

Contents

Preface — VII

Abbreviations — XII

List of Tables — XIII

List of Figures — XIV

1 Introduction — 1

2 Languages as Adaptive Systems — 9

- 2.1 The ontogenetic level: individual speakers — 11
- 2.2 The glossogenetic level: formation of dialects and languages — 13
- 2.3 The phylogenetic level: language evolution — 17
 - 2.3.1 The saltational account — 17
 - 2.3.2 The gradual account — 20
 - 2.3.3 Cultural evolution: beyond biological transmission — 21
- 2.4 Summary — 24

3 Language Change and Population Structure — 27

- 3.1 Populations and languages — 27
 - 3.1.1 Population drift — 32
 - 3.1.2 Language genealogy — 34
- 3.2 A note on “internal” and “external” factors of change — 37
- 3.3 Description, grouping, and explanation — 39
- 3.4 Summary — 40

4 Lexical Diversity across Languages of the World — 42

- 4.1 Introduction — 42
 - 4.1.1 Sampling from languages: parallel corpora — 43
- 4.2 Theoretical preliminaries: type and token frequencies — 44
- 4.3 Word entropy — 53
 - 4.3.1 Information, meaning, and natural language — 55
 - 4.3.2 Estimation methods — 62
- 4.4 Word entropies for 1217 languages — 70
- 4.5 Summary — 76

5	Descriptive Factors: Language “Internal” Effects — 78
5.1	Writing systems and scripts — 78
5.1.1	Scripts and unigram entropies — 82
5.2	Word-formation — 85
5.2.1	Analytic, synthetic and polysynthetic languages — 85
5.2.2	Isolating, agglutinative and fusional languages — 89
5.2.3	Entropy share and cross-linguistic entropy variance — 90
5.2.4	Inflectional morphology — 92
5.2.5	Derivation, compounds, contractions, and clitics — 96
5.2.6	Tone — 101
5.3	Register and style — 106
5.4	Summary — 109
6	Explanatory Factors: Language “External” Effects — 111
6.1	Population size — 111
6.2	Adult learner percentages — 115
6.3	Language status — 117
6.4	Summary — 119
7	Grouping Factors: Language Families and Areas — 121
7.1	Unigram entropy by family — 121
7.2	Unigram entropy by area — 124
7.3	Global patterns of entropy — 125
7.3.1	Latitude: the Low-Entropy-Belt — 125
7.3.2	Altitude: “high” entropy — 128
7.4	Summary — 129
8	Predicting Lexical Diversity: Statistical Models — 130
8.1	Multiple regression: combining explanatory factors — 130
8.2	Mixed-effects regression: controlling for non-independence — 133
8.2.1	Visualization by family — 136
8.2.2	Visualization by area — 137
8.2.3	Visualization by corpus — 137
8.3	Summary — 139
9	Explaining Diversity: Multiple Factors Interacting — 140
9.1	Explanation and grouping — 141
9.1.1	Simpson’s paradox in language typology — 142
9.2	Lexical diversity through time — 145
9.3	Multiple Factor Model — 150
9.4	Summary — 155

10	Further Problems and Caveats — 156
10.1	Language learning and lexical diversity — 157
10.1.1	Learnability vs. expressivity — 157
10.1.2	Learning in children and adults — 163
10.1.3	Esoteric and exoteric linguistic niches — 166
10.1.4	Small-scale multilingualism — 167
10.1.5	Are children better learners than adults? — 169
10.1.6	Learner age and exposure — 171
10.2	A note on grammaticalization — 172
10.3	Beyond words — 174
10.3.1	Complexity trade-offs and equi-complexity — 176
10.3.2	Entropy at different levels of language structure — 177
10.4	Summary — 181
11	Conclusions: Universality and Diversity — 183
12	Appendix A: Advanced Entropy Estimators — 186
13	Appendix B: Multiple Regression Assumptions — 189
14	Appendix C: Mixed-effects Regression Assumptions — 192
	Bibliography — 195
	Index — 215