3.71	a glimpse of	-0.34	the extent to which	3.49

2-gram association on the basis of lexical gravity exhibits a remarkable degree of stability in all datasets (Figure 5.98). In addition to the overall spoken/written distinction, several smaller groups are identified, mostly corresponding to smaller regional groups and potential epicenters. ALL finds less substantiation for spoken subclusters than for writing, but both separate datasets are highly compatible with the combined set and present further substantiation for clusters identified there. In all cases, the

African written varieties form clusters (but with UG allocated, as commonly, to the West African varieties), North America is separated within the Inner Circle, Asia is analyzed as both a stable group (with IND_{WRT} only being a part of this in writing), and HK+SIN is often identified as a distinct subgroup.

Separation by significant jump heights above average (Figure 5.99) indicates relevant splits at speech vs. writing at *k*=2 as well as at *k*=4 for ALL, separating spoken Inner vs. Outer Circle and written African varieties from all others. Successive splits above average are also indicated for 5 and 8 clusters, at which point IC is distinguished in both modes and African and Asian subclusters emerge, with IND, EA and NIG identified as separate cases. For SPK, *k*=5 (below average) and *k*=6 results in the same overall situation, with HK+SIN split off at the larger value. Writing requires *k*=6, at which point the same result as in ALL is achieved, with only IC subdivided by region.

NeighborNet analysis (Figure 5.100) also shows one or two clear clusters of IC varieties (with PHI somewhat connected to USA+CAN). The written African clusters are also revealed (and UG identified as in-between two groups), as is the relative distance of EA, IND and (to a lesser extent) NIG from the remaining varieties in speech. Similarity between HK and SIN is also only supportedin speech, while writing shows a more homogenous Asian group.

K-means clustering (Figure 5.101 and Table 5.53) indicates slightly more clusters for SPK and WRT and prefers a very fine-grained segmentation of the ALL data (*k*=7) as chosen in the hierarchical analysis above (*k*=2 could also be chosen given the 'elbow' rule, separating speech from writing). Still, results are highly compatible to those obtained before in all cases. While NIG_{SPK} and IND_{WRT} are clustered differently in ALL, and HK is separated within the WRT, the above findings can be taken as confirmed. SPK thus retrieves an IC cluster, unary EA, IND and NIG, allocating all remaining (Asian) varieties to a shared cluster. WRT indicates two IC cluster as well as two African clusters (but cf. the allocation of UG), placing the remaining varieties (without IND and HK) in a group of various regions, phases and backgrounds.

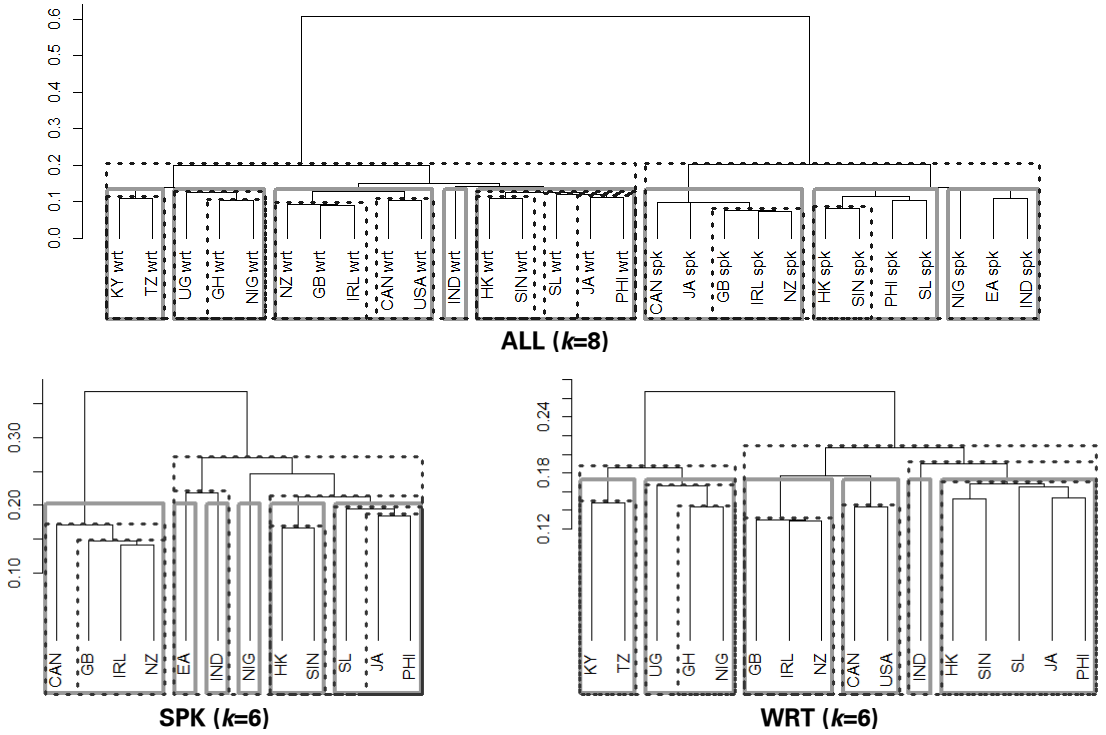


Figure 5.98: Hierarchical clustering results for lexical *g* 2-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

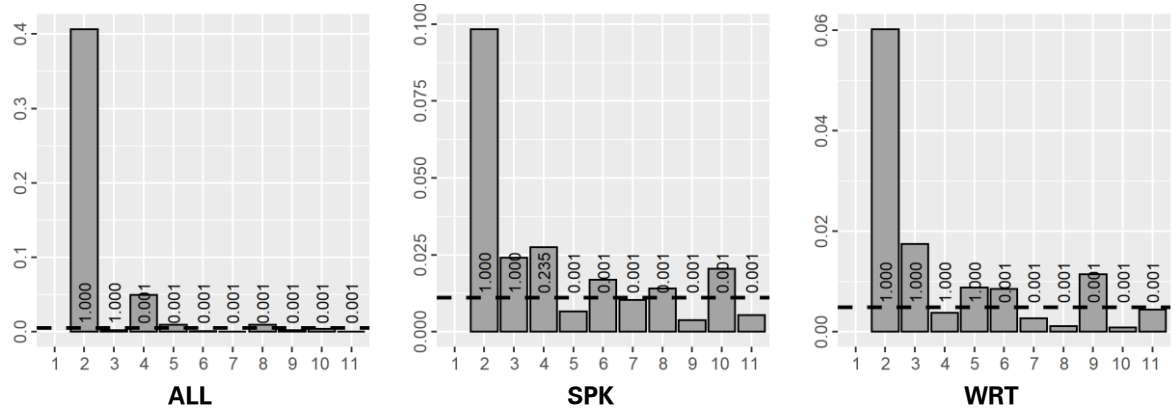


Figure 5.99: Jumps in node heights and respective *p*-values for lexical *g* 2-grams

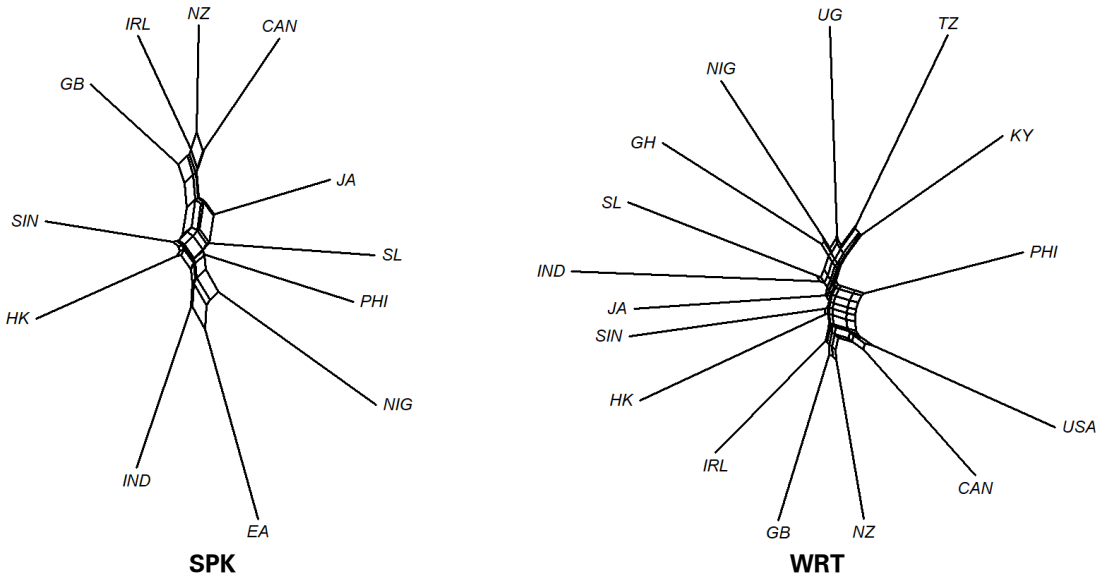


Figure 5.100: NeighborNets of the spoken and written data for lexical *g* 2-grams

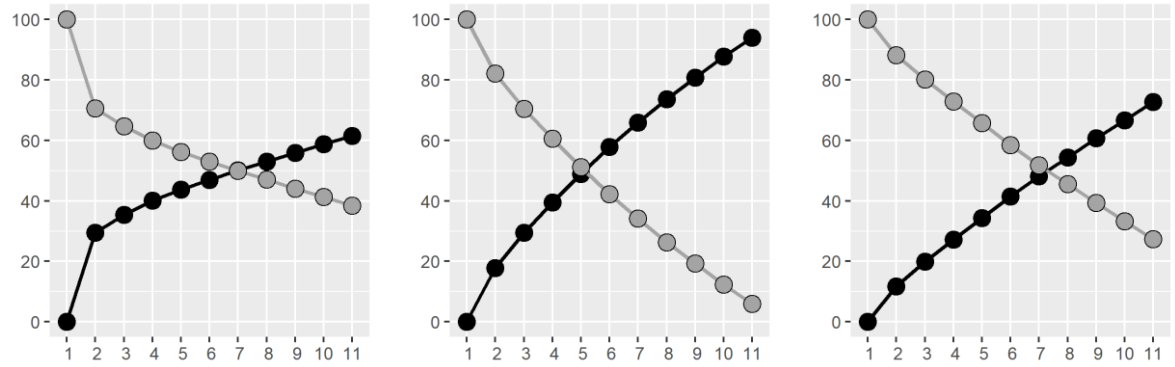


Figure 5.101: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *g* 2-grams

Table 5.53: K-means clustering results for specific values of *k* for lexical *g* 2-grams

	ALL (k=7)		SPK (k=5)		WRT (k=7)
1	KY, TZ, GH, NIG, UG $_{WRT}$	1	EA	1	IND
2	HK, IND, JA, PHI, SIN, SL $_{WRT}$	2	IND	2	HK
3	CAN, GB, IRL, NZ, USA $_{WRT}$	3	NIG	3	JA, PHI, SIN, SL
4	CAN, GB, IRL, NZ $_{SPK}$	4	HK, JA, PHI, SIN, SL	4	KY, TZ
5	HK, JA, PHI, SIN, SL $_{SPK}$	5	CAN, GB, IRL, NZ	5	GB, IRL, NZ
6	EA, IND $_{SPK}$			6	CAN, USA
7	NIG $_{SPK}$			7	GH, NIG, UG

Lexical gravity **3-gram** results are highly compatible with those obtained for 2-grams. For ALL, the spoken/written separation is supported by stable clusters (Figure 5.102), as are written African subclusters and an Asian group. While IC written groups are not entirely supported in writing, they emerge as an entirely stable cluster in speech, where a separation between IC and OC, the latter with a HK+SIN subcluster) is found. Except for the case of NIG, results for SPK are identical to 2-grams. Again, the data support an IC/OC distinction and within the latter a separateness of EA+IND and HK+SIN. In WRT, however, less support overall is found. While two IC clusters are found stable, the combined African group gives way to subclusters as in ALL (with the UG+GH+NIG cluster at AU=93), and no substantiation for the Asian group is discovered.

Significant jumps of above-average heights (Figure 5.103) are found for ALL between the isolation of IC from speech at *k*=3 and up to *k*=5, at which point a separation of written African varieties from all remaining varieties is found, while speech additionally splits off EA+IND from the OC group. About-average *k*=7 additionally separates North American written varieties and HK+SIN in speech, indicating some relative separateness (the former one however not found stable).

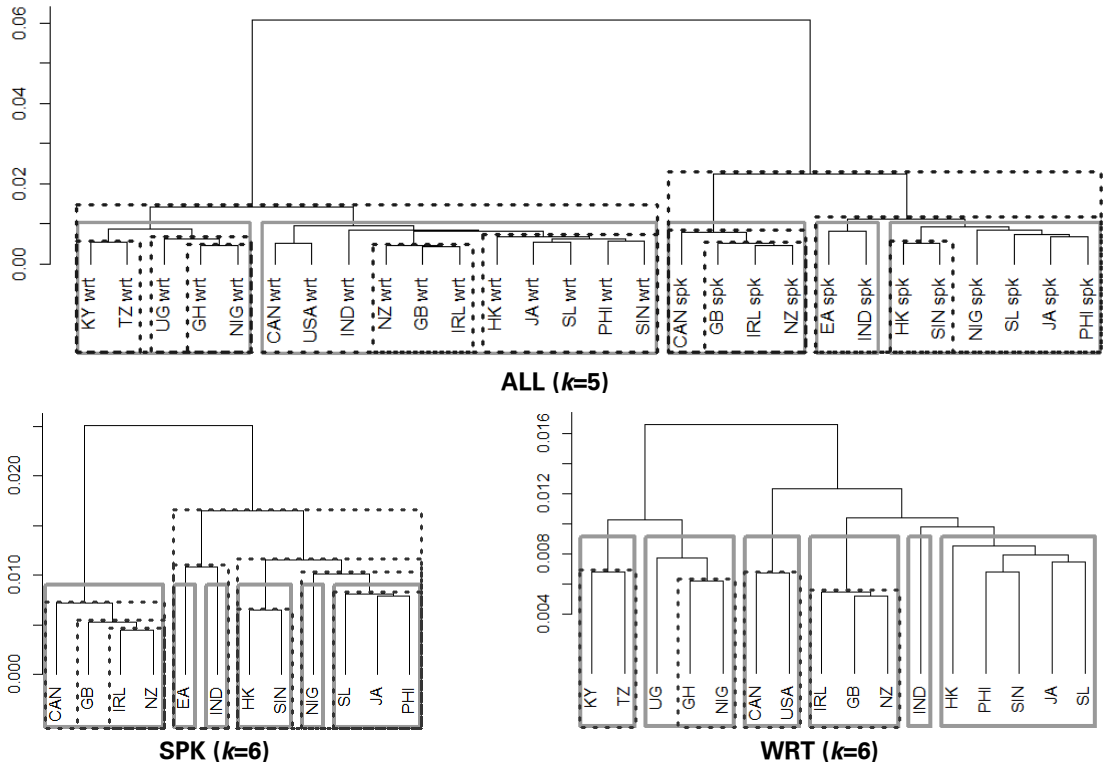


Figure 5.102: Hierarchical clustering results for lexical *g* 3-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

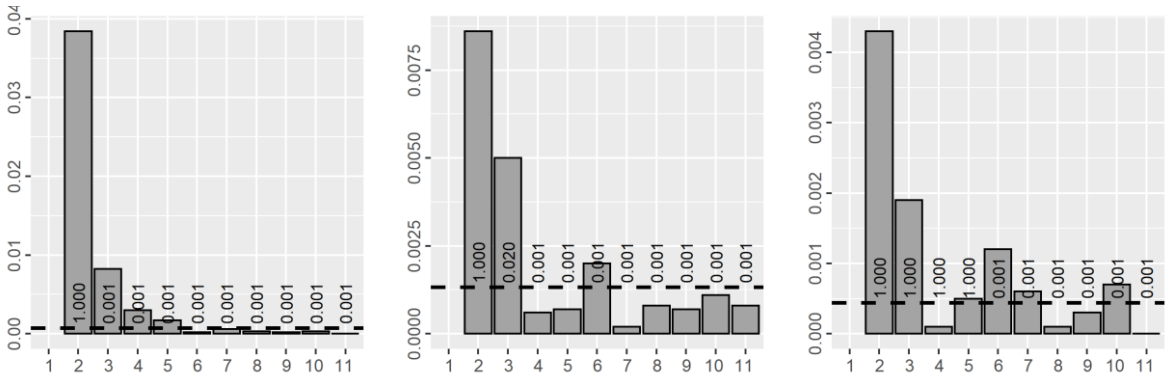


Figure 5.103: Jumps in node heights and respective *p*-values for lexical *g* 3-grams

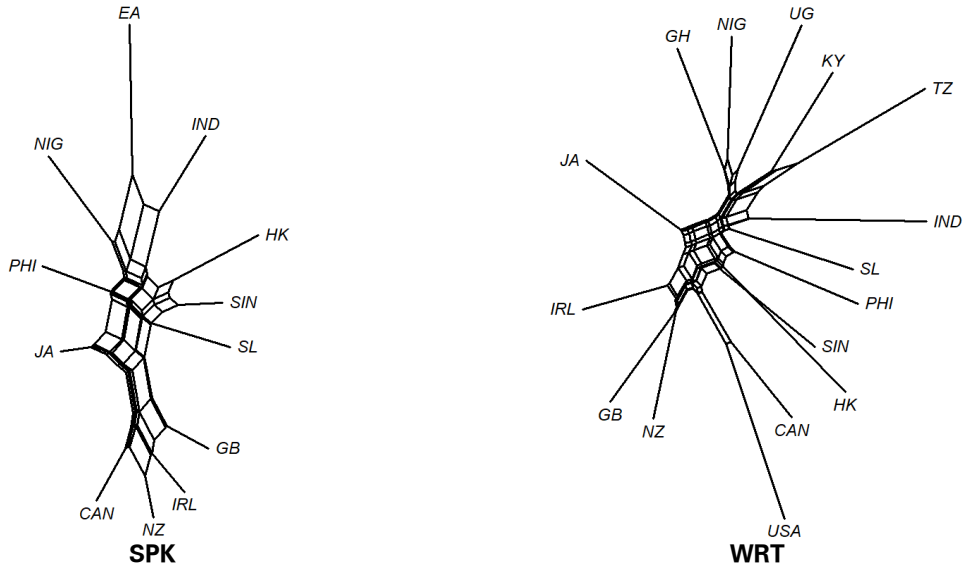


Figure 5.104: NeighborNets of the spoken and written data for lexical *g* 3-grams

SPK finds the first relevant split at three clusters, again separating IC and OC (with distinct EA+IND), while a case can also be made for *k*=6, at which point HK+SIN becomes a discrete cluster, the EA+IND group separates and NIG is also split off from the remaining OC varieties. WRT prefer six clusters, which results in segmentation by fine-grained regionality, with IND separated from the remaining Asian varieties.

The NeighborNets (Figure 5.104) for SPK indicate discreteness of IC, relative closeness of HK+SIN and the separation of mutually relatively distant EA, IND and NIG. For WRT, a separation of the African group can be identified, but KY+TZ also shares features with IND, and UG inhabits an intermediary position. WRT also supports regional separation within the IC cluster and provides some support for HK+SIN.

K-means (Figure 5.105 and Table 5.54) prefers larger clusters than for 2-grams, with *k*=3 for ALL supporting the separate spoken IC group, and *k*=4 further differentiating speech by analyzing a separate status of EA+IND and HK+SIN. Note, however, that a binary structure also appears indicated through the elbow rule, returning the usual spoken/written distinction. In WRT, *k*=6 results in almost the same clusters as above, with only IND+SL forming a distinct group and JA associated with one of the IC groups.

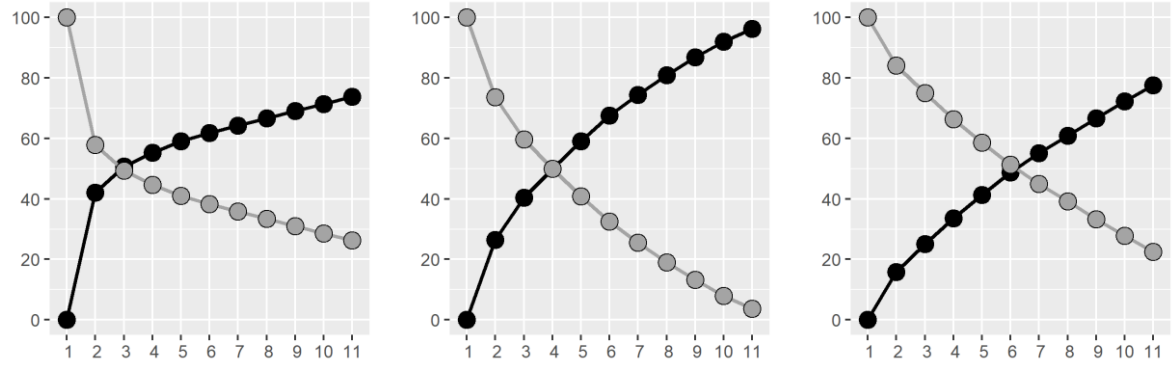


Figure 5.105: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *g* 3-grams

Table 5.54: K-means clustering results for specific values of *k* for lexical *g* 3-grams

	ALL (*k*=3)		**SPK (*k*=4)**		**WRT (*k*=6)**
1	All written corpus parts	1	EA, IND	1	IND, SL
2	CAN, GB, IRL, NZ $_{SPK}$	2	CAN, GB, IRL, NZ	2	CAN, USA
3	EA, HK, IND, JA, NIG, PHI, SIN, SL $_{SPK}$	3	JA, NIG, PHI, SL	3	KY, TZ
		4	HK, SIN	4	GB, IRL, JA, NZ
				5	HK, PHI, SIN,
				6	GH, NIG, UG

4-grams as the longest sequences show a profound loss of stable clusters (Figure 5.106). This affects the separate datasets in particular, where writing shows no stable groups at all and speech only finds support for a partial IC cluster vs. the remaining OC varieties after (the usual) EA, IND, NIG and (slightly less usual) HK+SIN are separated. Within ALL, this is partially reflected in a partition between spoken IC and OC and incomplete written IC clusters and PHI+SIN beyond a significant spoken/written distinction. It should be noted, however, that the non-EA+IND spoken cluster only barely misses significance at AU=94.

Significant jumps (Figure 5.107) are only found for ALL after *k*=4 and seem particularly indicated for *k*=5, segmenting written African and all remaining varieties and spoken IC vs. OC, with separation of EA+IND only at five clusters. SPK indicates significant jumps from *k*=3, at which point EA+IND is separated from all other OC varieties and an IC cluster. Finer segmentation is achieved at *k*=5, which splits the EA+IND group and separates HK+SIN from the remaining OC varieties. Slightly less indicated *k*=6 additionally provides indications towards a separate status of NIG. For WRT, *k*=3 separates roughly into CAN+USA, Africa (+IND+SL) and a remaining group of Asian and IC varieties plus JA. Slightly larger jumps of higher significance are found at *k*=6, which splits TZ and IND from the former group and separates the latter into the Asian and IC components.

NeighborNet analysis (Figure 5.108) of the SPK data indicates similar results as the HCA above and analyses at shorter sequence lengths, separating IC as well as EA and IND from the remaining varieties. NIG is, however, found less distinct than at shorter lengths. For WRT, a less coherent singular IC group is captured, which is better regarded as two separate clusters, and HK+SIN again inhabits an intermediary position. The African varieties (+IND) separate from the data, and UG sits between NIG and TZ.

Intersects for k-means clustering (Figure 5.109 and Table 5.55) indicate *k*=4 for ALL, isolating spoken IC and written African varieties from the respective remaining forms (with an elbow at *k*=2 and *k*=5 splitting off EA+IND$_{SPK}$). For SPK, IC is similarly distinguished from remaining varieties, with the only difference between equally indicated *k*=3 and *k*=4 found in either separate or combined EA and IND. For WRT, *k*=6 identifies IND+SL and generally results in identical clusters as sequences of length 3.

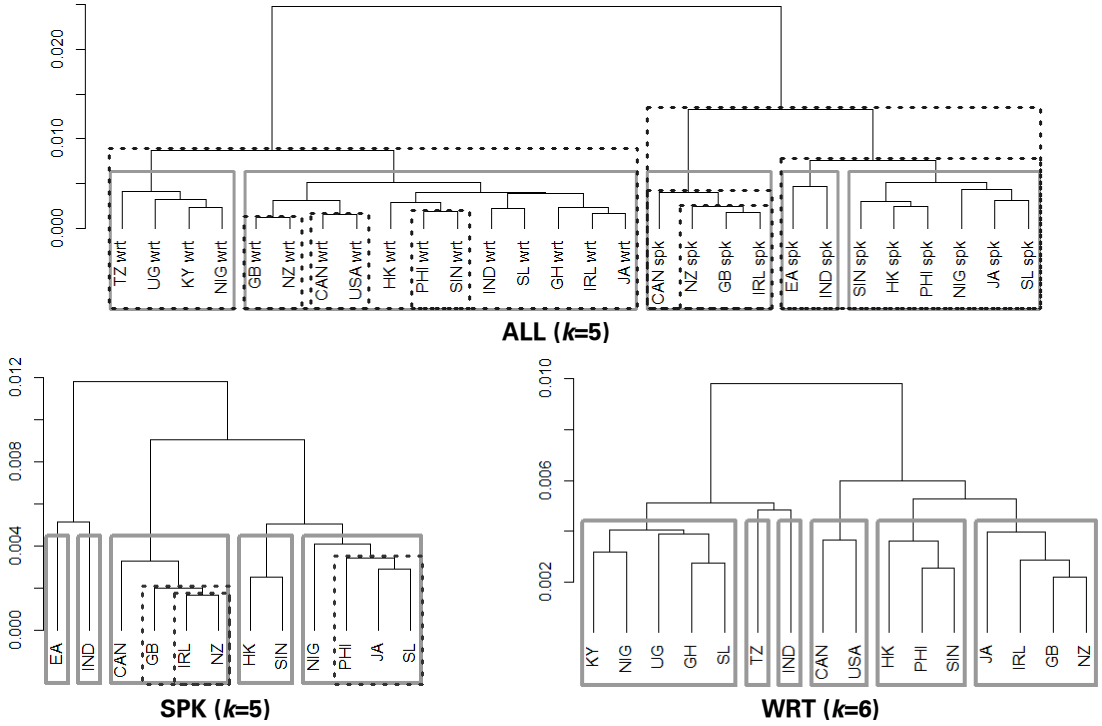


Figure 5.106: Hierarchical clustering results for lexical *g* 4-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

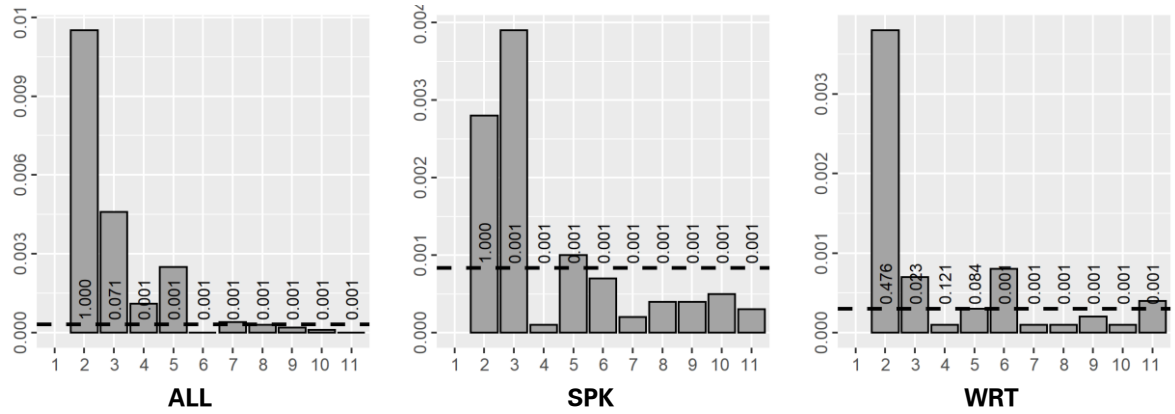


Figure 5.107: Jumps in node heights and respective *p*-values for lexical *g* 4-grams

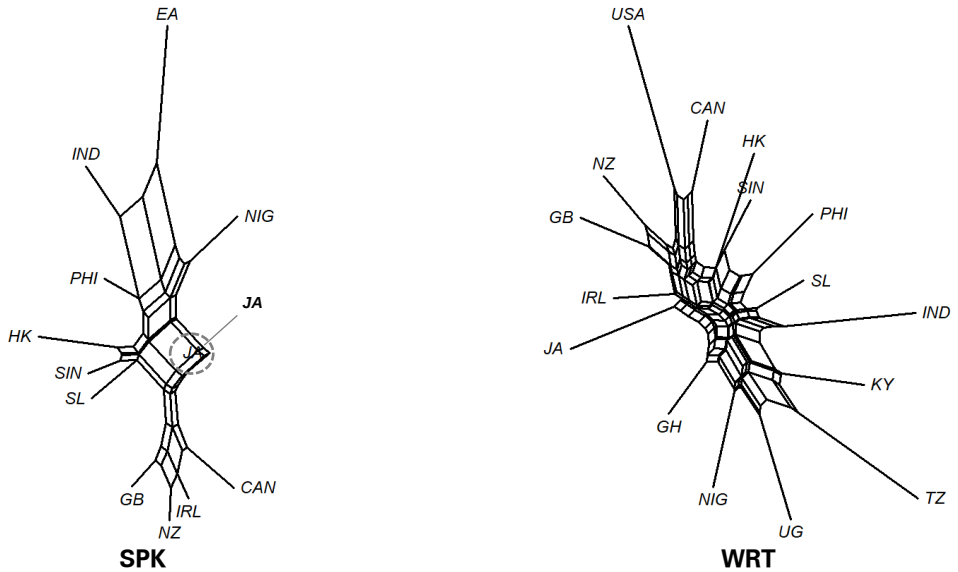


Figure 5.108: NeighborNets of the spoken and written data for lexical *g* 4-grams

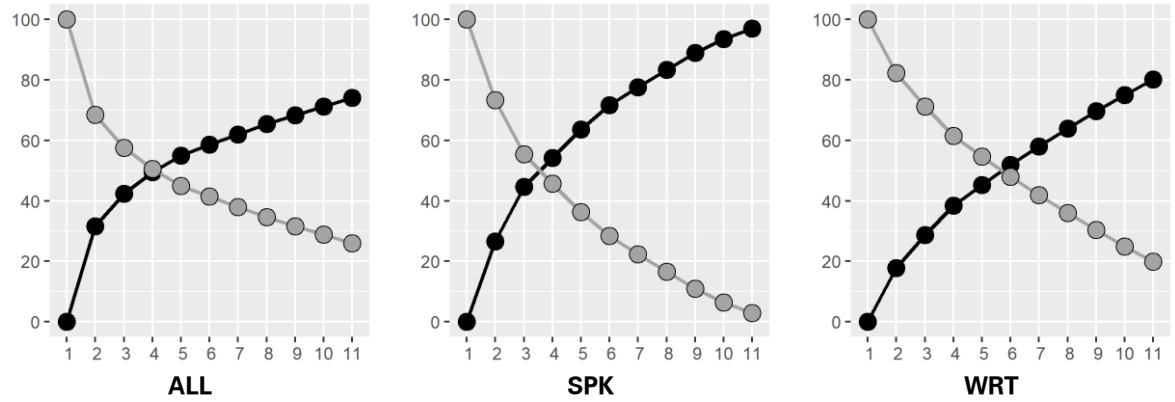


Figure 5.109: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *g* 4-grams

Table 5.55: K-means clustering results for specific values of *k* for lexical *g* 4-grams

	ALL (*k*=4)		SPK (*k*=4)		WRT (*k*=6)
1	CAN, GB, IRL, NZ $_{SPK}$	1	IND	1	IND, SL
2	EA, HK, IND, JA, NIG, PHI, SIN, SL $_{SPK}$	2	EA	2	CAN, USA
		3	HK, JA, PHI, SIN, SL	3	HK, PHI, SIN
3	CAN, GB, HK, IRL, JA, NZ, PHI, SIN, USA $_{WRT}$	4	CAN, GB, IRL, NZ	4	GB, IRL, JA, NZ
				5	KY, TZ
4	KY, TZ, GH, IND, NIG, SL, UG $_{WRT}$			6	GH, NIG, UG

5.3.5 Delta $P_{2|1}$

N-gram association values on the basis of Delta P display the measure's design to highlight 'forward' directionality: Most of the items in the top positions in Table 5.56 exhibit intuitively clear cases of a first item selecting a consecutive one, while the bottom items frequently represent more 'normal' strings of items not frequently apparent as collocations (but not clear cases of 'backward' directionality, either). An exception to a good separation between collocated and uncollocated forms can, however, be found in 4-grams, where mostly premodification of *be able to* is assigned top association values. Additionally, a few low-scoring items might be said to be relatively collocated sequences, such as *and of course the*, *in such a way*, *is made up of*. However, it could also be argued that in these cases 4-grams are too long and 3-grams contained therein better capture the collocation, e.g. *and of course* or *made up of*. In any case, top collocates are relatively often extended forms of shorter sequences. Highest association scores are often assigned to verb+particle combinations of similar use in speech as well as writing, which may explain similarities between the modes.

Table 5.56: Lexical *ΔP* *n*-grams with highest and lowest association scores

2-grams type	*ΔP*	3-grams type	*ΔP*	4-grams type	*ΔP*
Spoken					
according to	0.9589	depending on the	0.5302	should be able to	0.4187
able to	0.9292	dealing with the	0.5042	will be able to	0.4042
tends to	0.9261	according to the	0.4971	may be able to	0.4015
supposed to	0.9233	supposed to be	0.4928	won't be able to	0.3912
unable to	0.9101	behalf of the	0.4844	would be able to	0.3894
lined up	0.9039	be able to	0.4826	you'll be able to	0.3813
these the	-0.0540	think the the	-0.0244	and of course the	-0.0014
good the	-0.0546	the one the	-0.0251	two and a half	-0.0020
they the	-0.0553	know the the	-0.0252	one and a half	-0.0030
we the	-0.0560	to you the	-0.0303	and the and the	-0.0049
a the	-0.0573	so the the	-0.0342	and the the the	-0.0253
i the	-0.0577	the the the	-0.0419	the the the the	-0.0419
Written					
according to	0.9631	depending on the	0.5860	should be able to	0.4425
irrespective of	0.9497	accordance with the	0.5357	will be able to	0.4104
regardless of	0.9410	depend on the	0.5182	would be able to	0.3925
spite of	0.9386	according to the	0.5027	have been able to	0.3647
accounted for	0.9382	based on the	0.4931	did not want to	0.3596
unable to	0.9369	comply with the	0.4887	in accordance with the	0.3577
day the	-0.0632	to that of	-0.0180	in such a way	0.0413
should the	-0.0659	to study the	-0.0184	was found to be	0.0390
well the	-0.0680	be the first	-0.0194	the extent to which	0.0293
you the	-0.0684	be the most	-0.0209	from time to time	0.0180
will the	-0.0717	the way the	-0.0236	is not to say	0.0171
it the	-0.0731	the time the	-0.0256	is made up of	0.0095

Delta P **2-grams** are characterized by a relative absence of stable clusters (Figure 5.110). Beyond the spoken and written branches in ALL, only partial IC groups, African subclusters and a single HK+SIN cluster emerge as stable (dotted lines) in the different datasets: ALL reports the IC_{GB} groups, supported by SPK and only barely not (AU=94) by WRT. SPK further retrieves HK+SIN, and WRT indicates East and West African clusters, which emerge as one consistent African cluster in the written branch of ALL.

Segmentation by jump heights (Figure 5.111) only achieves significance after *k*=7 for ALL, which separates the spoken IC, HK+SIN, the 'usual suspects' EA, IND and NIG and the remaining OC varieties, while the written branch distinguishes Africa (plus IND) from all other varieties. Consecutively finer splits fragment the spoken branch (unary CAN and PHI) before separating IND_{WRT} from Africa, only afterwards retrieving the partial IC group. SPK indicates significance at *k*=6 but larger jumps after *k*=8. The coarser separation conforms to the analysis found for ALL, and finer segmentation similarly splits off CAN_{SPK} before separating EA and IND. For WRT, *k*=7 is indicated, identifying the two African clusters but nothing else of apparent consistency.

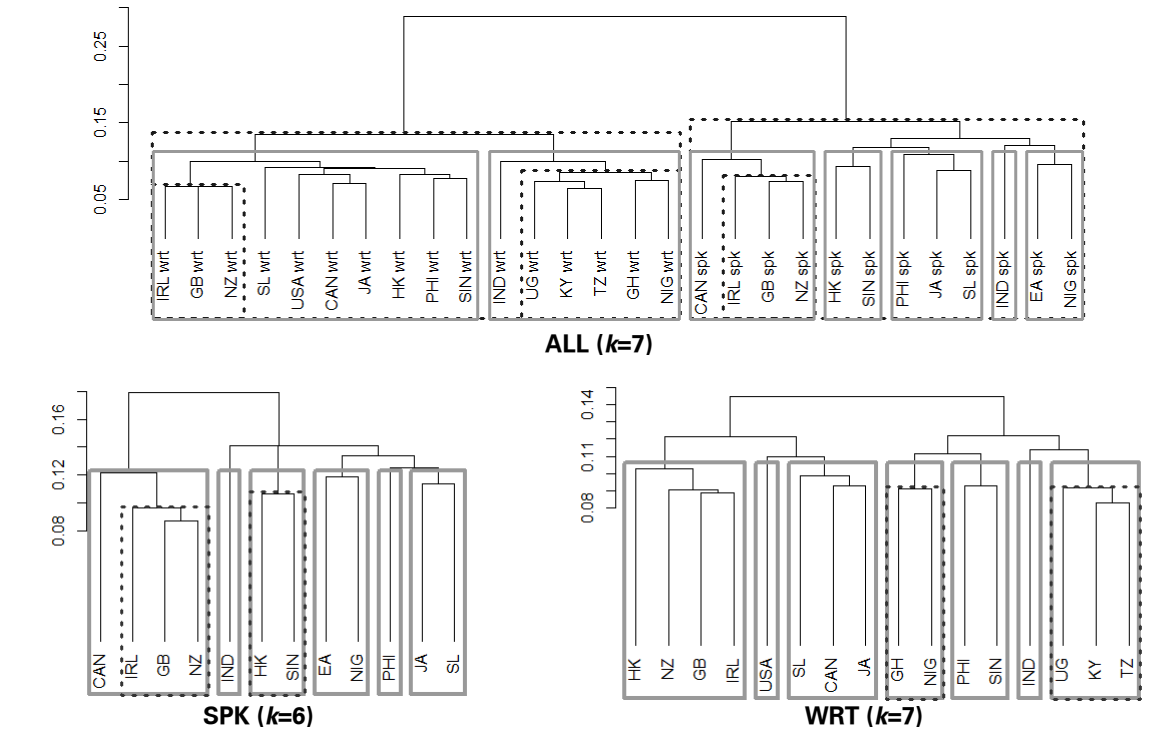


Figure 5.110: Hierarchical clustering results for lexical *ΔP* 2-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

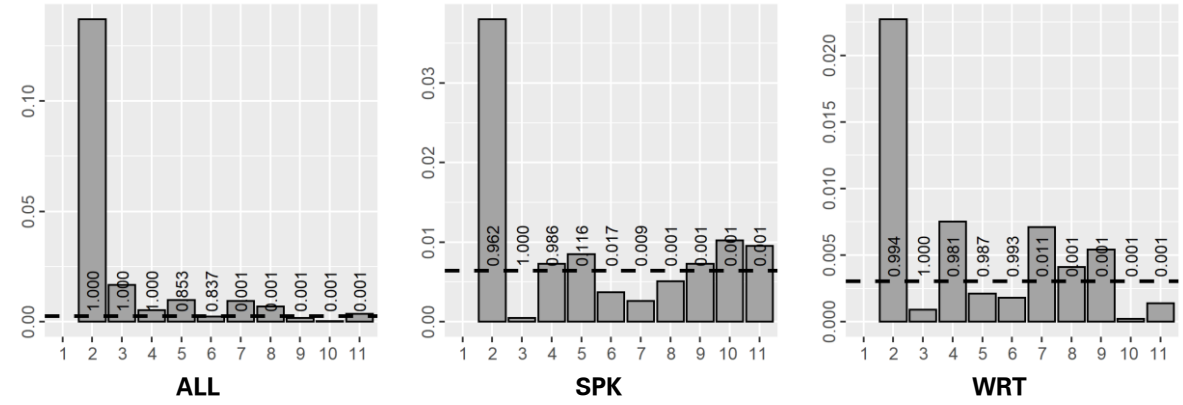


Figure 5.111: Jumps in node heights and respective *p*-values for lexical *ΔP* 2-grams

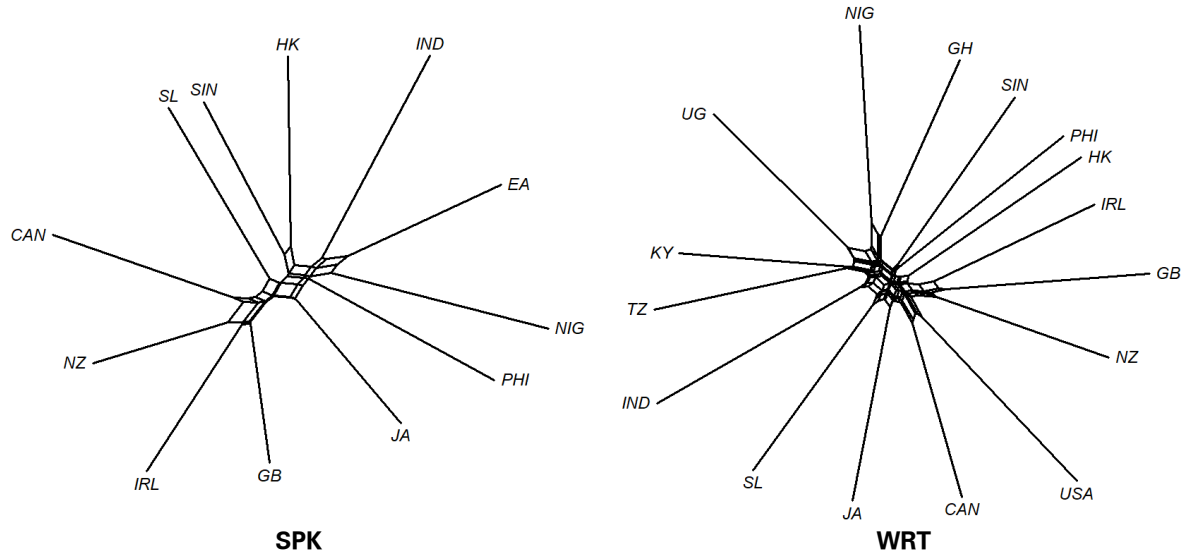


Figure 5.112: NeighborNets of the spoken and written data for lexical *ΔP* 2-grams

The NeighborNets in Figure 5.112 suggest a relatively weak separation of spoken IC as well as HK+SIN and almost equidistance of EA, IND and NIG. WRT shows high degrees of overall similarity in the data but still identifies some mutual distance of West and East African varieties to the remaining data. Furthermore, two related IC clusters can be confirmed. The NeighborNets thus indicate some similarity to previous datasets despite lack of stability and significance.

In accordance with the fine clustering indicated above, k-means analysis (Figure 5.113 and Table 5.57) also tends towards high values for *k*. This affects ALL in particular (*k*=10!), where results, particularly for speech, largely conform to the hierarchical analysis at a finer segmentation. The spoken and written datasets favor segmentation along largely identical lines as above (*k*=5 for SPK and *k*=8 for WRT respectively merging EA and IND or splitting off HK).

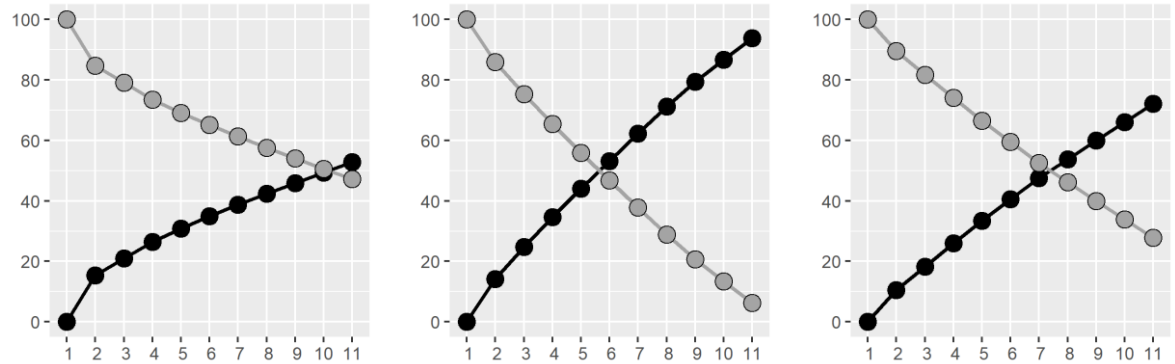


Figure 5.113: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *ΔP* 2-grams

Table 5.57: K-means clustering results for specific values of *k* for lexical *ΔP* 2-grams

	ALL (*k*=10)		**SPK (*k*=6)**		**WRT (*k*=7)**
1	EA, NIG $_{SPK}$	1	EA	1	JA, SL
2	IND $_{SPK}$	2	NIG	2	GB, HK, IRL, NZ
3	CAN, GB, IRL, NZ $_{SPK}$	3	CAN, GB, IRL, NZ	3	CAN, USA
4	JA, SL $_{SPK}$	4	JA, PHI, SL	4	GH, NIG
5	PHI $_{SPK}$	5	IND	5	IND
6	HK, SIN $_{SPK}$	6	HK, SIN	6	KY, TZ, UG
7	JA, PHI, SIN, SL $_{WRT}$			7	PHI, SIN
8	IND $_{WRT}$				
9	KY, TZ, GH, NIG, UG $_{WRT}$				
10	CAN, GB, HK, IRL, NZ, USA $_{WRT}$				

Delta P **3-grams** exhibit similar issues in finding stable clusters as shorter sequences, with stable groups retrieved only in speech (Figure 5.114). For ALL and SPK, this supports IC clusters (only partially in ALL), additionally separating SPK into IC and OC and partitioning off EA+IND within the latter group. Within the remaining OC

varieties, a HK+SIN and JA+SL (substantiated also within ALL) is detected. While relatively little substantiation is thus to be found in the present data, the common distinction of OC into the separate cases of EA and IND vs. HK+SIN and the remaining varieties thus appears the most warranted one. While no written stable clusters are detected, CAN+USA as well as the non-American IC+Asia group are reported as AU=93 in WRT.

