

Index

- accuracy *see* classification
actually, 6, 127–54
 adjusted estimates *see* estimates
 age, 5, 6, 7, 106, 129–39, 142–51, 164, 167, 186
 agency, 111, 122
 alignment, 55–9
 of parallel corpora, 328–32
 of recordings and transcriptions, 54–5
 alternation, 104–24, 194–205, 224, 292, 302,
 see also context, variable; envelope of
 variation; interchangeability
 animacy, 198, 200–1, 212–19, 234, 236–41
 annotation, 3–4, 8–13, 18, 42, 76–87, 98, 123,
 166–7, 188, 235, 274, 326, 343–9, *see also*
 tagging; parsing
 manual, 47–8, 57–9, 142, 145, 155, 244, 294,
 316–18, *see also* disambiguation
 types, 296–8, 311–13, 323–7, 335–49
 AntConc, 165, 274
 apparent-time change *see* change
 ARCHER (A Representative Corpus of
 Historical English Registers), 9, 195–8,
 299–315
 article, indefinite, 4, 17–22
 Audio BNC, 4, 49, 54–9
 auditory analysis, 4, 64–7
- bag-of-words approach, 5, 80, 137, 298
 balance *see* data
 baseline, 5, 11, 101–24, 135, 176, 200,
 205–6
 Bayes *see* likelihood, priors, posterior
 probability
 Bayesian statistics, 2, 8, 224–31, 240–6
be-passive *see* passive
 bias, 5, 18, 25, 39–41, 104, 257, 315
 selection bias, 41
 sparse data bias, 30, 36
 big data *see* data
 bigrams, 9, 21, 41, 52, 294, 305–6, 327, 333–48
 binomial regression *see* regression
- BNC (British National Corpus), 17–41, 54–5,
 75, 97–8, 127–8, *see also* Audio BNC
 BNC Baby, 82
 BNC 1994, 130–6, 142–5, 148–53
 BNC 2014, 130–5, 138–41
 BNCweb, 55
 bootstrapping, 182
 British National Corpus *see* BNC
 Brown Corpus, 52, 75, 76, 95
- categorical variables *see* variables
 change, 9, 42, 65, 129, 291, 299–318
 apparent-time change, 128–30, 138–41
 real-time change, 128–31, 136–41
 chunking, 9, 80, 310–15, 334
 C-index, 238–44, 251
 classification, 61, 80–3, 294–306, 328–37, *see*
 also classification tree; dendrogram;
 recursive partitioning
 accuracy, 80–2, 175–84, 298–306, 325,
 334–49
 supervised, 9, 293, 332
 classification tree, 7, 189, 195, 209–22, *see also*
 dendrogram; recursive partitioning
 pruning, 216
 visualization, 220
 cluster, 259–75, 283–4, *see also* hierarchical
 cluster analysis
 compactness, 261, 267, 278–82
 distinctness, 266–7, 275, 282
 goodness, 8, 266–72, 282–3
 clustered data structure *see* data structure
 clustering, 41–2, 80–92, 272, 278, 282
 of observations, 25–9, 155
 of texts *see* text clustering
 COCA (Corpus of Contemporary American
 English), 17–42, 97, 225, 234
 cognition *see* processes, cognitive
 COHA (Corpus of Historical American
 English), 9, 299–304
 collinearity, 293, 294–6, 298
 Common Ground, 346