Significant jump heights (Figure 5.115) does not support the spoken/written distinction in ALL, only returning significance after *k*=3 splits off the spoken IC varieties. Finer segmentation is indicated at *k*=5, separating the written data into African varieties (plus IND, as for 2-grams) and the rest and segmenting EA+IND from the remaining varieties in speech. Significant jumps in SPK are found at *k*=5, which confirms the IC group, a division between EA and IND and two remaining groups, one of which summarizes Southeast-Asian varieties. WRT displays significant jumps after *k*=5, at which point IND is separated from the African group, two IC clusters emerge and the remaining varieties are joined in one group of different phases but mostly coherently Asian regionality.

Visual inspection of the NeighborNets (Figure 5.116) also suggests separation of the IC spoken varieties and highlights relatively similar distances between EA and IND as well as NIG, contrary to the binary clustering results. For writing, the IC varieties remain relatively connected but CAN+USA presents itself as a distinct subgroup (and also IRL to some degree). Proximity of IND to the African group is visible but a IND+SL cluster could similarly be advised given relative proximity.

K-means clustering results (Figure 5.117 and Table 5.58) are highly congruent with the hierarchical analysis. For ALL, a coarser partition than for 2-grams appears advisable at *k*=6, segmenting the data identically to the hierarchical analysis with the exception of a separate IND+SL_{WRT} cluster. SPK results fully agree with the hierarchical analysis (and would also at *k*=4 in both analyses, merging clusters #1 and #3 below) while WRT additionally splits the African data by region and separates HK+SIN+SL from the 'Asian' cluster .

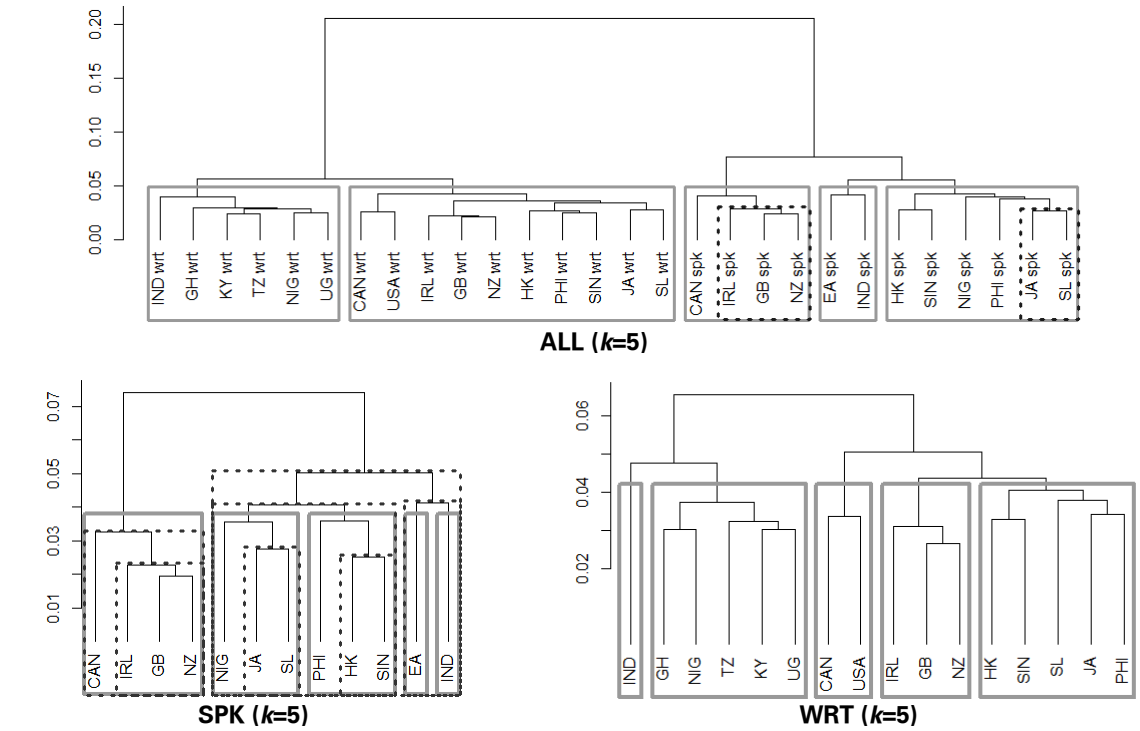


Figure 5.114: Hierarchical clustering results for lexical *ΔP* 3-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

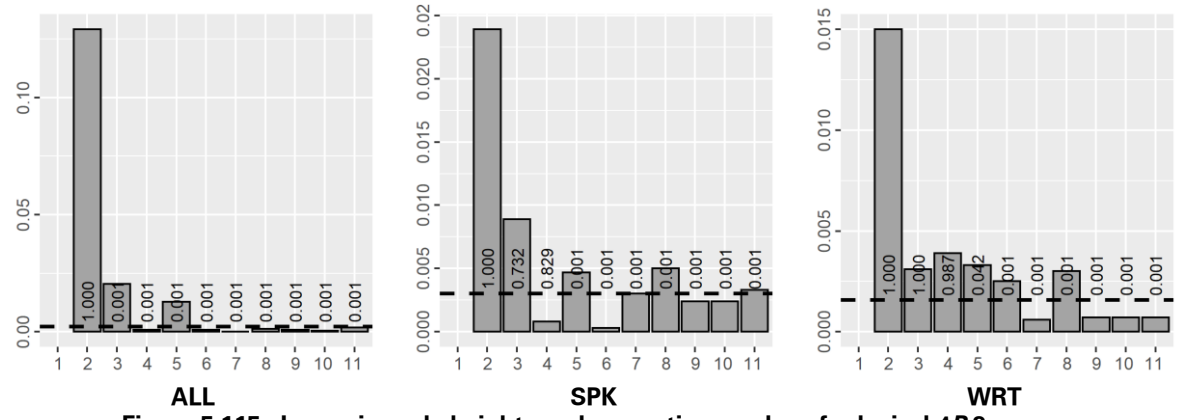


Figure 5.115: Jumps in node heights and respective *p*-values for lexical *ΔP* 3-grams

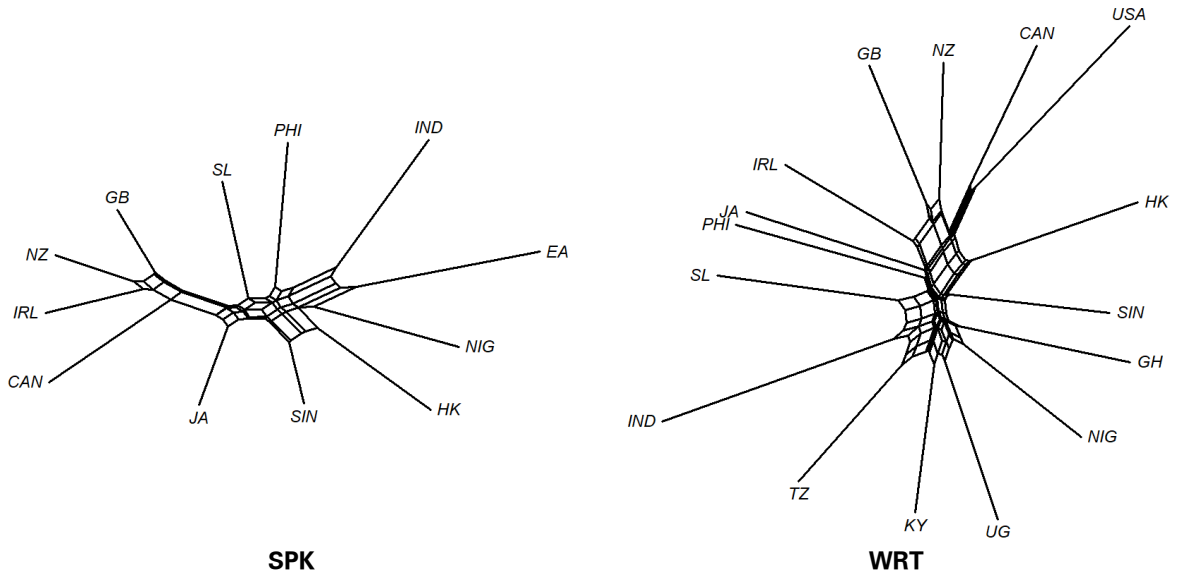


Figure 5.116: NeighborNets of the spoken and written data for lexical *ΔP* 3-grams

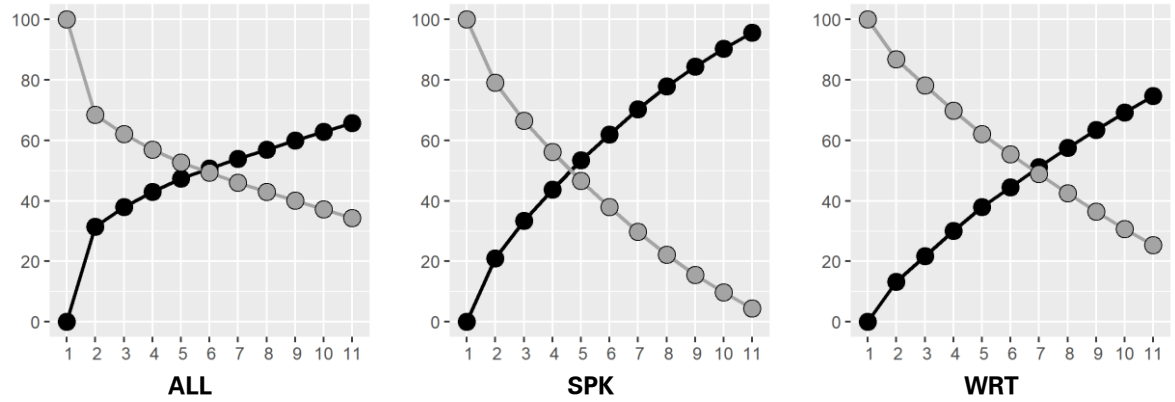


Figure 5.117: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *ΔP* 3-grams

Table 5.58: K-means clustering results for specific values of *k* for lexical *ΔP* 3-grams

	ALL (*k*=6)		SPK (*k*=5)		WRT (*k*=7)
1	CAN, GB, IRL, NZ $_{SPK}$	1	JA, NIG, SL	1	JA, PHI
2	EA, IND $_{SPK}$	2	CAN, GB, IRL, NZ	2	KY, TZ, UG
3	HK, JA, NIG, PHI, SIN, SL $_{SPK}$	3	HK, PHI, SIN	3	CAN, USA
		4	IND	4	GB, IRL, NZ
4	IND, SL $_{WRT}$	5	EA	5	GH, NIG
5	CAN, GB, HK, IRL, NZ, PHI, SIN, USA $_{WRT}$			6	IND
				7	HK, SIN, SL
6	KY, TZ, GH, JA, NIG, UG $_{WRT}$				

4-grams continue Delta P's trend towards ever sparser substantiation in terms of stable clusters (Figure 5.118), with even the spoken/written distinction again losing significance (but barely at AU=94). While WRT finds no stable groups, the written branch of ALL highlights a strange selection of varieties from different regions and phases. The next-larger cluster combines these to at least resemble a mostly African cluster (plus some potential norm providers), and misses significance at AU=93, together with the opposed mostly Asian cluster (AU=94), so that a vague distinction by Asia/Africa (and potential norm-providers) might conceivably be argued for. For SPK, only the separate dataset shows stable clusters, of which the above-mentioned IC_{GB} cluster constitutes a familiar group, supplemented by a combination of IND+PHI.

Significant jumps (Figure 5.119) are returned for ALL not until *k*=5, coinciding with the barely non-significant written groups above and separating the written branch by a vague Asia vs. Africa focus and speech by IC and OC with EA sectioned off. Overlap is even limited with the clusters indicated for SPK both in its finer (*k*=6) as well as the coarser (*k*=4) version. In both cases, however, IND+PHI and NIG+HK+SIN are returned and EA analyzed as distinct.

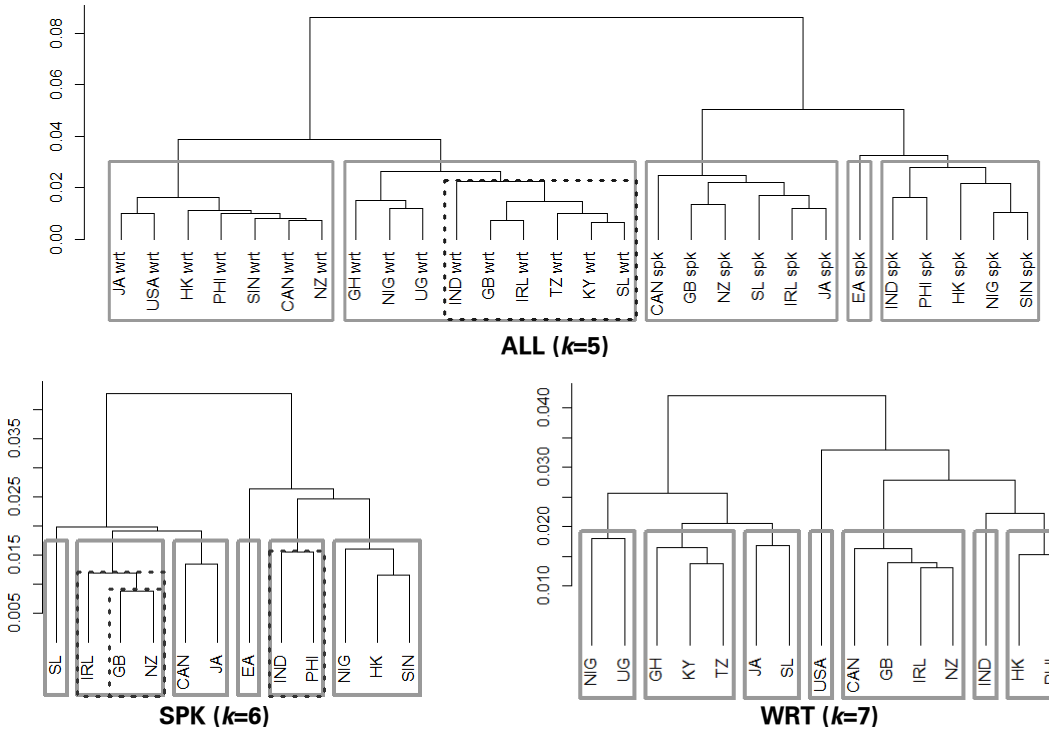


Figure 5.118: Hierarchical clustering results for lexical *ΔP* 4-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

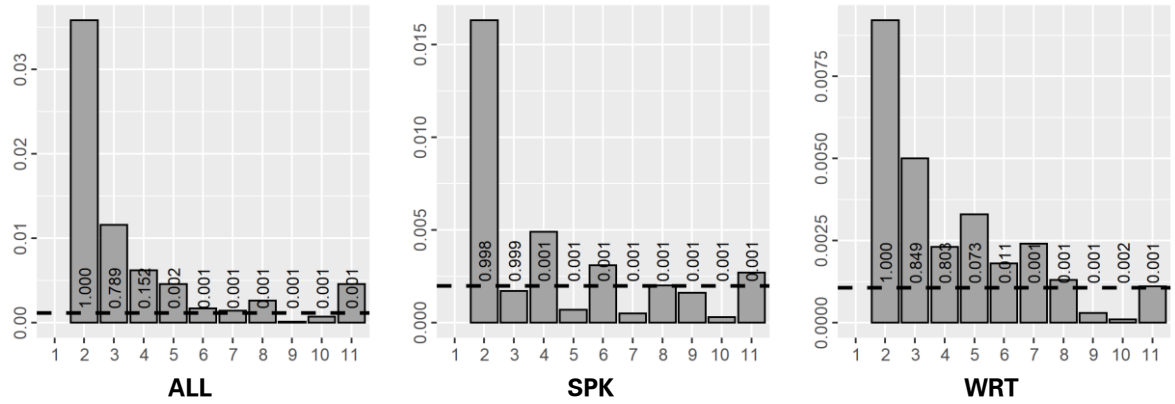


Figure 5.119: Jumps in node heights and respective *p*-values for lexical *ΔP* 4-grams

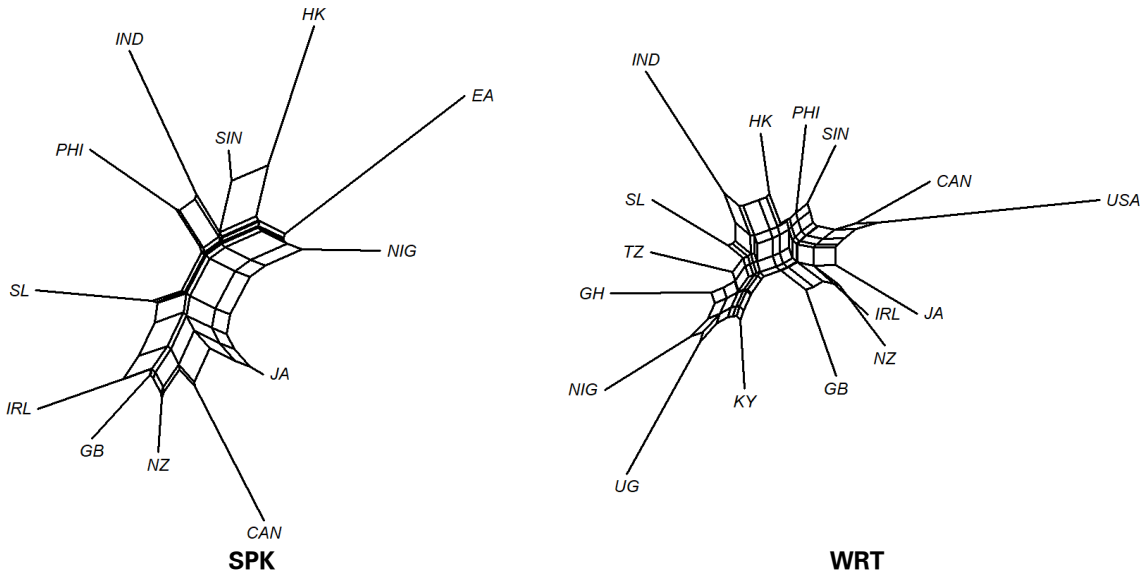


Figure 5.120: NeighborNets of the spoken and written data for lexical *ΔP* 4-grams

For WRT, *k*=7 surpasses the heights for the first significant jump at *k*=6, resulting in two African groups (but not aligned to East and West or phases, and merged with CAN+JA at *k*=6), a fragmented IC group and the HK+PHI+SIN cluster occasionally returned for *ΔP* at shorter lengths and in previous analyses.

NeighborNet analysis (Figure 5.120) indicates clearer IC clusters than the HCA (two in writing, interspersed by JA, and one in speech), similarity of SIN to HK (SPK) or PHI (WRT) and some proximity between IND and SL in writing. In speech, IND clusters with PHI instead of its usual combination with EA, which in turn produces a regionally consistent African cluster with NIG.

Again, consecutively fewer clusters than at shorter lengths are found by k-means (Figure 5.121 and Table 5.59). The results for speech agree between ALL at *k*=7 and SPK set at *k*=4, returning somewhat coherent groups like IC and HK+SIN, but each merged with an additional variety (SPK thus producing HCA clusters at *k*=4), and also analyzing EA and IND+PHI as separate. Results for WRT are similarly inconclusive and relatively different from the above except for the two heterogenous African groups and several Asian varieties within a shared cluster.

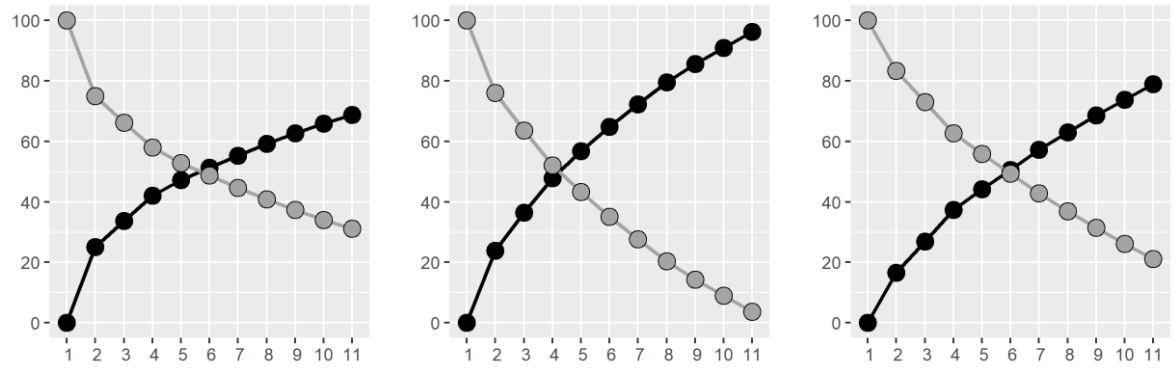


Figure 5.121: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for lexical *ΔP* 4-grams

Table 5.59: K-means clustering results for specific values of *k* for lexical *ΔP* 4-grams

	ALL (*k*=6)		**SPK (*k*=4)**		**WRT (*k*=6)**
1	CAN, GB, IRL, JA, NZ, SL $_{SPK}$	1	CAN, GB, IRL, JA, NZ, SL	1	USA
2	HK, NIG, SIN $_{SPK}$	2	EA	2	NIG, UG
3	EA$_{SPK}$	3	IND, PHI	3	KY, TZ, GH
4	IND, PHI$_{SPK}$	4	HK, NIG, SIN	4	IRL, JA
5	KY, TZ, GH, IND, NIG, SL, UG $_{WRT}$			5	CAN, GB, NZ
6	CAN, GB, HK, IRL, JA, NZ, PHI, SIN, USA $_{WRT}$			6	HK, IND, PHI, SIN, SL

5.4 Static-length POS-grams

As in the previous section, length distributions of variable-length POS-grams can be employed to ascertain whether the lengths selected for lexical *n*-grams also form good analytical categories for the evaluation of static-length POS-grams. In this regard, it firstly turns out that measures react differently to the change in base data: *MI* does not show overly large differences to the lexical data and still retrieves c. 90% of all tokens already at length 2 as well as c. 90% of types until length 3 (only very slightly less preferring 2-gram types in comparison to the lexical data). The other measures remain at c. 60% 2-gram tokens and approximate 85-90% with the inclusion of 3-gram tokens – or increase to about this level in case of *ΔP* (60% of tokens explained by 2-grams and 80% by the inclusion of 3-grams). At the level of types, all measures except *MI* only retrieve c.10% 2-grams (20% in *ΔP* in writing) or about 40% of types until length 3 (*ΔP* approaching 50% but *g* only 20-30% depending on mode). When 4-grams are included, consistently >95% of tokens are explained but type frequencies more frequently lie between 70-80%. Major exceptions concern the *MI* measure as discussed above, but also and *g*, which shows consistently below-average type frequencies at all three lengths, which summarily only explain c. 50% of types and furthermore present a relatively heterogenous distribution across varieties. While it can be argued that 5-grams form a somewhat frequent subset of types, reaching or surpassing the relative frequencies of 2-grams except in case of *MI*, it also must be acknowledged that 5-gram tokens only once (the lexical-gravity data) account for relevant shares of the data. For this reason and for the sake of comparability to static-length lexical sequences, they will be disregarded in the analysis to follow, but a trend towards longer sequences within the POS data is still recognizable.

Average frequencies and association statistics for the POS data are represented in Table 5.60. First of all, no major differences in tokens can be observed between the previous lexical static-length *n*-grams and the present set of POS-grams; the largest difference approaching 2% in the case of spoken 4-grams. Changes in token frequencies across lengths also follow the patterns found in the lexical data, in that uniform, linear decreases are affected in the token data. Type frequencies, however, attest to major diversification across lengths: While the average ICE component contains less

than 12,000 types in speech or writing, a five- to eightfold increase is found at length 3. The greatest absolute differences are however found between 3-and 4-grams.[80] Association scores only follow the uniform and linear changes observed in the lexical token data in about half of the POS-based cases but differences are generally smaller. The second half of cases is characterized by almost identical values across lengths. For types, association scores again produce consistent increases or decreases, but no clear shift can be observed either between either the two shorter or longer types of sequences: Greater differences can sometimes be observed between lengths 2 and 3 (spoken MI, g) but usually occur between 3-and 4-grams, thus also shifting this effect from the shorter to the longer types of sequences.

Table 5.60: Static-length POS *n*-gram average frequencies and association values

	Tokens						Types					
n	**Freq.**	***MI***	***t***	***G²***	***g***	***ΔP***	**Freq.**	***MI***	***t***	***G²***	***g***	***ΔP***
Spoken												
2	574,712	1.11	13.55	4,714.44	4.66	0.0874	11,329	-0.18	-7.29	26.10	-4.54	0.0139
3	517,018	1.10	13.50	4,734.52	4.64	0.0870	83,390	0.34	-2.73	648.06	0.52	0.0290
4	466,417	1.10	13.53	4,776.90	4.64	0.0868	219,383	0.75	4.35	2,191.32	2.88	0.0540
Written												
2	363,136	1.08	12.74	3,810.16	5.38	0.0951	10,220	0.61	-4.95	13.39	-4.51	0.0164
3	318,079	1.06	12.82	3,826.45	5.42	0.0933	50,541	0.59	-2.02	527.71	0.64	0.0374
4	279,094	1.05	12.95	3,880.36	5.44	0.0933	116,532	0.83	3.45	1,656.05	3.11	0.0617

Intersection of the variety-specific datasets removes much of the *n*-gram data in the process, and while some similarities exist to the results in the previous section, several differences can also be observed. Again, more sequences (Table 5.61) are found to be shared in speech rather than in writing, but the difference is lessened considerably, indicating fewer variety-specific patterns in the POS data. Generally, the merged POS data reflects much larger shares of the average component than previously observed for the lexical data. 2-grams, in particular, retain many of the types after merging (46% in speech and 39% in writing), while 3-grams still retain 19% or 20% and 4-grams 7% or 9%. This makes 4-grams a much more numerically relevant category in the POS data than in its lexical counterpart, where a retention of as little as 0.07% was observed. Moreover, both 3- as well as 4-grams show higher frequencies of shared items across varieties than sequences of length 2, even if the latter reflect a larger share of the average component. On the grounds of these frequencies,

[80] Given that the basis of the grammatical data lies mostly in a set of 137 POS tags, this large degree of heterogeneity appears remarkable.

an analysis of 5-grams might again be argued for. However, given the initial observation that 5-grams do not usually provide significant amounts of tokens, further observing the decrease in shared types between lengths 3 and 4, and additionally considering that even 4-grams only produces minor amounts of shared items in the lexical set, their evaluation does not appear sensible within the present analysis.

Table 5.61: POS *n*-gram type frequencies by length in the intersects of the variety-specific datasets

N-gram length	ALL	SPK	WRT
2	3,506	5,219	3,977
3	7,907	16,143	9,981
4	6,808	14,449	9,981

Merging of the separate varietal data was again checked for the number of mutually shared sequences as well as potential outliers which drastically reduce the number of shared sequences. Figure 5.122 indicates the dispersions of shared-type frequencies between any two varietal datasets. As before, outliers are only ever discovered in writing, but solely concern positive outliers: GB/NZ in case of 2- and 3-grams, and additionally IRL/NZ in 3- and 4-grams.

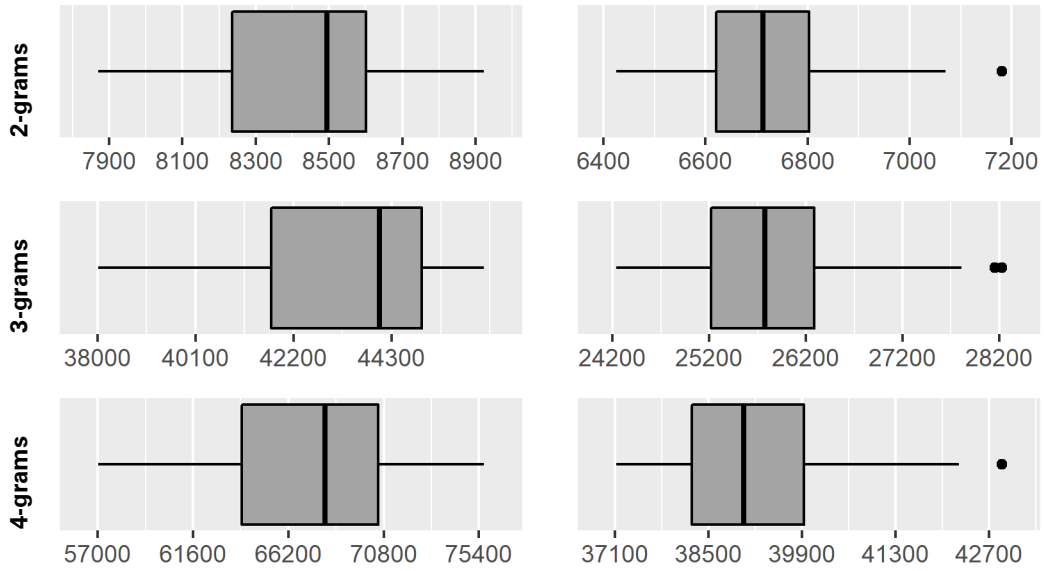


Figure 5.122: Number of shared POS 2-, 3- and 4-grams between any two datasets

5.4.1 MI-score

MI-score POS-grams of all lengths exhibit the same trends as also observed for dynamic-length sequences in that the rare types are consistently placed at the top of the association lists (Table 5.62). This concerns the less frequent POS tags (e.g. DAR for comparative after-determiners *more*, *less*, *fewer*, etc. or DDQV for wh-ever determiners *whichever*, *whatever*), items retained after the removal of word-specific POS tags (e.g. TO for infinitive marker *to*) but in particular so-called ditto tags assigned to "a sequence of similar tags, representing a sequence of words which for grammatical purposes are treated as a single unit." (UCREL 2007). This may extend to relatively frequent collocational sequences such as *in terms of* being annotated as a three-word 'general preposition' (`in_II31 terms_II32 of_II33`; UCREL 2007). Since these form relatively rare occurrences overall and particularly in relation to POS sequences, they (as well as combinations they are contained in) are highlighted by *MI*.

Table 5.62: POS *MI n*-grams with highest and lowest association scores

2-grams type	*MI*	3-grams type	*MI*	4-grams type	*MI*
Spoken					
the part	15.15	on the part	10.95	on the part of	9.10
the light	14.65	the part of	10.27	in the light of	8.08
even though	10.53	the light of	10.02	NN1 on the part	7.56
rather than	10.10	in the light	9.42	the part of AT	7.55
other than	10.09	as long as	7.80	as well as AT1	6.23
REX REX	9.34	as well as	7.80	DAR than MC NNO	5.66
RG NN1	-6.77	and AT and	-3.01	NN1 RR NN1 NN1	-1.65
AT1 AT	-6.81	AT and AT	-3.01	AT RR VVN NN1	-1.78
UH VVI	-7.09	VDI VBZ VVI	-3.02	AT AT AT AT	-1.86
of VBZ	-7.13	VM DD1 VVI	-3.15	and AT and AT	-1.95
AT VBZ	-7.31	NN1 VVI NN1	-3.51	NN1 AT AT AT	-1.98
PPIS1 AT	-7.68	VM NN1 VV0	-3.57	VM AT NN1 VVI	-2.57
the part	14.68	on the part	10.56	on the part of	8.60
even though	10.09	the part of	9.68	NN1 on the part	7.23
rather than	9.45	DDQV DDQV DDQV	9.40	the part of AT	7.04
DDQV DDQV	9.40	as long as	7.08	as well as AT1	5.66
other than	9.40	as soon as	7.07	as soon as PPHS1	5.36
PN1 PN2	8.65	as well as	7.07	VBM VVGK to VVI	5.34
AT VVZ	-6.59	NN1 JJ of	-3.16	to NN1 AT NN1	-1.71
of and	-6.59	to NN1 AT	-3.30	NN1 VVN NN1 NN1	-1.76
VM JJ	-6.60	NN2 NN1 NP1	-3.33	NN1 NN1 NN1 JJ	-1.81
in VVI	-6.83	to NN1 APPGE	-3.37	NP1 NN1 NN1 JJ	-1.93
VM NN2	-6.92	of RR NN1	-3.65	NN1 to NN1 AT	-2.19
AT AT	-8.73	AT1 NN2 NN1	-4.09	VM AT NN1 VVI	-2.96

MI score-based **2-grams** find substantial evidence for stability, albeit with surprising differences between the datasets (Figure 5.123). While ALL finds sufficient evidence for the spoken/written distinction, overall levels of difference between the modes (indicated by branch lengths) are very low, as well as the degree of internal differences (not much more diversity in speech). This finding reflects that *MI* has a relatively distinct focus as an association measure in contrast to others. In both the written branch of ALL as well as in WRT, two IC clusters and a PHI+SIN group are substantiated, with the addition of a GH+NIG cluster in case of ALL (and KY+TZ at AU=93). Differences in speech are very pronounced, with no substantiated clusters in ALL (the highest significance found for GB+IRL at AU=93) but IC stable at all levels in SPK, and almost all distinctions substantiated in OC (the height of HK found at AU=94).

Cutting of the ALL dendrogram is indicated after *k*=3 (Figure 5.124), splitting off the spoken IC data. Predominantly IC and African written varieties separate at *k*=4, and EA+IND and NIG split off in speech at *k*=6, with higher values sectioning off the Asian written varieties (but then also indicating separation of NZ from the spoken IC group). With the help of WRT and its indication of *k*=6, the finer segmentation of ALL finds backup, since both datasets then segment into two IC groups, one or two African clusters, the common HK+SIN+PHI group and IND+SL, with only JA found at varying positions. In the spoken data at the earliest significant *k*=4, the findings for ALL are essentially reproduced, with only IND becoming integrated into the Asian data and EA and NIG being found relatively separate from the remaining data.

NeighborNet analysis (Figure 5.125) supports the above IC/OC distinction in SPK, while WRT separates IC_{GB} and IC_{NA} but also produces a wider IC cluster incorporating HK+SIN and PHI. Writing also identifies an African group with internal East/West subdivision, as well as smaller IND+SL and HK+SIN clusters.

K-means (Figure 5.126 and Table 5.63) tends towards a very fine segmentation in ALL at *k*=7, mostly reproducing the hierarchical results for the spoken branch but rather separating the African and IND+SL groups in writing. The separate data are almost identical in results for WRT (*k*=8 splitting off UG), and for SPK only additionally indicating the difference of NZ to the remaining IC varieties found at finer resolutions of ALL in the HCA (*k*=6 further splits off IND).

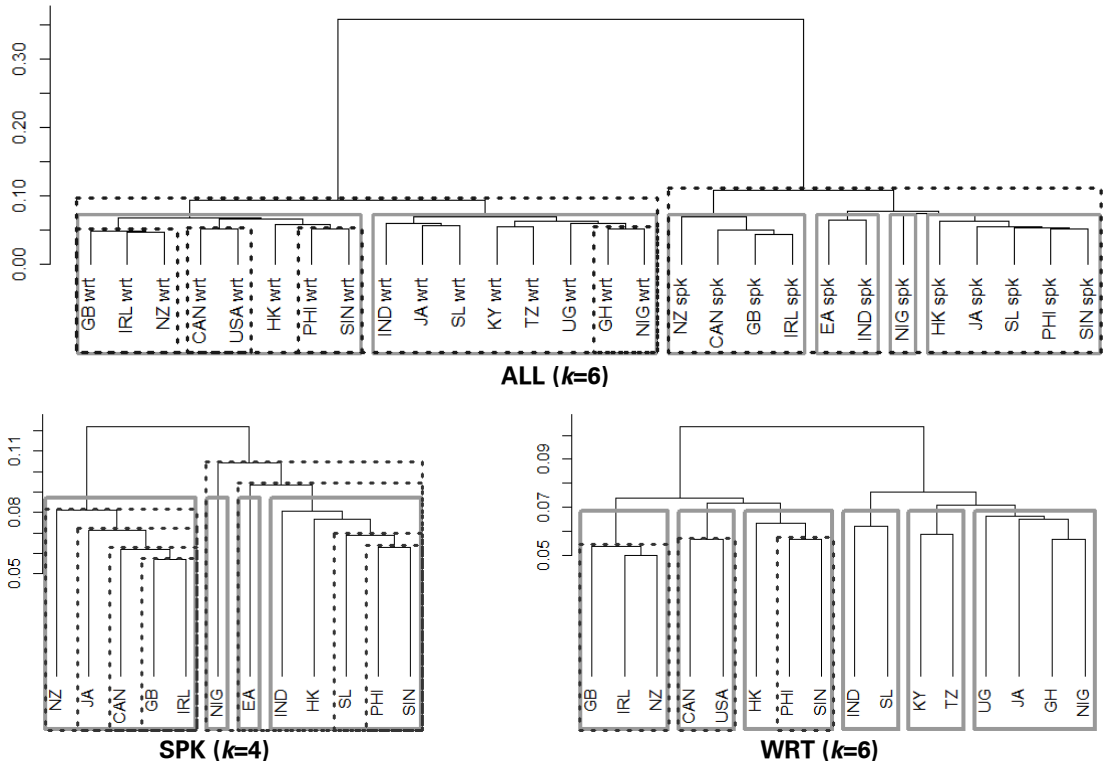


Figure 5.123: Hierarchical clustering results for POS *MI* 2-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

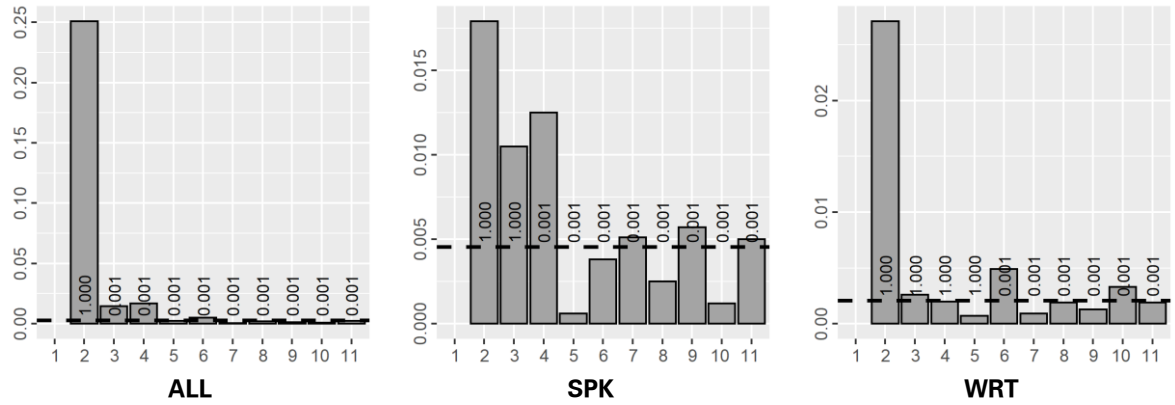


Figure 5.124: Jumps in node heights and respective *p*-values for POS *MI* 2-grams

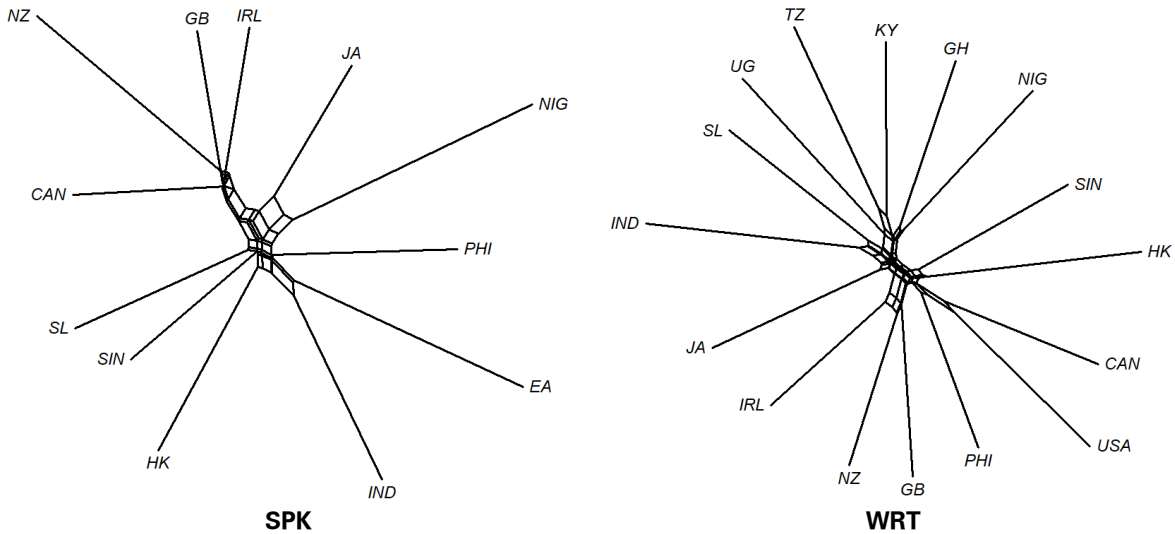


Figure 5.125: NeighborNets of the spoken and written data for POS *MI* 2-grams

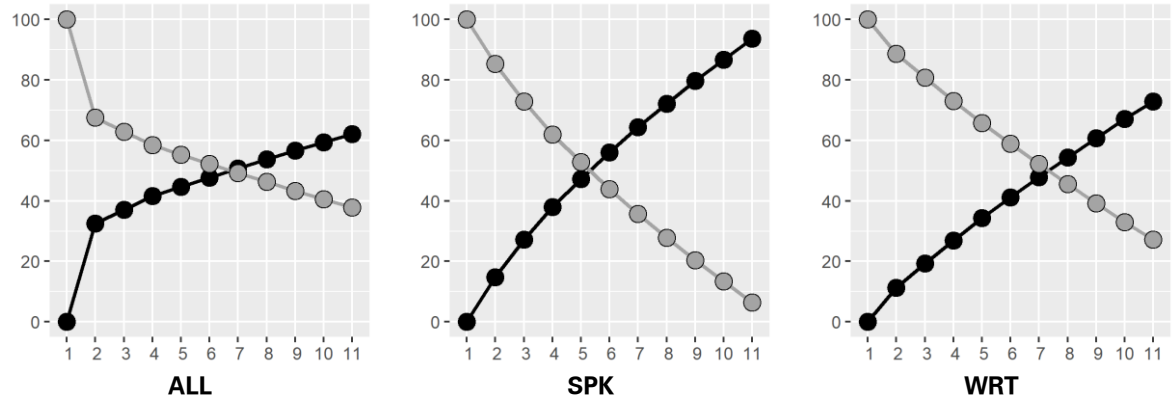


Figure 5.126: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *MI* 2-grams

Table 5.63: K-means clustering results for specific values of *k* for POS *MI* 2-grams

	ALL (*k*=7)		SPK (*k*=5)		WRT (*k*=7)
1	EA, NIG_{SPK}	1	EA	1	GB, IRL, NZ
2	HK, IND, JA, PHI, SIN, SL_{SPK}	2	CAN, GB, IRL, JA	2	IND, SL
3	CAN, GB, IRL, NZ_{SPK}	3	NIG	3	CAN, USA
4	IND, SL_{WRT}	4	HK, IND, PHI, SIN, SL	4	HK
5	KY, TZ, GH, NIG, UG_{WRT}	5	NZ	5	JA, PHI, SIN
6	CAN, HK, PHI, USA_{WRT}			6	GH, NIG, UG
7	GB, IRL, JA, NZ, SIN_{WRT}			7	KY, TZ

With the change from 2- to **3-grams**, differences within the written data appear to lessen in contrast to speech, indicated by relatively shorter branch lengths in ALL. However, the same clusters as before are found stable in ALL (the two written sub-branches at AU=92), and even further ones emerge with the ICE-EA data, the remaining African varieties and IND+JA+SL (Figure 5.127). In the spoken branch, too, more stability is detected, particularly in case of the IC/OC distinction, but the two OC groups also each achieve a moderately high AU=92. The separate data indicate almost entirely identical clusters, with only $EA+IND_{SPK}$ found to align slightly closer with the remaining OC group, and some OC clusters lower in significance ($UG+GH+NIG_{WRT}$ found at AU=94).

The first large significant jumps (Figure 5.128) are found in SPK at *k*=4, splitting off NIG and EA from the OC and IC data, and *k*=6 in writing, separating the data mostly by regionality/epicentricity (with the exception of UG and JA allocated to GH+NIG). While ALL shows the first significant jump at *k*=3, bisecting the spoken branch into IC/OC, *k*=5 is the final one above average height, further indicating a separation of EA+IND from spoken OC and splitting writing coarsely between IC and Africa, with the Southeast Asian varieties clustered with the former and the remaining ones with the latter.

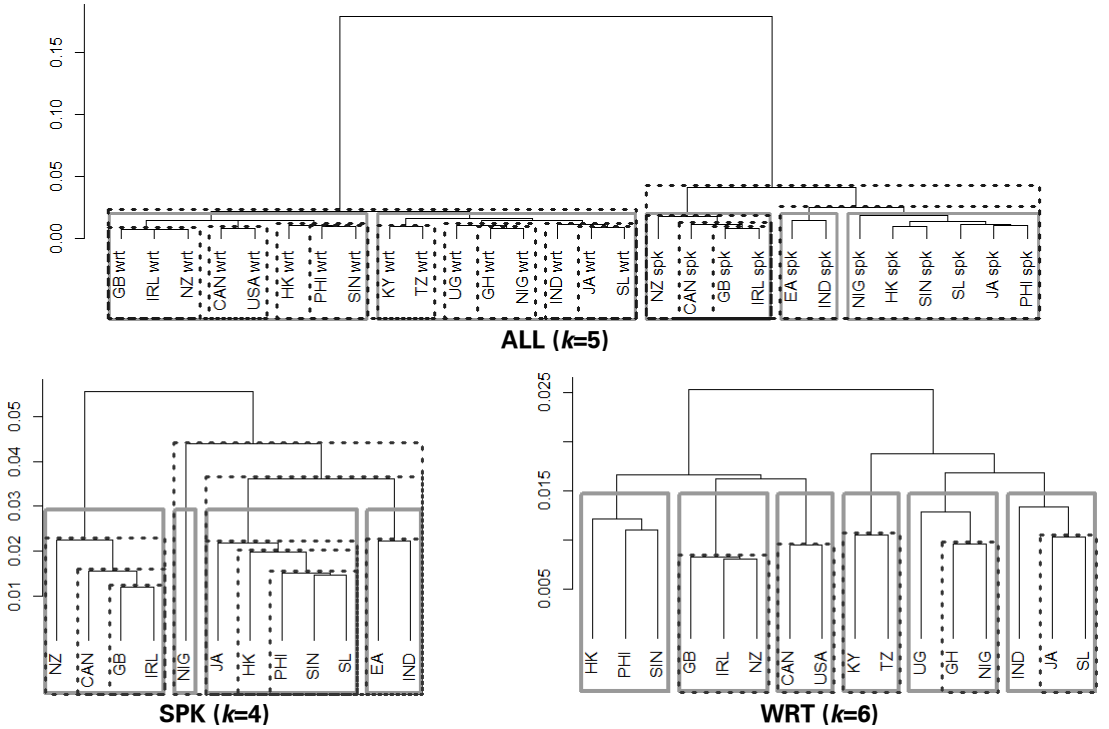


Figure 5.127: Hierarchical clustering results for POS *MI* 3-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

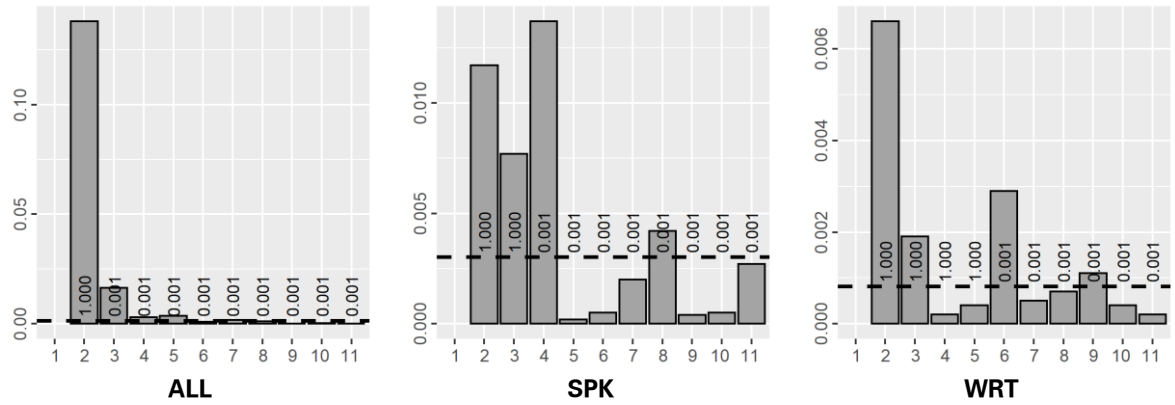


Figure 5.128: Jumps in node heights and respective *p*-values for POS *MI* 3-grams

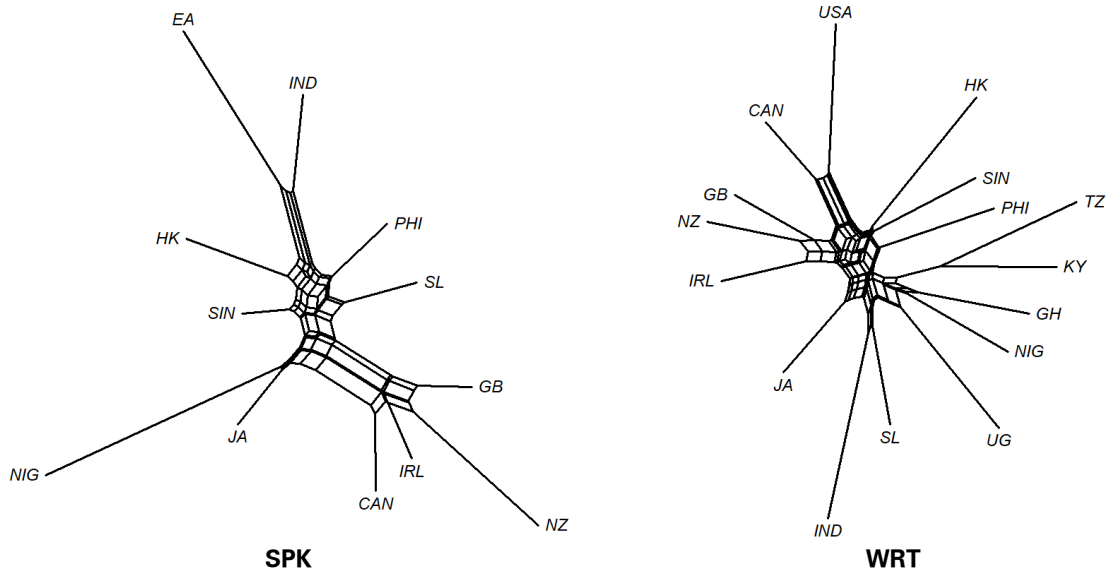


Figure 5.129: NeighborNets of the spoken and written data for POS *MI* 3-grams

Inspection of the NeighborNets (Figure 5.129) reveals a clear separation of IC and EA+IND from the remaining data (and proximity of JA+NIG) in SPK. WRT instead favors two related IC clusters, and supports further regional groups in the form of HK+SIN, IND+SL and two related African clusters. UG is, however, found at some distance from either subcluster, but more so from its regionally proximal varieties.

K-means (Figure 5.130 and Table 5.64) converges earlier for 3-grams than observed for shorter sequences, greatly so for ALL and slightly for the separate sets. It indicates only *k*=2 for ALL (i.e. the spoken/written distinction), substantiating a distinct IC spoken group at *k*=3. In SPK at *k*=4, this is extended to NIG and EA+IND clusters discrete from the remaining OC, and WRT at *k*=6 indicates the exact same clusters as found in the hierarchical data, again a regional interpretation with the exception of JA and, to a lesser degree, UG.

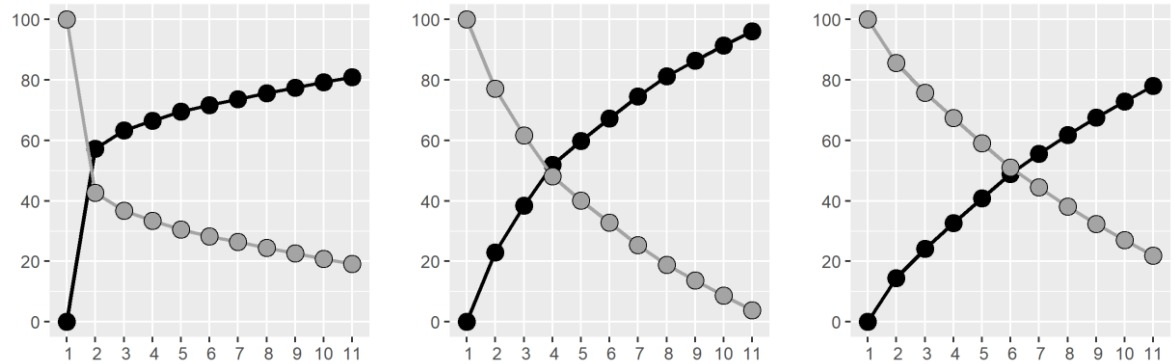


Figure 5.130: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *MI* 3-grams

Table 5.64: K-means clustering results for specific values of *k* for POS *MI* 3-grams

	ALL (*k*=3)		SPK (*k*=4)		WRT (*k*=6)
1	All spoken corpus parts	1	HK, JA, PHI, SIN, SL	1	GH, NIG, UG
2	All written corpus parts	2	EA, IND	2	IND, JA, SL
		3	CAN, GB, IRL, NZ	3	KY, TZ
		4	NIG	4	CAN, USA
				5	GB, IRL, NZ
				6	HK, PHI, SIN

Hierarchical analysis of *MI*-based **4-grams** shows the greatest extent of stability of all lengths (Figure 5.131). While ALL finds less evidence in writing, far more clusters are substantiated in the spoken branch. Written subclusters run along regional/epicentral lines, while in speech a primary segmentation into IC and OC is found. However, the latter group also finds substantiated subclusters which are neither entirely in line with a regional nor evolutionary perspective. SPK, and to a somewhat lesser extent

also WRT, retrieve significant clusters very similar to those found at shorter lengths: Again, a spoken IC/OC distinction, a separateness of EA+IND and a remaining group exhibiting a stepwise patterning are found, with a subsection of mostly Asian varieties. For writing, stable clusters follow a regional perspective, with the exception of IND not being allocated to neighboring varieties and only separate African clusters but not the overall group being substantiated.

Significant jumps (Figure 5.132) for SPK are indicated after *k*=4, which supports the separation of EA+IND and NIG from the remaining OC varieties in addition to retrieving the IC group (the next-largest jump at *k*=5 further subdivides EA and IND). For WRT, significant jumps are only discovered after *k*=6, at which point some regional clusters but also a mixed cluster of Asian (without HK+SIN and IND) and African varieties emerges. This separates only at the less-indicated *k*=7 and splits off UG from the West African varieties at *k*=8, while *k*=9 separates PHI from a group to which it is not well-suited from either a regional or evolutionary standpoint. Since both segmentations agree with stable clusters, the better interpretative choice may be seen to lie in one of the latter segmentations, i.e. finer and mostly regional groups. While many of these clusters can also be seen to emerge in ALL at higher values of *k*, the more strongly indicated ones (up to *k*=4) segment writing and speech, and distinguish IC vs. OC and EA within the latter.

NeighborNet analysis (Figure 5.133) supports most of the clusters above. SPK retrieves one IC cluster and the separateness of EA+IND, while WRT shows differentiation within the IC group and rather differentiates two separate African clusters from the data. It also identifies similarity of HK+SIN as well as IND+SL.

K-means (Figure 5.134 and Table 5.65) converges at identical values as for 3-grams, indicating identical results for ALL and SPK. ALL thus again only indicates a binary separation into speech and writing (at *k*=3 again isolating spoken IC), while SPK identifies EA+IND and NIG as separate from IC and the remaining (Asia+JA) data. WRT however indicates partially different clusters, retrieving two IC groups and KY+TZ, but failing to isolate the second African cluster, instead merging varieties from various regions and states of institutionalization. That said, the familiar HK+SIN cluster is clearly identified.

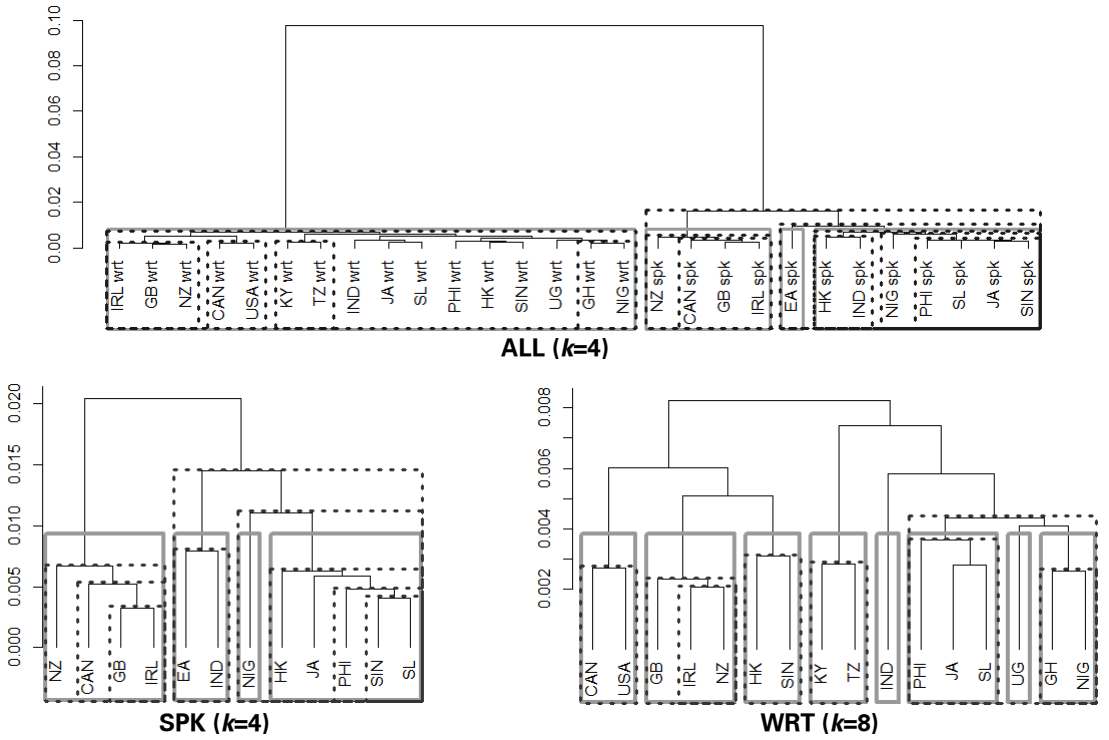


Figure 5.131: Hierarchical clustering results for POS *MI* 4-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

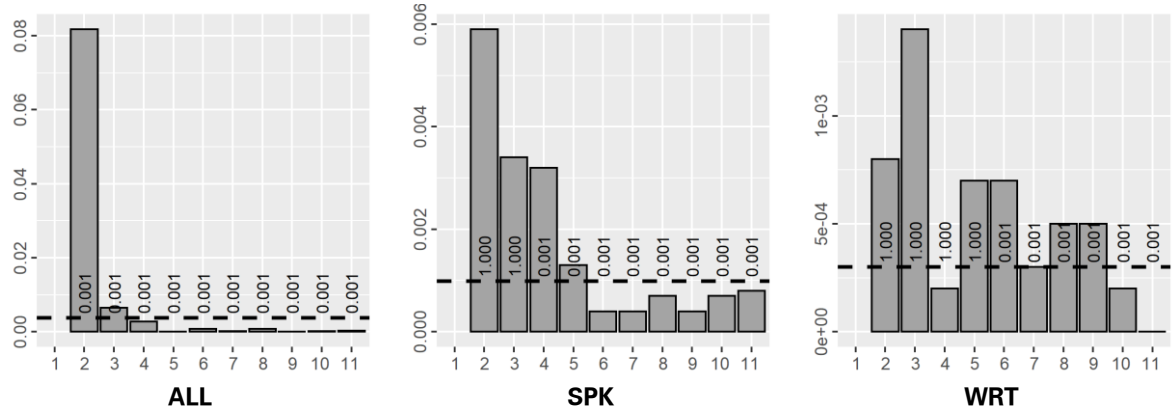


Figure 5.132: Jumps in node heights and respective *p*-values for POS *MI* 4-grams

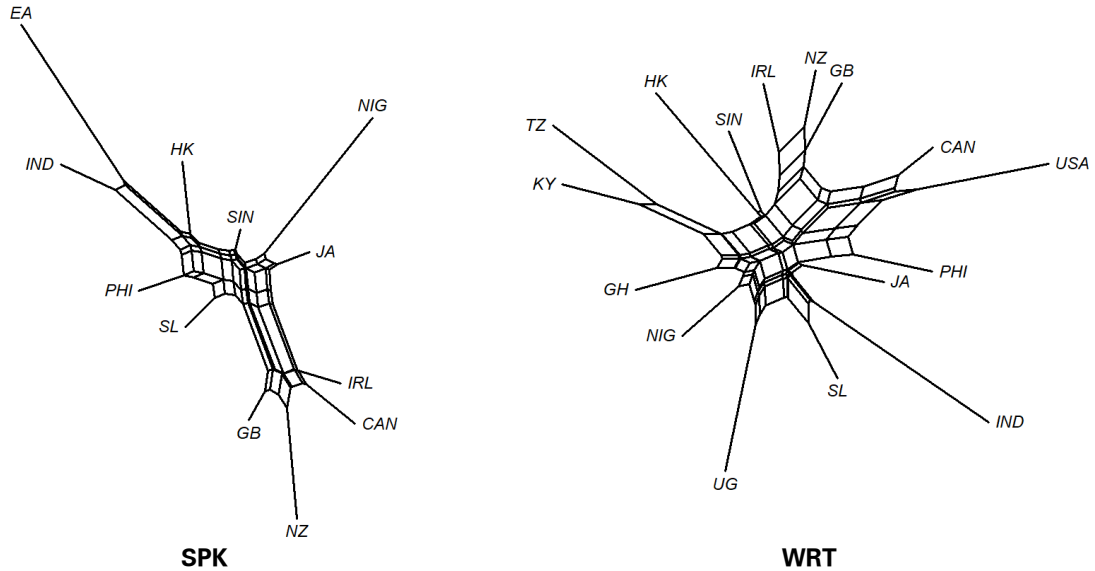


Figure 5.133: NeighborNets of the spoken and written data for POS *MI* 4-grams

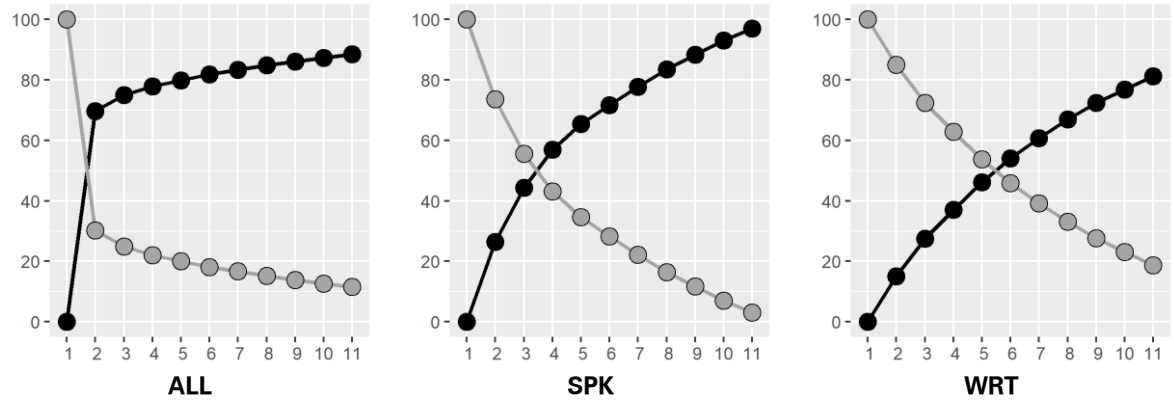


Figure 5.134: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *MI* 4-grams

Table 5.65: K-means clustering results for specific values of *k* for POS *MI* 4-grams

	ALL (*k*=3)		SPK (*k*=4)		WRT (*k*=6)
1	All spoken corpus parts	1	CAN, GB, IRL, NZ	1	HK, SIN
2	All written corpus parts	2	HK, JA, PHI, SIN, SL	2	GB, IRL, NZ
		3	EA, IND	3	CAN, USA
		4	NIG	4	KY, TZ
				5	GH, JA, NIG, PHI, SL, UG
				6	IND

5.4.2 T-score

POS-grams using the t-score show an interesting facet of the measure in the lowest-associated items (Table 5.66). Contrary to what might be expected, these are not constituted of the words retained in the data in their lexical form (and thus comparatively infrequent) and even appear to contain lexical forms less frequently than the most strongly collocated sequences. The lowest-ranking items do however instantiate some strange and rare overall forms, such as multiple consecutive articles, articles after nouns or verbs before nouns, and singular after plural nouns. Some of these may well result from errors in the automatic annotation and certainly appear to be rare sequences. Yet, it remains noteworthy, given the measure's bias towards individually frequent forms, that the t-score calculates sequences with these frequent tags as lesser-collocated items than ones with actual lexical forms. Apart from this observation, *t* shows relatively similar changes of association scores over lengths as observed for lexical *n*-grams, in that positive values face a smaller reduction than the converse increase within the lowest association scores. Reductions in the top association values are somewhat more linear, however, compared to lexical *t*-based sequences.

Table 5.66: POS *t n*-grams with highest and lowest association scores

2-grams type	*t*	3-grams type	*t*	4-grams type	*t*
Spoken					
AT NN1	94.14	AT NN1 of	81.15	AT NN1 of AT	68.92
to VVI	82.30	JJ NN1 of	72.50	NN1 of AT NN1	68.92
JJ NN1	76.84	of AT NN1	69.30	of AT NN1 of	68.92
NN1 of	68.15	in AT NN1	67.60	in AT NN1 of	67.78
VM VVI	59.66	AT1 NN1 of	63.89	AT JJ NN1 of	65.12
AT1 NN1	59.63	AT JJ NN1	63.60	AT1 JJ NN1 of	64.39
AT and	-275.10	DD1 NN1 VVI	-139.71	VDD AT NN1 VVI	-74.78
AT1 AT	-298.34	NN1 VVI to	-144.76	NN1 AT JJ AT	-78.19
UH VVI	-308.96	NN1 VVI that	-147.45	NN1 AT AT AT	-80.29
NN1 VVI	-309.83	VM NN1 VV0	-149.69	NN1 RR NN1 NN1	-88.21
PPIS1 AT	-388.99	NN1 VVI and	-160.22	and AT and AT	-89.08
AT VBZ	-400.29	NN1 VVI NN1	-187.75	VM AT NN1 VVI	-107.63
AT NN1	76.53	AT NN1 of	70.73	AT NN1 of AT	61.02
to VVI	71.51	JJ NN1 of	66.19	NN1 of AT NN1	61.02
JJ NN1	67.45	of AT NN1	59.06	of AT NN1 of	61.02
NN1 of	64.94	AT JJ NN1	56.42	AT JJ NN1 of	59.26
JJ NN2	54.33	AT1 JJ NN1	55.95	AT1 JJ NN1 of	58.95
AT JJ	45.38	in AT NN1	55.66	in AT NN1 of	58.75
NN1 JJ	-276.01	NN2 NN1 NP1	-179.89	VVI into NN1 AT	-111.42
AT VVN	-276.90	JJ AT JJ	-181.74	NN1 NN1 JJ to	-114.66
NN1 VVI	-305.96	JJ AT NN2	-195.08	NN1 NN1 AT JJ	-118.55
NN1 AT	-337.59	NN1 NN1 AT	-200.52	NP1 NN1 NN1 JJ	-127.52
JJ AT	-408.86	RGQ JJ AT	-202.30	NN1 NN1 NN1 JJ	-134.31
AT AT	-993.21	to NN1 AT	-205.08	NN1 to NN1 AT	-136.49

2-grams exhibit relatively little substantiation in the separate datasets (Figure 5.135), erratic cluster structures and, perhaps most curiously, more homogeneity within ALL's spoken than written branch. Only writing shows similar stable clusters in both datasets (HK+IND+USA at AU=94 in ALL), but none of these are consistent with regional, epicentral or evolutionary perspectives. The single substantiated IND+SL cluster in SPK (additionally GB+JA+IRL+PHI is found at AU=93) is not confirmed in ALL, which instead separates by the tripartite segmentation into IC (albeit without CAN and only at AU=92), EA+IND and a group of remaining OC varieties.

Significant jumps (Figure 5.136) are only detected for very large numbers of clusters in the separate SPK and WRT datasets, demanding k=6 or even k=9 for speech and k=9 for writing. Both the coarser and finer segmentations for speech return very strange clusters overall, and only the relatively complete Asian (or otherwise IND+SL and HK+SIN subclusters) appeal to a linguistic interpretation, while separation of EA and NIG from the remaining data is always indicated. In WRT, only counterintuitive and highly heterogenous groups emerge (cf. above) as well as many unary nodes.

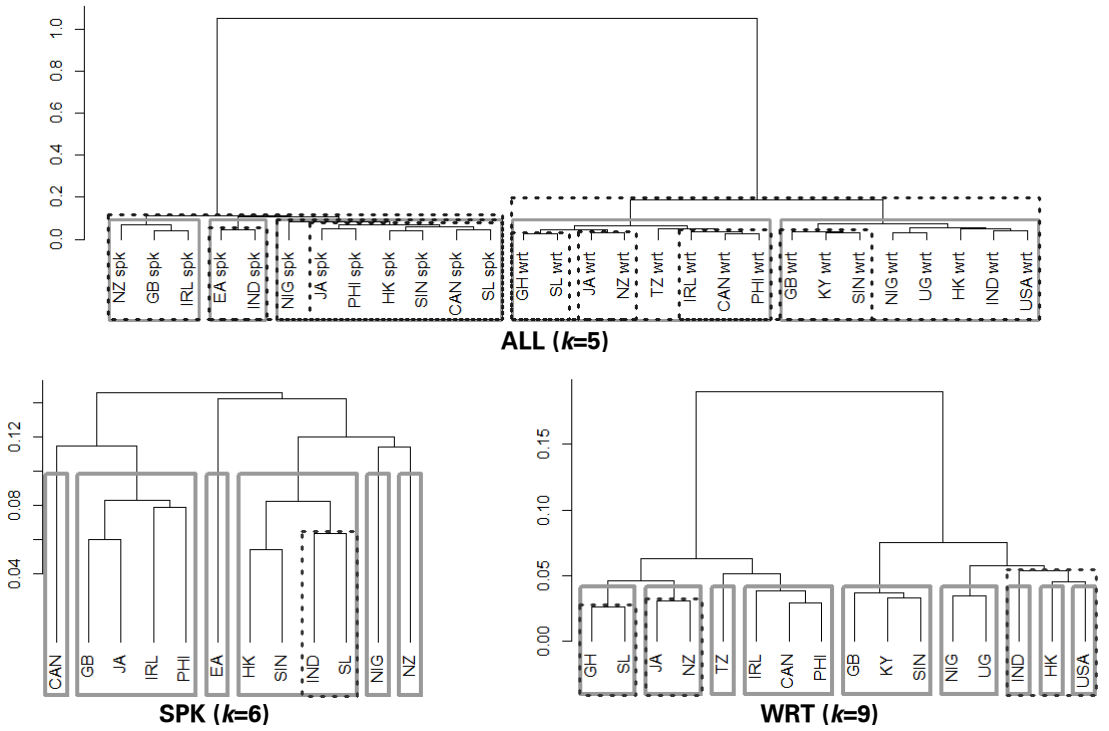


Figure 5.135: Hierarchical clustering results for POS *t* 2-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

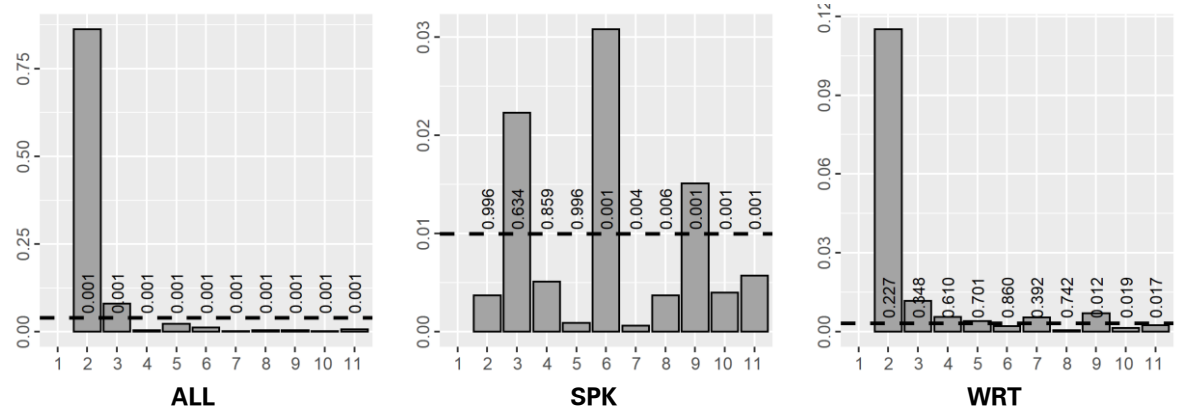


Figure 5.136: Jumps in node heights and respective *p*-values for POS *t* 2-grams

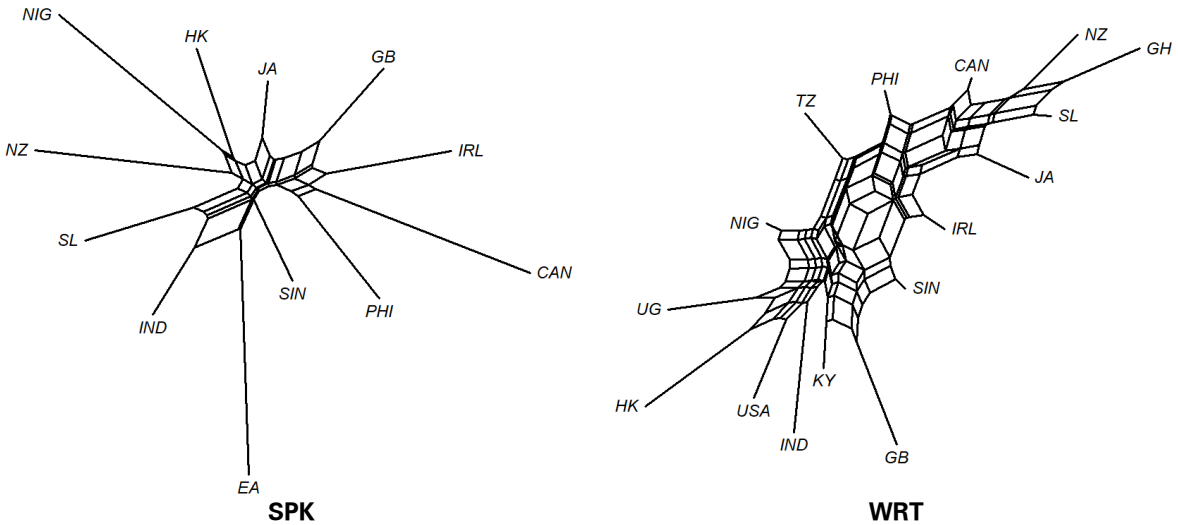


Figure 5.137: NeighborNets of the spoken and written data for POS *t* 2-grams

For ALL, anything over *k*=2 shows relatively small jumps. At *k*=5, speech segments by the tripartite structure discussed above but writing presents no meaningful clusters. At very high values (*k*=20), the written clusters mimic those for the separate data, but speech at this point fully fragments beyond GB+IRL.

The NeighborNets (Figure 5.137) reinstate many of these strange combinations and generally reflect strong heterogeneity. Yet, they also show some connection between spoken GB, IRL and CAN (with PHI close to CAN) and indicate that a IND+SL group may be similarly valid as EA+IND. For writing, the method does little more than provide another representation of the counterintuitive results, and no effect of regional proximity, direction of norm orientation or degree of institutionalization becomes apparent.

K-means clustering (Figure 5.138 and Table 5.67) indicates far fewer clusters than the HCA. For ALL, these reinforce speech vs. writing at *k*=2, while at *k*=3 producing mostly similar (strange) groups to the HCA. SPK distinguishes very roughly between IC (-NZ) and Asia, separating EA and NIG but also NZ. WRT at *k*=4 produces clusters of little value except for a mostly African cluster (lacking GH) and identification of proximity of SIN to GB+IRL as well as of HK and IND to USA. For *k*=3, UG is added to cluster #4 while #1 and #2 merge. The 'elbow' indicates *k*=2, retrieving ALL's cluster #3 and thus indicating its stability, but none of these provide intuitively better results.

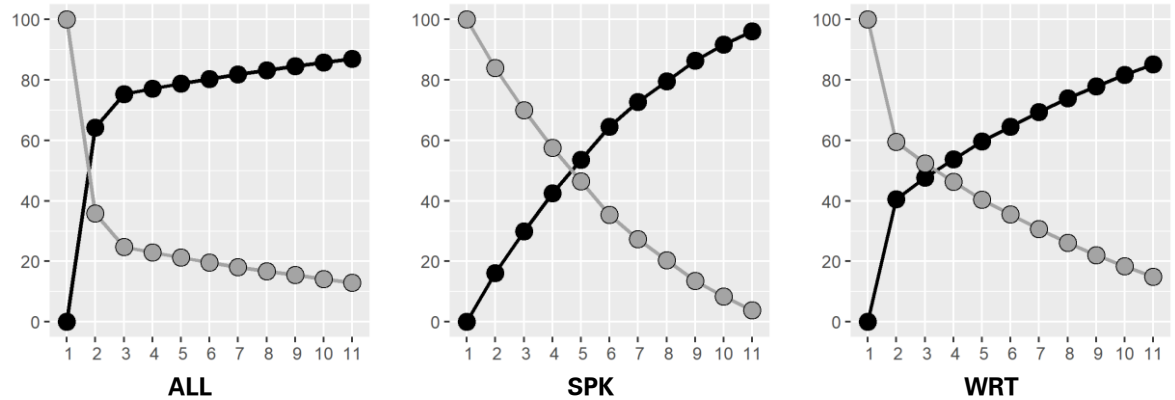


Figure 5.138: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *t* 2-grams

Table 5.67: K-means clustering results for specific values of *k* for POS *t* 2-grams

	ALL (*k*=3)		**SPK (*k*=5)**		**WRT (*k*=3)**
1	All spoken corpus parts	1	HK, IND, JA, SIN, SL	1	KY, TZ, NIG, UG
2	KY, TZ, GB, HK, IND, IRL, NIG, SIN, UG, USA $_{WRT}$	2	CAN, GB, IRL, PHI	2	GB, IRL, SIN
		3	NZ	3	CAN, GH, JA, NZ, PHI, SL
3	CAN, GH, JA, NZ, PHI, SL $_{WRT}$	4	EA	4	HK, IND, USA
		5	NIG		

At the level of 3-grams, *t*'s erratic behavior changes to the detection of more regular structures (Figure 5.139). ALL presents stable spoken/written separation and returns to the more usual finding of more differences in speech. A separation of the spoken branch into IC and OC (at AU=94) and further binary segmentation of the latter into EA+IND and the remaining varieties (also at AU=94) is supported in part by SPK. In both types of written data, the same clusters are supported (IND+JA+SL_{WRT} in ALL at AU=94), often returning regional and epicentral clusters (with the usual exception of GH+NIG+UG), but also finding the latter African group stably aligned with IND+JA+SL.

Significant jumps (Figure 5.140) are indicated at lower cluster numbers than for 2-grams. In case of SPK, this results in *k*=4 indicating mostly the same tripartite separation as above, with only NIG split off from the remaining OC. For writing, *k*=6 appears a better choice than the very low jumps at *k*=5, also resulting in largely the same clusters found stable before and thus mostly in line with a regional view (but note the isolation of HK). For ALL, the spoken/written distinction the only resolution above average height, but at *k*=5, the structures of SPK are mostly repeated (completely at *k*=6 completely). Larger values, e.g. the somewhat elevated jump heights at *k*=9 coincide with successive fragmentation of the spoken branch (particularly NZ and NIG split off) but parallel emergence of a written IC cluster and small but familiar KY+TZ and HK+PHI+SIN.

The NeighborNets in Figure 5.141) support one coherent IC cluster in SPK but separate the group in WRT. SPK further indicates EA+IND, while WRT supports IND+SL. Further support can be found for SIN+HK_{SPK} and a similar HK+SIN+PHI_{WRT}. The African data is found relatively heterogenous, and the KY+TZ_{WRT} cluster diverges strongly from the group while it appears to share several features with IC_{GB} and JA. At the same time, it also indicates great distances to these varieties.

K-means clustering (Figure 5.142 and Table 5.68) for ALL only indicates a binary split, and SPK and WRT each favor *k*=4. This results in a segmentation of SPK identical to that obtained from the hierarchical approach. For WRT, however, even clusters previously found stable are broken up and rearranged. The IC varieties are found together with JA and SIN (not entirely without precedent), but only a HK+PHI and an ICE-EA group remain in addition to a heterogenous group of remaining varieties.

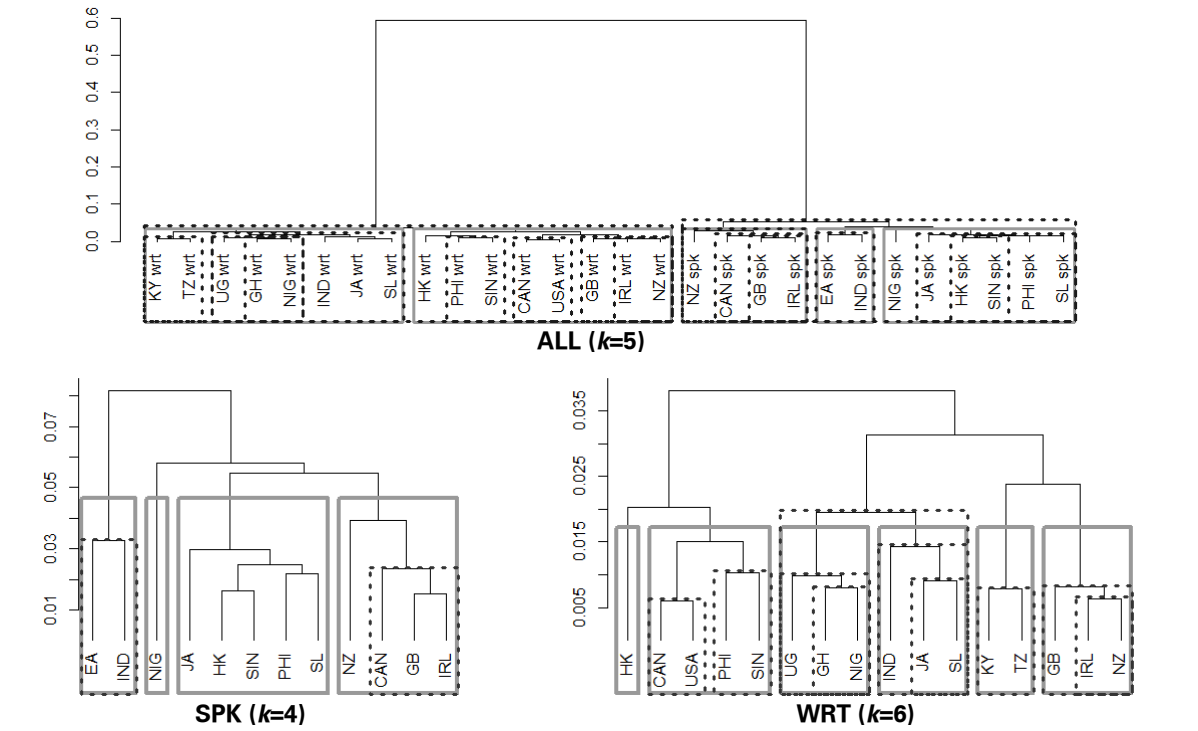


Figure 5.139: Hierarchical clustering results for POS *t* 3-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

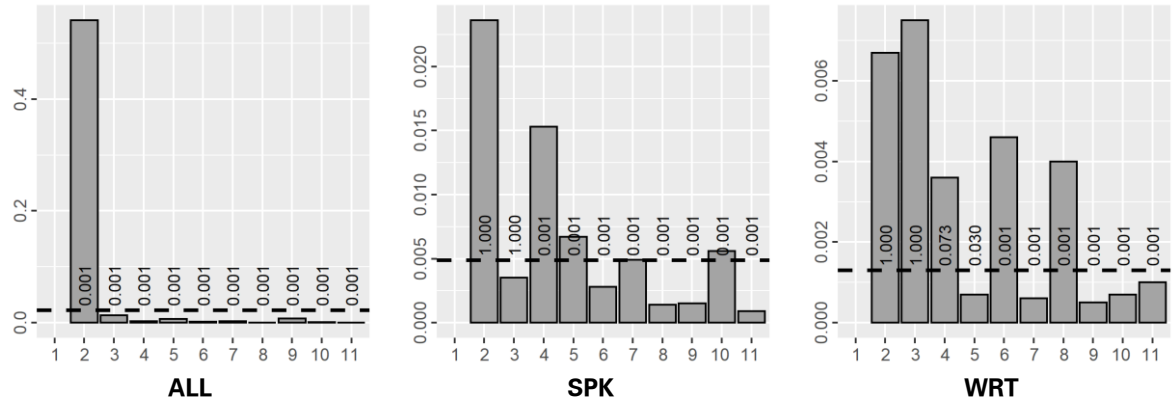


Figure 5.140: Jumps in node heights and respective *p*-values for POS *t* 3-grams

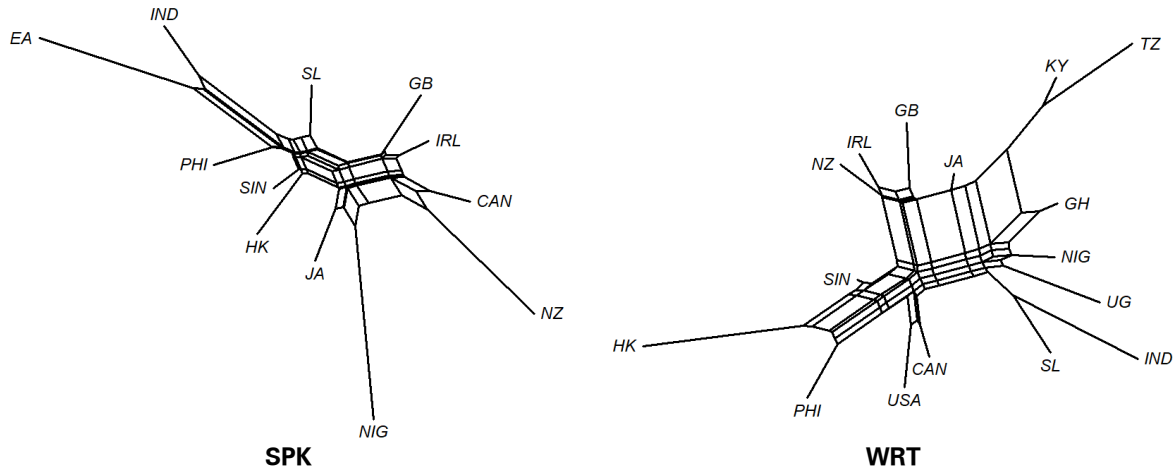


Figure 5.141: NeighborNets of the spoken and written data for POS *t* 3-grams

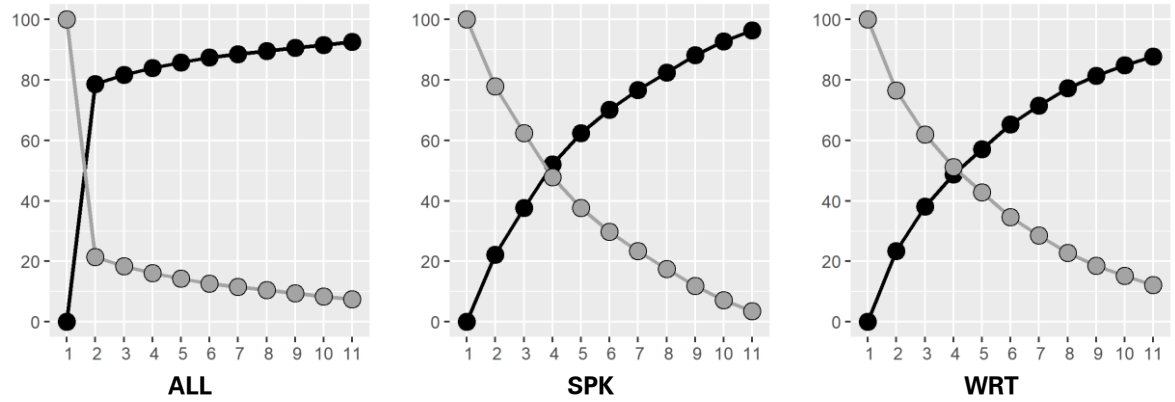


Figure 5.142: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *t* 3-grams

Table 5.68: K-means clustering results for specific values of *k* for POS *t* 3-grams

	ALL (*k*=2)		SPK (*k*=4)		WRT (*k*=4)
1	All spoken corpus parts	1	EA, IND	1	HK, PHI
2	All written corpus parts	2	HK, JA, PHI, SIN, SL	2	CAN, GB, IRL, JA, NZ, SIN, USA
		3	CAN, GB, IRL, NZ	3	KY, TZ
		4	NIG	4	GH, IND, NIG, SL, UG

4-grams in ALL again return to the detection of greater differences within the written branch observed for the shortest sequences (Figure 5.143). These mainly result from a binary opposition within this branch, which however is not found to be stable. Subclusters are found for two IC and African groups in addition to JA+SL. Results for WRT are identical except for a combination of most African varieties with IND, SL and JA additionally found in ALL. Speech also identifies the same clusters in both datasets (the two clusters not highlighted missing significance by fractions). Both spoken sets thus roughly separate EA+IND from Asia(+NZ) and (remaining) IC+JA+NIG.