354 Index

- comparability, 11, 22–31, 75–6, 95–7, 155
- complement clause
 - finite, 300, 308–10
 - non-finite, 300, 304, 308–10, 313
- compounds, 306, 316
- computational linguistics, 1–11, 291, 294, 298, 316
- concessive clauses, 272
- confidence interval *see* interval
- confusion matrix, 175–6
- conjunctions, 273–4
- connected speech *see* speech
- consonant *see* onset consonant
- constructions, participial, 308–11
- context, variable, 166, *see also* envelope of variation; interchangeability
- contractions, 92–6, 302, 305
- contrastive linguistics, 323–4
- corpus *see also* annotation
 - corpus-based approach, 6, 9, 48, 127, 141–3, 153–6, 243, 260, 291–3, 299–304, 323
 - corpus-driven approach, 5, 9, 263, 291–3, 296, 304–15
 - interface, 128, 133–5, 136–41, 153–6
 - multilingual *see* parallel
 - parallel, 8, 323, 328
 - spoken, 4, 47–53, 130
 - translation *see* parallel
- Corpus of Contemporary American English *see* COCA
- Corpus of Historical American English *see* COHA
- CQPweb, 128, 133–6
- cross-validation, 295, 298, 304, 334–42
- data *see also* data structure
 - balance, 22, 77, 146, 154, 339
 - big data, 3, 11, 13, 18, 41–2, 295
 - cleaning, 21, 58, 111–15, 122, 165–6
 - homogenization, 164
 - metadata, 3, 5–6, 17–18, 25–33, 37, 41, 61, 75–7, 127–56
 - multidimensional, 265–7, 272
 - quality, 3, 13, 39–42, 55, 58
 - sparsity, 7–8, 30, 36, 61, 97, 229, 244–6, 299–301, *see also* bias
 - verification, 6, 41–2
- data structure, 130–3
 - clustered, 7, 28
 - hierarchical, 5, 11, 145–8, 156, 172, 190
 - multilevel *see* hierarchical
- data-driven *see* corpus-driven approach
- dendrogram, 259–66, *see also* classification tree
- dependency length, minimum, 315
- dependent variables *see* variables
- design effects, 152
- dimensionality reduction, 77, 80, 87–8, 265
- disambiguation, 40, 123, 141, 274
- discourse marker *see actually*
- discourse particles *see* particles, modal
- distance, 6, 88–92, 232–40, 261–72, 278–85
 - matrix, 87, 266
- distance matrix, 264
- distribution, 59–62, *see also* Zipfian
 - distribution
 - negative binomial, 139
 - Poisson, 137
 - skewed, 144, 150, 188
- down-sampling *see* sampling
- economy, 66
- editorials *see* press editorials
- effects, 53, 61–2, 168, 202, 227–9, 238–42, 250–4, *see also* design effects
 - fixed, 172–9, 199, 240–4, 255
 - mixed *see* mixed-effects model
 - random, 172–9, 188, 199
- end weight, 201
- envelope of variation, 104, 294, 313, *see also* alternation; context, variable; interchangeability
- estimates, adjusted, 30–1
- estimation, 26, 96, 137–8, 145, 155, 224
- etymology, 38
- exclusion criteria, 23, 41, 130
- exemplar model *see* model
- existential *there*, 306–8, 317
- exploratory research, 259, 284, 293, 308
- factor analysis, 80–1, 295, *see also* multi-dimensional scaling
- feature weight, 295–8, 305–6
- finite complement clause *see* complement clause
- fixed effects *see* effects
- forward model selection, 173–4, *see also* model, selection
- frequency, 5, 23–4, 62–4, 66, 87–90, 102–4, 167, 332
 - normalized, 5, 101–6, 128, 135, 137, 198
- profile *see* profile
- token frequency, 17, 33–8, 60
- type frequency, 38, 40
- vector, 327–8
- frequentist statistics, 224, 225, 226–7, 231
- Fuzzy Tree Fragment (FTF), 111–13
- gender, 6, 7, 61, 129, 131, 142, 164, 167
- genitive, *of*, 194–222