Significant jumps (Figure 5.144) are discovered for SPK at *k*=5, providing counterindication against associating NIG and NZ with the clusters observed above, and additionally highlighting the separation of EA+IND. It should be noted that *k*=9 could also be seen as indicated, which isolates almost all varieties except for GB+IRL, HK+SIN and PHI+SL. For WRT, *k*=5 indicates the first significant separation, combining the North American data with Southeast Asia as well as the remaining Asian varieties with parts of the African data. A somewhat less indicated *k*=7 would harmonize more consistently with stable clusters by emphasizing the Southeast Asian (PHI+HK+SIN) group, two African clusters (with UG again strangely placed) but also indicating separateness of IND. For ALL, nothing beyond the separation of spoken and written modes is strongly supported.

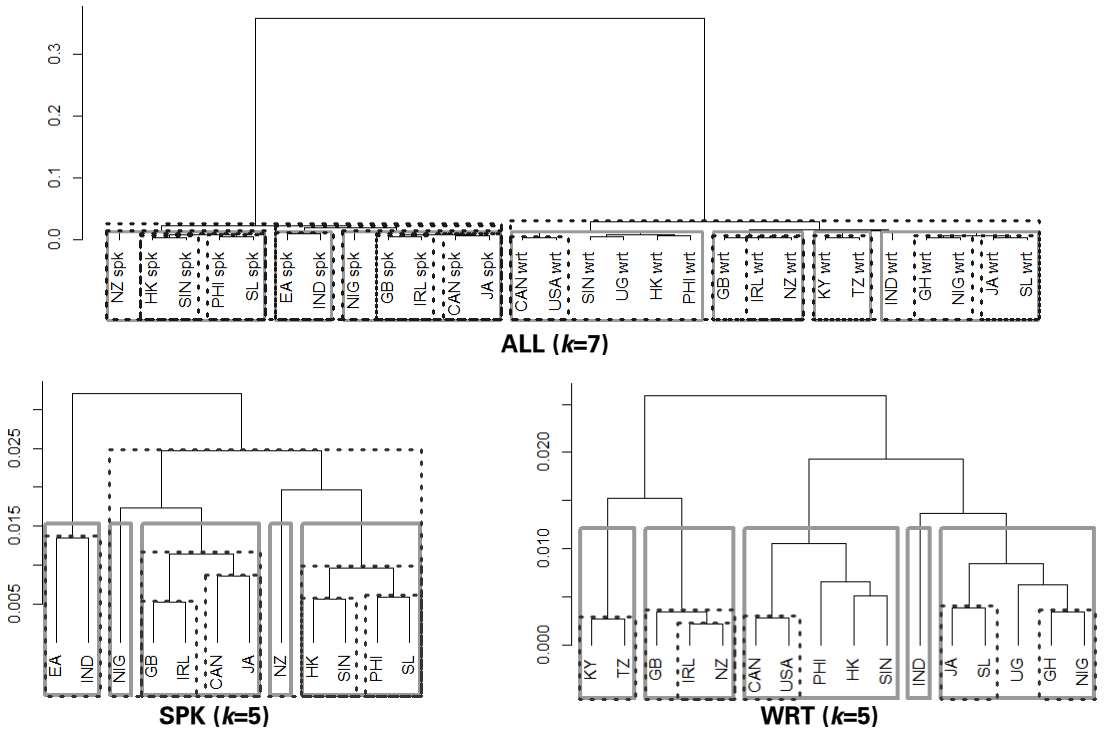


Figure 5.143: Hierarchical clustering results for POS *t* 4-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

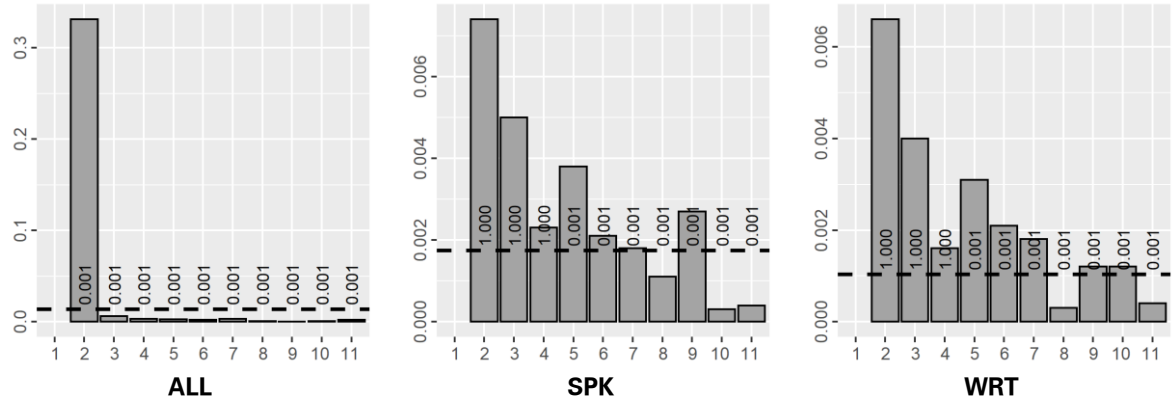


Figure 5.144: Jumps in node heights and respective *p*-values for POS *t* 4-grams

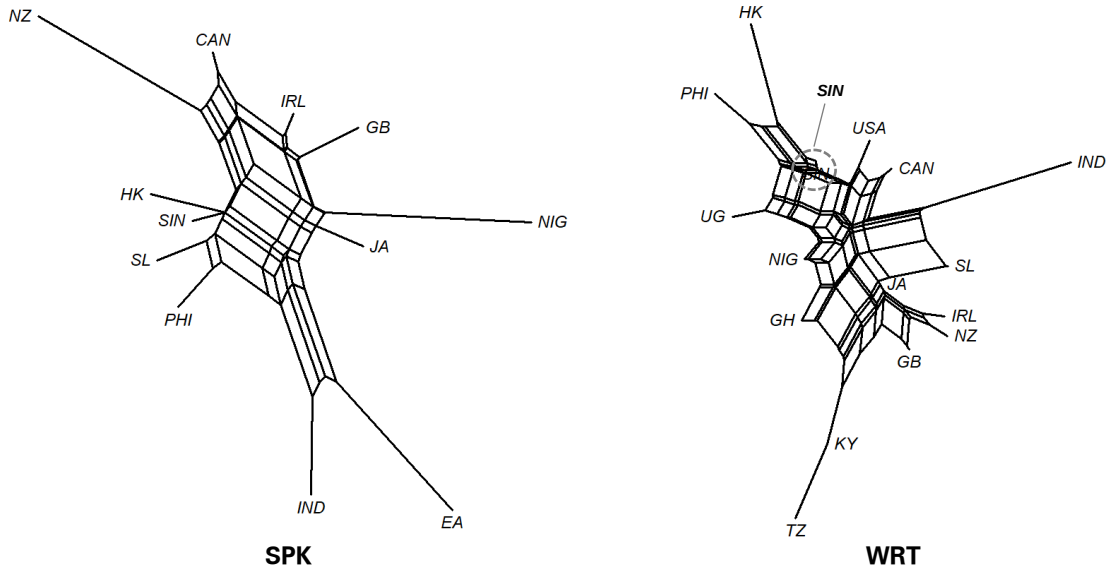


Figure 5.145: NeighborNets of the spoken and written data for POS *t* 4-grams

The NeighborNets in Figure 5.145 imply that the difference of NZ to the overall IC set may not be as drastic as indicated above. They moreover show HK+SIN in speech but less in writing and a clear written KY+TZ cluster but not the second African group. IND and SL diverge strongly from the remaining written data but are also mutually relatively different. SPK in turn retrieves EA+IND in speech.

K-means favors the same cluster numbers (Figure 5.146 and Table 5.69) as for 3-grams. For ALL, this again results in the spoken/written separation only. For speech, the same clustering is obtained as would be at the same height within the hierarchical approach (k=4), including separation of NZ from the Asian cluster. For writing (k=4), however, altogether different results are obtained, and those varieties removed from most others in the NeighborNet analysis are highlighted.

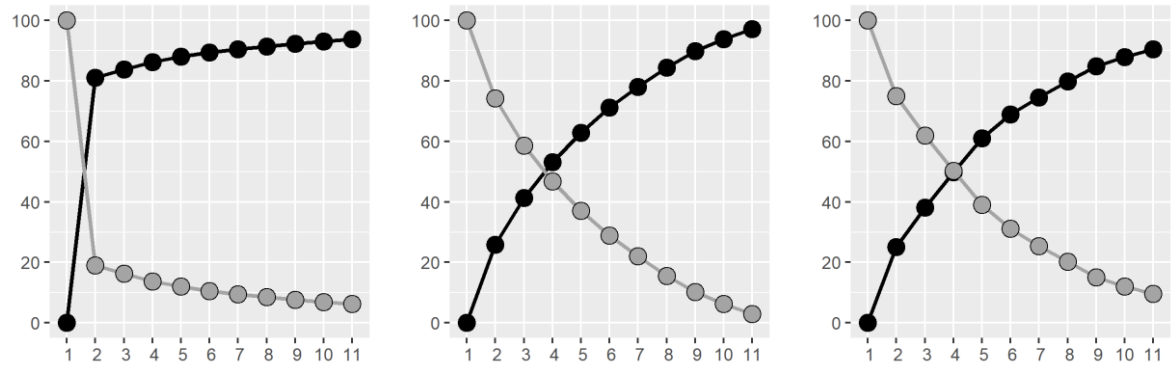


Figure 5.146: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *t* 4-grams

Table 5.69: K-means clustering results for specific values of *k* for POS *t* 4-grams

	ALL (*k*=2)		SPK (*k*=4)		WRT (*k*=4)
1	All written corpus parts	1	EA, IND	1	KY, TZ, GH
2	All spoken corpus parts	2	CAN, GB, IRL, JA, NIG	2	HK, PHI
		3	HK, PHI, SIN, SL	3	IND
		4	NZ	4	CAN, GB, IRL, JA, NIG, NZ, SIN, SL, UG, USA

5.4.3 Log Likelihood

As already observed within the lexical data, several of both the top and bottom G^2-based n-grams (Table 5.70) are constituted from the same types as in case of the t-score. This effect is particularly pronounced for 2-grams but levels off for longer sequences, and overall it appears to be weaker than within the lexical data. Another shared tendency with the t-score can be observed in G^2 assigning the highest association scores to identical sequences within both modes. Moreover, the measure's

previously-observed tendency towards excessive positive scores also expresses itself again, in contrast to some strongly negative values in case of *t*. However, G^2 values also reveal a general preference towards assigning 'outlying' values if compared to the lexical data, since positive as well as negative values are much more increased by the change to POS annotation than observed for the t-score.

Table 5.70: POS G^2 *n*-grams with highest and lowest association scores

2-grams type	G^2	3-grams type	G^2	4-grams type	G^2
Spoken					
to VVI	50856.20	AT NN1 of	31132.51	to VVI to VVI	34280.43
AT NN1	40103.78	VVGK to VVI	28043.32	to VVI AT NN1	30479.38
VM VVI	23160.15	to VVI RP	26866.96	AT NN1 to VVI	30314.92
JJ NN1	23079.81	JK to VVI	26804.34	VM VVI to VVI	25048.41
NN1 of	22161.24	VVN to VVI	26524.08	JJ NN1 to VVI	24640.27
PPIS1 VV0	20686.42	VMK to VVI	26210.45	to VVI JJ NN1	24457.35
PPY NN1	-4788.43	to NN1 JJ	-5629.20	NN1 NN1 JJ to	-3281.44
NN2 NN1	-4813.23	VVI NN1 JJ	-5691.36	NN1 NN1 NN1 AT	-3300.40
NN1 AT	-7117.91	VBZ NN1 JJ	-5905.26	NN1 to NN1 AT	-3310.64
NN1 VVI	-7330.62	NN2 NN1 AT	-5965.57	NN1 NN1 NN1 JJ	-3747.63
NN1 JJ	-8459.61	RR NN1 VV0	-6035.56	NN1 AT AT AT	-3853.91
RR NN1	-8630.86	NN2 NN1 JJ	-6636.42	NN1 RR NN1 NN1	-4155.34
to VVI	42029.78	VVN to VVI	22385.52	to VVI to VVI	28103.21
AT NN1	24742.37	AT NN1 of	22061.19	to VVI AT NN1	22500.44
NN1 of	19380.02	JK to VVI	21828.00	AT NN1 to VVI	22258.89
JJ NN1	17046.27	to VVI RP	21670.34	to VVI NN1 of	19810.82
VM VBI	14296.00	to VVI APPGE	21453.23	JJ NN1 to VVI	19693.53
VM VVI	13050.64	order to VVI	21381.95	to VVI JJ NN1	19604.44
RR NN1	-6246.90	NN2 NN1 NN1	-8789.16	NN1 to NN1 AT	-5904.36
AT AT	-6403.98	NN2 NN1 NP1	-8802.03	NN1 NN1 NN2 NN1	-6167.97
JJ AT	-7012.73	to NN1 AT	-8858.81	NN1 NN2 NN1 NN1	-6167.97
NN2 NN1	-11873.40	to NN1 JJ	-9335.64	NN1 NN1 JJ to	-7094.06
NN1 AT	-14584.01	NN1 NN1 AT	-10144.46	NP1 NN1 NN1 JJ	-7710.03
NN1 JJ	-15537.66	NN1 NN1 JJ	-10621.28	NN1 NN1 NN1 JJ	-8982.49

Log-likelihood POS **2-grams** exhibit an abundance of stable clusters in all datasets (Figure 5.147), and identical stable clusters and even branch structures between ALL and the separate datasets. However, stability of the spoken branch fails to arise, indicating a less reliable overall separation of the two modes. All spoken data identify EA as not within a stable group, and separates IND+PHI and (to a lesser extent) NZ from two larger structures partially agreeing with an Asia vs. IC separation (AU values of 93 and 91 as well as 94 and 92, in ALL and SPK, respectively). Writing tends towards smaller stable clusters, finding a predominantly African alongside a less intuitive PHI+SL cluster as well as two IC groups (found at AU=93 for the cluster these share with HK, SIN and UG).

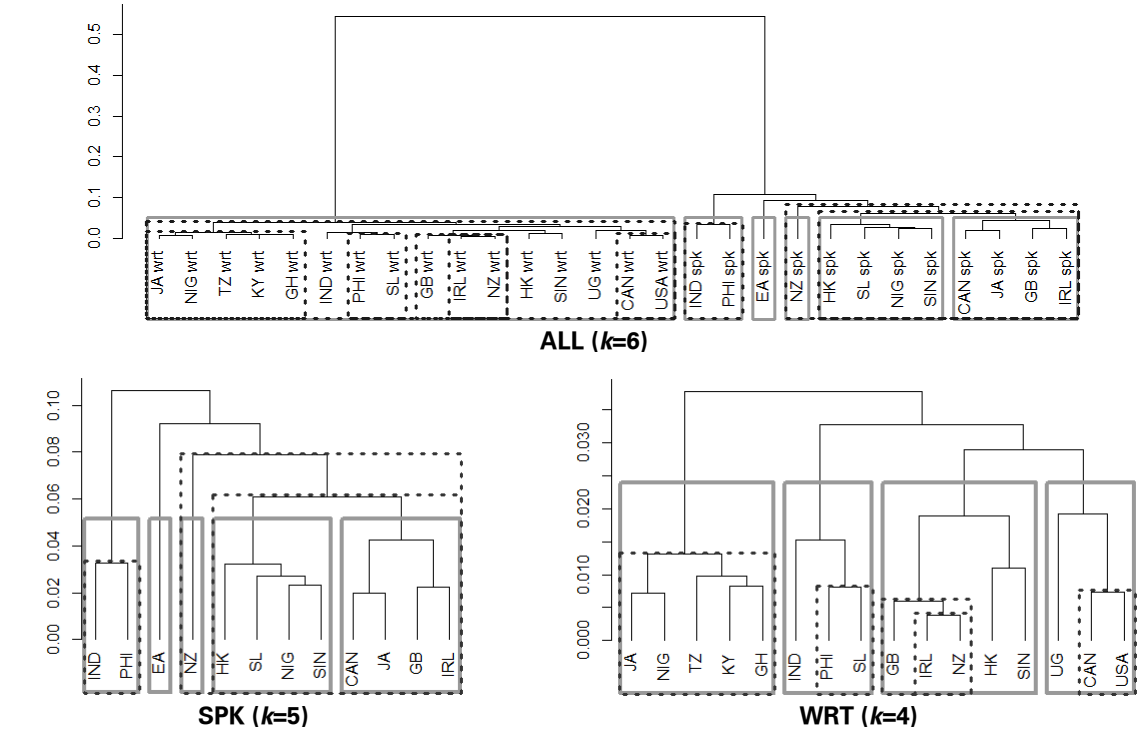


Figure 5.147: Hierarchical clustering results for POS G^2 2-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

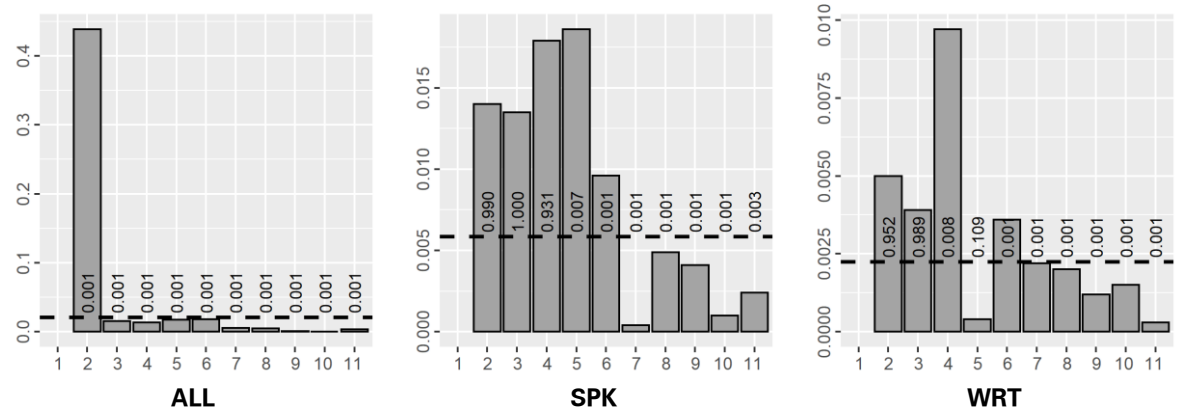


Figure 5.148: Jumps in node heights and respective *p*-values for POS G^2 2-grams

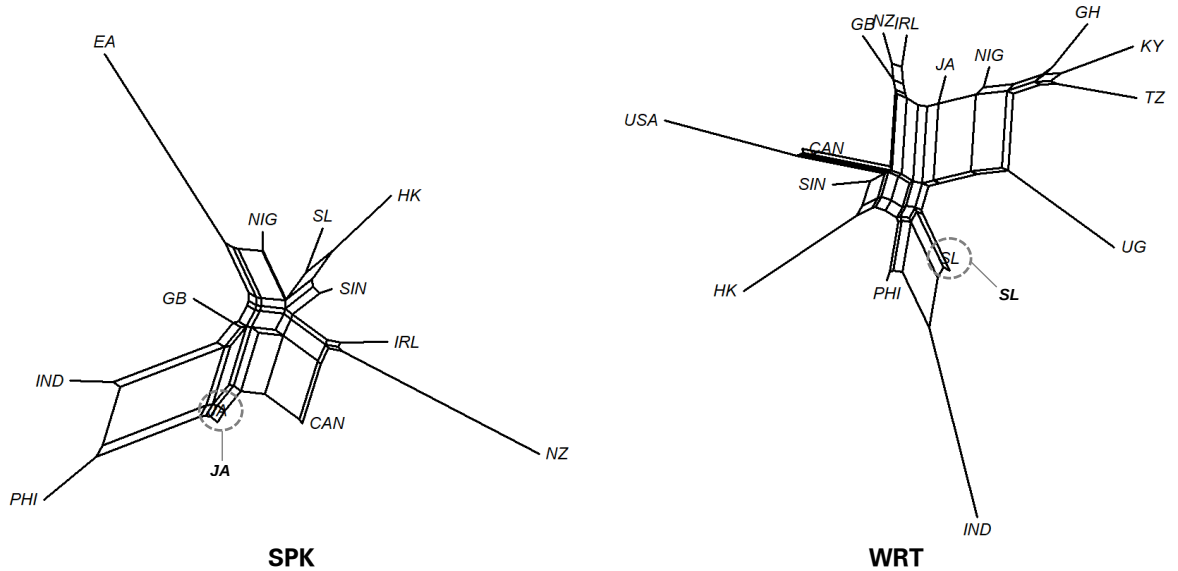


Figure 5.149: NeighborNets of the spoken and written data for POS G^2 2-grams

In terms of significant jumps (Figure 5.148), SPK at *k*=5 highlights the distinctiveness of IND, PHI, EA and also NZ from the two groups discussed above. For WRT at *k*=4, the African cluster is reasserted, but all other remaining clusters are extended. While combining SL (and PHI) with IND reconstitutes a South Asian group, assigning HK+SIN to IC_{GB} and UG to IC_{NA} runs counter to expectations and previous findings. For ALL, the largest jump after the spoken/written separation at *k*=6 equals the solution obtained in SPK but only produces a homogenous written cluster.

NeighborNet analysis (Figure 5.149) confirms large distances of IND+PHI, EA and also NZ to the remaining SPK data. It also reflects the relatively meaningless structures found above and only weakly or partially supports some typical groups (IRL+NZ+CAN, SIN+HK+SL, NIG+EA). WRT indicates two IC groups, which are found at relatively large distances. It also identifies a partial African cluster, which shares some features with IC_{GB} and JA, as well as a relatively heterogenous IND+SL+PHI group.

K-means (Figure 5.150 and Table 5.71) presents similar issues in terms of meaningful clusters. For ALL, only *k*=2 finds support, while *k*=4 for SPK singles out HK (merged with #2 at *k*=3) and EA as well as IND+JA+PHI (also discovered in ALL at *k*=3). WRT at *k*=4 also isolates HK and IND+PHI (+SL), and with the exception of SIN assigns to the two IC groups the same varieties as in the hierarchical approach.

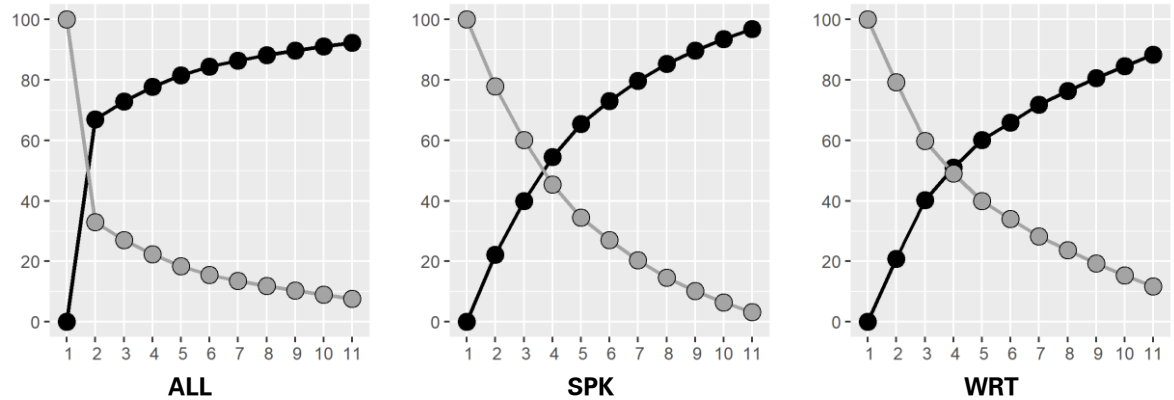


Figure 5.150: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS G^2 2-grams

Table 5.71: K-means clustering results for specific values of *k* for POS G^2 2-grams

	ALL (*k*=2)		**SPK (*k*=4)**		**WRT (*k*=4)**
1	All spoken corpus parts	1	IND, JA, PHI	1	HK
2	All written corpus parts	2	CAN, GB, IRL, NIG, NZ, SIN, SL	2	KY, TZ, GB, GH, IRL, JA, NIG, NZ
		3	HK	3	CAN, SIN, UG, USA
		4	EA	4	IND, PHI, SL

Despite many stable clusters overall, POS **3-grams** show the same lack of stability of ALL's spoken branch as 2-grams (Figure 5.151). EA_{SPK} even intrudes into the written set, albeit at some distance from its cluster of mostly African and British-epicentral IC varieties. While both types of written data otherwise retrieve the same stable groups (the GH+KY+TZ subcluster not in the same hierarchy), speech only finds two stable subclusters (CAN+JA and NIG+SIN) in ALL. Speech further isolates EA, HK and a IND+PHI group, and identifies substructures in the remaining data, within which some IC (with SL but without CAN) separate from (late phase 3 to) phase 4 OC.

Despite the intrusion of EA_{SPK} into writing, significant jumps (Figure 5.152) still identify a binary separation of ALL. Finer segmentation is indicated for k=8, isolating EA_{SPK} (already after k=4) and IND_{WRT} in the written branch and thus leaving more African (-UG) and Asian groups (+IC_{NA}/IC_{GB}, respectively). In the spoken branch, IND, HK and IC (-CAN) are isolated from the remaining varieties. The distinctness of EA, HK and IC (again without CAN, but with SL) is also supported in SPK, the remaining group splitting off IND+PHI. WRT segments as ALL would at finer resolutions, but clusters again only vaguely resemble meaningful categories: a EA+GH group, two IC clusters (one joined with JA+NIG) and a larger mostly Asian group.

The NeighborNets (Figure 5.153) retrieve a strong heterogeneity within SPK but support the removal of CAN from the IC group and IND+PHI and furthermore indicate the regional EA+NIG (instead of EA+IND) and HK+SIN. WRT shows somewhat clearer clusters of unusual structure: It retrieves a partial African group (without UG, and NIG at elevated distance) and finds indication towards HK+SIN and IND+SL. It also supports the allocation of IC_{GB} and IC_{NA} to mostly African or Asian varieties, respectively.

K-means (Figure 5.154 and Table 5.72) supports only a binary split in ALL, but retrieves speech vs. writing (without EA, which is isolated with HK at k=3). SPK intersects at k=4, separating EA and HK from IND+PHI (plus GB) and the remaining varieties. Following the (slight) elbow at k=5 produces minimally more meaningful clusters by assigning the IC varieties into their separate group. WRT at k=3 establishes the same Africa+IC_{GB} vs. Asia+IC_{NA} clusters retrieved for the hierarchical analysis at k=3, with only HK instead of IND found separate. Following the elbow at k=5 produces more meaningful clusters by assigning IND+SL and KY+TZ+GH to separate clusters.

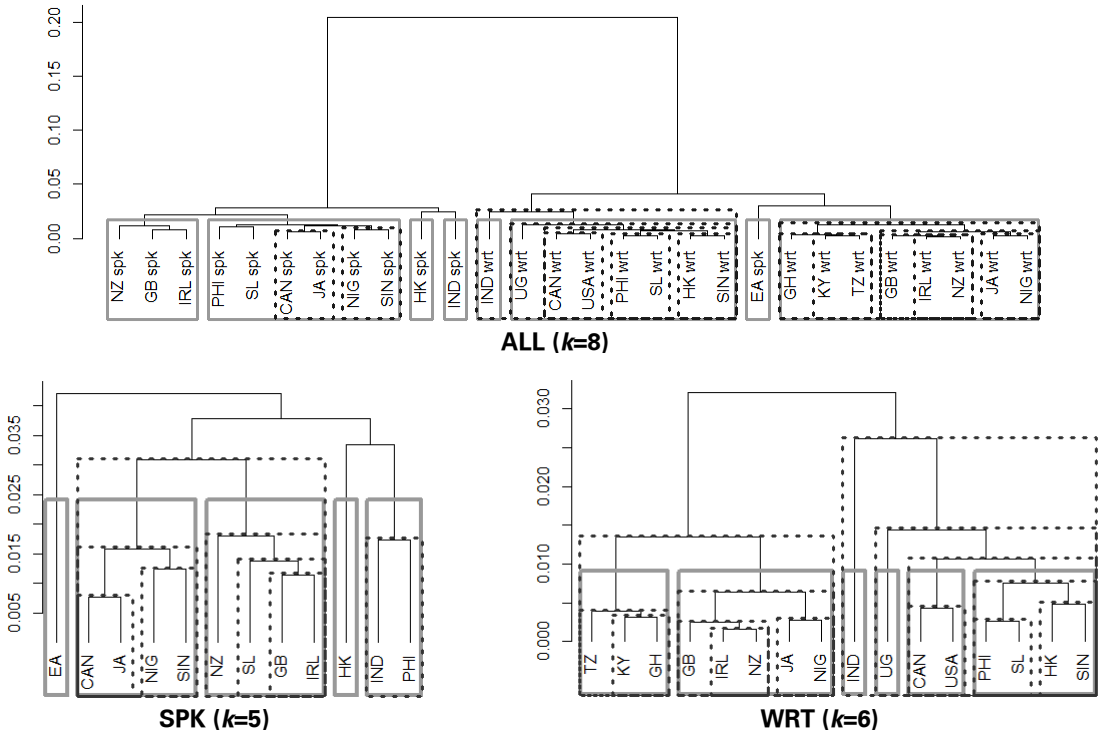


Figure 5.151: Hierarchical clustering results for POS G^2 3-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

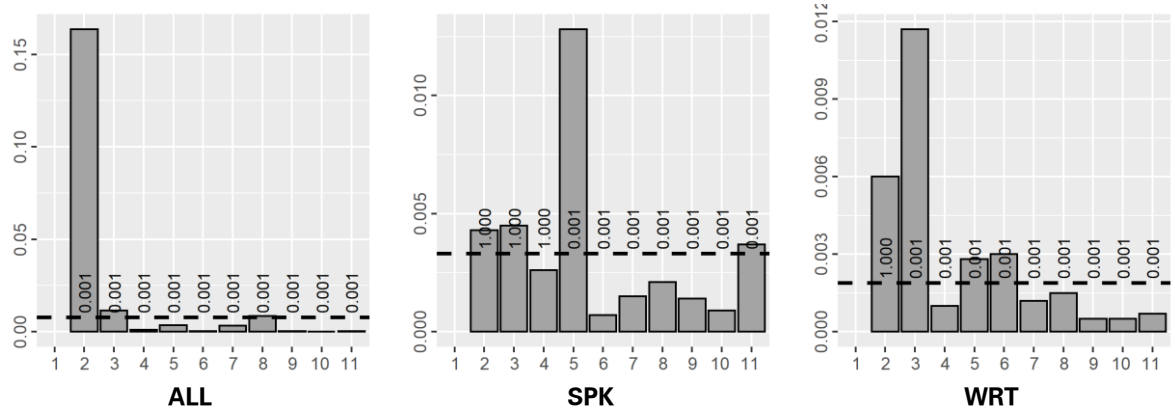


Figure 5.152: Jumps in node heights and respective *p*-values for POS G^2 3-grams

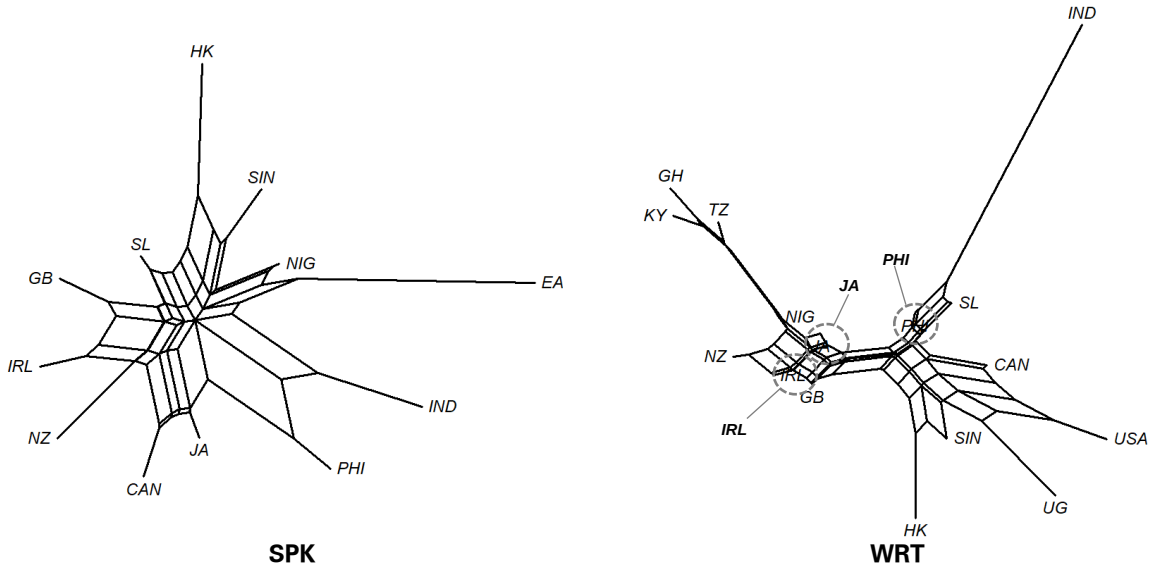


Figure 5.153: NeighborNets of the spoken and written data for POS G^2 3-grams

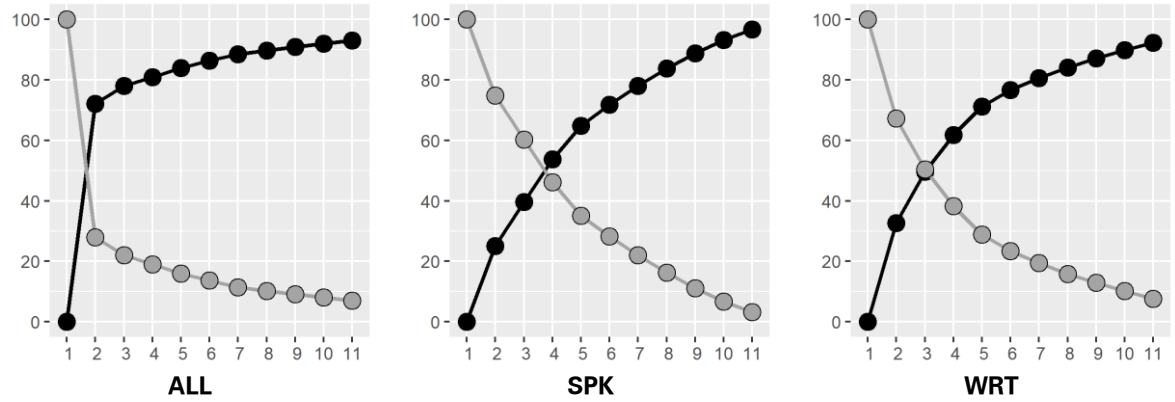


Figure 5.154: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS G^2 3-grams

Table 5.72: K-means clustering results for specific values of *k* for POS G^2 3-grams

	ALL (*k*=2)		SPK (*k*=4)		WRT (*k*=3)
1	All spoken corpus parts	1	JA, NIG, SIN	1	HK
2	All written corpus parts	2	CAN, GB, IRL, NZ, SL	2	KY, TZ, GB, GH, IRL, JA, NIG, NZ
		3	EA	3	CAN, IND, PHI, SIN, SL, UG, USA
		4	IND, PHI		
		5	HK		

After the instability of the spoken branch observed above, only **4-grams** manifest a clear separation of the spoken and written modes as well as stable clusters for both branches (Figure 5.155). Written varieties exhibit relatively similar stable clusters in both sets of data, at the coarsest level distinguishing African varieties (-UG) plus British-branch IC from Asian varieties with North American IC (as for 3-grams). Finer clusters identify recurring but incomplete regional clusters KY+TZ+GH, GB+IRL+NZ and HK+SIN, while additionally JA+NIG and PHI+SL are observed. IND is found at large distances to its cluster in both sets, and CAN and USA are clustered closely only once, while being grouped with either PHI+SL or HK+SIN in ALL. Results for speech are conflicting and not frequently informative: While IND is twice found inside stable clusters with SL and IC (minus CAN in SPK), stranger CAN+JA and NIG+SIN emerge in both sets, and otherwise the datasets mismatch, e.g. analyzing EA and HK as either separate or part of a larger heterogenous group.

After the spoken/written distinction in ALL, significant jumps (Figure 5.156) only reach average heights again at *k*=6, partitioning speech into three groups approximating IC vs. OC, but with a separate IND+SL(+GB). Results for writing are analogous to those discussed above, i.e. British IC+Africa (-UG), North American IC+Asia and a separate IND. Similarly indicated *k*=10 further splits off unary EA_{SPK}, HK_{SPK} and UG_{WRT} as well as $GH+KY+TZ_{WRT}$.

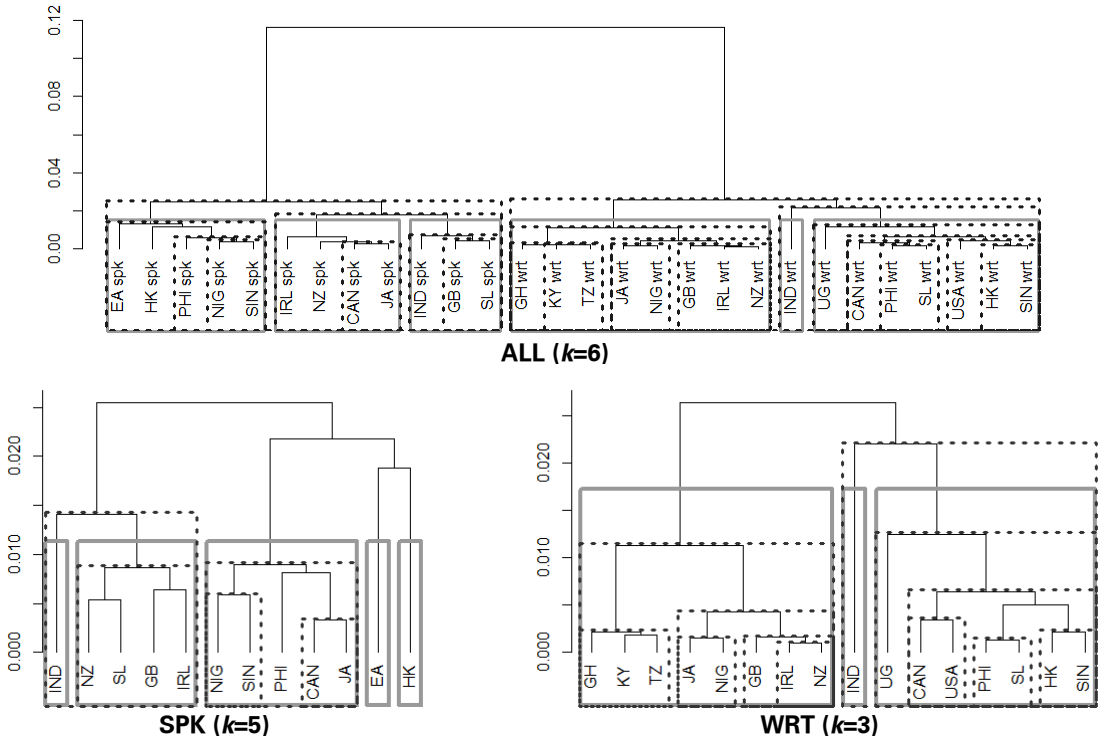


Figure 5.155: Hierarchical clustering results for POS G^2 4-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

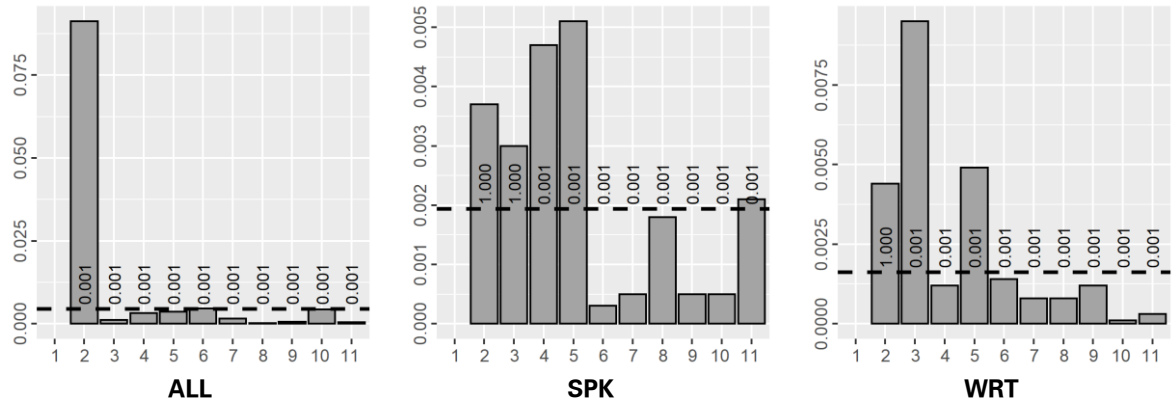


Figure 5.156: Jumps in node heights and respective *p*-values for POS G^2 4-grams

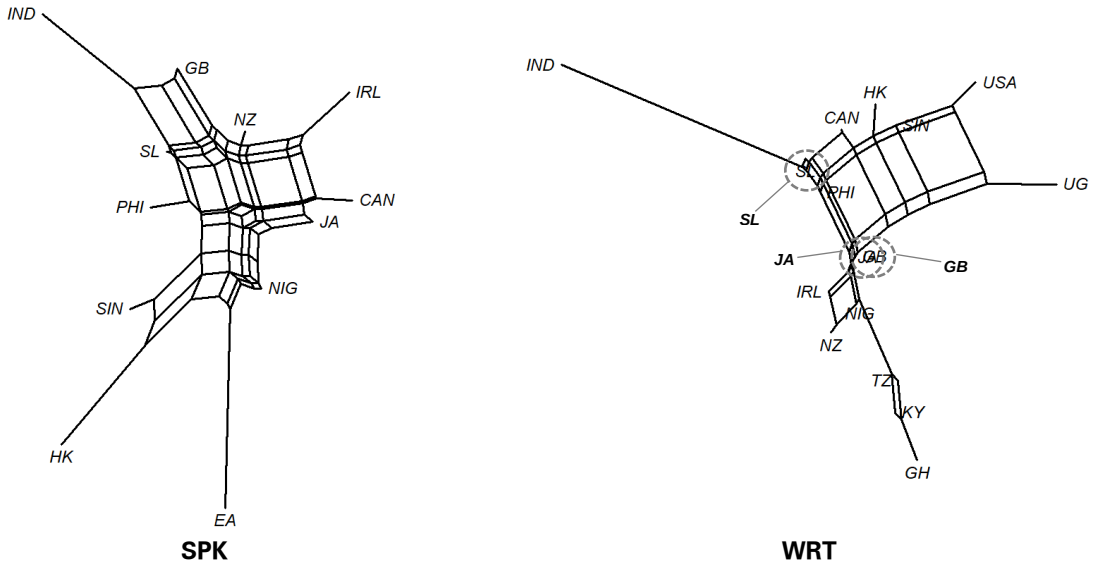


Figure 5.157: NeighborNets of the spoken and written data for POS G^2 4-grams

SPK shows the earliest significant jump at *k*=4, separating EA and HK from IC+IND+SL and the remaining OC, but also indicates *k*=5, which splits off IND. WRT clearly prefers *k*=3, mirroring the segmentation of ALL. Finer *k*=5 splits off UG and GH+KY+TZ.

NeighborNet analysis (Figure 5.157) also shows few groups clearly (or meaningfully) delineated. The similarity of spoken IND+SL with most IC is identified as well as HK+SIN. Writing, too, brings out some similarity of IND+SL and supports the general separation into Asia and Africa with their respective IC subclusters found in the hierarchical analysis. Most of the African varieties follow a pattern of incremental dissimilarity to the remaining data.

K-means clustering (Figure 5.158 and Table 5.73) indicates speech vs. writing for ALL, isolating HK+EA as the next-most distinct at *k*=3. The same EA+HK is separated from two further very heterogenous groups in SPK, analogous to the dendrogram at *k*=3. WRT returns identical results to the hierarchical approach (*k*=3), with only HK replacing the position of unary IND.

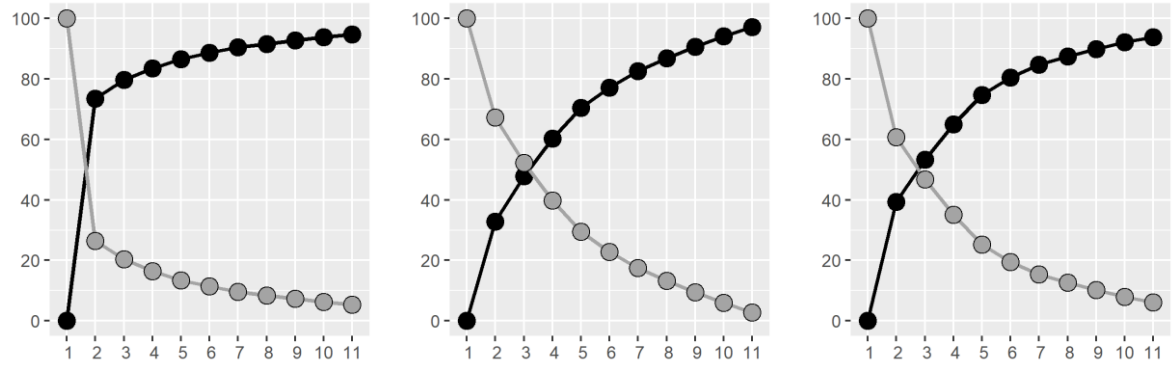


Figure 5.158: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS G^2 4-grams

Table 5.73: K-means clustering results for specific values of *k* for POS G^2 4-grams

	ALL (*k*=2)		**SPK (*k*=3)**		**WRT (*k*=3)**
1	All spoken corpus parts	1	GB, IND, IRL, JA, NZ	1	HK
2	All written corpus parts	2	CAN, NIG, PHI, SIN, SL	2	CAN, IND, PHI, SIN, SL, UG, USA
		3	EA, HK	3	KY, TZ, GB, GH, IRL, JA, NIG, NZ

5.4.4 Lexical Gravity

Lexical gravity POS-grams show a remarkable stability of the types assigned top association scores (Table 5.74), both over the various lengths for each separate dataset as also, if less, between datasets. Top collocates are from relatively frequent POS tags, predominantly forming noun phrase components or prepositional phrases. The lowest collocates are more heterogenous and only rarely overlap between different lengths and modes. While rare constituents (lexical items and infrequent tags) constitute the majority of sequences in the bottom ranks, longer sequences in these rows contain ever more purely lexical sequences. Thus, a sequence like *on the part of* scores among the lowest 4-grams in both modes, but also not at all infrequent items are found in the lowest ranks, such as AT1 VVN (e.g. *a given*), PPIS2 and (*we and*) or RRQ VM PPIS1 VVI (e.g. *when will I start*).

Table 5.74: POS *g n*-grams with highest and lowest association scores

2-grams		**3-grams**		**4-grams**	
type	***g***	**type**	***g***	**type**	***g***
Spoken					
AT NN1	11.17	AT NN1 of	10.88	AT JJ NN1 of	10.31
JJ NN1	11.12	JJ NN1 of	10.85	in AT NN1 of	10.25
NN1 of	10.59	AT JJ NN1	10.17	AT NN1 of AT	10.24
to VVI	10.39	in AT NN1	10.08	NN1 of AT NN1	10.24
PPY VV0	9.69	of AT NN1	10.07	of AT NN1 of	10.24
JJ NN2	9.64	AT NN1 NN1	10.07	AT NN1 NN1 of	10.24
AT1 VVN	-10.92	VH0 PPHS2 VVN	-4.75	on the part of	-1.64
PPIS2 and	-11.09	VM DD1 VBI	-4.77	XX PPIS2 VDI XX	-1.77
of VM	-11.09	in support of	-4.87	NN1 VBZ such that	-2.00
APPGE and	-11.28	VDI VBZ VVI	-5.30	NN1 rather than VVI	-2.19
of VBZ	-11.42	VBZ such that	-6.47	DD2 of PPY PNQS	-2.22
PPIS1 AT	-11.67	VM EX VBI	-7.20	in the light of	-3.01
Written					
JJ NN1	12.22	JJ NN1 of	11.96	AT JJ NN1 of	11.35
NN1 of	11.70	AT NN1 of	11.67	JJ NN1 of AT	11.28
AT NN1	11.64	AT JJ NN1	11.18	JJ NN1 NN1 of	11.20
JJ NN2	11.28	JJ NN1 NN1	10.95	AT NN1 of AT	11.08
to VVI	10.92	NN1 of AT	10.81	NN1 of AT NN1	11.08
AT JJ	10.14	JJ NN1 and	10.81	of AT NN1 of	11.08
AT AT	-10.09	VM PPH1 VBI	-4.95	XX VBN JK to	0.23
AT VVZ	-10.14	VDZ PPHS1 VVI	-5.40	in addition to JJ	0.04
VM NN2	-10.65	VM PPIS1 VDI	-5.66	in addition to NN1	0.02
VM and	-10.88	VM DB VVI	-5.91	AT1 MC NNT1 NN1	-0.41
PPIS1 VVN	-10.88	VDZ DD1 VVI	-6.08	on the part of	-1.67
in VVI	-11.53	VM MD VVI	-7.68	RRQ VM PPIS1 VVI	-2.39

Stable clusters for lexical gravity POS **2-grams** are detected more frequently in the written than the spoken parts of the data (Figure 5.159). While speech reports the same PHI+SL cluster in both ALL and SPK, the partial spoken IC cluster (-NZ) misses significance slightly at AU=93 in ALL, where additionally EA+NIG is found at AU=94. In SPK, the stepwise pattern in OC misses significance by fractions at the EA and SL heights. Clusters in WRT are identical between both datasets, identifying an African cluster (but cf. the common allocation of UG to GH+NIG), two IC groups and IND+SL.

Significant jumps (Figure 5.160) for ALL are found only after NZ_{SPK} is split off from the other spoken varieties due to its relatively difference to other varieties. Above-average jumps are identified until *k*=6, at which point speech identifies JA+(remaining) IC and separates the African spoken varieties from OC. The written branch, however, only distinguishes the African varieties from all others. The separate datasets require relatively large values for *k*, finding the first significant jump for SPK at *k*=5 to segment the data exactly as in ALL's spoken branch. WRT at *k*=7 (*k*=6 is also significant, but at much smaller jump heights) presents a much finer separation overlapping with most of the clusters found stable above and supporting a separation into two clusters each for IC, Asia and Africa, with JA remaining as a unary node.

Inspection of the NeighborNet graphs (Figure 5.161) suggests a more homogenous IC group within SPK, but NZ is still found to be relatively distinct. The OC group appears highly heterogenous, and only HK, SIN (and SL) are relatively close within the network. WRT suggests two weakly related IC clusters, with JA in an intermediary position to the OC varieties. PHI is found to be similar to USA and CAN, and WRT furthermore retrieves small distances for HK+SIN and IND+SL. It also delimits an African group from the data, within which UG is found slightly more similar to KY and TZ than the West African varieties.

K-means clustering (Figure 5.162 and Table 5.75) supports the relatively fine partitioning suggested above. ALL indicates *k*=5, leading to the same clusters as would be found in the hierarchical approach at this resolution, except for a merger of written IND+SL with the African varieties. SPK at *k*=5 produces identical results to those above, and WRT is only different in that HK takes the place of unary JA.

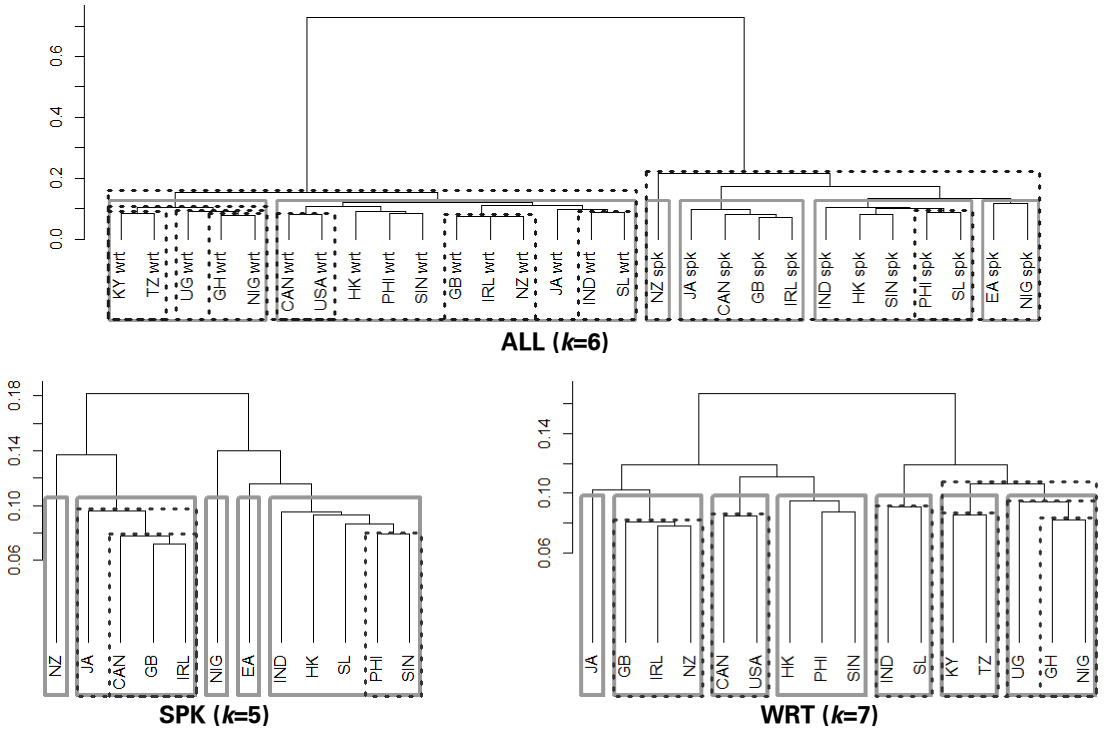


Figure 5.159: Hierarchical clustering results for POS *g* 2-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

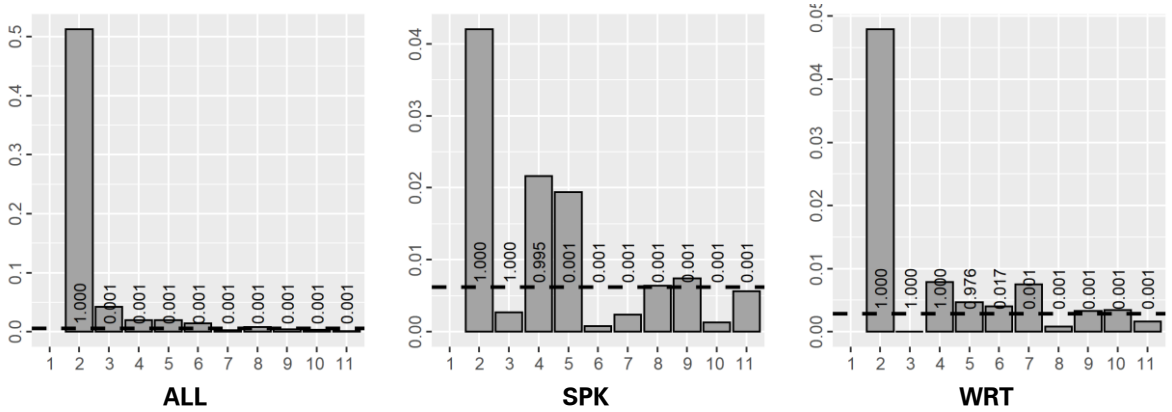


Figure 5.160: Jumps in node heights and respective *p*-values for POS *g* 2-grams

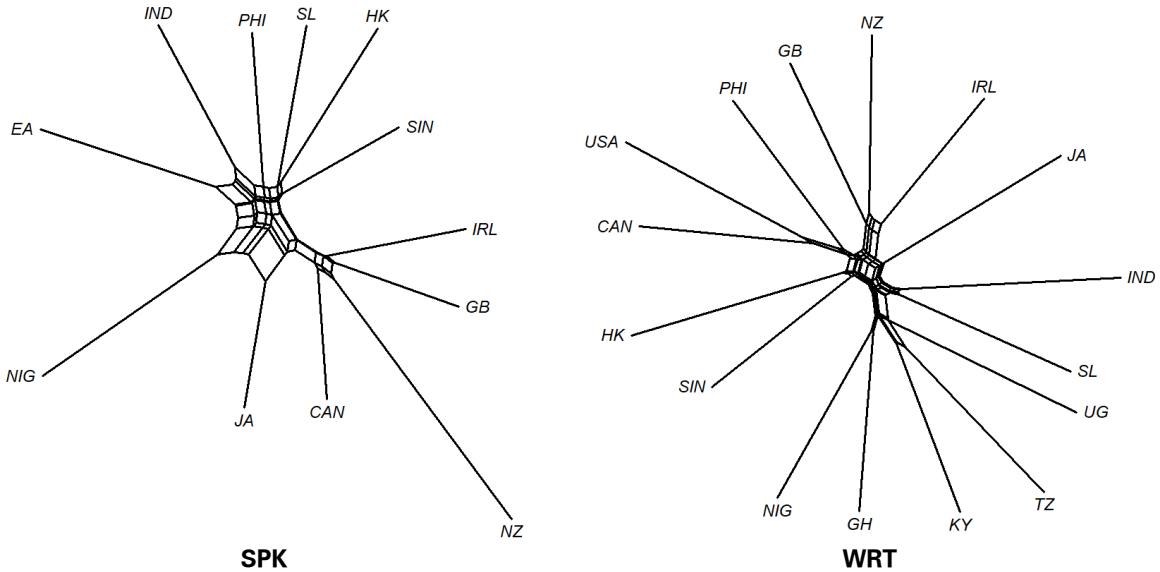


Figure 5.161: NeighborNets of the spoken and written data for POS *g* 2-grams

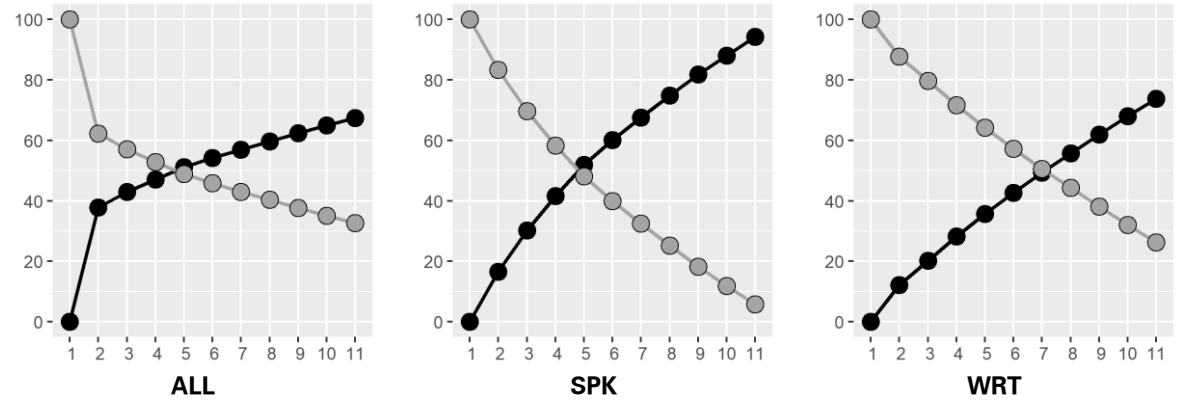


Figure 5.162: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *g* 2-grams

Table 5.75: K-means clustering results for specific values of *k* for POS *g* 2-grams

	ALL (*k*=5)		SPK (*k*=5)		WRT (*k*=7)
1	KY, TZ, GH, IND, NIG, SL, UG $_{WRT}$	1	HK, IND, PHI, SIN, SL	1	HK
2	CAN, GB, HK, IRL, JA, NZ, PHI, SIN, USA $_{WRT}$	2	NIG	2	GB, IRL, NZ
3	NZ $_{SPK}$	3	CAN, GB, IRL, JA	3	GH, NIG, UG
4	EA, HK, IND, NIG, PHI, SIN, SL $_{SPK}$	4	EA	4	CAN, USA
5	CAN, GB, IRL, JA $_{SPK}$	5	NZ	5	IND, SL
				6	JA, PHI, SIN
				7	KY, TZ

POS **3-grams** show an increase in stable clusters (Figure 5.163) and more meaningful fine-grained hierarchies. While SPK produces an almost identical dendrogram as for 2-grams, NZ is no longer separated from other IC, which in turn is contrasted against the entirety of OC, with NIG and EA most dissimilar to all others. Further subclusters emerge in HK+SIN and PHI+SL (lacking the usual IND). The WRT dendrogram also accords with that for shorter sequences, identifying Africa and in particular the ICE-EA varieties. Additionally, CAN+USA and an Asian group (-IND) are detected. ALL' written branch also identifies the African group, PHI+SIN and CAN+USA, and also the spoken branch is largely identical to SPK with the exception of EA and IND allocated to a separate cluster and JA moving from the IC to the remaining OC group.

Significant jumps (Figure 5.164) require at least *k*=3 for ALL, separating IC within speech. Finer *k*=6 and *k*=8 are also supported, distinguishing the written African cluster as well as EA+IND in speech, but also isolating NZ_{SPK}. At *k*=8, NIG_{SPK} is split off, but so is written HK+CAN+USA. This also isolates NZ, NIG, and EA, but leaves IND with the Asian data, and creates the same two IC sets as above. Even finer *k*=8 removes JA from the remaining IC set and separates the Asian group into IND and two binary groups. Writing requires *k*=7, which leads to small regional/epicentral groups, but excludes HK and IND from a mostly Asian set.

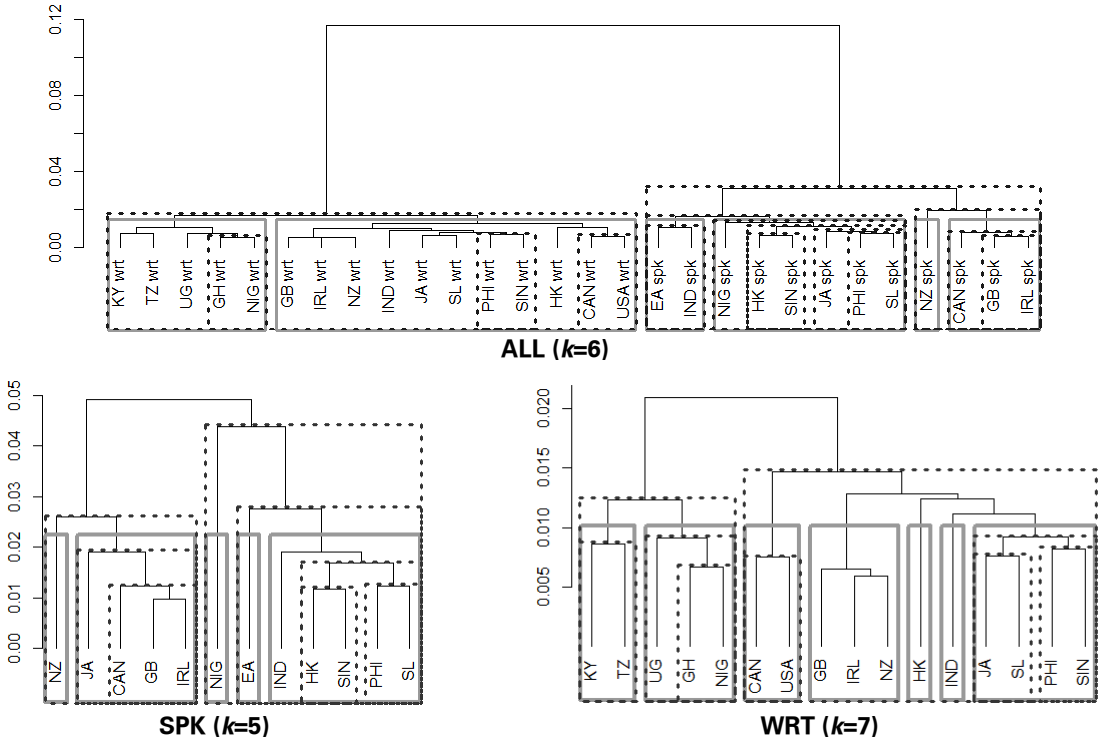


Figure 5.163: Hierarchical clustering results for POS *g* 3-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

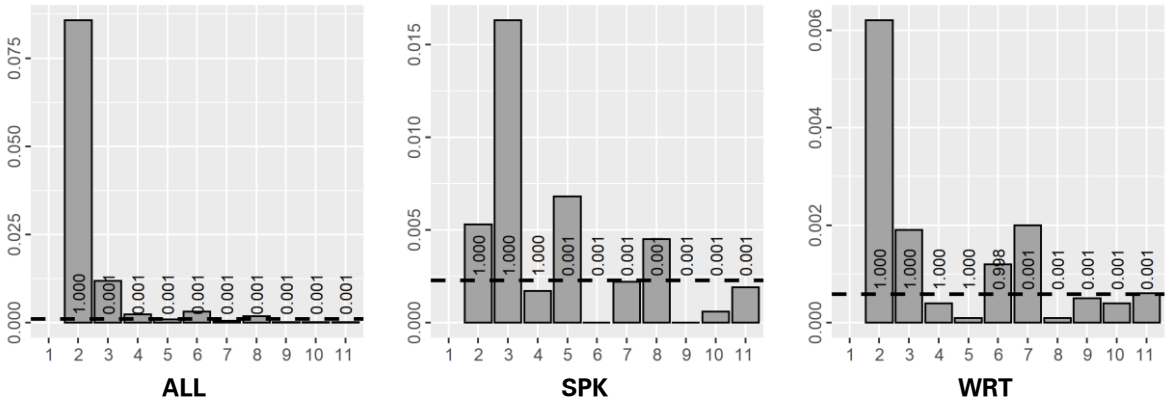


Figure 5.164: Jumps in node heights and respective *p*-values for POS *g* 3-grams

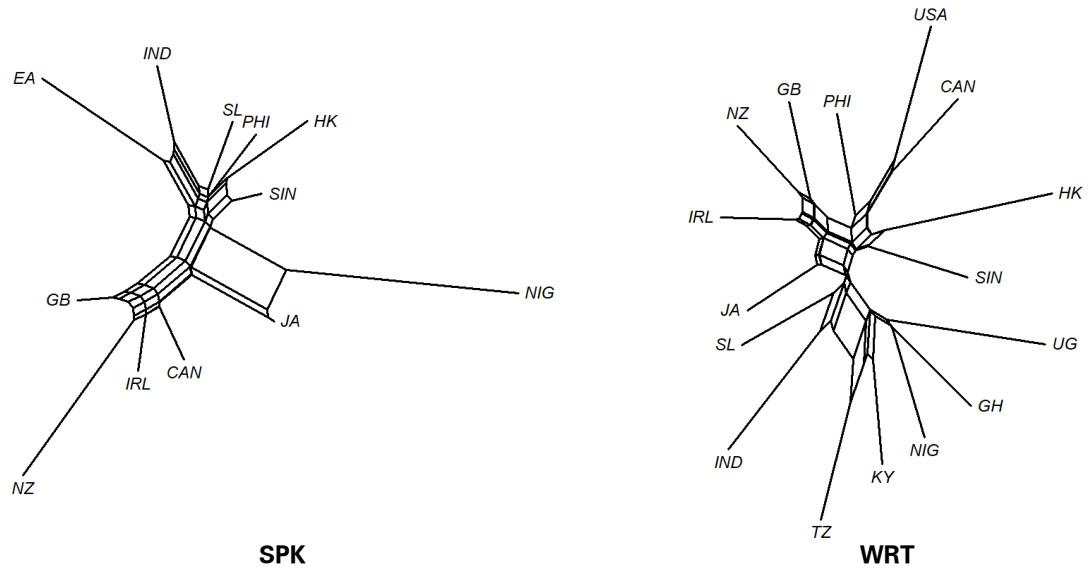


Figure 5.165: NeighborNets of the spoken and written data for POS *g* 3-grams

The NeighborNets (Figure 5.165) identify spoken IC but separate NZ. SPK also removes EA+IND and NIG+JA from the remaining OC data. IC in writing instead produces two mostly unrelated clusters (with PHI and HK+SIN relatively close to IC_{NA}). Within the African group, KY+TZ diverge most strongly, which appears to hinge on several shared features of TZ and IND, which itself shows elevated similarity to SL.

K-means (Figure 5.166 and Table 5.76) indicates only the binary spoken/written split for ALL (*k*=3 also separating IC_{SPK}). For SPK at *k*=4, the large distances of EA and NIG lead to them being isolated from the two larger groups identifiable in the NeighborNet above. WRT indicates *k*=6, which supports the African and IC clusters (+JA), but has the Asian data conform more closely to regional proximity.

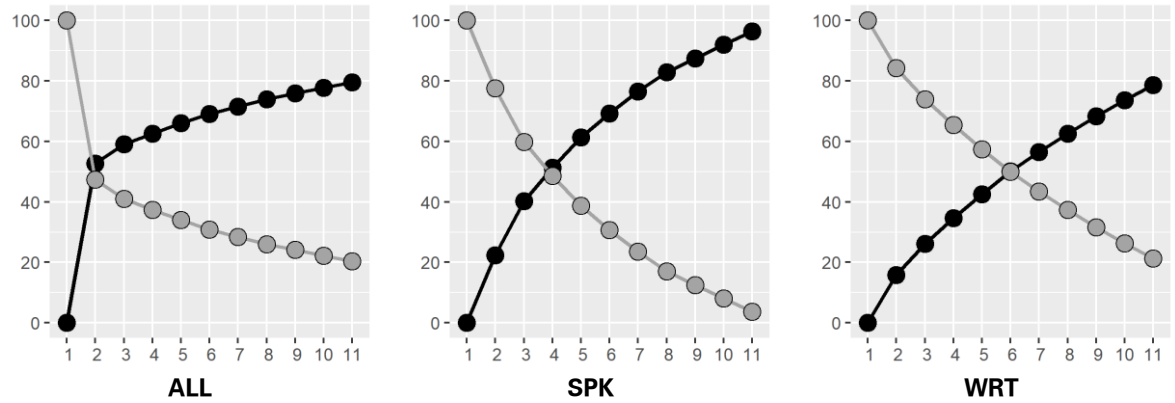


Figure 5.166: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *g* 3-grams

Table 5.76: K-means clustering results for specific values of *k* for POS *g* 3-grams

	ALL (*k*=2)		SPK (*k*=4)		WRT (*k*=6)
1	All spoken corpus parts	1	EA	1	IND, SL
2	All written corpus parts	2	CAN, GB, IRL, JA, NZ	2	GH, NIG, UG
		3	HK, IND, PHI, SIN, SL	3	HK, PHI, SIN
		4	NIG	4	GB, IRL, JA, NZ
				5	CAN, USA
				6	KY, TZ

The longest *g*-based sequences, i.e. **4-grams**, show remarkable similarities to their shorter forms, even if somewhat lower levels of substantiation are achieved (Figure 5.167). WRT only finds stable African subclusters (-UG) as well as CAN+USA, while ALL additionally assigns HK to the latter group (the predominantly Asian cluster including HK in WRT achieving AU=94) and also identifies GB+NZ, IRL+JA as well as PHI+SIN (fractionally missing significance in WRT). SPK finds a substantiated IC vs OC split with several HK+SIN, PHI+SL and JA+NIG subclusters. In ALL, only the GB+IRL and HK+SIN subclusters achieve significance.

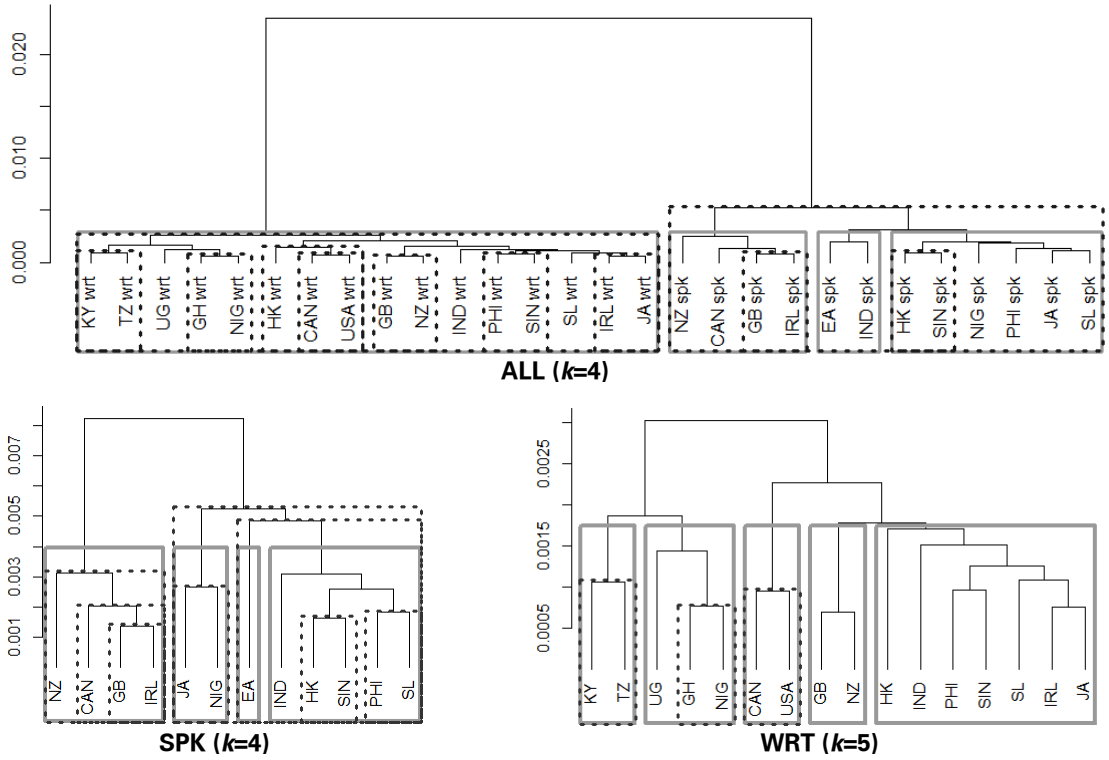


Figure 5.167: Hierarchical clustering results for POS *g* 4-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

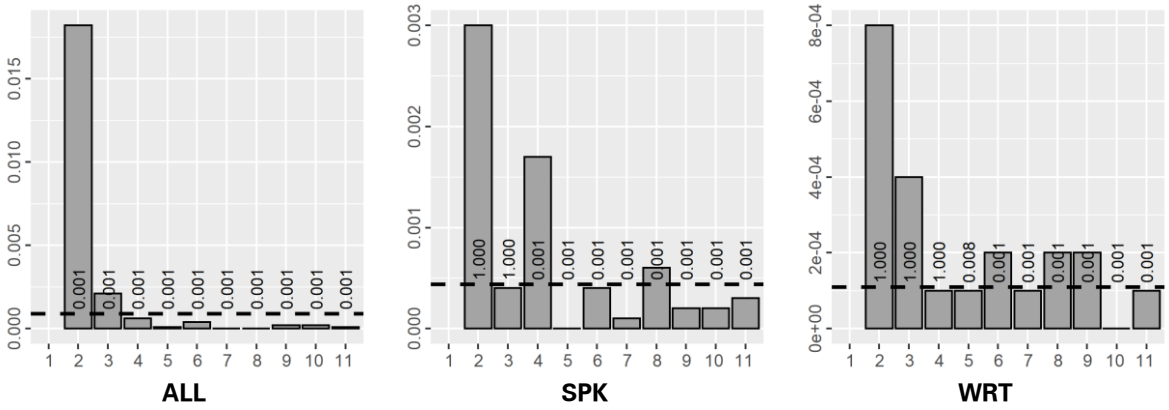


Figure 5.168: Jumps in node heights and respective *p*-values for POS *g* 4-grams

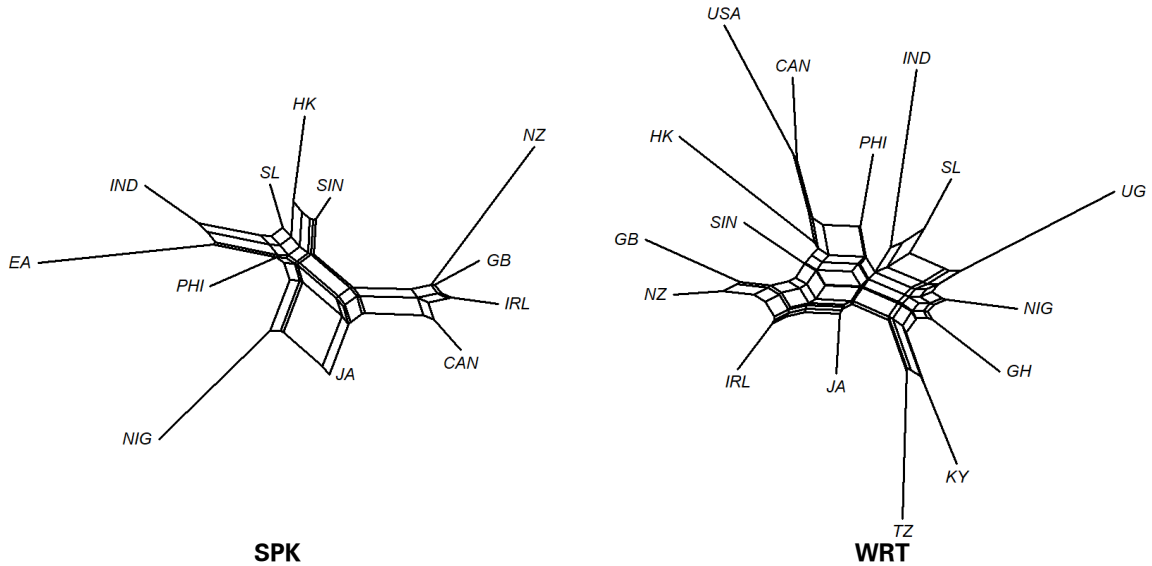


Figure 5.169: NeighborNets of the spoken and written data for POS *g* 4-grams

Significant jumps (Figure 5.168) are found above average for ALL until *k*=3 and separate speech into IC and OC. Finer *k*=4 additionally separates EA+IND from the remaining spoken OC data. SPK strongly indicates *k*=4, resulting in partitions for the IC and Asian data as well as a JA+NIG and unary EA. WRT only achieves significance at *k*=5, identifying Asia, two African and two IC clusters (-IRL). Higher jumps are found at *k*=6 and *k*=8, which successively split off HK, IND and UG from their clusters.

Inspection of the NeighborNets (Figure 5.169) again shows a relatively homogenous IC spoken cluster and familiar HK+SIN and EA+IND clusters as well as mutual distance of JA and NIG to the other varieties. Writing confirms two IC clusters (PHI again found close to USA+CAN), but HK and SIN is less supported while IND+SL and particularly the African varieties are more clearly identifiable (with UG again closer to NIG and GH).

K-means clustering (Figure 5.170 and Table 5.77) again only supports a binary split of ALL (*k*=3 again separating spoken IC). WRT at *k*=5, however, finds a very different solution: While the EA and two IC clusters reemerge (and IRL is no longer missing) and the HK+SIN subcluster as well as a separateness of IND are highlighted, the remaining varieties are merged into a group of heterogenous regions and evolutionary stages.

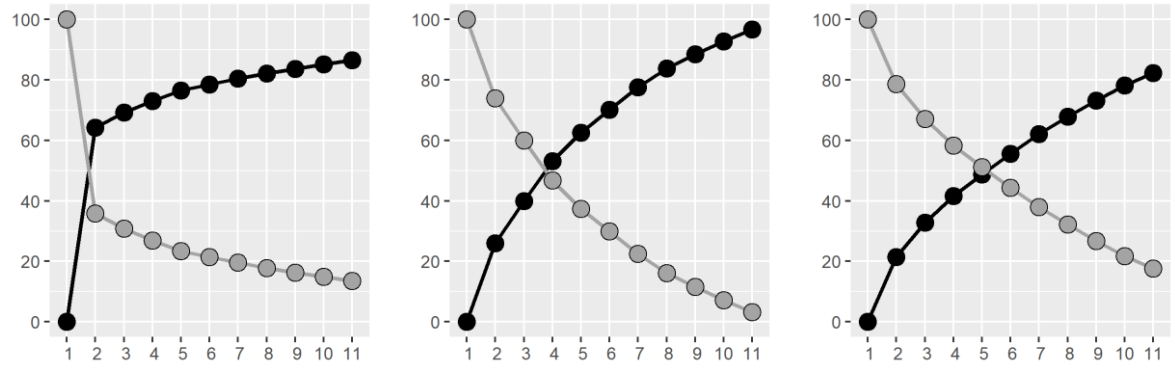


Figure 5.170: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *g* 4-grams

Table 5.77: K-means clustering results for specific values of *k* for POS *g* 4-grams

	ALL (*k*=3)		**SPK (*k*=4)**		**WRT (*k*=5)**
1	All spoken corpus parts	1	JA, NIG	1	GH, JA, NIG, PHI, SL, UG
2	All written corpus parts	2	HK, IND, PHI, SIN, SL	2	GB, IRL, NZ
		3	CAN, GB, IRL, NZ	3	HK, SIN
		4	EA	4	CAN, USA
				5	IND
				6	KY, TZ

5.4.5 Delta $P_{2|1}$

Delta P POS-gram association values (Table 5.78) again display the directionality of the measure, particularly in such items as *rather than*, *accordance with* or *depending on* AT NN1 (e.g. the *choice*). Lowest association scores are derived for sequences in which the first constituents are items frequently associated with a wide range of choices, such as RL (e.g. *alongside*), RG (e.g. *very*, *too*), PPHS (e.g. *she*) or MC (cardinal numbers), as well as for sequences of mostly nouns (here: NN1, NN2) stemming from very long compounds, complex nouns in ditransitive constructions or potentially repetitions or errors in the data. Interestingly, 2-grams are mostly concretely lexical items, but this effect lessens for longer sequences. Also, most of the top collocates are shared between speech and writing, often found even at the same table ranks, and top 4-grams contain several items also found at the shorter lengths, even if no clear tendency of continuous expansion of 2-grams to 4-grams can be discerned in the present (small) dataset.

Table 5.78: POS *ΔP* *n*-grams with highest and lowest association scores

2-grams type	*ΔP*	3-grams type	*ΔP*	4-grams type	*ΔP*
Spoken					
rather than	0.9992	the part of	0.7965	the part of AT	0.5956
apart from	0.9963	JK to VVI	0.7162	depending on AT NN1	0.5597
far as	0.9956	RPK to VVI	0.7160	apart from AT NN1	0.5373
long as	0.9955	VMK to VVI	0.7071	away from AT NN1	0.5363
touch with	0.9948	the light of	0.6589	on the part of	0.5323
line with	0.9948	order to VVI	0.6535	terms of AT NN1	0.5208
PPY NN1	-0.1388	VBZ NN1 JJ	-0.0850	NN2 NN1 NN1 RR	-0.0613
PPHS1 NN1	-0.1388	AT1 RR NN1	-0.0851	VBZ NN1 NN1 NN1	-0.0625
VM NN1	-0.1389	RR NN1 NN1	-0.0851	VVN NN1 NN1 NN1	-0.0631
PPHS2 NN1	-0.1392	VVN NN1 JJ	-0.0859	NN2 NN1 NN1 NN1	-0.0633
PPIS2 NN1	-0.1395	NN2 NN1 JJ	-0.0862	UH NN1 UH NN1	-0.0633
RG NN1	-0.1403	PPY NN1 NN1	-0.0876	NN1 RR NN1 NN1	-0.0669
Written					
rather than	0.9987	the part of	0.7986	the part of AT	0.5939
away from	0.9949	JK to VVI	0.7250	depending on AT NN1	0.5420
apart from	0.9949	RPK to VVI	0.7248	on the part of	0.5337
depending on	0.9928	VVNK to VVI	0.7248	away from AT NN1	0.5259
accordance with	0.9927	VMK to VVI	0.7131	JK to VVI AT	0.5052
far as	0.9890	VVGK to VVI	0.7072	VMK to VVI AT	0.4973
PPHO1 NN1	-0.1836	VVN NN1 NN1	-0.1363	NN1 NN1 NN2 NN1	-0.1042
RL NN1	-0.1839	that NN1 NN1	-0.1371	NN1 NN2 NN1 NN1	-0.1042
PPY NN1	-0.1849	VBDZ NN1 NN1	-0.1383	VV0 NN1 NN1 NN1	-0.1068
RR NN1	-0.1876	VBR NN1 NN1	-0.1385	VVI NN1 NN1 NN1	-0.1087
NN2 NN1	-0.1884	RR NN1 NN1	-0.1431	to NN1 NN1 NN1	-0.1153
VM NN1	-0.1897	NN2 NN1 NN1	-0.1435	MC NN1 NN1 NN1	-0.1160

As in the case of the lexical sequences, *ΔP*-based **2-grams** only cluster into very few stable constellations (Figure 5.171). While ALL's spoken and written branches are found stable overall, relatively little substructures are detected. For speech, only a GB+IRL+HK+SIN cluster is obtained in both datasets, while writing only shows a (single) stable East African cluster in WRT (a written NZ+GB+IRL subcluster is found at AU=93 in ALL, and HK+SIN barely misses significance in WRT at AU=94). All further evaluations should thus be seen in the light of this low level of overall substantiation.

Analysis of jump heights (Figure 5.172) retrieves relatively fine clusters for all but ALL. There, *k*=3 provides the first significant jump, separating NZ+CAN+PHI from the spoken branch. Beyond this, *k*=6 and *k*=7 show above-average jumps but only add similarly meaningless clusters by partitioning off USA+CAN+PHI from the remaining written data, isolating EA and NZ into unary nodes and a with some varieties from Asia and the British Isles. SPK also reports much of this cluster (-IND), but apart from a JA+NIG group fragments into only unary nodes at the first significant (*k*=8) partition. WRT also requires *k*=8, at which point two IC clusters remain (USA+CAN+PHI like in ALL) as well as the stable East African cluster and two binary groups (HK+SIN and JA+SL).

NeighborNet interpretation (Figure 5.173) suggests a more homogenous IC spoken group but otherwise finds strong heterogeneity within the data, with only some similarity supported for PHI+JA+SL. Of the typical clusters, writing only truly identifies a North American IC group (+PHI), an East African cluster and HK+SIN.

K-means clusters (Figure 5.174 and Table 5.79) also exhibit the tendencies towards fragmentation observed above. In ALL, an elbow at *k*=2 separates speech and writing, but relevant intersects are only found at *k*=6. At this level, some of the (stranger) clusters from the hierarchical analysis are repeated, such as written CAN+PHI+USA and spoken GB+HK+IRL+SIN (+IND) as well as unary EA, but the remaining (above unstable) structure is differently presented as JA+NIG+PHI+SL_{SPK} and CAN+NZ_{SPK}. SPK identifies almost identical structures, only isolating IND from the GB+IRL+HK+SIN cluster found stable before, and WRT at *k*=7 identifies the same structures as the dendrogram at that height, thus merging GH with HK+SIN.