- genitive, Saxon, 194–222, 306–10
 genre, 3–5, 18, 20, 24–31, 61, 107, 118, 164–5, 262, 294
get-passive *see* passive
 Gini index, 88, 210, 217
 Google Books Ngrams, 4, 17–42
 Google Books Ngram Viewer, 23

help + (to)-infinitive, 224–5, 230–46
 hierarchical cluster analysis, 80, 266, *see also*
 dendrogram, classification tree
 hierarchical data structure *see* data structure
h-initial lexemes, 17–42
 homogenization *see* data
horror aequi, 232–40, 257
 hypothesis testing, 137, 142, 176, 216, 225, 237, 241, 246
 null hypothesis testing, 224
 hypothesis-driven approach *see* corpus-based approach
 hypothesis-generating methods, 308

 Iambic Reversal Rule *see* stress shift
 ICE (International Corpus of English), 5, 7, 8, 75–7, 82–97, 101, 164, 262, 274
 ICECUP, 106, 112
 ICE-GB, 87, 95–6, 112–13, 120–2, 164–6
 ICETree, 88
 markup conventions, 165
 idiom principle, 66
 importance of variables *see* variables
 indefinite article *see* article
 independent variables *see* variables
 inference, statistical, 26, 155
 infinitive *see to-infinitive*
 information criteria, 173, 189, 246
-ing form, 300–3, 308, 316–17
 interchangeability, 195–9, 221, *see also* alternation; context, variable; envelope of variation
 interface *see* corpus
 International Corpus of English *see* ICE
 interpretability, 97, 346
 interval
 confidence interval, 6, 91–2, 115–17, 135–42, 146–52, 231, 241–4, 255, 334–42
 credible interval, 231, 244, 255
 uncertainty interval, 24–32, 35–6, 40, 276
 Wilson score interval, 110, 115–20

 lemma, 235, 326–7, 332–48
 lexical storage *see* storage
 Lightside, 293–5, 304, 317
 likelihood, 226–30, 237–41, 244–6, *see also* maximum likelihood statistics
 LLC (London-Lund Corpus), 81
 LOB (London-Oslo-Bergen) corpus, 79, 81
 log odds, 199, 206, 220, 230–1, 333
 log-odds ratios, 206
 logarithm *see* transformation, logarithmic
 logistic regression *see* regression
 London-Lund Corpus *see* LLC
 London-Oslo-Bergen corpus *see* LOB corpus

 machine learning, 2–3, 8, 80, 88, 291–8, 316
 manual annotation *see* annotation
 manual linguistic review, 41, 83, 97, 112–14, 123, *see also* annotation, manual
 Markov Chain Monte Carlo *see* MCMC algorithm
 MAT (Multidimensional Analysis Tagger), 87
 matrix *see* distance
 maximum likelihood statistics, 224, 237–46, 255
 MCMC algorithm, 229–31, 241, 246, 276
 MDA (Multidimensional Analysis) *see* multidimensional scaling
 MDS *see* multidimensional scaling
 metadata *see* data
 metrical principles *see* Principle of Rhythmic Alternation (PRA)
 metrical wellformedness *see* Principle of Rhythmic Alternation (PRA)
 minimum dependency length *see* dependency length
 mixed-effects model, 7, 163–4, 172–82, 187–91, 224
 modal auxiliaries, 79, 92–4, 301–2, 311
 modal expressions, 339–40
 modal particles *see* particles
 model, 7, 62–4, 88–92, 199–210, 224–6, 237–46, 255, 332–42, *see also* regression
 exemplar model, 56
 language model, 306, 317
 mixed-effects *see* mixed-effects model
 Poisson model, 275–7
 selection, 172, 189–90
 usage-based model, 66
 monofactorial analysis, 167–70, 185–7, 190
 monograms, 5, 77, 81, 294, *see also* unigrams
 multidimensional analysis (MDA) *see* MAT, multidimensional scaling
 multidimensional data *see* data
 multidimensional scaling, 5, 8, 11, 80, 87–92, 259–66, 277–84
 multifactorial analysis, 163–91, 194–222, 224–46, 259–85
 multilevel data structure *see* data structure, hierarchical