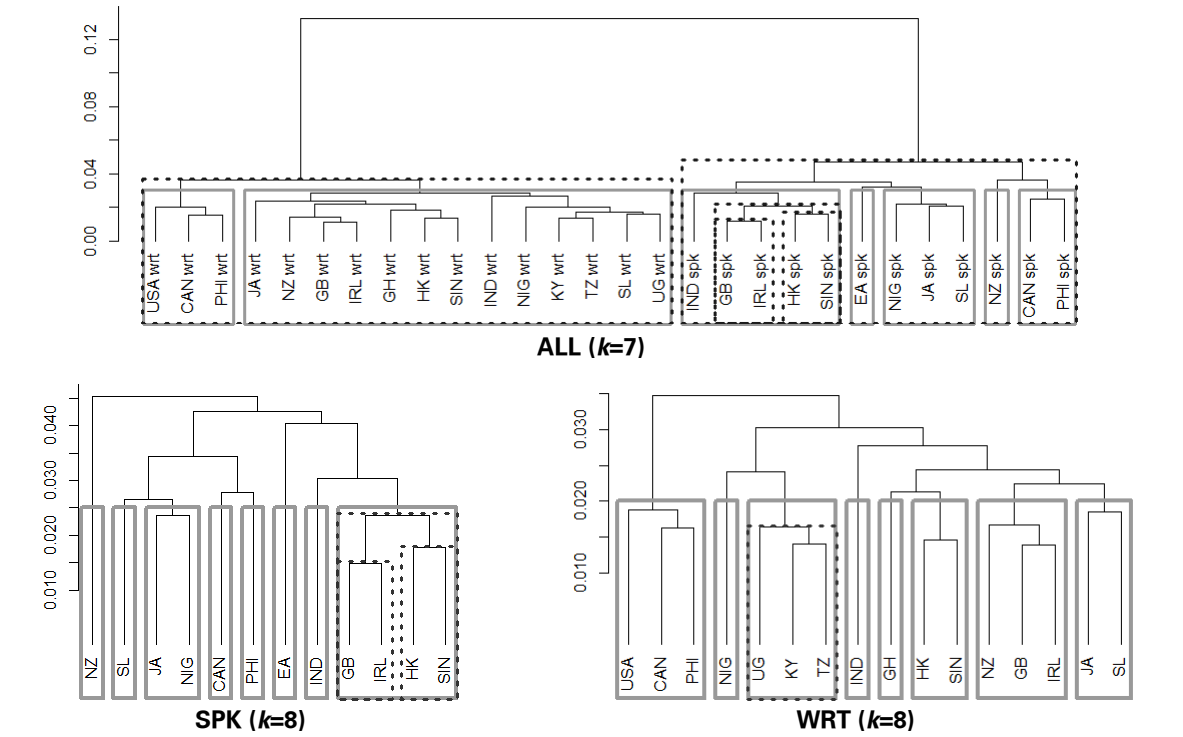


Figure 5.171: Hierarchical clustering results for POS *ΔP* 2-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

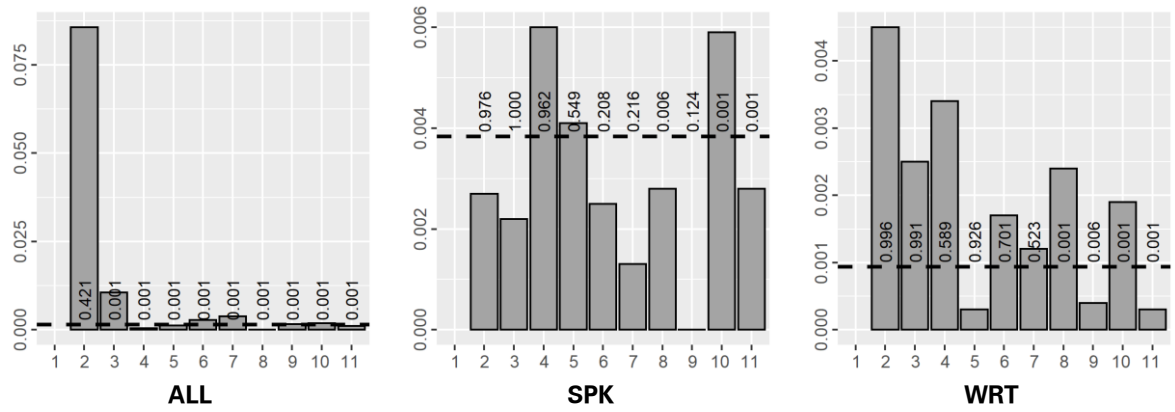


Figure 5.172: Jumps in node heights and respective *p*-values for POS *ΔP* 2-grams

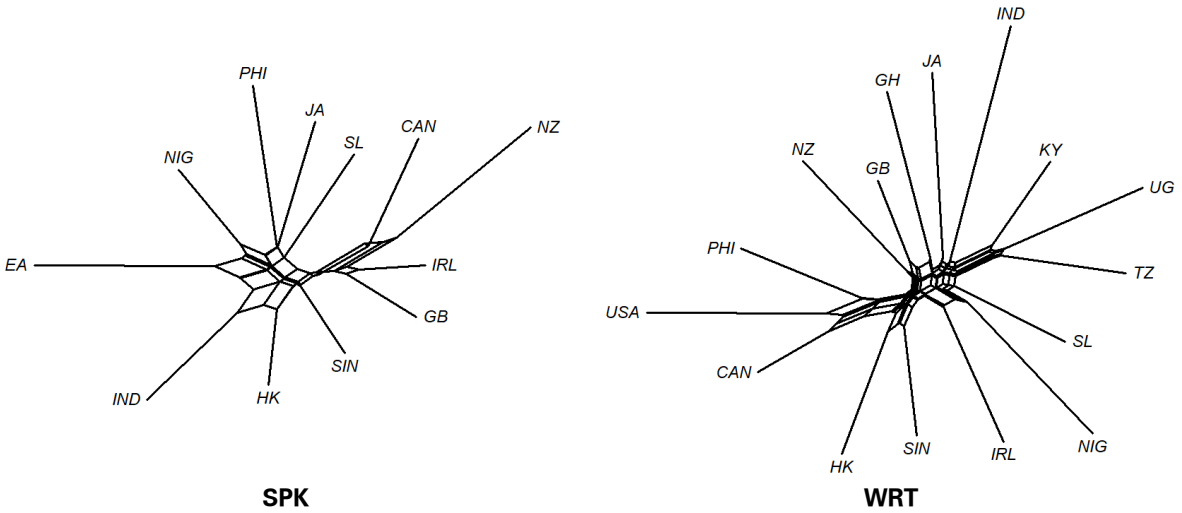


Figure 5.173: NeighborNets of the spoken and written data for POS *ΔP* 2-grams

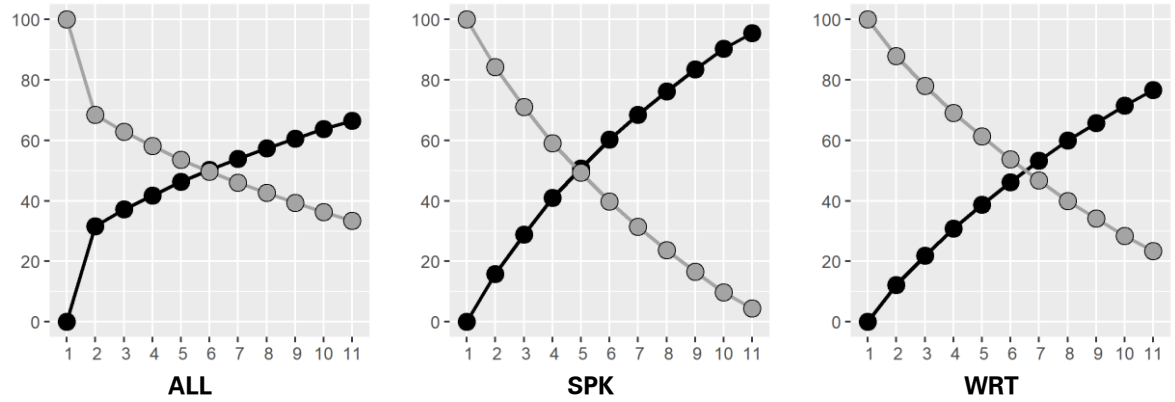


Figure 5.174: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *ΔP* 2-grams

Table 5.79: K-means clustering results for specific values of *k* for POS *ΔP* 2-grams

	ALL (*k*=6)		SPK (*k*=5)		WRT (*k*=7)
1	KY, TZ, GB, GH, HK, IND, IRL, JA, NIG, NZ, SIN, SL, UG $_{WRT}$	1	JA, NIG, PHI, SL	1	JA, SL
		2	EA	2	CAN, PHI, USA
		3	GB, HK, IRL, SIN	3	KY, TZ, UG
2	CAN, PHI, USA $_{WRT}$	4	CAN, NZ	4	GB, IRL, NZ
3	GB, HK, IND, IRL, SIN $_{SPK}$	5	IND	5	NIG
4	JA, NIG, PHI, SL $_{SPK}$			6	IND
5	EA $_{SPK}$			7	GH, HK, SIN
6	CAN, NZ $_{SPK}$				

3-grams continue the tendency of *ΔP*-based sequences to find relatively little substantiation (Figure 5.175). This becomes particularly evident in speech, where the spoken branch overall does not reach sufficient AU values and only substantiates GB+IRL and HK+SIN subclusters (also the only clusters found stable in SPK), supplemented by a barely insignificant partial Asian JA+NIG+PHI+SL. Writing overall achieves significant stability, and substantiates two internal IC clusters, JA+SL, PHI+SIN and KY+TZ in both datasets, but no larger structures can be detected.

Segmentation by jump heights (Figure 5.176) only shows significant results after *k*=3 in ALL, splitting off spoken IND+EA, consecutive *k*=4 additionally sectioning off spoken IC. Further clusters emerge at the above-average jumps for *k*=6, which separates EA and IND and identifies an African (+IND) cluster in writing. SPK clearly indicates *k*=4, also resulting in unary EA and IND and an IC vs. remaining OC cluster (like in ALL). For WRT, *k*=7 provides the first significant jumps, resulting in clusters strongly oriented around those found stable above, but extending PHI+SIN by HK and KY+TZ by UG, and further identifying unary IND as well as GH+NIG.

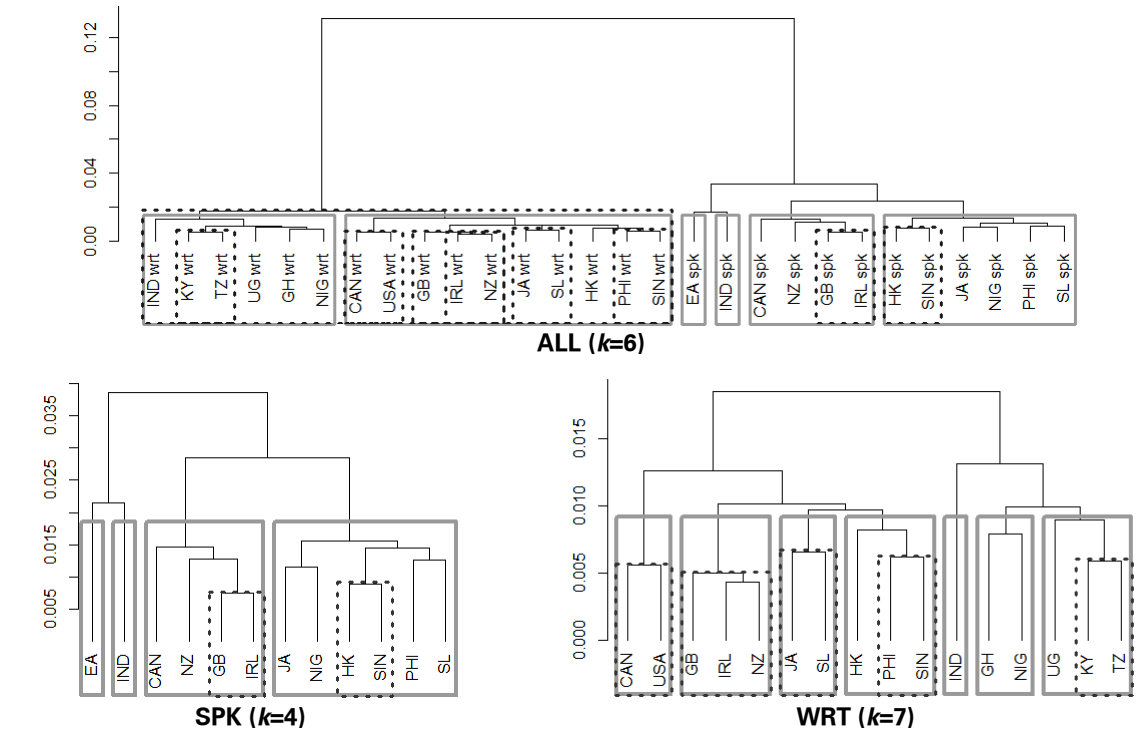


Figure 5.175: Hierarchical clustering results for POS *ΔP* 3-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

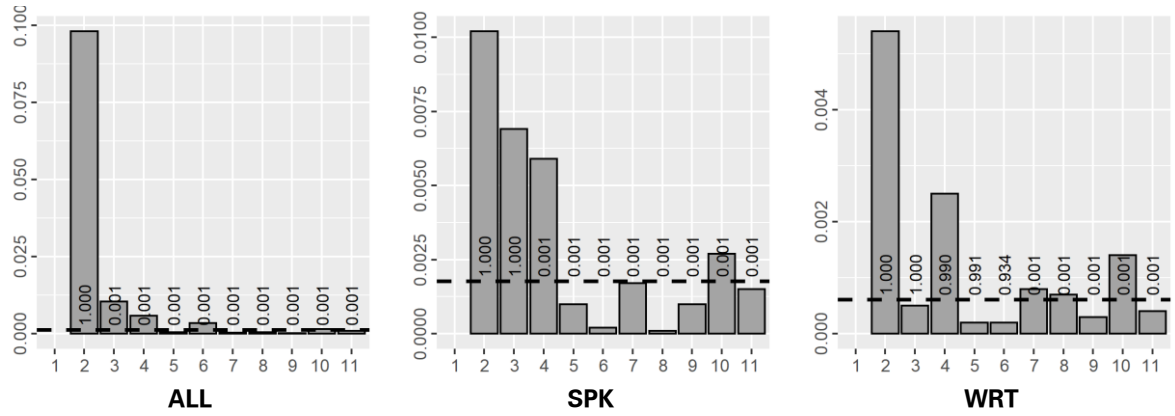


Figure 5.176: Jumps in node heights and respective *p*-values for POS *ΔP* 3-grams

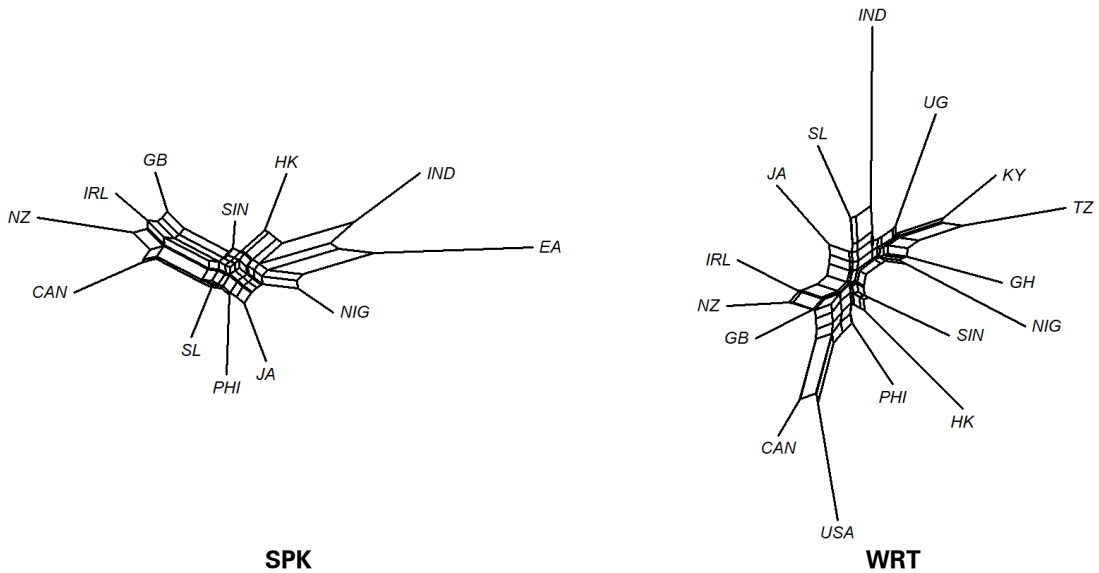


Figure 5.177: NeighborNets of the spoken and written data for POS *ΔP* 3-grams

The NeighborNets (Figure 5.177) strongly separate spoken IC and find large distances of IND, EA and (less so) NIG to the remaining data. The other varieties broadly differentiate into two groups of HK and SIN as well as SL, PHI and JA. For writing, CAN+USA are relatively different from the other IC varieties, but many shared features still exist. Africa is both separate from other OC varieties but also internally separated into the EA component and West Africa.

K-means analysis (Figure 5.178 and Table 5.80) of ALL identifies EA_{SPK} as the one variety most strongly influencing the instability of the spoken branch, and merges the variety with the written corpus parts. Only at *k*=3 is the distinctness of IND_{SPK} also brought out through a separate cluster together with EA_{SPK}. SPK at *k*=4 also supports distance of EA and IND from the IC and OC clusters as found in the hierarchical analysis. WRT meanwhile favors *k*=6, resulting in almost the same structure as observed above, with only the JA+SL cluster dissolved through the merger of JA with IC and SL with IND.

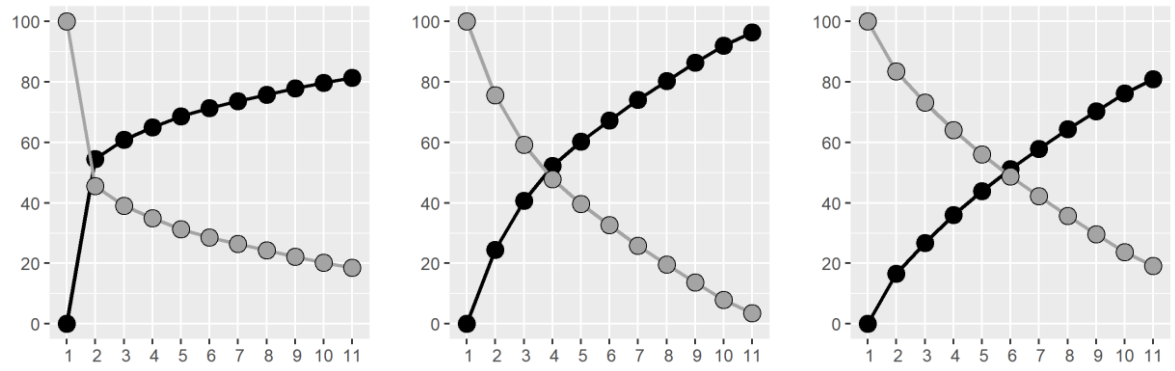


Figure 5.178: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *ΔP* 3-grams

Table 5.80: K-means clustering results for specific values of *k* for POS *ΔP* 3-grams

	ALL (*k*=2)		**SPK (*k*=4)**		**WRT (*k*=6)**
1	All spoken corpus parts -EA	1	EA	1	CAN, USA
		2	IND	2	KY, TZ, UG
2	All written corpus parts +EA	3	CAN, GB, IRL, NZ	3	IND, SL
		4	HK, JA, NIG, PHI, SIN, SL	4	GH, NIG
				5	HK, PHI, SIN
				6	GB, IRL, JA, NZ

4-grams further substantiate the allocation of EA_{SPK} to the written varieties hinted at before by an unstable spoken branch and retrieved through k-means (Figure 5.179). While joined at relatively high distances, the variety is still stably clustered with the written branch. Except for a partial IC group and PHI+SL, no further clusters emerge in ALL. SPK finds the same structures, but also merges PHI+SL with CAN+JA, and furthermore clearly indicates the difference of EA to the remaining varieties. For writing, clusters in both ALL and WRT retrieve two IC groups, an incomplete African and a mostly consistent Asian cluster.

Significant jumps (Figure 5.180) are again only found beyond the binary segmentation in ALL, and EA_{SPK} is split off at k=3. Jumps slightly above average heights are found up to k=7, when the written branch partitions IC, HK+IND and the remaining OC while the spoken branch identifies the stable (but incomplete) African cluster, CAN+USA and a mixed Asia+British-epicentral IC cluster. SPK clearly indicates k=4, at which point, similarly to the written branch above, HK+IND, IC (-CAN) and EA are differentiated from the remaining varieties. Jumps in WRT become significant at k=6, which delineates the stable African (-NIG) and CAN+USA clusters found before, but splits off IND, HK and the (remaining) OC varieties from a group of three Asian varieties joined by NIG and JA.

NeighborNet analysis (Figure 5.181) for SPK also captures the great difference of EA as well as IRL+GB+NZ from the other varieties. It further indicates relatively large distances of IND and HK to the remaining data but also similarities between EA and IND as well NIG. For writing, two IC clusters can be observed, which are distinct but also, together with HK+SIN, still mutually dissimilar from many other varieties. IND+SL and the African group (least clearly NIG) are also identified, which improves on some of the less relatable aspects of the dendrogram above.

K-means (Figure 5.182 and Table 5.81) also analyzes the spoken/written separation of ALL confounded through EA_{SPK}, which is isolated at k=3. SPK produces the same clusters as found in the dendrograms (clusters #2 and #4 merging at k=3), but the solution for WRT differs: There, some meaningful clusters are detected in the (incomplete) African group and IC_{NA}, and furthermore IND+SL and HK+PHI+SIN emerge, but IC_{GB} is joined by NIG and JA.

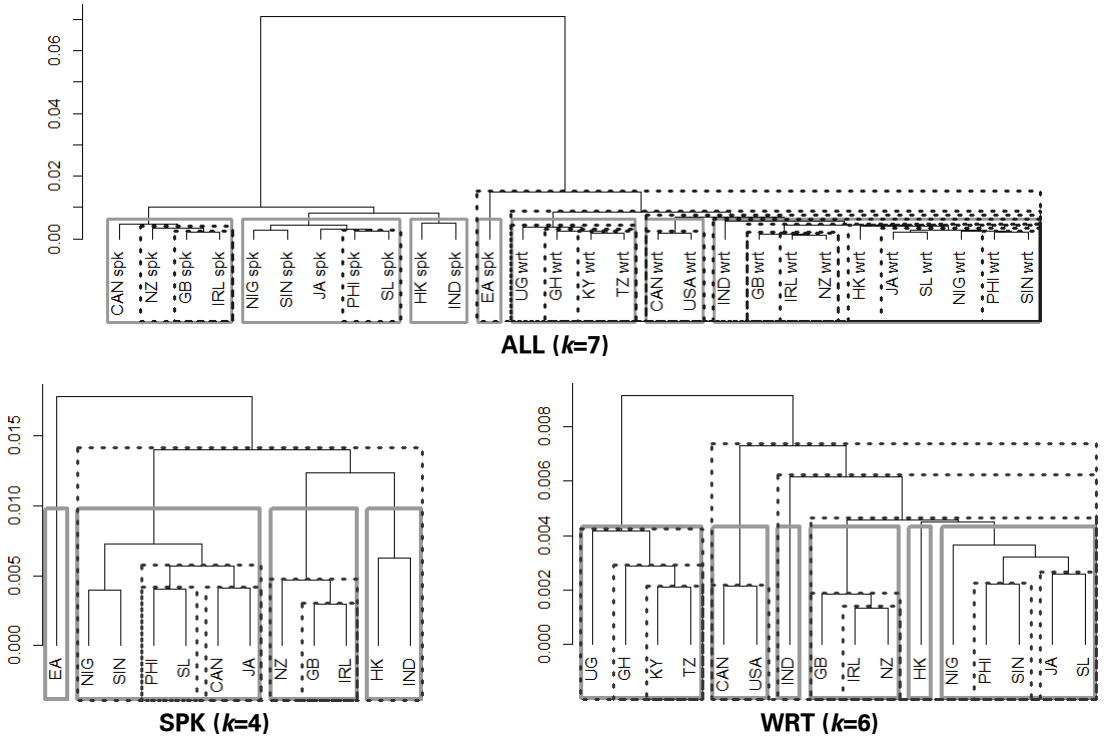


Figure 5.179: Hierarchical clustering results for POS *ΔP* 4-grams; rectangles indicate significant clusters (AU≥95) identified by pvclust (black dotted lines) or through jumps in node height (gray solid lines)

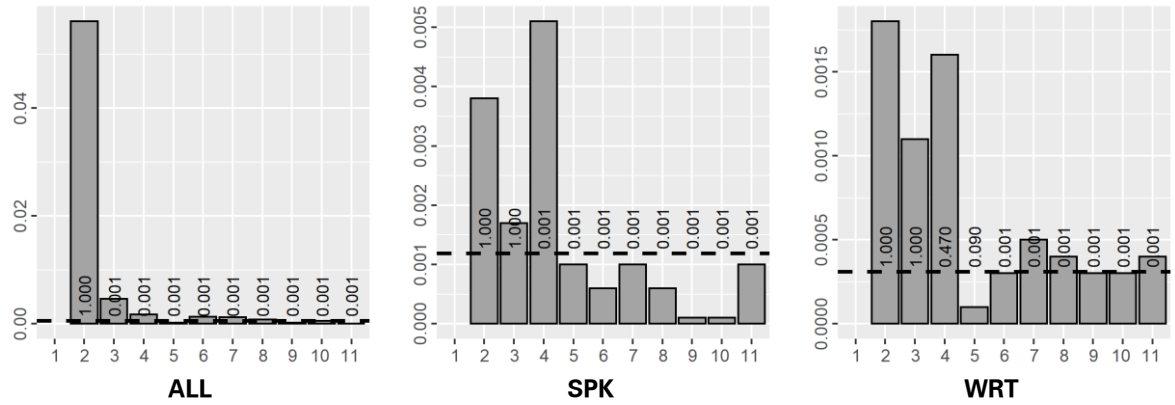


Figure 5.180: Jumps in node heights and respective *p*-values for POS *ΔP* 4-grams

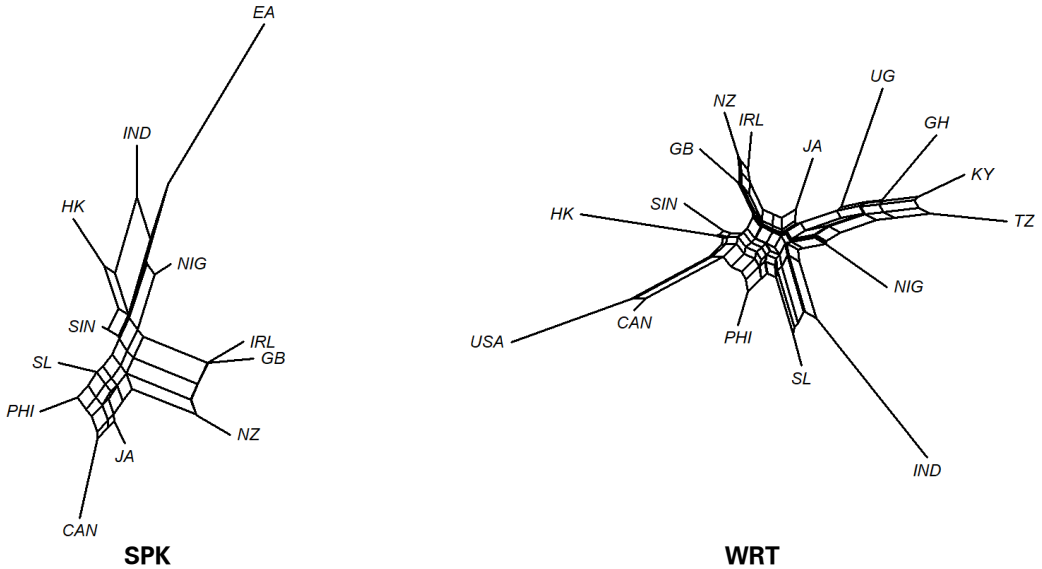


Figure 5.181: NeighborNets of the spoken and written data for POS *ΔP* 4-grams

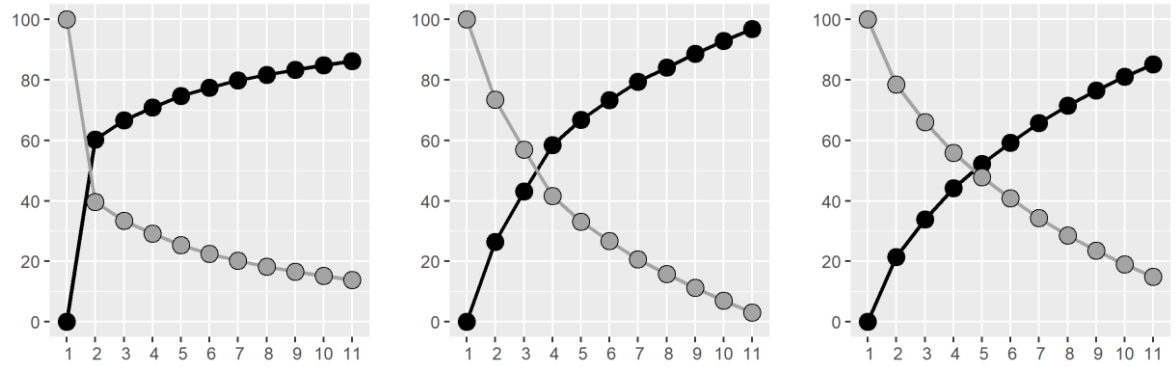


Figure 5.182: Percent variability explained (black) and within-cluster variation (gray) against number of k-means clusters for POS *ΔP* 4-grams

Table 5.81: K-means clustering results for specific values of *k* for POS *ΔP* 4-grams

	ALL (*k*=2)		**SPK (*k*=4)**		**WRT (*k*=5)**
1	All spoken corpus parts -EA	1	EA	1	KY, TZ, GH, UG
		2	GB, IRL, NZ	2	IND, SL
2	All written corpus parts +EA	3	CAN, JA, NIG, PHI, SIN, SL	3	CAN, USA
				4	HK, PHI, SIN
		4	HK, IND	5	GB, IRL, JA, NIG, NZ

6 Discussion and Evaluation

The aim of the present analysis has been to study whether preference patterns for the same sets of lexical or grammatical sequences consistently reflect language-external classifications of Englishes. A particular emphasis was placed on three major modeling perspectives on the global situation of English: simple binary segmentation along the Inner/Outer Circle dichotomy, regional proximity as potential aspects of epicentricity and historic-cultural similarities, or categorizations informed by a evolutionary dynamics. The study focused on lexical and grammatical *n*-grams, since they represent habitual language use and a close approximation to the common core of English. All the while, they remain variety-neutral and offer both ease of extraction as well as high methodological sophistication. Lexical *n*-grams and grammatical POS-grams were extracted either at dynamic lengths informed by the data, or at static lengths derived from the lengths preferred within the dynamic approach (which resulted in sequences of 2, 3, and 4 units). Association preferences were evaluated on the basis of five different measures of both traditional (MI-score, t-score, log-likelihood G^2) as well as more innovative designs (lexical gravity *g*, Delta P ΔP). Statistical analysis was carried out with the aid of four clustering methods: Major types of methods encompassed hierarchical, k-means and NeighborNet clustering, while hierarchical clustering, moreover, attempted to reconcile bottom-up most-substantiated clusters (`pvclust`) with an enforced top-down segmentation into as few clusters as possible.

After the conclusion of the analysis of 40 distinct datasets,[81] each of which evaluated on the grounds of four clustering methods, the question remains which general findings can be abstracted from this wealth of data. It appears only fitting that the data-driven approach is continued as much as possible even in this final step, and thus a triangulative approach will be embraced: Groups of varieties will first be established by agreement between clustering methods within each individual dataset, resulting in 40 sets of varietal clusters results. These can be evaluated globally for general tendencies across the entire data, which will be the analytical focus of Section 6.1. Following the assessment of the general patterns within the data, Section 6.2 will tease apart

[81] Two modes × two types of base data × four sequence lengths × five association measures.

the overall results by individual variables (mode, base data, length, measure). After the discussion and evaluation of findings, Chapter 7 will conclude the analysis and provide an outlook on the potential for further research.

6.1 Clusters of World Englishes

Evaluation of the 160 different perspectives on the data (four methods per dataset) will proceed in the following fashion: First, major clustering results for each of the 40 analyses will be determined by considering the agreement of methods on similar clusters. In this, a cluster will be considered a major result if it is retrieved by the majority of clustering methods (two for the hierarchical approach plus k-means and NeighborNet), which means that three out of four methods need to be in agreement. However, please note that this is not as clear-cut a procedure as it may seem, since clusters are frequently found to coincide only in parts. Thus, a group of varieties will also be regarded as a major cluster in case it is frequently found as a subcluster of more heterogenous larger structures (e.g. if two varieties cluster with various others across methods). Even so, many relevant clusters would be missed if only these major groups were considered. This is why, in addition to major clusters, minor clusters will be included in the discussion below in case only two methods retrieve a similar group and heterogeneity between methods is too pronounced.

Tables 6.1 and 6.2 (lexical and POS sequences, respectively) represent the overall results of the individual analyses and will be explained in more detail in the following. Several symbols and shorthand forms are required in order to economically present these results: The plus sign is used to represent the frequent merger of two (groups of) varieties within a cluster, the forward slash indicates two competing (similarly substantiated) perspectives on the same set of data, and commas separate related findings within a line (particularly found in case of groups from a common region). Isolated single variety labels within one line indicate separateness of the unary node ('outliers') just as groups of labels indicate clusteredness (i.e. separateness of a group). Brackets reflect a degree of uncertainty and are employed to signal minor results and demarcate them from major clusters. They can also combine with the other symbols in the table to convey lower confidence in a merger (+) or competing alternative (/). Lastly, spaces denote a degree of subclustering within groups of varieties. Cluster groups are

presented in the same order as much as possible for each analysis, going from coarser segmentation (binary splits) and outliers (if present) in the first two rows to frequently established smaller clusters in consecutive rows, which are ordered by their degrees of usual substantiation.

By way of illustration, a line such as '`KY+TZ + UG + GH+NIG`' (*MI*-based lexical written 3-grams) should be understood in the following way: An overarching African cluster shows relatively strong regional differentiation into KY+TZ as well as GH+NIG, while UG is found in varying constellations with any of the other varieties. Note that no brackets are present in this example, and that as such both the overall cluster as well as its subclusters are deemed stable across methods. As a second example contextualizing all symbols, consider the groups for the lexical spoken *MI* data at dynamic-length *n*: There, clearly a group of Inner Circle/phase 5 varieties is substantiated but also frequently joined by JA at some distance (`GB+IRL+NZ+CAN + JA`). A coherent overall group of all remaining varieties cannot be supported (unlike for 2-grams). The second row indicates that EA as well as IND (and NIG to a lesser extent) separate stongly from the remaining data, providing cases of outliers which need to be observed for the consecutive evaluation of smaller clusters. Somewhat less secure subclusters are detected in IND+SL as well as HK+SIN+PHI. These often occur together in some permutation as a joint cluster, which is why the merger itself (the plus sign) is not bracketed: `(HK+SIN+PHI) + (IND+SL)`.

Table 6.1: Clusters in the lexical data by agreement of clustering methods
(Codes: '+': mergers, '()': lower confidence, '/': alternatives, ',': related subclusters)

		MI	*t*	G^2	*g*	*ΔP*
SPEECH	n	▪ GB+IRL+NZ+CAN + JA ▪ EA, IND, (NIG) ▪ (HK+SIN+PHI) + (IND+SL)	▪ GB+IRL+NZ+CAN, all others ▪ EA+IND, (NIG) ▪ (HK+SIN), (JA+IND+PHI+SL)	▪ (GB+IRL+NZ+CAN) ▪ EA+IND ▪ (HK+SIN)	▪ GB+IRL+NZ(+)CAN ▪ EA+IND ▪ HK+SIN, JA+PHI+SL(+NIG)	▪ GB+IRL+NZ(+)CAN ▪ EA, IND ▪ (HK+SIN)
	2	▪ GB+IRL+NZ+CAN, all others ▪ EA, IND, NIG ▪ HK+SIN, JA+PHI(+SL)	▪ GB+IRL (+NZ+CAN) ▪ EA+IND/NIG	▪ (GB+IRL+CAN) (+NZ) ▪ EA+IND, (NIG)	▪ GB+IRL+NZ+CAN, (all others) ▪ EA+IND, NIG ▪ HK+SIN, JA+PHI+SL	▪ GB+IRL+NZ(+)CAN ▪ EA(+)NIG, IND ▪ HK+SIN
	3	▪ GB+IRL+NZ+CAN, all others ▪ EA, IND, NIG ▪ JA+PHI+SL(+SIN+HK)	▪ GB (+) IRL+NZ+CAN (+JA) ▪ EA+IND/NIG ▪ HK+SIN	▪ (GB+IRL+NZ) (+CAN) ▪ EA+IND, (JA+NIG)	▪ GB+IRL+NZ+CAN, (all others) ▪ (EA+IND) ▪ HK+SIN	▪ GB+IRL+NZ (+) CAN, (all others) ▪ IND, EA, (NIG) ▪ (JA+SL)
	4	▪ GB+IRL+NZ+CAN EA(+)IND ▪ (HK + IND+PHI)	▪ GB (+) IRL+NZ+CAN (+) JA, all others ▪ EA+IND	▪ (GB+IRL+NZ+CAN), (PHI_{SPK}+USA_{WRT}) ▪ (EA+IND) ▪ HK+SIN	▪ GB+IRL+NZ+CAN, (all others) ▪ EA, IND ▪ (JA+PHI+SL)	▪ (GB+IRL+NZ) / GB+IRL+NZ + JA+SL ▪ EA
WRITING	n	▪ GB+IRL+NZ+CAN+USA KY+TZ, GH+NIG+UG ▪ IND+SL, HK+SIN, JA+PHI	▪ GB+IRL+NZ (+) CAN+USA ▪ KY+TZ (+) GH+NIG(+)UG	▪ GB+IRL+NZ, CAN+USA (+PHI) ▪ KY+TZ+UG + IND(+SL) ▪ GH+NIG (+JA)	▪ GB+IRL+NZ, CAN+USA ▪ (KY+TZ) + (NIG+GH+UG)	▪ GB+IRL+NZ, (CAN+USA) ▪ (KY+TZ+UG+GH+NIG)
	2	▪ GB+IRL+NZ, CAN+USA ▪ KY+TZ (+) GH+NIG + UG ▪ (HK+SIN)	▪ GB+IRL+NZ (+) CAN+USA ▪ KY+TZ/UG + GH+NIG ▪ JA+SIN (+IND+SL)	▪ (GB+IRL+NZ+CAN+USA), (CAN+USA+PHI/SIN/SL) ▪ KY+TZ (+UG/IND), GH+NIG (+JA)	▪ GB+IRL+NZ, CAN+USA (+PHI) ▪ KY+TZ + GH+NIG+UG ▪ (HK/SIN+) (JA+PHI+SL)	▪ GB+IRL+NZ +HK, (CAN+USA) ▪ KY+TZ+UG, GH+NIG
	3	▪ GB+IRL+NZ, CAN+USA ▪ KY+TZ + UG + GH+NIG ▪ (HK+SIN)	▪ GB+IRL+NZ, CAN+USA ▪ KY+TZ (+) UG, GH+NIG	▪ KY+TZ (+UG) + HK+IND(+SL)	▪ GB+IRL+NZ, (CAN+USA) ▪ KY+TZ + GH+NIG+UG ▪ (IND+SL)	▪ GB+IRL+NZ, CAN+USA ▪ EA(+)IND, (NIG) ▪ KY+TZ+UG (+) GH+NIG
	4	▪ (KY+TZ+UG) + (GH+NIG)	▪ (GB+IRL+NZ + HK) (CAN+USA+SIN) ▪ (KY+TZ+UG) + IND, (GH+NIG)	▪ GB+SIN, CAN+USA+PHI ▪ (TZ+KY/IND+HK+SL)	▪ (GB+IRL+NZ,) CAN+USA ▪ KY/TZ/UG + GH+NIG ▪ PHI+SIN (+HK)	▪ GB+IRL+NZ+CAN / GB+IRL+NZ + JA+SL ▪ (NIG+UG), (GH+KY+TZ)

Table 6.2: Clusters in the POS data by agreement of clustering methods
(Codes: '+': mergers, '()': lower confidence, '/': alternatives, ',': related subclusters)

		MI	*t*	*G²*	*g*	*ΔP*
SPEECH	n	▪ GB+IRL+NZ+CAN ▪ EA, (IND, NIG)	▪ (GB+IRL+NZ+CAN) ▪ EA (+IND)	▪ (GB+IRL+NZ+SL), (CAN+JA) ▪ EA (/ EA + GH+EA_{wrt}) ▪ NIG+SIN	▪ (GB+IRL+NZ+CAN) ▪ (EA_{SPK}+WRT) ▪ (IND+PHI) (+JA)	▪ (GB+IRL+NZ+CAN) ▪ EA, IND, (EA_{SPK}+WRT)
	2	▪ GB+IRL+NZ+CAN ▪ EA, NIG, IND ▪ (PHI+SIN) (+SL)	▪ (GB+IRL) (+) CAN+PHI ▪ EA, NIG, NZ	▪ EA, IND (+) PHI, NZ	▪ (GB+IRL+CAN) ▪ EA, NIG ▪ PHI+SL	▪ GB+IRL + HK+SIN ▪ (EA, IND, NZ)
	3	▪ GB+IRL+NZ+CAN, all others ▪ EA+IND, NIG	▪ (GB+IRL+NZ+CAN), ▪ EA+IND, NIG ▪ HK+SIN	▪ (GB+IRL+NZ+CAN +SL), CAN+JA ▪ EA, HK, (IND), NIG+SIN ▪ IND+PHI	▪ GB+IRL+NZ+CAN, all others ▪ EA, NIG, (IND) ▪ HK+SIN, PHI+SL	▪ GB+IRL(+NZ+CAN) ▪ EA, IND ▪ HK+SIN
	4	▪ GB+IRL+NZ+CAN, all others ▪ EA+IND, NIG ▪ HK+JA+PHI+SIN+SL	▪ (GB+IRL+CAN)+JA ▪ EA+IND ▪ HK+SIN, PHI+SL	▪ (GB+IRL+NZ) / (GB + JA + IND+SL) ▪ (EA, HK)	▪ GB+IRL+NZ+CAN, (all others) ▪ (EA) ▪ HK+SIN (+PHI+IND+SL)	▪ GB+IRL+NZ, (CAN+JA) ▪ EA, (EA_{SPK}+WRT) ▪ PHI+SL
WRITING	n	▪ (GB+IRL+NZ+CAN+USA) ▪ TZ+UG+KY+GH+NIG	▪ (GB+IRL+NZ), (CAN+USA) ▪ KY+TZ (+UG) + GH+NIG ▪ HK+SIN	▪ (GB+IRL+NZ + JA+NIG) ▪ KY+TZ+GH ▪ HK+SIN, PHI+SL	▪ GB+IRL+NZ+CAN+HK/SIN ▪ KY+TZ	▪ GB+IRL+NZ, CAN+USA+PHI ▪ KY+TZ+UG ▪ IND
	2	▪ GB+IRL+NZ, CAN+USA ▪ (KY+TZ) + GH+NIG (+UG) ▪ PHI+SIN, (IND+SL)	▪ (CAN+PHI), NZ+JA ▪ GH+SL	▪ (GB+IRL+NZ), (CAN+USA) ▪ KY+TZ+GH+NIG + JA ▪ SL+PHI/IND	▪ GB+IRL+NZ, CAN+USA ▪ KY+TZ + UG + GH+NIG ▪ PHI+SIN, IND+SL	▪ GB+IRL+NZ, CAN+USA+PHI ▪ (NIG, IND) ▪ KY+TZ+UG
	3	▪ GB+IRL+NZ, CAN+USA ▪ (KY+TZ) + GH+NIG (+UG) ▪ IND+SL+JA, (HK+) PHI+SIN	▪ GB+IRL+NZ, (CAN+USA) ▪ KY+TZ, GH+NIG (+UG) ▪ IND+SL+JA (+KY+TZ+GH+NIG+UG) ▪ PHI+SIN	▪ (GB+IRL+NZ+ KY+TZ+GH+NIG+UG), (CAN+USA+HK+SIN+IND +PHI+SL) ▪ KY+TZ+GH ▪ PHI+SL, HK+SIN	▪ (GB+IRL+NZ), CAN+USA ▪ KY+TZ + GH+NIG + UG ▪ PHI+SIN+HK ▪ (IND+SL)	▪ GB+IRL+NZ, CAN+USA ▪ KY+TZ(+UG) ▪ (JA+SL), PHI+SIN
	4	▪ GB+IRL+NZ, CAN+USA ▪ IND ▪ KY+TZ, GH+NIG ▪ HK+SIN	▪ GB+IRL+NZ, (CAN+USA) ▪ KY+TZ, (GH+NIG) ▪ (JA+SL)	▪ GB+IRL+NZ+JA+ KY+TZ+GH+NIG, CAN+USA+HK+SIN+PHI +SL+UG ▪ IND	▪ (GB+IRL+NZ), CAN+USA ▪ KY+TZ, (GH+NIG) ▪ (PHI+SL)	▪ GB+IRL+NZ (+HK+SIN+PHI+IND+SL), CAN+USA ▪ KY+TZ+GH+UG ▪ IND+SL, HK+SIN(+PHI)

6.1.1 Binary splits and Outliers

While the individual findings presented within Tables 6.1 and 6.2 make apparent that there is heterogeneity between individual analyses, they also highlight a ubiquity of connecting lines and thus allow for generalizing across the diversity of results. One finding of almost universal applicability is the identification of the Inner Circle (IC) varieties, in speech as well as in writing. In the latter case, the singular group more commonly subdivides by region or aspects of epicentricity, i.e. GB+IRL+NZ vs. USA+CAN. These two written IC branches, labelled IC_{GB} and IC_{NA} hereafter, can sometimes even be found at a moderate distance from each other. Even then, however, the IC groups are typically found to be more mutually similar than either cluster is found to remaining Outer Circle varieties. As such, the IC varieties are found in isolation from other varieties, but occasional mergers with further, usually individual varieties can be observed. If they occur, these most commonly associate PHI with $IC_{NA,}$ or JA with various IC Englishes. Either variety is more frequently attested in other constellations across the study, but PHI only switches between either the IC or an Asian context, while JA varies more freely. The colonial legacy of PHI with USA may be informative for the former case, while the variance observed within the latter could conversely be a result of JA representing the only Caribbean variety in the ICE data. Regional categorization will only be explored in the following section, so further evaluation will be postponed until then.

Given the binary separation of the IC group in writing, it seems reasonable to assume that such distinctions could also surface in the spoken data if it contained material for USA. This seems plausible in light of CAN_{SPK} exhibiting a degree of separateness from the remaining spoken IC group (lexical G^2 3-grams; POS G^2 n- & 3-grams, POS t 2-grams, ΔP 4-grams) which is reminiscent of the IC_{GB} vs. IC_{NA} split in writing. Furthermore, the proximity of CAN to PHI (t, ΔP) mirrors the IC_{NA}+PHI merger in writing both in kind as well as frequency.[82] The data for NZ show a similar redcreased readiness to integrate into the spoken IC_{GB} cluster, and GB+IRL are usually the first to form a cluster. For speech, this is particularly prevalent among POS 2-grams, but a

[82] The fact that the merger with PHI is encountered more frequently in the POS data may furthermore be taken to indicate an underlying structural similarity usually glossed over in more topic-dependent lexical sequences.

minor tendency towards separating NZ and clustering it with OC varieties is also observable in writing. While divergence between the IC_{NA} and IC_{GB} branches thus finds some moderate support even in speech, separation of the NZ data is much more ambiguous but would have been fascinating to explore with the aid of comparable data for neighboring Australia. After all, what is hinted at in the present data might well be a reflection of regional similarities within the Austro-Oceanic region and/or shared Australian English epicentral influences.

In stark contrast to the Inner Circle, a coherent Outer Circle group is only rarely encountered and thus only infrequently noted explicitly in Tables 6.1 and 6.2 (as 'all others' to conserve space over naming all varieties individually). If at all, such a group is only detected in speech but not in writing. It thus appears to depend at least in part on the lower overall number of spoken components, in that fewer data points more easily combine into a group than in the case of more diverse data. Writing produces no results compatible with the coarse OC label, and instead appears to prefer a finer segmentation for the data like the separation of the IC group observed above. The establishment of a coherent spoken OC cluster appears most challenged by the frequent exclusion of EA_{SPK}, which is almost unanimously found at great distances to the remaining data, even to the extent of clustering with written varieties in a few of the POS analyses.[83] The case of EA_{SPK} thus presents a curious outlier, which further analyses delving into the more detailed structures of the dataset will further explore. At the present moment, it should be noted that the strong separation of the component is only observable in speech, while it typically clusters with NIG (itself sometimes identified as a weaker type of 'outlier') in writing. A similar constraint applies in case of IND, which is also found at some elevated distances to the remaining varieties, potentially clustering with EA. What unites these three cases is that their spoken clusters diverge strongly from those obtained in writing. This is in sharp contrast to virtually all other varieties, which emerge in mostly similar or even identical clusters, which gives some reason to doubt the cluster allocations of these outlier components. The clusteredness of EA and IND (and occasionally NIG) should, therefore, not be regarded as a clear indication of their similarity. Indeed, it is likely that they cluster more through

[83] This is, of course, only observable in the ALL data, and thus one out of three datasets within each analysis. As such, it can never become a major finding, but is still included in Tables 6.1 and 6.2 as a minor result.

their individual large distances to other Englishes than on the grounds of actual mutual similarity. That mutual similarity of these varieties is low can also be observed from the fact that they only enter into binary combinations when other varieties have long coalesced into larger structures. As such, they only merge because they are found dissimilar to both the IC and the (remaining) OC varieties.

Surveying the initial major splits in the data has revealed a strong separation of the IC varieties from all remaining data. Yet, a truly binary division of the dataset is not supported given that the remaining varieties do not systematically coalesce into a joint cluster, defeating a clear Inner vs. Outer Circle perspective. Moreover, the written data often shows a regional separation within the Inner Circle, which also finds reflection in tendencies observable in the spoken data. As such, a regional interpretation emerges as the most sensible second analytical framework and will be explored below.

6.1.2 Regions and Regional Imbalances

The internal differentiation within the Inner Circle observed above shows that in addition to a substantial degree of homogeneity between IC varieties, there is also an element of divergence. This becomes most apparent in case of writing and the additional number of varieties contained in this dataset. The structure of the internal differentiation of the IC group is consistent with a regional and/or epicentral perspective. The Outer Circle, on the other hand, only rarely forms a substantiated or group clearly differentiates itself from the IC varieties. As observed above, this suggests that the substructures observable within the Outer Circle are an even more relevant level of analysis than witnessed for the IC group, while 'outlier' varieties may provide a confounding factor particularly in the smaller spoken dataset. Thus, a more fine-grained interpretation of common subgroups appears necessary, which usually results in a regional perspective on the data.

Beyond the regions observed within the IC group, the OC varieties covered in the ICE data stem from two major regions, Africa and Asia, with a single Carribbean variety represented by JA. However, the degree to which these broader regions are represented varies considerably across speech and writing: Writing equally represents Africa and Asia with five varieties each, of which KY and TZ stem from the same ICE-EA

component. Speech, however, only accounts for two of the African varieties (EA and NIG), and thus equal representation in writing gives way to an imbalance in speech which only affects the African data.

Africa

For the African varieities, the differences in data availability and structure noted above lead to drastically diverging results between modes. As far as writing is concerned, clusters typically identify the African region relatively well on the broadest level and establish it as distinct from other groups of varieties. On a finer level, separation into two related groups of African varieties can be observed, with an internal structure mostly consistent with an East-West African separation. There is a caveat to this in that UG is often incorrectly placed with the West instead of East African varieties; much less frequently, it is even entirely removed from the African varieties. Surprising as this finding may be, it should not be taken as greater evidence against regional segmentation within the African data: Firstly, it needs to be recalled that the two East African written varieties both stem from the ICE-EA data. This component has been shown to deviate strongly from other ICE components in terms of its data sampling and corpus structure, but particularly with regards to its markup (cf. Chapter 4 for details). It is likely that this increases similarities between the KY and TZ data beyond the purely linguistic level while simultaneously removing these varieties from more regularly compiled and annotated components. This, in turn, would lead to UG adopting a more intermediate position between linguistically close East African varieties and similarly annotated West African data. A middle ground between two poles can quickly lead to variant clusterings in methods which enforce a clear binary segmentation at each step. As such, the NeighborNet analysis may be particularly informative here. Indeed, this method reveals that other methods exacerbate the separation of the data. Instead, the NeighborNets often retrieve UG as a clearly African variety placed in an intermediate position between East and West.

In contrast to the relatively clear-cut results for the African written varieties, spoken results pose larger challenges for meaningful interpretation. With only two relevant varieties available in the data (EA and NIG), generalizations on the broader region would be on somewhat shaky ground even if these clustered consistently. Unfortunately, one of these two is EA_{SPK}, a component which has repeatedly been found to

be the most problematic of all within the present study: In almost all analyses, EA_{SPK} is isolated from the other spoken varieties, even to the extent of being merged with the written data (in the hierarchical analysis of the ALL dataset). In turn, clusters of only EA and NIG are virtually absent from the data. Therefore, while it appears that no larger degree of similarity between the African spoken varieties can be established on the basis of the present data, there are issues with the location of either variety that allow for a potential alternative explanation: Certainly, the EA component can be seen to behave strangely even in the written data, where an influenc of corpus structure and annotation can be felt. This effect will be exacerbated in speech, where in addition to all other divergences, the missing notation of speech units in ICE-EA (cf. Chapter 4) is in very strong deviance from other components and most certainly leads to differences from other data beyond the purely linguistic. It is not unreasonable to assume that a lack of speech unit information can have severe consequences, and it appears that this may even have come to overrule any similarities that may otherwise exist between the EA data and NIG (and are documented in writing). Indeed, there is evidence for divergence of NIG from the remaining data: The variety is often found outside of other substantiated clusters, in heterogenous contexts or at elevated distances to the remaining data – if never to the extent observed for EA.[84] Thus, while there is no evidence for similarity between spoken EA and NIG in the present data, there is also no evidence for systematic patterns of NIG with other varieties. It seems plausible to assume that EA and NIG would indeed form clusters if the corpus structure of EA did not interfere as strongly as apparently is the case.

Asia

The Asian data is simultaneously more and less challenging to interpret than its African counterpart. With all varieties equally represented in the spoken and written data, results for the modes can be clearly contrasted, unlike the issues encountered with the African data. However, this only leads to a dual establishment of the Asian data as a comparatively heterogenous group. In contrast to the IC and African groups identified above, the Asian varieties lack substantiated identification as a joint regional group clearly distinct from the African or IC centers. Instead, the Asian group is found in a state of flux: A few relatively frequent subgroups can be observed, and there definitely

[84] Please consult individual analyses for details.

is some recurrent mutual proximity between individual Asian varieties or pairs thereof. Most of these combinations are, however, found to be too fluid within or across methods, failing to meet the required bootstrap probabilities in the former case or the criterion of inter-method stability in the latter. The overall group thus appears fragile and mutable: Variation within the group appears to be of only little systematicity, while individual varieties are often found to detach from the larger group and merge with non-Asian varieties. A semblance of a combined Asian group consequently appears to more strongly depend on the clear identification achieved for other regional centers than on pronounced mutual similarity among the Asian varieties.

Given the overall mutability of the Asian group discussed above, regional subclusters usually do not reach the levels of substantiation found for either the African or IC subgroups. Still, two regional subgroups can be established with relative confidence in all or parts of the data. Most prominently, this concerns a HK+SIN subcluster, which can frequently be substantiated in either mode and type of base data, even if it is often found as part of a larger structure incorporating another Asian variety. These latter varieties can stem from a diverse pool of Englishes, but there appears to be a slight (and usually unsubstantiated) preference for PHI. If the latter variety is not found as part of this cluster, it tends to associate, in decreasing frequency, with SL or IC_{NA} (see above). HK+SIN is, however, itself infrequently allocated to the IC group, and particularly to IC_{NA}, but this is not supported strongly by the data.[85] A second regional IND+SL group only manifests in writing (most frequently visually identified in the NeighborNets), but is not highly consistent even there. In the best of cases, these subgroups result in a partition of the Asian data which reflects a regional South vs. South-East separation. More typically, however, some version of deviance from this pattern is observed. In particular, PHI often merges with SL (instead of HK+SIN), which in itself shows a tendency to combine with the Caribbean JA, while IND occasionally merges with (East) African varieties.

In speech, only the HK+SIN cluster remains, while the IND+SL cluster fails to achieve the moderate support it finds in writing. It appears that the cluster is more directly disrupted by IND, which separates from the remaining Asian data and merges

[85] A converse allocation of African clusters to IC_{GB} is also observable but remains even rarer.

with EA_{SPK}. The results of this process are largely identical to those observed in writing in case the IND+SL cluster fails to materialize: SL_{SPK} enters into combinations with various other varieties, of which the association with PHI shows the greatest systematicity (followed by the association with JA). The spoken data thus sees the Asian group disintegrating into heterogenous and unsubstantiated clusters even more easily than the written data, but it appears that this is triggered by the isolation of IND more than any other processes within the data.

From the perspective of the Asian data, the tendency of IND to combine with the EA data is slightly more curious than the other way around: For EA, issues with corpus structure and content become evident already within the corpus manual. IND, on the other hand, appears much more like a regular ICE component, even if spoken samples seem to reflect some slightly formal interview-style interactions between university staff and students with a somewhat limited pool of frequent phrases (e.g. introductions with "May I know your name and address", "May I know your good name"). It seems probable that these types of interactions predominate in earlier components (EA and IND being the first within their respective regions), given the low availability of data in the late 1990s. This might, in turn, explain some degree of similarity between the data. In any case, this appears no less a suitable explanation for the similarity of the data than the numerically minor Indian dispora to East Africa. Most likely, however, it is an effect of the clustering methods themselves: In light of the limited homogeneity within the Asian data but generally great difference of the IC varieties to any OC English, even minor variation can lead to IND being placed with the African group. The NeighborNet method, which is more suited to the multidimensional nature of such specific problems, usually shows IND in relative proximity to SL in the written mode but simultaneously retrieves relatively large distances of IND to the remaining Asian data. Distance from an otherwise inconsistent Asian group may, in turn, be enough for other clustering methods to allocate IND to some other cluster. Since the IC varieties are almost universally the first to separate from the other data, clustering to the African data is the more likely option (even if rare cases of IC+IND clusters do exist). Inherent variability and instability is in turn consistent with the observation that any of these combinations are only weakly substantiated, in stark contrast to the almost universal stability found within other larger clusters.

The case of Jamaica

A final confounding case frequently presents itself in the shape of the Jamaican data, which changes cluster allocations apparently freely, occurring with varieties from different regions and evolutionary phases. Somewhat elevated frequencies can be observed for its combinations with varieties as diverse as the IC Englishes (or branches thereof), SL and NIG (the latter particularly in the NeighborNet analyses). Providing reasons for this patterning is challenging given the lack of regionally compatible Caribbean varieties within the present data. It could be expected that North American epicentral influences find JA more closely allocated to IC varieties than others, and indeed this pattern is the slightly most numerous one over JA+SL and the least frequent JA+NIG group. Yet, within the regional analysis carried out above, JA remains the odd one out from the theoretical viewpoint, and the data seems to reflect this.

6.1.3 Evolutionary Perspectives

Readers may notice that a regional interpretation appears preferred over one on the grounds of the Dynamic Model. Individual analyses have more frequently identified clusters in terms around regional groups than any parallel structures derived from stages of increasing nativization and endonormativity, and as such it was the first perspective adopted in this final evaluation. This may have arisen to some extent from more intuitively available labels and clearer dividing lines between regional categories. Phase assessments, in contrast, are of a less precise nature, given that clusters need not establish themselves neatly along the major phase boundaries of the Dynamic Model (phases 2, 3, 4, 5). Instead, many other thresholds might conceivably allow for sensible segmentation of the data, such as lumping some phases together (e.g. phases 2-3) or partitioning within others (e.g. 'late phase 3'). Worse still, the numbers of sensible divisions could even diverge across analyses, impeding on the definition of consistent interpretative categories. Theoretical issues aside, however, it is also true that categrization of the cluster findings along any cline of the Dynamic Model only rarely presented itself as a viable alternative: While it is true that frequently some parts of the data were accessible to an evolutionary interpretation, this perspective could almost never account for all varieties within any individual study, and findings were inconsistent across analyses.

Given the above caveats, some of the previously discussed regional clusters also exhibit shared evolutionary stages. In particular, the IC data consistently represent phase 5 varieties, while the written African varieties partition somewhat neatly into groups of varieties more (GH, NIG) and less (KY, TZ) advanced along the varietal cline (late phase 3 and early-mid phase phase 3, respectively). From this perspective, UG can be regarded as assuming an intermediate position between the more and less advanced African varieties. An evolutionary perspective could also provide an explanation for the fragmentation of the Asian data, since the respective varieties cover a wide span of the evolutionary cycle from early phase 3 (HK, PHI) to late phase 4 (SIN).

Unfortunately, such an explanation produces more issues than it helps to address: In the case of the Asian data, the outer bounds of the wide varietal cline described before (early phase 3 to late phase 4) are marked by exactly those Asian varieties most commonly found in close association (HK, SIN, and sometimes PHI), which results in a group of varieties of widely diverging phases. At the same time, those Asian varieties more similar with respect to their phases (IND, SL) are often assigned into separate clusters. For the Asian context, explanation along the Dynamic Model thus breaks down entirely. While the African data, by contrast, mirrors the evolutionary cline, it needs to be noted that the respetive threshold appears arbitrary, lumping together a variety likely to be stuck in very early phase 3 (TZ) with one more firmly within nativization (KY). The second mid-phase 3 variety (UG) is in turn more commonly assigned to varieties on their way towards endocentricity (GH, NIG). Finally, even the seeming conspicuousness of the phase 5 group is curtailed by its frequent segmentation into two distinct clusters in writing and indications towards related processes in speech.

A reversal of the above analytical process, i.e. scrutinizing whether varieties of similar stages achieve compatible cluster allocations, presents itself as just as unfruitful: Groups of phase 4 varieties – SIN, JA and (less reliably) IND – are almost entirely absent from the data. The same holds true for nativizing varieties (phase 3), which cannot be observed to form consistent early- (TZ, HK, PHI; EA_{SPK}) or mid- (SL, UG, KY) stage 3 clusters. Only a late-phase 3 GH+NIG group is well defined, but IND as another variety between phases 3 and 4 is neither consistently clustered with this group nor any phase 4 varieties. Similarly, frequently observed smaller groups often conflict with any sensible phase boundaries within the data. As such, the frequent combinations of

IND or NIG (less frequently) with EA within the spoken data each describe a merger of varieties more advanced than EA. This occurs at the expense of even a single spoken EA+HK (or maybe PHI) cluster. Moreover, PHI is found close to USA/CAN, merging varieties of widely diverging phases.

Given the above evaluation of patterns within the data against phase estimates and corresponding groups of similarly-advanced varieties, it appears only at first glance that the Dynamic Model seems to hold. As soon as a more fine-grained perspective is adopted and the analysis proceeds beyond the most general separations in the data, clusters neither conform to predictions of the basis of the model, nor do any phase estimates reflect consistently in the language data as obtained in the present analysis.

6.2 Factors of Cluster Variation

The previous section presented findings generalized across several parameters within the present study (base data, measures, lengths) and at most covered differences by mode. This was only possible because similar groups of varieties emerged with relative consistency across all variables. Differences to the general pattern were rather found to be gradual or only concerned specific combinations of values for the variables under scrutiny, but no systematic deviation from the general pattern could be established for any individual variable. Of course, this does not preclude some finer aspects of variation across these factors. The following sections will aim at a discussion, in turn, of the effects of all variables underlying the general findings above.

6.2.1 Cluster Variation across Types of Base Data

Discussion of the clusters obtained from the two separate types of base texts, i.e. the lexical and POS-annotated data, has largely proceeded in tandem, and no major differences have been addressed. Given the greater diversity of actual lexical choices over grammatical categories and the possible reflection of topics within lexical sequences, overall similarity of the datasets appears as a strange finding. However, it is not the case that no differences were observed between the two types of data, only that these are gradual and quantitative rather than categorical and qualitative. That is, highly similar cluster structures were produced within both types of base data, but frequencies and degrees of substantiation may be different.

While the typical clusters as captured in Tables 6.1 and 6.2 are largely similar, there is a (slight) tendency towards minor (i.e. less inter-method substantiated) clusters in the POS data, at the expense of major groups. This results from the fact that several of the POS-based analyses faced challenges to a meaningful interpretation of the clustering results due to great degrees of variation between the individual clustering methods. Confoundingly, however, the grammatical data are usually evaluated more favorably within the `pvclust`-based stability assessments, which retrieve larger numbers of substantiated groups than within the lexical data (note that this only ever applies to the hierarchical analysis). However, this may be expected given that significant results are usually more easily obtained in larger versions of similar datasets, even if the same fundamental processes are at work in both. While individual methods thus retrieve high degrees of substantiation, the concurrent variability across methods rather suggests that the choice of method has the strongest effect on the results. Thus, linguistic differences, even though substantiated within a particular clustering approach, are less clearly retrieved in the overall analysis. As such, the POS data should be understood to indicate more overall homogeneity among varieties than its lexical counterpart. After all, this should also not be a surprising finding given that the common grammatical structure is what makes all varieties Englishes, while indigenous lexical choices prevail. The two datasets thus present themselves overall as largely complementary perspectives on the same linguistic reality, even if this is expressed in very different ways.

In terms of concrete clustering results, increased homogeneity within the POS data leads to some loss of the IC-OC distinction clearly observable in the lexical data. This becomes visible through more bracketed clusters of this type in Tables 6.1 and 6.2, allocations of individual OC varieties to IC clusters, as well as a more frequent isolation of single IC varieties (cf. in particular the results for spoken *t* and G^2). The lessening of ties within the IC group leads to a more regular merger of PHI with IC_{NA}, and similarly SIN is also more frequently associated with (written) IC. None of these combinations are strongly systematic, however, and other varieties are variously found in similar positions. In particular, written G^2 sequences often identify greater similarities of the IC_{GB} group to African varieties but conversely of IC_{NA} to Asian ones. While further actual linguistic processes may lie underneath these individual occurrences, generalization of any larger trends does not appear warranted. Instead, within the present

data, these appear to be more likely caused by lessened overall distances between varieties within the POS data.

Beyond the larger patterns discussed above, some even less systematic findings concerning individual clusters leave room for some speculation on larger processes at work. In particular, the HK+SIN cluster appears to emerge more regularly in the lexical than POS spoken data but, which may be understood as an effect of the shared Chinese substate manifesting itself through concrete lexical choices (e.g. discourse markers). A similar tendency can be detected for the African varieties (only observable in writing due to data availability) These are marginally more frequently found in coherent groups in the lexical data while the POS data shows more regional separation into East and West. The most striking difference is, however, encountered in case of the 'outlier' variety of EA_{SPK}. While this component is identified as relatively separate from all other varieties in almost all analyses, it is with the change to POS that its unique situation is fully highlighted. The lexical analyses usually report EA as a curious element of the spoken corpus parts, accentuating great dissimilarity to all other varieties. However, on the level of POS, EA_{SPK} is repeatedly found to form clusters with (parts of) the written data, if in highly diverse configurations. This effect is absent from the lexical data except for a single case (k-means clustering of *MI* 4-grams). The POS data, however, reports it in a multitude of cases, such as in hierarchical clustering of G^2, *g* and ΔP n-grams, G^2 and ΔP 3-grams and ΔP 4-grams. Furthermore, EA_{SPK} is frequently isolated within k-means clustering and also separated early during the identification of significant jump heights. While it is true that this could be taken as an index of actual linguistic similariy to the written norm, it is curious that no other spoken varieties, even similarly exonormative ones, display similar patterns (except for a single case of PHI_{SPK} merging with USA_{WRT} in lexical spoken G^2 4-grams). As such, this occurrence rather gives further credit to the argument of corpus discrepancies and the status of EA_{SPK} as an outlier in the present study: As laid out above, the ICE-EA spoken data diverges strongly from the usual markup. In particular, spoken texts often lack speech unit markup, which is likely to impact association scores by introducing additional cotext and in turn render the data more dissimilar to other spoken varieties. The mergers, unsystematic as they are, also cannot be found sensible from a linguistic perspective: If the EA_{SPK} actually were linguistically close to the written norm, combinations with

written KY or TZ should be the default result of the spoken form crossing modes. That this is not a systematic finding should be understood as indication against linguistic similarities and for a major influence of corpus effects.

6.2.2 Clusters Variation across Speech and Writing

Some features of the spoken-written distinction have already been addressed in previous sections. These concerned differences within the regional clusters obtained in speech and writing as well as the odd allocation of EA_{SPK} (and, once, PHI_{SPK}) to written varieties in the POS data. In both cases, however, effects of the data at hand proved to be more informative than actual linguistic differences. Thus, the spoken and written modes actually returned mostly compatible results, as has been discussed above. The present section instead rather focuses on the overall distinction of the spoken and written modes, and furthermore addresses gradual differences in the clarity by which they are identified. As such, the present section is largely confined to those parts of the analysis which scrutinize the combined spoken-plus-written datasets (ALL) instead of the separate modal data. Please recall that these always built on the intersect of the two modes. While the individual spoken or written datasets can indeed contain different sequence types, the ALL datasets cannot, and instead differentiate modes solely on grounds of diverging association of an otherwise identical set of sequences. Even given this approach, however, it should not come as a surprise that speech and writing separate clearly in all but a few exceptional cases: Despite resting on the same types of sequences, different cotexts of sequence constituents in speech and writing are likely to result in divergent association values.[86]

In the vast majority of cases, a clear-cut binary separation of speech and writing is obtained despite the identical set of types underlying the data. This can be seen in cluster stability assessments of each major branch or in k-means repeatedly favoring only a binary segmentation of ALL. Yet, the two groups are not equally substantiated: Primarily, stability assessments on the basis of `pvclust` within the spoken branch are usually less favorable than for writing, indicating more variability within the branch. It

[86] It may, instead, rather be surprising that it might only take a lack of speech unit markup in EA for the spoken-written division to hold much less clearly. However, please recall that this only occurs in the POS dataset, which already lessens overall variability within the data, and thus makes such an effect more easily obtainable.

seems likely that this is a result of the 'outlier' varieties, which are only truly found in the spoken data (EA, IND). Their separation from the other OC varieties causes the fragmentation of groups substantiated in writing (IND+SL or the African cluster), which may, in turn, leave other varieties stranded (cf. the case of NIG in Section 6.1.2). Ambiguous situations quickly result in varying cluster allocations during resampling, and thus lead to lower degrees of confidence. In contrast, the IC varieties emerge as even more clearly clustered. This can even extend to the point that a tripartite separation (IC_{SPK}, OC_{SPK}, writing) is preferred over the binary spoken-written distinction by Greenacre's test or k-means variances. This situation leads to fewer individual spoken than written clusters being supported, and can even result in the entire spoken branch failing to be substantiated. While writing is not always found to be stable either, it is substantiation for the spoken branch which usually ceases to occur first (cf. lexical *MI* 4-grams, all G^2 sequences, ΔP 3- and 4-grams, and POS *MI* n-grams, all G^2 sequences except 4-grams, *g* n-grams, and all ΔP except 2-grams).

Findings less immediately related to the clusters themselves can be gleaned from the distances as well as the speed of separation of the two datasets. This is particularly readily observable in the hierarchical analysis. In all but a few individual outlier cases, the spoken part of the data is found at greater internal distances (i.e. more heterogeneity). Within the ALL data, the spoken branch is almost unanimously found more heterogenous (longer branches) and differences usually increase with sequence lengths.[87] In case EA_{SPK} switches modes, written branch lengths greatly increase, revealing that the variety is not strictly well-placed. Beyond the ALL data, more heterogeneity between spoken varieties is also reflected in greater spans of the respective distance values on the dendrograms. Building on these greater differences in speech, segmentation by largest jumps consistently subdivides the spoken branch of the combined data first. Similarly, k-means clustering repeatedly produces spoken subgroups in ALL and much less frequently written ones.

Overall, previous analyses have maintained more actual linguistic similarity than difference of the spoken and written modes. Divergence has instead been attributed to

[87] But contrast lexical G^2 2-grams and all ΔP sequences, which show only slightly different branch lengths and less clear cases of increasing difference. POS *t* 2- and 4-grams indicate more written heterogeneity, while POS G^2 not does not retrieve a consistent structure.

the effects of individual corpora as well as lower availability of data for speech. Speech and writing usually clearly diverge even given identical sequence types, but heterogeneity is larger in speech, partially due to effects of outliers. The switch from lexical to POS data furthermore brings about a quantitative decrease of the overall distinctness of the spoken and written mode, which triggers the component least representative of the spoken mode (EA) to cross modes on occasion.

6.2.3 Cluster Variation across Measures

In the present study, collocational preferences in World Englishes were calculated on the basis of five distinct association measures. In turn, these were either contrasted directly across static-length sequences of 2, 3 and 4 units in length or otherwise informed the best cut-off points for the definition of dynamic-length *n*-grams. As such, the question remains how the measure-specific results compare to one another. Before continuing, however, please note that the present study can only describe results obtained from the application of the various measures. The purpose was to apply a diverse set of measures to the same varietal data and use all findings to triangulate the most sensible clusters. More precise performance estimates would instead require comparisons against previously defined benchmarks. Still, after the conclusion of the analysis, better- and worse-performing association statistics can be distinguished by contrasting the clusters obtained by each individual measure against the triangulated findings across all measures.

In direct comparison of all measures, G^2 and ΔP present themselves as those simultaneously producing the least systematic clusters within their respective datasets as well as arriving at results least consistent with those of other measures. Thus, they appear to underperform within the present framework. G^2 diverts most strongly from the majority of results, never retrieves a significant spoken/written distinction in the lexical data, and only manages this once in the POS data (at length 4). ΔP fares better but still only manages to produce the spoken/written distinction at length 2 in both types of data as well as for dynamic-length lexical *n*-grams. Even then, it finds speech only slightly more heterogenous than writing. In terms of concrete subclusters, the two measures produce the most heterogenous results across lengths and datasets. This results in some of the largest proportions of minor, i.e. less inter-method substantiated, clusters in case of G^2, while findings for ΔP are only infrequently found to

be stable by `pvclust`. Results for G^2 and ΔP could still be put to use in triangulation with findings of other measures. But evaluated on their own, they appear as the least reliable of the five statistics. In a study resting on any of the two measures alone, misleading results might be the consequence, and in particular G^2 fails to earn a clear recommendation.

The other three measures produce better findings: Results are more consistent within each measure and mesh well with those of others. This is true even for *MI*, which may be astonishing given the limitation of requiring of one of the most restrictive threshold values for dynamic *n*-grams. Still, the measure produces consistent findings across all variables and rarely fails to substantiate the distinction of speech and writing. It does, however, fall short of other measures in terms of the frequency of substantiated subclusters. This is particularly apparent in the lexical data, where its threshold value leads to the exclusion of many sequence types, which impacts reliability within the resampling-based stability assessment of `pvclust`. Another well-performing measure is the *t-score*, which mostly produces sensible clusters in line with the overall analysis. It struggles, however, at the length of (particularly lexical) 4-grams and, like G^2, fails to retrieve the IC clusters in the POS-based data. The final measure, lexical gravity *g*, overall fares best at all lengths except for dynamic-length approaches (particularly in the POS data). In the latter case, the effect of its threshold value for the selection of relevant bigrams is the limiting factor, which leads to an exceptionally large loss of underlying bigram types (c. 99%). This, in turn, results in the generation of less consistent clusters which, moreover, frequently fail to reach substantiation. Still, it presents itself as the best performer of all in the fixed-length approach and could be relied upon if only a single measure needed to be selected.

6.2.4 Cluster Variation across Sequence Lengths

The final variable concerns the length of sequences. This variable shows some worse-performing cases but overall is characterized by more homo- than heterogeneity across cases. The greatest differences can be found between *n*-grams produced by the dynamic-length approach as opposed to static-length sequences. The former may have the merit of being particularly well-adapted to any specific (varietal) set of data, but this also results in the formation of a more heterogenous set of sequences and a concomitant reduced overlap of types. Furthermore, they require the definition of

threshold values for the selection of relevant bigrams, so as not to generate sequences starting in mutually repulsing items (cf. Chapter 4). While overall cluster results for dynamic sequences are still coherent with static *n*-grams, limited data availability due to the loss of underlying bigram types usually leads to lower levels of substantiation. *MI* in particular always retrieves only few stable clusters, which can be seen as a direct consequence of the very low numbers of retained items. Still, the effect is not strictly linear, and it is not only fewer base bigrams which lead to less substantiated clusters. This can be witnessed for the G^2 *and* ΔP measures, whose threshold values are the laxest of all, resulting in the largest set of sequences. Apparently, however, these are of a less reliable nature, since these two measures still produce low numbers of stable clusters (G^2 performing somewhat better for POS). Yet, at least within the lexical spoken and POS written data, ΔP arguably produces the most sensible results within this measure, even if substantiation levels are low.

Of the static-length sequences (2-, 3-, 4-grams), the longest sequences frequently retrieve the lowest numbers of typical clusters, are not assessed favorably by `pvclust` or produce nonsensical results such as IC_{GB}+HK and IC_{NA}+SIN (lexical *t*-score) or IC_{GB}+JA+Africa-UG (POS G^2). However, *MI*-based POS 4-gram results may arguably be the best out relatively homogenous findings across all lengths for this measure. Shorter sequences (2 and 3 units) generally fare best, except for G^2's strange and erratic results particularly in speech. Sequences of these lengths most commonly detect either the most or second-most number of typical clusters. 2- and 3-gram results also produce strongly supported clusters, which is an area in which particularly dynamic-length *n*-grams score low. There is a slight tendency towards 3-grams faring better in POS data and 2-grams in the lexical data, coinciding with the fact that average dynamic-length sequences are longer in the grammatical data. On the whole, both 2- and 3-grams produce good results, and generally outperform both the dynamic-length sequences as well as 4-grams in terms of the clusters obtained.

7 Conclusion and Outlook

The present study attempted a strictly data-driven evaluation of models of World Englishes. Methodologically, diverging degrees of association within lexical and grammatical *n*-grams were chosen as the linguistic basis on which to estimate similarities and differences between World Englishes. To this end, sequences of both dynamic and statics lengths were generated from homogenized components of the *International Corpus of English* in both its regular lexical format as well as a POS-annotated version. Analysis of collocational preference was carried out by applying five association measures of both traditional (*MI, t,* G^2) as well as more innovative designs (*g*, ΔP) to the respective datasets. On the basis of these association patterns, groups of varieties exhibiting similar association profiles were established through various clustering techniques. After a consistent application of bottom-up techniques for both sequence extraction/generation as well as statistical analysis, agreement between methods was assessed through triangulation of their findings. The variety clusters thus obtained were in turn contrasted to expectations derived from extra-linguistic assessments informed by major language-externally grounded models, particularly phases within the Dynamic Model and degrees of regional proximity.

Clustering results were found to systematically support a regional and/or cultural perspective onto the underlying data over an interpretation resting on the Dynamic Model. The latter could only rarely explain even parts of the data, and almost never their entirety. Those cases in which it fit on the data at hand, a regional explanation accounted for at least the same but usually even larger shares of the data. Not only was it often found impossible to explain the typical clusters with a sensible segmentation of the evolutionary cline, but some of the more typical clusters were even found to consist of varieties from very different evolutionary stages. Therefore, it appears that related colonial histories, epicentral effects, shared substrate languages and the impact of the media are much more reliable predictors of collocational similarities between varieties than comparable stages of post-colonial identity formation. This effect was found to be more pronounced in the written data, which frequently retrieved an African cluster with a relatively clear internal regional differentiation into East and West

African varieties. The Inner Circle varieties were sometimes found as a coherent group, which rather implies cultural than regional closeness (in the sense of 'the Western world'). However, while they consistently separate from the remaining varieties, there is at least as much confidence in a regional subpattern to their overall group. As such, the larger group commonly diverged into a North American branch and a British branch, to the latter of which NZ was regularly assigned. In contrast to the previous clusters, Asia, however, emerged more through exclusion from the previous clusters than by actual similarity. Internal configurations are diverse, and only one subcluster could be established with greater confidence (HK+SIN), while a second IND+SL group emerges with much lower systematicity. This also indicates limitations of the regional perspective, since HK+SIN appears more appropriately explained by similar substrates than a vague Southeast-Asia region. Similar caveats apply in case of the African data, in which UG does not cluster in a strongly consistent way with the EA varieties, but it was noted that anomalies within the very old EA component may also lead to increased distance of the East African varieties to the combined African group. Fitting the regional perspective, the single Carribbean variety JA was found stranded throughout the analysis, clustering relatively freely with diverse other points of data. Additionally, the occasional mergers of PHI with IC_{NA} also indicate less pronounced patterns of regional similarity.

Effects of data availability and data annotation have also been put forward as the main reason for the divergence of results observed between the spoken and written modes. At first glance, the spoken data appear to differ strongly from a regional perspective, since the most pronounced result was often found in the separation of an internally relatively homogenous IC cluster from the much more heterogenous group of OC Englishes. As such, it appears that a relatively 'traditional' ENL/ESL distinction (from the perspective of World Englishes research) is upheld. However, upon closer inspection, it was found that the structure of the spoken dataset presents the more probable explanation than any actual strict division between IC and OC. Primarily, the spoken data suffers from the lack of four regional components. This impacts the African data most strongly, which further fragment upon the isolation of the EA corpus on which has been argued to be effects of its missing annotation of speech units. As such, the African cluster observable in speech can never truly be established by the

data, while similar degrees of mutability of the Asian data is observed as in writing. However, finer effects within the dataset were found to reflect patterns of proximity like in the case of writing: The strong deviance of EA from the remaining data was found to trigger isolation of the NIG component as well, which, while not retrieving mutual similarity, still indicates mutual difference from the other data for two regionally proximal varieties. Similarly, the more homogenous IC group was also found to show a stepwise pattern of similarity, with NZ less similar to GB+IRL, and CAN occasionally merging with PHI and as such reflecting similar processes in writing. It was thus reasoned that clusters of spoken varieties would more directly reflect those found in writing if material were available for USA, GH, UG as well as each EA variety.

Further variables underlying the present study were also discussed but provided only gradual differences and overall clearly less systematic findings. The measure of lexical gravity *g* was found to serve the present analytical purposes best as long as static-length sequences were studied. As such, it appears that its inclusion of type frequencies actually provides a valuable source of information for the distinction of habitual patterns in varieties of English. However, in the case of the dynamic approach to sequence lengths, it performed worse than other measures, which was reasoned to be a result of the exceptionally large loss of sequences incurred through its high threshold value. Consistently strange clusters were retrieved by the established log-likelihood measure G^2, which would have mislead the analysis had it been the only statistic. Generally, however, static lengths were found to produce more reliable findings, while dynamic sequences were found to perform worse in the present framework. Their variety-specific generation of *n*-grams results in too diverse varietal datasets, which in turn reduces the number of types in the intersect between components. While their sequences may be more informative qualitatively, they perform worse in a quanitative framework. Of the static-length sequences, 2- and 3-grams most consistently produced the typical cluters as well as a large number of findings. Shorter sequences fared minimally better in the more diverse lexical data, while longer sequences produced better results in the POS data. Thus, type diversity appears to correlate with sequence lengths, so that less diverse data is better studied with longer sequences (which produce more diverse types through their lengths).

The present analysis has demonstrated that *n*-grams can and do differentiate, if on a relatively coarse level, successfully and consistently between World Englishes on both the concretely lexical as well as on the more grammatical level of POS patterns. It may be true that *n*-grams are difficult to apply as a suitable tool for small-scale comparisons of varieties, appearing "too fine-grained and volatile" there (Gries & Mukherjee 2010: 541). But if the global picture is considered, they are very well suited to revealing relevant distributional characteristics of World Englishes. While differences between the lexical and POS data can be observed in some details, they are not as comprehensive as might be expected. In the present analysis, lexical and grammatical sequences produced similar typical clusters in relatively different ways: Lexical *n*-grams more quickly produced typical clusters but struggled with method-internal validation, while clusters on the basis of POS-grams reached significance much more easily but diverged more strongly across methods. As such, compatible results were produced within either dataset, but slightly more overall similarity was discovered on the basis of the grammatical data. Bernaisch & Koch (2016: 118–119) note a similar division between fine-grained and coarse-grained linguistic objects, and it is noteworthy that a similar situation is obtained on the basis of lexical vs. grammatical *n*-grams. However, this also means that the topic dependence of *n*-grams is not as strong as is often conceived (Gries & Mukherjee 2010: 541), since lexical sequences fare about as well as grammatical ones and can indeed be applied beyond genre classifications and for the distinction of varieties on a more general level. A particular case in point is the consistent allocation of one of the oldest Outer Circle components (EA) with some of the most recent (GH, NIG), for which similar topics are virtually unthinkable.

N-grams emerge as an object which allows the consistent evaluation evaluation of degrees of similarity and difference in an entirely data-driven way. In the majority of cases, different measures and methods point towards compatible findings and lend strong support to the data-driven identification of groups. It is, however, also true that the present analysis benefitted from its somewhat data-mining inspired approach, in that arriving at clear results was aided significantly by the availability of triangulation between an occasionally heterogenous range of results. While the Dynamic Model could not be found to apply with any greater consistency, degrees of (supra-)regional proximity accounted well for almost all cases. On the smaller regional level, phases of

the Dynamic Model might be reflected in the data, but this never exceeds the explanatory potential of a competing fine-grained perspective on proximity. There are indications towards a primary separation of the Inner from the Outer Circle varieties. However, this appears rather as a matter of degree than as a clear-cut division, with regional separation also being strongly supported by the written data and foreshadowing for less limited sets of spoken data. Therefore, the present study concludes with a clear assessment of regional similarity as the prime predictor for general differences across all World Englishes covered in the ICE data, and the evaluation of strictly data-driven techniques as a suitable tool for their analysis.

Going Forward

The present study has brought to the fore relations of similarity and difference between World Englishes from a bird's-eye perspective onto lexical and grammatical co-occurrence patterns. Still, several points have remained unaddressed, concerning both the methodological process in hindsight as well as more general implications for the study of World Englishes.

Methodologically, as extensively as empirical *n*-grams were analyzed in the preceding chapters, the present study has also had to leave several issues only vaguely approached, reserving them as a potential subject for further study. This is particularly evident in the introductory paragraphs to each analysis, which only discuss relatively general distributional characteristics of each dataset at hand and present a small selection of top and bottom *n*-grams from which each is constituted. By no means can this be regarded as exhaustive, and more in-depth evaluations of the qualitative nature of the common core thus identified should be evaluated in future studies. The purpose of the present study has been placed rather on a practical application of co-occurrence statistics onto parallel data for the purpose of quantitative contrastive distinction of groups of varieties. It refrained from further qualitative, and thus less immediately comparable analyses. Future studies might want to trace which the formation of the intersect of the varieties, potentially discriminating stages of (qualitative) overlap between varieties or distinguishing 'regions' of the common core in which only a subset of varieties show greatest qualitative similarities. Similarly, a description of the syntactic properties and pragmatic functions of these common-core sequences present a worthwhile subject for further research.

Theoretically, a major implication for the field of World Englishes and its theory formation comes from the limited evidence for evolutionary stages within the preceding analysis. Following previous studies of World Englishes (cf. Chapter 3) provided the reasonable expectation of an explanative significance of the Dynamic Model, which however did not hold. It might be that any such effects are overshadowed by other variables in the present case. Still, the apparent lack of impact of exo-/endonormative forces and institutionalization both on the overall level as well as within the parts of the data raises questions about the generalizability of the Dynamic Model for actual linguistic outcomes (which were not strictly claimed by its original author, cf. also Schneider 2014 and Schneider 2017). As it stands, evidence from five different association measures in two different types of linguistic data consistently points away from the Dynamic Model as a helpful predictor for linguistic outcomes on this, admittedly general, level of description. Recent discussions of the Dynamic Model have criticized its "main focus […] almost exclusively on colonization as the driving force behind English" (Deshors 2018b: 5) instead of more fine-grained processes within the speech communities. This echoes authors such as Pennycook (2010: 684–685), who caution against putting too much faith into the concept of 'national varieties', since "states-centric pluralities [might] reproduce the very linguistics they need to escape in order to deal with globalized linguascapes." As such, recent improvements of the Dynamic Model have attempted to account for variety-internal heterogeneity and for varying extents of linguistic diversity within regionally defined varieties (cf. Buschfeld & Kautzsch 2017, Buschfeld et al. 2018). At the present time, however, these extensions of the Dynamic Model remain more desideratum than reality. Moreover, the presently available data, even in more recent corpus projects such as GlowbE (Davies & Fuchs 2015), are not designed to facilitate either clear distinctions of variety-internal evolutionary differences (i.e. a diachronic approach), much less allow for their comparison across a wide range of different national contexts.

Further issues of the data within the present study concern both the limited availability of spoken components as well as deviations of individual components form the ICE annotation standards. The lower number of spoken components resulted in less substantiated patterns, which would have produced misleading results if the written data were not available. While some of the missing components will become available

in due time (GH, UG), others will not (EA) or not in a manner strictly comparable to other components (USA).[88] Data on further Carribbean varieties would also have beneficial in hindsight, since JA remained a geographical outlier in an analysis mostly returning proximity as the main explanatory parameter. It thus may have introduced more of a confounding factor than actually greatly contribute to the analysis. The effect of limited availability of spoken components was compounded by the outlier status of the EA data, which was shown to consistently deviate from other varieties, but not in any outwardly sensible manner. The form of markup application in EA was found to provide the most probable reason for the strange patterns observed for this component (particularly the missing speech unit information). Furthermore, it was surmised that lower availability of material during the compilation of the earliest Outer Circle corpora could account for some of the heterogeneity, which would also explain why IND formed less stable clusters in speech than writing. The possibility of systematic flaws in the data is certainly a troubling thought, and further evaluation appears mandated. While studies occasionally make note of particular oddities or errors within their corpora, it appears that more systematic studies are warranted of the data we as linguists apply on a daily basis. One such attempt is found in Gries (2010a), who successfully replicates register classifications within the BNC through bigram attraction values. It is hoped that more such critical evaluation of the internal homogeneity of corpus resources, their reliability and systematicity will ensue.

At the close of this study of lexical and grammatical sequences in World Englishes, it is the belief of the author that the findings and insights will facilitate further studies in terms of the most successful choices of data selection, sequence generation and evaluation methods, as well as provide valuable context in both methodological impact as well as quantitative frames of reference for studies to come.

[88] Spoken USA data was only available through the *Santa Barbara Corpus of Spoken American English* (SBCSAE), and was not included due to its divergence in form and size. Its inclusion, does, however, appear more tempting in hindsight given that the reproduction of two separate IC clusters in speech was found to be impeded by the lack of spoken USA data. Still, the introduction of a spoken component of differing design would have compounded the observed loss of shared data. For studies less focused onto the overall bird's-eye perspective, though, inclusion of the SBCSAE data would present intruiging prospects.

References

Aitchison, Jean. 2012. *Words in the Mind*: *An Introduction to the Mental Lexicon.* Malden, MA: Wiley-Blackwell.

Altenberg, B. 1990. Speech as linear composition. In G. Caie, Haastrup, K. Jacobsen, A. L. J. E. Nielsen, J. Sevaldsen, H. Specht & A. Zettersten (eds.), *Proceedings from the Fourth Nordic Conference for English Studies* (1), 133–143. University of Copenhagen: Department of English.

Altenberg, B. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. P. Cowie (ed.), *Phraseology*: *Theory, Analysis, and Applications*, 101–124. Oxford: Clarendon.

Altenberg, Bengt. 1991. Amplifier collocations in spoken English. In Stig Johansson & Anna-Brita Stenström (eds.), *English Computer Corpora*: *Selected Papers and Research Guide*, 127–147. Berlin: Mouton de Gruyter.

Bailey, Richard W. 1991. *Images of English.* Cambridge: Cambridge University Press.