356 Index

- multilingual corpus *see* corpus, parallel
 multinomial regression *see* regression
 mutual substitution *see* substitution
- natural Language Processing (NLP), 9, 323
 negative binomial distribution *see* distribution
 negative binomial regression *see* regression
 NeighborNet, 8, 11, 261–3, 265–6, 279, 282
 nested data structure *see* data structure,
 hierarchical
n-grams, 9, 77, 80–2, 97, 325–7, 332–4,
 336–48, *see also* Google Books Ngrams;
 unigrams; bigrams; trigrams; skipgrams
 NLP *see* natural language processing
 non-finite complement clause *see* complement
 clause
 normalization of spelling, 87, 293, 297, 299,
 see also VARD 2
 normalized frequency *see* frequency
 numerals, 56–64
- of*-genitive *see* genitive
 onset consonant, 17–18
 open choice principle, 66
 opinion pieces *see* press editorials
 orthography, 49, *see also* normalizat^{on} of
 spelling; transcription; VARD 2
 outlier, 89–90, 140, 182, 236
 out-of-bag error, 183
 overconfidence, 27
 overfitting, 182, 217, 244, 295–6
 overuse metric, 293, 297, 304
- paradigm shift, 1
 parallel corpus *see* corpus
 parsing, 235, 297, 331
 participle *see* constructions, participial
 particles, modal, 9, 323–6, 328–48
 partitioning *see* recursive partitioning
 part-of-speech tagging *see* tagging
 passive
 be-passive, 105–24
 get-passive, 101–24, 300–1
 progressive passive, 300–1
 pauses, 7, 163–90
 per million words (pmw) *see* frequency,
 normalized
 phonemic transcription *see* transcription
 phonological prominence *see* prominence
 phrasal verbs, 313–17
 pmw *see* frequency, normalized
 Poisson *see* distribution, regression, model
 POS tagging *see* tagging, part-of-speech
 posterior probability, 226–31, 241–6, 255
 posteriors *see* posterior probability
- pragmatic variables *see* variables
 precision, 145–53, 297, 300–4
 prediction, 175–84, 296
 predictor variables *see* variables
 preference *see* selectional preference
 press editorials, 5, 75–7, 83–91, 95–6
 primary sampling units *see* sampling
 principal component analysis, 80, 295
 Principle of Rhythmic Alternation (PRA), 4,
 46–67, 237
 priors, 225–30, 240–6, 255–7
 flat, 227–30, 241, 255
 informative, 8, 225, 229, 240, 243–4, 255
 non-informative *see* flat
 weak, 225, 227, 241, 255
 processes, cognitive, 48–50, 65–7
 profile, 9, 81–93, 96–7, 138–40
 progressive, 300–1, 306–11, 316, *see also*
 passive
 prominence, phonological, 19, 58
 pronouns, 79, 90–6, 309, 346
 pruning *see* classification tree
- query, restricted, 133, 143
- random effects *see* effects
 random forest, 7, 87–97, 163–4, 182–90,
 194–5, 216–22
 ratio variables *see* variables
 real-time change *see* change
 recursive partitioning, 177, 216
 reduction *see* dimensionality reduction
 register, 2, 5, 75–98, 107, 120, 122, 298–9
 regression, 6, 7, 11, 172, 194–5, 209–22,
 237–46, 284
 binomial, 7
 logistic, 7, 62–4, 196–205, 209–22, 229–32,
 294, 298
 multinomial, 7, 205–22
 negative binomial, 139
 Poisson, 137
 visualization, 220
 relativizers, 194, 300, 308
 replication crisis, 10, 225, 231
 reproducibility, 246, *see also* replication crisis
 restricted query *see* query
 Rhythm Rule *see* stress shift
- sample, 76, 332
 size, 26, 35–9, 116, 142–8, 235, 243–6, 295
 size planning, 142
 sampling, 5, 11, 25, 77, 84–7, 95, 141
 down-sampling, 6, 41, 128, 141–56
 primary sampling units, 26, 142, 146,
 148, 155