Bartsch, Sabine. 2004. *Structural and Functional Properties of Collocations in English*: *A Corpus Study of Lexical and Pragmatic Constrainsts on Lexical Co-occurrence.* Tübingen: Narr.

Bartsch, Sabine & Stefan Evert. 2014. Towards a Firthian notion of collocation. In Andrea Abel & Lothar Lemnitzer (eds.), *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, 48–61. Mannheim: Institut für Deutsche Sprache Mannheim.

Bernaisch, Tobias. 2012. Attitudes towards Englishes in Sri Lanka. *World Englishes* 31(3). 279–291.

Bernaisch, Tobias. 2015. *The Lexis and Lexicogrammar of Sri Lankan English* (Varieties of English Around the World). Amsterdam/Philadelphia: John Benjamins.

Bernaisch, Tobias, Stefan T. Gries & Joybrato Mukherjee. 2014. The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide* 35(1). 7–31.

Bernaisch, Tobias & Christopher Koch. 2016. Attitudes towards Englishes in India. *World Englishes* 35(1). 118–132.

Bernaisch, Tobias & Claudia Lange. 2012. The typology of focus marking in South Asian Englishes. *Indian Linguistics* 73(1-4). 1–18.

Biber, Douglas. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5(4). 257–269.

Biber, Douglas. 1993. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities* 26. 331–345.

Biber, Douglas. 2004a. Conversation text types: A multi-dimensional analysis. In Gérald Purnelle, Cédrick Fairon & Anne Dister (eds.), *Le poids des mots: Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data*, 15–34. Louvain: Presses universitaires de Louvain.

Biber, Douglas & Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26(3). 263–286.

Biber, Douglas, S. Conrad & Viviana Cortes. 2004b. 'If you look at…': Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3). 371–405.

Biber, Douglas, Stig Johansson, Geoffrey Leech, S. Conrad & E. Finegan (eds.) (1999). *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education.

Biber, Douglas & Randi Reppen (eds.) (2015). *The Cambridge Handbook of English Corpus Linguistics.* Cambridge: Cambridge University Press.

Bolton, Kingsley. 2006a. Varieties of World Englishes. In Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson (eds.), *The Handbook of World Englishes* (Blackwell handbooks in linguistics), 289–312. Malden, MA: Blackwell.

Bolton, Kingsley. 2006b. World Englishes. In Kingsley Bolton & Braj B. Kachru (eds.), *World Englishes*: *Critical Concepts in Linguistics* (1), 186–216. London: Routledge.

Brutt-Griffler, Janina. 2002. *World English: A Study of its Development.* Clevedon: Multilingual Matters.

Buschfeld, Sarah & Alexander Kautzsch. 2017. Towards an integrated approach to postcolonial and non-postcolonial Englishes. *World Englishes* 36(1). 104–126.

Buschfeld, Sarah, Alexander Kautzsch & Edgar W. Schneider. 2018. From colonial dynamism to current transnationalism: A unified view on postcolonial and non-postcolonial Englishes. In Sandra C. Deshors (ed.), *Modeling World Englishes*:

Assessing the Interplay of Emancipation and Globalization of ESL Varieties (Varieties of English around the world (VEAW) G61), 15–44. Amsterdam, Philadelphia: John Benjamins.

Butler, Susan. 1997. Corpus of English in Southeast Asia: Implications for a regional dictionary. In Maria L. S. Bautista (ed.), *English is an Asian Language: The Philippine Context*, 103–124. Manila: Macquarie Library.

Cameron, D. 1995. *Verbal Hygiene.* London: Routledge.

Church, Kenneth W. William A. Gale, Patrick Hanks & Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik (ed.), *Lexical acquisition*: *Exploiting Online Resources to Build a Lexicon*, 115–164. Hillsdale, NJ: Lawrence Erlbaum.

Church, Kenneth W. & Patrick Hanks. 1990. Word association norms, Mutual Information, and lexicography. *Computational Linguistics* 16(1). 22–29.

Collins, Peter. 2012. Singular agreement in there-existentials: An intervarietal corpus-based study. *English World-Wide* 33(1). 53–68.

Cortes, Viviana. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23(4). 397–423.

Crossley, Scott A. & Max Louwerse. 2007. Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics* 12(4). 453–478.

Crystal, David. 1997. *English as a Global Language.* Cambridge: Cambridge University Press.

Daudaravičius, Vidas & Rūta Marcinkevičienė. 2004. Gravity Counts for the boundaries of collocations. *International Journal of Corpus Linguistics* 9(2). 321–348.

Davies, Mark. 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries. https://www.english-corpora.org/glowbe/ (15 December, 2020).

Davies, Mark & Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide* 36(1). 1–28.

de Klerk, V. 1999. Black South African English: where to from where? *World Englishes* 18(2). 311–324.

Deshors, Sandra C. (ed.) (2018a). *Modeling World Englishes*: *Assessing the Interplay of Emancipation and Globalization of ESL Varieties* (Varieties of English around the world (VEAW) G61). Amsterdam, Philadelphia: John Benjamins.

Deshors, Sandra C. 2018b. Modeling World Englishes in the 21st century: A thematic introduction. In Sandra C. Deshors (ed.), *Modeling World Englishes*: *Assessing the Interplay of Emancipation and Globalization of ESL Varieties* (Varieties of English around the world (VEAW) G61), 1–14. Amsterdam, Philadelphia: John Benjamins.

Dunning, Ted E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1). 61–74.

Edwards, Alison & Samantha Laporte. 2015. Outer and expanding circle Englishes: The competing roles of norm orientation and proficiency levels. *English World-Wide* 36(2). 135–169.

Ellis, Nick C. 2003. Constructions, Chunking, and Connectionism: The Emergence of Second Language Structure. In Catherine J. Doughty & Michael H. Long (eds.), *The Handbook of Second Language Acquisition*, 63–103. Malden, MA: Blackwell.

Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 188–221.

Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20(1). 29–62.

Evert, Stefan. 2004. *The Statistics of Word Cooccurrences*: *Word Pairs and Collocations.* Stuttgart: University of Stuttgart Dissertation.

Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics*: *An International Handbook*, 1212–1248. Berlin: Mouton de Gruyter.

Ferraresi, A. & Stefan T. Gries. 2011. *Type and (?) token frequencies in measures of collocational strength*: *Lexical gravity vs. a few classics* (Corpus Linguistics 2011). University of Birmingham, UK.

Filppula, Markku, Juhani Klemola & Devyani Sharma (eds.) (2017). *The Oxford Handbook of World Englishes.* Oxford: Oxford University Press.

Firth, John R. 1957/1968. A Synopsis of Linguistic Theory, 1930-1955. In F. R. Palmer (ed.), *Selected Papers of J. R. Firth*, 168–205. London: Longmans, Green & co. Ltd.

Firth, John R. 1957. *Papers in Linguistics*: *1934-1951.* London: Oxford University Press.

Fitzpatrick, Tess & Andy Barfield. 2009. *Lexical Processing in Second Language Learners*: *Papers and Perspectives in Honour of Paul Meara* (Second Language Acquisition 3). Clevedon: Channel View Publications.

Fletcher, William H. 2003-2010. Phrases in English (PIE). http://phrasesinenglish.org/ (15 December, 2020).

Francis, W. N. & H. Kucera. 1964. Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with digital computers. korpus.uib.no/icame/manuals/brown/index.htm.

Glass, Cordula. 2019. *Collocations, Creativity and Constructions*: *A Usage-based Study of Collocations in Language Attainment* (Multilingualism and language teaching 6). Tübingen: Narr Francke Attempto.

Gomes Matos, F. de. 1998. Learner's pronunciation rights. *Braz-TESOL Newsletter*(September). 14–15.

Görlach, M. 1988. The development of Standard Englishes. In M. Görlach (ed.), *Studies in the History of the English Language*, 9–64. Heidelberg: Carl Winter.

Götz, Sandra & Marco Schilk. 2011. Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and learner English of advanced German learners. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring Second-Language Varieties of English and Learner Englishes*: *Bridging a paradigm gap*. (2nd proofs), 79–100. Amsterdam/Philadelphia: John Benjamins.

Greenacre, Michael. 2011. A simple permutation test for clusteredness. *Barcelona GSE Working Paper Series*(555).

Greenbaum, Sidney. 1988. A proposal for an international computerized corpus of English. *World Englishes* 7(3). 315.

Greenbaum, Sidney (ed.) (1996a). *Comparing English Worldwide*: *The International Corpus of English.* Oxford: Clarendon.

Greenbaum, Sidney. 1996b. Introducing ICE. In Sidney Greenbaum (ed.), *Comparing English Worldwide*: *The International Corpus of English*, 3–12. Oxford: Clarendon.

Gries, Stefan T. 2008a. Phraseology and linguistic theory: a brief survey. In Sylviane Granger & Fanny Meunier (eds.), *Phraseology*: *An Interdisciplinary Perspective*, 3–25. Amsterdam: John Benjamins.

Gries, Stefan T. 2008b. *Statistik für Sprachwissenschaftler.* Göttingen: Vandenhoeck & Ruprecht.

Gries, Stefan T. 2010a. Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. In Michaela Mahlberg, Victorina González-Díaz & Catherine Smith (eds.), *Proceedings of Corpus Linguistics 2009*.

Gries, Stefan T. 2013. 50-something years of work on collocations: What is or should be next … *International Journal of Corpus Linguistics* 18(1). 137–166.

Gries, Stefan T. 2018a. 7.Zur Identifikation von Mehrwortausdrücken: ein Algorithmus, seine Validierung und weiterführende Überlegungen. In Angelika Wöllstein, Peter Gallmann, Mechthild Habermann & Manfred Krifka (eds.), *Grammatiktheorie und Empirie in der germanistischen Linguistik*, 225–240. Berlin, Boston: de Gruyter.

Gries, Stefan T. & Tobias Bernaisch. 2016. Exploring epicentres empirically. *English World-Wide* 37(1). 1–25.

Gries, Stefan T. Tobias Bernaisch & Benedikt Heller. 2018b. A corpus-linguistic account of the history of the genitive alternation in Singapore English. In Sandra C. Deshors (ed.), *Modeling World Englishes*: *Assessing the Interplay of Emancipation and Globalization of ESL Varieties* (Varieties of English around the world (VEAW) G61), 245–279. Amsterdam, Philadelphia: John Benjamins.

Gries, Stefan T. & Philip Durrant. to appear. Analyzing co-occurrence data. In Magali Paquot & Stefan T. Gries (eds.), *Practical Handbook of Corpus Linguistics*. Berlin, New York: Springer Gabler.

Gries, Stefan T. & Nick C. Ellis. 2015. Statistical Measures for Usage-Based Linguistics. *Language Learning* 65(S1). 228–255.

Gries, Stefan T. B. Hampe & D. Schönefeld. 2010b. Converging evidence II: More on the association of verbs and constructions. In John Newman & Sally Rice

(eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*, 59–72. Stanford, CA: CSLI.

Gries, Stefan T. & Martin Hilpert. 2008. The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora* 3(1). 59–81.

Gries, Stefan T. & Joybrato Mukherjee. 2010. Lexical gravity across varieties of English: An ICE-based study of n -grams in Asian Englishes. *International Journal of Corpus Linguistics* 15(4). 520–548.

Gries, Stefan T. John Newman & Cyrus Shaoul. 2011. N-grams and the clustering of registers. *Empirical Language Research* 5(1).

Güldenring, Barbara. 2020. *Emotion Metaphors in New Englishes*: *A Corpus-Based Study of Emotion Concepts in Institutionalized Second-Language Varieties of English.* Marburg: Philipps-Universität.

Gupta, Anthea F. 1997. Colonisation, migration, and functions of English. In Edgar W. Schneider (ed.), *Englishes around the World* (Varieties of English Around the World), 47. Amsterdam: John Benjamins.

Gut, Ulrike. 2011. Studying structural innovations in New English varieties. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring Second-Language Varieties of English and Learner Englishes*: *Bridging a paradigm gap*. (2nd proofs), 101–124. Amsterdam/Philadelphia: John Benjamins.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie & Yorick Wilks. 2006. A Closer Look at Skip-gram Modelling. In, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA).

Halliday, M. A. K. 1991. Corpus studies and probabilistic grammar. In Karin Aijmer & Bengt Altenberg (eds.), *English Corpus Linguistics*: *Studies in Honour of Jan Svartvik*, 30–43. London: Longman.

Halliday, Michael A. K. 2006. Written language, standard language, global language. In Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson (eds.), *The Handbook of World Englishes* (Blackwell handbooks in linguistics), 349–365. Malden, MA: Blackwell.

Heller, Benedikt, Tobias Bernaisch & Stefan T. Gries. 2017. Empirical perspectives on two potential epicenters: The genitive alternation in Asian Englishes. *ICAME Journal* 41(1). 111–144.

Hickey, Raymond (ed.) (2016). *Sociolinguistics in Ireland.* Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Hoey, Michael. 2005. *Lexical Priming*: *A New Theory of Words and Language.* London: Routledge.

Huber, Magnus. 2014. Stylistic and sociolinguistic variation in Schneider's Nativization Phase: T-affrication and relativization in Ghanaian English. In Sarah Buschfeld, Thomas Hoffmann, Magnus Huber & Alexander Kautzsch (eds.), *The Evolution of Englishes* (G49), 86–106. Amsterdam: John Benjamins.

Hudson, R. 1996. *Sociolinguistics,* 2nd edn. Cambridge: Cambridge University Press.

Hudson-Ettle, Diana & Josef Schmied. 1999. *Manual to accompany the East African component of the International Corpus of English (ICE-EA)*: *Background information, coding conventions and lists of source texts.* Department of English, Chemnitz University of Technology.

Hundt, Marianne. 2013. The diversification of English: Old, new and emerging epicentres. In D. Schreier & Marianne Hundt (eds.), *English as a Contact Language*, 182–203. Cambridge: Cambridge University Press.

Hundt, Marianne. 2018. It is time that this *(should) be studied* across a range of Englishes: A global trip around mandative subjunctives. In Sandra C. Deshors (ed.), *Modeling World Englishes*: *Assessing the Interplay of Emancipation and Globalization of ESL Varieties* (Varieties of English around the world (VEAW) G61), 217–244. Amsterdam, Philadelphia: John Benjamins.

Hundt, Marianne, Andrea Sand & Rainer Siemund. 1998. Manual of information to accompany The Freiburg - LOB Corpus of British English. korpus.uib.no/icame/manuals/flob/index.htm (15 December, 2020).

Hundt, Marianne, Andrea Sand & Paul Skandera. 1999. Manual of information to accompany The Freiburg - Brown Corpus of American English. http://clu.uni.no/icame/manuals/frown/index.htm (15 December, 2020).

Hunston, Susan. 2002. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Hunston, Susan. 2014. Pattern grammar in context. In Thomas Herbst, Hans-Jörg Schmid & Susen Faulhaber (eds.), *Constructions Collocations Patterns*, 99–119. Berlin: Walter de Gruyter.

Hunston, Susan & G. Francis. 2000. *Pattern Grammar*: *A Corpus-Driven Approach to the Lexical Grammar of English.* Amsterdam: John Benjamins.

Jain, A. K. & R. C. Dubes. 1988. *Algorithms for Clustering Data.* Englewood Cliffs, N.J: Prentice Hall.

Jelinek, Frederick. 1990. Self-organized language modeling for speech recognition. In Alex Waibel & Kai-Fu Lee (eds.), *Readings in Speech Recognition*, 450–506. San Mateo, CA: Morgan Kaufmann.

Jenkins, Jennifer. 2014. *English as a Lingua Franca in the International University. The Politics of Academic English Language Policy.* London: Routledge.

Jenkins, Jennifer. 2015. *Global Englishes*: *A Resource Book for Students,* 3rd edn. London, New York: Routledge.

Johansson, Stig. 2011. Corpus, lexis, discourse: A tribute to John Sinclair. In Thomas Herbst, Susen Faulhaber & Peter Uhrig (eds.), *The Phraseological View of Language*: *A Tribute to John Sinclair*, 17–26. Berlin: De Gruyter Mouton.

Johansson, Stig, Geoffrey Leech & Helen Goodluck. 1978. Manual of information to accompany The Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers. http://clu.uni.no/icame/manuals/lob/index.htm (15 December, 2020).

Jones, Susan & John M. Sinclair. 1974. English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie* 24(1). 15–61.

Jurafsky, Daniel & James H. Martin. 2009. *Speech and Language Processing*: *An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics.* Upper Saddle River, NJ: Prentice Hall.

Kachru, Braj B. 1991. Liberation linguistics and the Quirk concern. *English Today* 25. 3–13.

Kachru, Braj B. 1992a. Models for non-native Englishes. In Braj B. Kachru (ed.), *The Other Tongue*: *English across Cultures*, 2nd edn. 48–74. Urbana, IL: University of Illinois Press.

Kachru, Braj B. 1992b. Teaching World Englishes. In Braj B. Kachru (ed.), *The Other Tongue*: *English across Cultures*, 2nd edn. 355–366. Urbana, IL: University of Illinois Press.

Kachru, Braj B. 1998. English as an Asian Language. *Links & Letters* 5(8). 89–108.

Kachru, Braj B. Yamuna Kachru & Cecil L. Nelson. 2006. Introduction: The world of World Englishes. In Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson (eds.), *The Handbook of World Englishes* (Blackwell handbooks in linguistics), 1–16. Malden, MA: Blackwell.

Kachru, Braj B. & Larry E. Smith. 1985. Editorial. *World Englishes* 4. 209–212.

Kallen, Jeffrey L. & John M. Kirk. 2008. *ICE-Ireland: A User's Guide*: *Documentation to accompany the Ireland Component of the International Corpus of English (ICE-Ireland).* Belfast: Cló Ollscoil na Banríona.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics.* London: Longman.

King, Robert D. 2006a. First Steps: Wales and Ireland. In Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson (eds.), *The Handbook of World Englishes* (Blackwell handbooks in linguistics), 30–40. Malden, MA: Blackwell.

King, Robert D. 2006b. The Beginnings. In Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson (eds.), *The Handbook of World Englishes* (Blackwell handbooks in linguistics), 19–29. Malden, MA: Blackwell.

Kirk, John & Gerald Nelson. 2018. The International Corpus of English project: A progress report. *World Englishes* 37(4). 697–716.

Kjellmer, Göran. 1991. A mint of phrases. In Karin Aijmer & Bengt Altenberg (eds.), *English Corpus Linguistics*: *Studies in Honour of Jan Svartvik*, 111–127. London: Longman.

Koch, Christopher & Tobias Bernaisch. 2013. Verb complementation in South Asian English(es): The range and frequency of "new" ditransitives. In Gisle Andersen & Kristin Bech (eds.), *English Corpus Linguistics*: *Variation in Time, Space and Genre*. Selected papers from ICAME 32, 69–89. Amsterdam: Rodopi.

Koch, Christopher, Claudia Lange & Sven Leuckert. 2016. "This hair-style called as 'duck tail'": The 'intrusive as'-construction in South Asian varieties of English and Learner Englishes. *International Journal of Learner Corpus Research* 2(2). 151–176.

Kortmann, Bernd & Kerstin Lunkenheimer (eds.) (2013). *The Mouton World Atlas of Variation in English.* Berlin: Mouton de Gruyter.

Kortmann, Bernd & Edgar W. Schneider (eds.) (2008). *A Handbook of Varieties of English.* Berlin: Mouton de Gruyter.

Kreyer, Rolf. 2013. *The Nature of Rules, Regularities and Units in Language*: *A Network Model of the Language System and of Language Use* (Cognitive Linguistics Research (CLR) 51). Berlin: De Gruyter Mouton.

Lange, Claudia. 2012. *The Syntax of Spoken Indian English.* Amsterdam: John Benjamins.

Lehr, Andrea. 1996. *Kollokationen und Maschinenlesbare Korpora*: *Ein Operationales Analysemodell zum Aufbau Lexikalischer Netze.* Tübingen: Max Niemeyer.

Leitner, Gerhard. 1992. English as a pluricentric language. In Michael Clyne (ed.), *Pluricentric Languages*: *Differeng Norms in Different Nations*, 179–237. Berlin: Mouton de Gruyter.

Lenko-Szymanska, Agnieszka. 2012. The role of conventionalized language in the acquisition and use of articles by Polish EFL learners. In Yukio Tono, Yuji Kawaguchi & Makoto Minegishi (eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research* (Tokyo University of Foreign Studies), 83–104. Amsterdam: John Benjamins.

Levshina, Natalia. 2015. *How to do Linguistics with R*: *Data Exploration and Statistical Analysis.* Amsterdam, Philadelphia: John Benjamins.

Lindquist, Hans. 2009. *Corpus Linguistics and the Description of English.* Edinburgh: Edinburgh University Press.

Loureiro-Porto, Lucía. 2017. ICE vs GloWbE: Big data and corpus compilation. *World Englishes* 36(3). 448–470.

Louw, B. 1993. Irony in the text or insincerity in the writer?: The diagnostic potential of semantic prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.), *Text and Technology*: *In Honour of John Sinclair*. Amsterdam: John Benjamins.

Lunkenheimer, Kerstin. 2013. Typological profile: L2 varieties. In Bernd Kortmann & Kerstin Lunkenheimer (eds.), *The Mouton World Atlas of Variation in English*, 845–872. Berlin: Mouton de Gruyter.

Mahboob, Ahmar. 2017. Understanding language variation: Implications of the NNEST lens for TESOL teacher education programs. In Dios Martínez Agudo, Juan de (ed.), *Native and Non-Native Speakers in English Language Teaching*: *Implications and challenges for teacher education*, 13–32. Boston: De Gruyter Mouton.

Mair, Christian. 2007. British English/American English grammar: Convergence in writing–divergence in speech. *Anglia* 125(1). 84–100.

Mair, Christian. 2013. The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34(3). 253–278.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mason, O. 1997. The weight of words: An investigation of lexical gravity. In B. Lewandowska-Temaszcyk & P. J. Melia (eds.), *PALC'97: Practical Applications in Language Corpora*, 361–375. Lodz: Lodz University.

Mason, O. 1999. Parameters of collocation: The word in the centre of gravity. In J. Kirk (ed.), *Corpora Galore*: *Analyses and Techniques in Describing English*, 267–280. Amsterdam: Rodopi.

McArthur, Tom. 1987. The English languages? *English Today* 11. 9–13.

McArthur, Tom. 1998. *The English Languages.* Cambridge: Cambridge University Press.

McEnery, Tony & Costas Gabrielatos. 2006. English Corpus Linguistics. In Bas Aarts & April McMahon (eds.), *The Handbook of English Linguistics*, 33–71. Malden, MA: Blackwell.

McEnery, Tony, Yukio Tono & Richard Xiao. 2006. *Corpus-Based Language Studies: An Advanced Resource Book.* London: Routledge.

Meierkord, Christiane. 2012. From English as a lingua franca to Interactions across Englishes. In Christiane Meierkord (ed.), *Interactions across Englishes*, 12–47. Cambridge: Cambridge University Press.

Mencken, Henry L. [1919] 1921. *Preface to the First Edition. The American Language: An Inquiry into the Development of English in the United States,* 2nd edn. New York: Alfred A. Knopf.

Mesthrie, Rajend. 2006. Contact Linguistics and World Englishes. In Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson (eds.), *The Handbook of World Englishes* (Blackwell handbooks in linguistics), 273–288. Malden, MA: Blackwell.

Mesthrie, Rajend & R. Bhatt. 2008. *World Englishes: The Study of New Linguistic Varieties.* Cambridge: Cambridge University Press.

Michelbacher, Lukas, Stefan Evert & Hinrich Schütze. 2011. Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory* 7(2). 175.

Milroy, Lesley. 2001. Language ideologies and the consequences of standardisation. *Journal of Sociolinguistics* 5(4). 530–555.

Mittmann, Brigitta. 2011. Prefabs in spoken English. In Thomas Herbst, Susen Faulhaber & Peter Uhrig (eds.), *The Phraseological View of Language*: *A Tribute to John Sinclair*, 197–210. Berlin: De Gruyter Mouton.

Moag, R. F. 1992. The life cycle of non-native Englishes: a case study. In Braj B. Kachru (ed.), *The Other Tongue*: *English across Cultures*, 2nd edn. 233–252. Urbana, IL: University of Illinois Press.

Moisl, Hermann. 2015. *Cluster Analysis for Corpus Linguistics* (Quantitative linguistics 66). Berlin, Boston: De Gruyter Mouton.

Mufwene, Salikoko S. 1996. The founder principle in Creole genesis. *Diachronica* 13(1). 83–134.

Mufwene, Salikoko S. 2001. *The Ecology of Language Evolution.* Cambridge: Cambridge University Press.

Mufwene, Salikoko S. 2002. Colonization, globalization, and the future of languages in the twenty-first century. *International Journal on Multicultural Societies* 4(2). 162–193.

Mufwene, Salikoko S. 2005. Language evolution: the population genetics way. In Günther Hauska (ed.), *Gene, Sprachen und ihre Evolution*, 30–52. Regensburg: Universitätsverlag Regensburg.

Mufwene, Salikoko S. 2010. Globalization, Global English, and World English(es): Myths and Facts. In Nikolas Coupland (ed.), *The Handbook of Language and Globalization*, 29–55. Malden, MA: Blackwell.

Mugglestone, L. 2003. *'Talking Proper': The Rise of Accent as Social Symbol.* Oxford: Oxford University Press.

Mukherjee, Joybrato. 2007. Steady states in the evolution of New Englishes: present-day Indian English as an equilibrium. *Journal of English Linguistics* 35(2). 157–187.

Mukherjee, Joybrato & Stefan T. Gries. 2009. Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide* 30(1). 27–51.

Mukherjee, Joybrato & Marianne Hundt (eds.) (2011). *Exploring Second-Language Varieties of English and Learner Englishes*: *Bridging a paradigm gap*. (2nd proofs). Amsterdam/Philadelphia: John Benjamins.

Murtagh, Fionn & Pierre Legendre. 2014. Ward's Hierarchical Agglomerative Clustering method: Which algorithms implement Ward's criterion? *Journal of Classification* 31(3). 274–295.

Nattinger, J. R. & J. S. DeCarrico. 1992. *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Nelson, Gerald. 2002a. ICE Markup Manual for Spoken Texts. https://www.ice-corpora.uzh.ch/en/manuals.html (15 December, 2020).

Nelson, Gerald. 2002b. ICE Markup Manual for Written Texts. https://www.ice-corpora.uzh.ch/en/manuals.html (15 December, 2020).

Nelson, Gerald. 2015. Response to Davies and Fuchs. *English World-Wide* 36(1). 38–40.

Nesselhauf, Nadja. 2009. Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties. *English World-Wide* 30(1). 1–26.

O'Donnel, Matthew B. 2011. The Adjusted Frequency List: A method to produce clustersensitive frequency lists. *ICAME Journal* 35. 135–169.

Palmer, Harold E. 1933. *Second Interim Report on English Collocations*: *Submitted to the Tenth Annual Conference of English Teachers, under the auspices of the Institute for Research in English Teaching*. Tokyo: Kaitakusha.

Pawley, Andrew & Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (eds.), *Language and Communication*, 191–226. London: Longman.

Pennycook, Alastair. 1998. *English and the Discourses of Colonialism.* London: Routledge.

Pennycook, Alastair. 2010. The future of Englishes. One, more or none? In Andy Kirkpatrick (ed.), *The Routledge Handbook of World Englishes*, 673–687. London [u.a.]: Routledge.

Platt, J. H. Weber & M. Ho. 1984. *The New Englishes.* London: Routledge and Kegan Paul.

Quirk, Randolph. 1962. *The Use of English.* London: Longman.

Quirk, Randolph. 1990. Language varieties and standard language. *English Today* 21. 3–10.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, Jan Svartvik & David Crystal. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Quirk, Randolph, Jan Svartvik, Geoffrey Leech & Sidney Greenbaum. 1972. *A Grammar of Contemporary English.* Harlow: Longman.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*: *Version 3.5.3.* Vienna: R Foundation for Statistical Computing.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*: *Version 3.6.3.* Vienna: R Foundation for Statistical Computing.

Renouf, Antoinette & John M. Sinclair. 1991. Collocational frameworks in English. In Karin Aijmer & Bengt Altenberg (eds.), *English Corpus Linguistics*: *Studies in Honour of Jan Svartvik*, 128–182. London: Longman.

Römer, Ute & Rainer Schulze. 2009. Introduction: Zooming in. In Ute Römer & Rainer Schulze (eds.), *Exploring the Lexis-Grammar Interface* (Studies in Corpus Linguistics 35), 1–11. Amsterdam: John Benjamins.

Sand, Andrea. 2008. Angloversals?: Concord and interrogatives in contact varieties of English. In Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta & Minna Korhonen (eds.), *The Dynamics of Linguistic Variation*: *Corpus Evidence on English Past and Present*, 183–202. Amsterdam, Philadelphia: John Benjamins.

Sarstedt, Marko & Erik Mooi. 2014. Cluster Analysis. In Marko Sarstedt & Erik Mooi (eds.), *A Concise Guide to Market Research* (Springer Texts in Business and Economics), 273–324. Berlin: Springer.

Schilk, Marco. 2006. Collocations in Indian English: A corpus-based sample analysis. *Anglia* 124(2). 276–316.

Schilk, Marco. 2011. *Structural Nativization in Indian English Lexicogrammar.* Amsterdam: John Benjamins.

Schliep, K. P. 2011. Phangorn: Phylogenetic analysis in R. *Bioinformatics (Oxford, England)* 27(4). 592–593.

Schliep, Klaus, Alastair J. Potts, David A. Morrison & Guido W. Grimm. 2017. Intertwining phylogenetic trees and networks. *Methods in Ecology and Evolution* 8(10). 1212–1220.

Schmid, Hans-Jörg. 2000. *English Abstract Nouns as Conceptual Shells*: *From Corpus to Cognition.* Berlin: Mouton de Gruyter.

Schneider, Edgar W. 2004. How to trace structural nativization: Particle verbs in world Englishes. *English World-Wide* 23(2). 227–249.

Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the World.* Cambridge/New York: Cambridge University Press.

Schneider, Edgar W. 2010. Developmental patterns of English: similar or different? In Andy Kirkpatrick (ed.), *The Routledge Handbook of World Englishes*, 372–384. London [u.a.]: Routledge.

Schneider, Edgar W. 2012. Exploring the interface between world Englishes and second language acquisition-and implications for English as a Lingua Franca. *Journal of English as a Lingua Franca* 1(1). 57–91.

Schneider, Edgar W. 2014. New reflections on the evolutionary dynamics of world Englishes. *World Englishes* 33(1). 9–32.

Schneider, Edgar W. 2017. Models of English in the World. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford Handbook of World Englishes*, 35–60. Oxford: Oxford University Press.

Schneider, Ulrike. 2018. ΔP as a measure of collocation strength. *Corpus Linguistics and Linguistic Theory* 0(0).

Sedlatschek, Andreas. 2009. *Contemporary Indian English: Variation and Change* (Varieties of English Around the World 38). Amsterdam [u.a.]: John Benjamins.

Seidlhofer, Barbara. 1999. Double standards: teacher education in the Expanding Circle. *World Englishes* 18(2). 233–245.

Seidlhofer, Barbara. 2001. Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics* 11(2). 133–158.

Seif, George. 2018. The 5 clustering algorithms data scientists need to know. https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 (15 December, 2020).

Sharma, Devyani. 2009. Typological diversity in New Englishes. *English World-Wide* 30(2). 170–195.

Shastri, S. V. C. T. Patilkulkarni & Geeta S. Shastri. 1986. Manual of information to accompany The Kolhapur Corpus of Indian English, for use with digital computers. http://clu.uni.no/icame/manuals/kolhapur/index.htm (15 December, 2020).

Simon, H. F. 1953. Two substantival complexes in Standard Chinese. *Bulletin of the School of Oriental and African Studies* 15(2). 327–355.

Simpson, R. 2004. Stylistic features of academic speech: the role of formulaic expressions. In U. Connor & T.A. Upton (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*, 37–64. Amsterdam: John Benjamins.

Simpson-Vlach, Rita & Nick C. Ellis. 2010. An academic formulas list: New methods in Phraseology research. *Applied Linguistics* 31(4). 487–512.

Sinclair, John M. 1966. Beginning the study of lexis. In Charles E. Bazell, John C. Catford, Michael A. K. Halliday & Robert H. Robins (eds.), *In Memory of J. R. Firth*, 410–430. London: Longman.

Sinclair, John M. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, John M. 2000. Lexical grammar. *Naujoji Metodologija* 24. 191–203.

Sinclair, John M. 2004. *Trust the Text*: *Language, Corpus and Discourse.* London: Routledge.

Sinclair, John M. Susan Jones & Robert Daley. 2004 [1970]. *English Collocation Studies*: *The OSTI Report*. Reprint including a new interview with J. M. Sinclair.

Smith, Stephen P. & Richard Dubes. 1980. Stability of a hierarchical clustering. *Pattern Recognition* 12(3). 177–187.

Strevens, P. 1992. English as an international language: directions in the 1990s. In Braj B. Kachru (ed.), *The Other Tongue*: *English across Cultures*, 2nd edn.. Urbana, IL: University of Illinois Press.

Strevens, Peter. 1980. *Teaching English as an International Language.* London: Pergamon.

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2(2). 23–55.

Stubbs, Michael. 2002. Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7(2). 215–244.

Stubbs, Michael. 2007. An example of frequent English phraseology: distributions, structures and functions. In Roberta Facchinetti (ed.), *Corpus Linguistics 25 Years on*, 89–105. Amsterdam, New York: Rodopi.

Stubbs, Michael & I. Barth. 2003. Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language* 10(1). 61–104.

Suzuki, Ryota & Hidetoshi Shimodaira. 2006. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics (Oxford, England)* 22(12). 1540–1542.

Suzuki, Ryota & Hidetoshi Shimodaira. 2015. *pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version*.

Szmrecsanyi, Benedikt & Bernd Kortmann. 2009. Vernacular universals and angloversals in a typological perspective. In Markku Filppula, Juhani Klemola & Heli Paulasto (eds.), *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*, 33–53. London: Routledge.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work.* Amsterdam: John Benjamins.

Trudgill, Peter. 1986. *Dialects in contact.* Oxford, New York: Blackwell.

Trudgill, Peter. 1999. Standard English: What it isn't. In T. Bex & R. Watts (eds.), *Standard English. The Widening Debate*, 117–128. London: Routledge.

Trudgill, Peter. 2004. *New-Dialect Formation: The Inevitability of Colonial Englishes.* Edinburgh: Edinburgh University Press.

Trudgill, Peter & Jean Hannah. 1982. *International English: A Guide to the Varieties of Standard English.* New York: Edward Arnold.

UCREL. 2007. UCREL CLAWS7 Tagset. http://ucrel.lancs.ac.uk/claws7tags.html (15 December, 2020).

van der Wouden, Ton. 1997. *Negative Contexts*: *Collocation, Polarity and Multiple Negation.* London: Routledge.

van Lancker-Sidtis, Diana & Gail Rallon. 2004. Tracking the incidence of formulaic expressions in everyday speech: Methods for classification and verification. *Language & Communication* 24(3). 207–240.

Wahl, Alexander. 2015. Intonation unit boundaries and the storage of bigrams. *Review of Cognitive Linguistics* 13(1). 191–219.

Wahl, Alexander & Stefan T. Gries. 2018. Multi-word expressions: A novel computational approach to their bottom-up statistical extraction. In Pascual Cantos-Gómez & Moisés Almela-Sánchez (eds.), *Lexical Collocation Analysis*: *Advances and Applications*, 85–109. Berlin, New York: Springer.

Warren, Beatrice. 2005. A Model of idiomaticity. *Nordic Journal of English Studies* 4(1). 35–54.

Werner, Valentin. 2013. Temporal adverbials and the present perfect/past tense alternation. *English World-Wide* 34(2). 202–240.

Widdowson, Henry G. 1994. The ownership of English. *TESOL Quarterly* 28(2). 377–389.

Williams, A. 2007. Non-standard English and education. In D. Britain (ed.), *Language in the British Isles*, 2nd edn. 401–416. Cambridge: Cambridge University Press.

Wood, David. 2010. *Perspectives on Formulaic Language*: *Acquisition and Communication.* London: Continuum.

Wray, Alison. 1998. Protolanguage as a holistic systemfor social interaction. *Language & Communication* 18. 47–67.

Wray, Alison. 2002. *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Wray, Alison & Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20. 1–28.

Xiao, Richard. 2015. Collocation. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 106–124. Cambridge: Cambridge University Press.

Yano, Yasukata. 2001. World Englishes in 2000 and beyond. *World Englishes* 20(2). 119–131.

Zelterman, Daniel. 2015. *Applied Multivariate Statistics with R.* Cham: Springer International Publishing.

Appendix A: Digital Material

Please find the *R* code files and a digital copy (PDF) of the study on the accompanying CD-ROM.

Appendix B: Summary of the Study in German

Die vorliegende Studie unternimmt eine datengesteuerte Analyse von Mehrworteinheiten lexikalischer (*n*-grams) und grammatischer Form (POS *n*-grams, kurz POS-grams; von *part of speech*, DE Wortart) in verschiedenen national definierten Erst- und Zweitsprachenvarietäten des Englischen weltweit. Das zentrale Anliegen der Arbeit ist es, zu ergründen, ob auf Basis des routinehaften Sprachgebrauchs, wie er durch *n*-grams und POS-grams abgebildet wird, systematische Variation zwischen den Varietäten belegbar ist.

N-grams stellen dabei rein empirische Mehrworteinheiten dar, die zumeist anhand ihrer Länge definiert werden: So sind 3-grams beispielweise Sequenzen aus exakt drei Einheiten, hier also Wörter oder Wortartkategorien. Im Rahmen der vorgestellten Methodik wird jedoch auch eine Definition entlang sprachstatistischer Werte vorgenommen, welche in *n*-grams variabler Längen resultiert. Grundlage für die Wahl von *n*-grams ist dabei, dass sie aufgrund ihrer empirischen Natur besonders varietätenneutrale Beschreibungseinheiten darstellen, andererseits aber diverse fortschrittliche Methoden zu ihrer Erfassung und Analyse bereitstellen. Gleichzeitig argumentiert die Arbeit, dass *n*-grams den *Common Core* (Quirk et al. 1985: 16) des Englischen in besonderer Weise abbilden, also jene lexikalische und grammatische Grundlage des weltweiten Gebrauchs des Englischen, die es zulässt, trotz regionaler Heterogenität von einer einzigen Sprache zu sprechen (Englishes). Dies steht im Gegensatz zur sonst häufig vertretenen Auffassung, dass lokale Eigenheiten qualitativ auffällig und ‚markiert' sind. Die vorliegende Studie argumentiert hingegen, dass *n*-grams den gewohnheitsmäßen Sprachgebrauch abbilden, indem sie Frequenzunterschiede zwischen konkurrierenden Kombinationsmustern quantitativ-statistisch erfassen. Methodisch bedient sich die Arbeit insgesamt fünf etablierter sowie innovativer Assoziationsmaße zur Quantifizierung der Bindungsstärke zwischen den Konstituenten der Mehrworteinheiten. Sie erweitert diese auf Zweiworteinheiten bezogenen Methoden dahingehend, dass auch Sequenzen mit $n>2$ Konstituenten erfasst werden können, und analysiert

auf Basis dieser die Passung der variationistischen Modelle auf die gewonnenen Sprachdaten unter Zuhilfenahme verschiedener Techniken der Clusteranalyse. Die Datenbasis für die Analyse liefert dabei das *International Corpus of English* (ICE), welches die umfassendste Sammlung verlässlicher, vergleichbarer Sprachdaten aus nationalen Varietäten des Englischen darstellt. Verschiedene Assoziationsmaße mit unterschiedlichen bekannten Stärken (*MI*-score, *t*-score, log-likelihood) sowie innovative Methoden (lexical gravity, Delta P) erlauben dabei eine Charakterisierung der linguistischen Formen dieser Konvergenz und Divergenz.

Theoretisch greift die Arbeit auf verschiedene gängige Modelle zur Beschreibung von *World Englishes* zurück und gleicht die gewonnen Erkenntnisse zur Hypothesenüberprüfung mit diesen ab. Zentrale Modelle zur Beschreibung und Verortung von Varietäten des Englischen zeigen sich in drei Formen: 1) Klassische dreigliedrige Modelle unterscheiden (genetische) Muttersprachler, Zweitsprachler und Fremdsprachenlerner oder spiegeln diese Einteilung mit anderen Worten wider. 2) Regionale Beschreibungen und die Epizentrumstheorie (Hundt 2013) liefern Gruppierungen, die im Rahmen der vorhandenen Daten primär auf gegenseitige räumliche Nähe hinauslaufen. 3) Evolutionsmodelle wie vorrangig Schneiders (2007, 2014) *Dynamic Model* hingegen beschreiben die Entwicklung postkolonialer Varietäten als einen Prozess der Identitätskonstruktion, der sich anhand gesellschaftlicher Effekte in aufeinander folgende Schritte einteilen lässt, und der mit sprachlicher Eigenständigkeit einhergeht. Die Arbeit leitet aus diesen Modellen ab, dass sprachliche Ähnlichkeiten bei Anwendbarkeit der genannten Modelle deren hauptsächlichen Gruppierungen entsprechen müssten: Für klassische dreigliedrige Modelle ergäbe sich eine binäre Unterscheidung von mutter- vs. zweitsprachlicher Sprachverwendung. Regionale oder kulturelle Nähe sollte hingegen zu vorrangig proximitätsbasierten Gruppen (z.B. Afrika, Asien) führen, während Evolutionsmodelle eine Einteilbarkeit entlang der ihnen zugrunde liegenden Gradienten nahelegen (z.B. eher exo- oder endonormativ orientierte Varietäten).

Die Auswertung der extrahierten sprachlichen Sequenzen und ihrer Assoziationsstatistiken geschieht über Methoden der Clusteranalyse. Diese ermöglichen es, Muster in großen und multivariaten Daten zu erkennen, die sich nur aus der Kombination einer Vielzahl von Parametern ergeben. In ihren Grundzügen vergleichen diese Methoden alle *n* binären Objektpaare in den Daten (hier: Varietäten) auf ihre Ähnlichkeit. Das

Objektpaar, das die geringste interne Heterogenität zeigt, wird im Anschluss zu einem eigenen Objekt verschmolzen, worauf sich der Analyseprozess mit den verbliebenen *n*-1 Objektpaaren wiederholt, bis im letzten Schritt alle Objekte nach ihren Ähnlichkeiten verortet sind. Weil jedoch Clusteranalysen die Tendenz besitzen, selbst in zufälligen Daten Muster zu erkennen, werden die Ergebnisse verschiedener Clusterverfahren trianguliert. Zum Einsatz kommen die hierarchische Clusteranalyse (vgl. etwa Moisl 2015), k-means (vgl. Moisl 2015, Sarstedt & Mooi 2014) und phylogenetisches Clustering durch den NeighborNet-Algorithmus (Schliep 2011, Schliep et al. 2017). Die Segmentierung der hierarchischen Baumstrukturen (Dendrogramme) wird zudem empirisch durch Werkzeuge des randomisierten *Resamplings* (wiederholte Zufallsvariationen der Daten) unterstützt, um ein datengesteuertes Vorgehen sicherzustellen.

Die Ergebnisse der Arbeit zeigen, dass eine Beschreibung der sprachlichen Ähnlichkeiten zwischen den betrachteten Varietäten am besten entlang regionaler Kategorien gelingt. Zwar trennen sich gerade in den gesprochenen Daten die (genetisch) muttersprachlichen Varietäten oft deutlich von den übrigen Daten ab, jedoch zeigt sich gerade in den geschriebenen Daten eine Trennung zwischen nordamerikanischen Varietäten (USA, Kanada) und jenen mit stärkerer Bindung zum britischen Epizentrum (Irland, Neuseeland). Trotz des Fehlens von gesprochenen Daten zu den USA deuten sich ähnliche interne Unterscheidungen auch in den gesprochenen Daten an. Es deutet sich demnach bereits hier eine regionale Unterscheidung an, die darüber hinaus insbesondere in den afrikanischen Daten weitere Unterstützung findet. Diese grenzen sich zumeist deutlich von den übrigen Varietäten ab und zeigen häufige interne Unterscheidung in Ost- und Westafrika. Die asiatischen Daten entsprechen dieser Hypothese allerdings am wenigsten und fragmentieren häufig. Es ergeben sich deshalb weniger verlässliche und generell kleinere Cluster, zumeist in einer Verbindung von Hong Kong und Singapur sowie seltener Indien und Sri Lanka. Am wenigsten spiegeln die Daten eine Einteilung nach Schneiders *Dynamic Model* wider. Zwar entsprechen manche der Cluster auch Gruppen auf Basis dieses Modells, doch die Deckung innerhalb der gesamten Daten ist generell schlechter als bei einer proximitätsbasierten Analyse. Während letztere häufig große Teile der Daten sowohl in groben als auch feineren Einteilungen sinnvoll beschreiben, treten einige Gruppen, die auf Basis des *Dynamic Model* zu erwarten wären, überhaupt nicht aus den Daten hervor, etwa alle eher

exonormativ orientierten Varietäten. Während für eine vollständige Bewertung zunehmender sprachlicher Eigenständigkeit diachrone Daten von Nöten wären, schließt die Arbeit dennoch mit einer Priorisierung von räumlicher und kultureller Nähe als Erklärung für sprachliche Ähnlichkeit entlang von hochfrequentem Sprachgebrauch.