- secondary sampling units, 142, 146, 155
 simple random down-sampling, 143, 150, 152, 153
 stratified random down-sampling, 143, 150, 152, 153
 structured down-sampling, 148–50, 152, 153
 subsampling, 102, 112, 115–20, 123, 154
 theory, 141–3, 152
 uneven proportion subsampling, 102, 115
 variation, 35–6, 137
 Saxon genitive *see* genitive
 secondary sampling units *see* sampling
 selectional preference, 46, 47, 50
 selection bias *see* bias
 selection of variables *see* variables
 sensitivity analysis, 4, 11, 31–3, 241, 255, 334–40
 simulation study, 145–8
 situational characteristics, 5, 75, 76, 83, 95, 97
 skewed distribution *see* distribution
 skipgrams, 327, 338–48
 sociolinguistic categories, 129–48, 154
 sparse data bias *see* bias
 speaker variation *see* variation
 speech corpora *see* corpus, spoken
 speech, connected, 46
 spoken corpus *see* corpus
 statistical inference *see* inference
 storage, lexical, 64
 stress, 38–9
 clash, 46–7, 52–4, 58–61, 236–7, 243–4
 lapse, 46, 236–7, 243–4
 shift, 46–66
 study design, 127–8, 145–54, *see also* design
 effects
 subsampling *see* sampling
 substitution, mutual, 6, 105, 108, 123
 supervised approach, 298
 supervised classification *see* classification
 Support Vector Machine (SVM), 9, 221, 333–7
 tagging, 80–1, 97, 111–12, 299
 part-of-speech (POS), 54, 77, 87, 296–7
 text categories, 30–1, 63
 text clustering, 77, 80–2
 text types *see* categories
 there *see* existential there
 Thirteen-Men Rule *see* stress shift
 to-infinitive, 224, 227–44, 261, 303, 308–13
 token frequency *see* frequency
 transcription, 49–59, 165
 orthographic, 4, 49–54, 65
 phonemic, 49, 54–9, 65
 transformation, logarithmic, 63, 87, 97, 135
 translation, 9, 323–35, 347–8, *see also* corpus,
 parallel
 trigrams, 9, 294, 305–6, 327, 338–48
 type frequency *see* frequency
 uncertainty interval *see* interval
 uncertainty, statistical, 29, 33, 135, 139, 155
 unigrams, 9, 327, 332–49, *see also* monograms
 usage-based model *see* model
 validity, 18, 24–5, 41–2
 VARD 2, 299, 317, *see also* normalization of
 spelling
 variables, 6–8, 87, 194–222, 234–7, 263, 275,
 278–83, *see also* alternation; context
 categorical, 166–70, 194, 212
 dependent, 62, 106–8, 121–4, 172–82,
 184–90, 196–218
 importance, 184–6, 217, 340–1
 independent, 63, 104, 166–70, 172–82,
 184–90
 pragmatic, 163
 predictor, 88, 183, 194–222
 ratio, 167
 selection, 81, 199, 246
 variation, 2, 18–20, 76, 80–1, 104, 194–5,
 224–6, 234, 272–80, 343–6
 between speakers, 136–40, 154, 172, 190,
 see also sampling; sociolinguistic
 categories
 by age *see* age
 by gender *see* gender
 by register *see* register
 by situation *see* situational characteristics
 by variety *see* variety
 variationism, 2, 129, 195–6
 variety, 17–42, 89–91, 96, 121–2, 177, 185,
 275–6
 vector *see* frequency vector
 verification *see* data
 visualization, 8, 87–8, 130, 220, 262–6,
 284–5, *see also* regression; classification
 tree
 WAVE (Mouton World Atlas of Varieties of
 English), 261
 weight *see* end weight; feature weight
 word order, 46–7, 309–14, 317
 Zipfian distribution, 59, 87, 188