Index

A Representative Corpus of Historical English Registers (ARCHER), 103, 232, 258, 265–7, 269, 272
Academic Journal Research Corpus (AJRC), 85, 105–12
Academic Word List (AWL), 24
academic writing
in the BNC
corpus design, 210, 211
distribution considerations, 212–15
documentation, 202–3, 215–16
domain description, 203–5
evaluation, 209–10
motivation for, 201–2
operational domain, 208–9
in the COCA
corpus design, 211–12
distribution considerations, 212–15
documentation, 202–3, 215–16
domain description, 203–5
motivation for, 201–2
operational domain, 205–8
accuracy, 128, 146, 152, 157–61, 165, 167–70, 176, 217, 220
defined, 123
see also linguistic parameter estimation,
precision, representativeness,
evaluation of
audiences for this book, 13–15, 17
consumers of corpus research, 15
corpus analysis, 14
corpus designers/builders, 14
Auslan corpus, 46
authenticity, 2–3, 8, 10, 11, 21
see also corpus, defined
balance, 21, 31, 32, 33, 34, 36, 37–9, 42, 43, 49–50, 59, 63, 209, 210, 220
see also corpus, size, proportionality
bias, 37, 123, 175, 176, 181, 220
defined, 31
selection, 98, 100, 103, 104, 112, 158–9, 161, 162, 169, 172, 174, 189, 196, 222
see also, coverage bias, selection bias
biographies, 84
British Academic Written English (BAWE), 153, 154, 251
British National Corpus (BNC), 7, 31, 34, 99, 141, 142–5, 149, 150–1, 201–17, 236, 252
see also academic writing, in the BNC
Brown corpus, 2, 8, 34, 45, 63, 98, 103, 237, 259, 262, 263–7, 269
Central Limit Theorem, 130
Civil Judgments Corpora, 256
Cohen’s d, 164, 165–73, 214, 220
certainty, 35–6, 130–2, 153–4, 182, 191, 197, 220
see also distribution considerations,
precision, probability level
construct
operationalizing, 91–2
cooking shows, 73, 79, 83, 95, 96, 104
corpus
defined, 7–11, 220
definitions of, 2–4
see also evaluation, representativeness,
sample, text
Corpus Brasileiro de Variação de Registro (CBVR), 44
corpus design, 220
steps for achieving and evaluating, 178–82
see also corpus, domain considerations,
distribution considerations, evaluation, representativeness
corpus linguistics, 3–7
Corpus of Contemporary American English (COCA), 7, 24, 30–1, 146, 201–16, 227, 258, 260, 261–3, 265
see also academic writing, in the COCA

280
Index

Corpus of Early English Correspondence (CEEC), 247, 260, 265, 269
Corpus of Historical American English (COHA), 229, 258, 265, 267
Corpus of L1 Arabic, L1 Chinese, and L1 English Student Writing, 26
Corpus of Online Registers of English (CORE), 48, 125, 126, 127, 128, 135, 141, 145, 148–9, 153, 235, 263, 265, 267, 269
Corpus of Personal Blog Posts, 25, 60
Corpus of Recent Yelp Restaurant Reviews (CRYRR), 182–91, 199–201
corpus design, 188, 190
documentation, 190–1
domain description, 183–5
evaluation of, 188
motivation for, 182–3
operational domain, 185–7
Corpus of YouTube Vlogs (CYTV), 199–201
corpus design, 195, 197–9
distribution considerations, 197
documentation, 199
domain description, 192–3
evaluation of, 194–5, 196
motivation for, 191–2
operational domain, 185–7
coverage bias. See bias, coverage
demographic variables, 88, 220
see also domain considerations, situational variables, stratification distribution. See distribution considerations, linguistic distribution, linguistic parameter estimation
distribution considerations, 12, 18, 55, 57, 60–2, 63, 64, 98, 122–55, 157–8, 160–73, 180, 181–2, 212–17
defined, 54, 220
see also evaluation, domain considerations, representativeness, research questions, sample size
documentation, 14, 20, 58, 178, 202–3, 205, 218, 221
see also evaluation
domain, 3–7, 10–13, 53–7, 221
see also domain considerations, population domain boundaries. See domain description, domain boundaries
domain considerations, 54, 55, 58–62, 63, 64, 68–121, 122, 125, 157–160–1, 172–4, 180
defined, 12, 221
see also domain description, operational domain, representativeness, research questions, sampling unit, sampling methods
domain description, 59, 70–91, 113, 120, 221
a priori (new corpora), 180
domain boundaries, 86–8, 221
domain-internal categories, 88–91, 221
for academic journal articles, 105–8
for academic writing, 203–5
for blogs, 192–3
for online restaurant reviews, 183–5
methods for, 81–6
post hoc (existing corpora), 195
sources of information, 81–6
domain generalizations. See generalization
EF-Cambridge Open Language Database (EFCAMDAT), 22–3
evaluation, 6, 7, 10, 11–15, 35, 70, 71, 92, 112, 119, 129, 134, 180, 181, 221
corpus to operational domain, 104–5
operational domain to domain, 97–8
see also corpus, distribution considerations, documentation, domain considerations, representativeness, evaluation of
fiction, 44, 45, 212, 229, 233
late nineteenth-century novels, 74, 95
First Year Business and Engineering Textbook Corpus (FYBETC), 253, 268
Freiburg English Dialect (FRED) corpus, 238, 258, 265
generalization, 3, 5–6, 7, 11, 12, 13, 39–40, 41, 56, 57, 102, 160, 221
see also domain considerations, distribution considerations, linguistic parameter estimation, representativeness
Global Web-Based English (GloWE) corpus, 48, 230, 263, 265, 267, 269
grant proposals, 73, 77, 82, 84, 94–5, 97, 98, 105, 209
Intelligent Web-Based (iWeb) corpus, 9
International Corpus of English (ICE), 244, 260, 265, 268, 269
International Corpus of Learner English (ICLE), 243, 245, 260, 265, 269
Iowa State University Research Article (ISURA), 47, 51
job interviews, 73, 80, 83–4, 89–90, 95, 97, 103–4, 105
journals, scholarly, 105–12, 201–16
see also Academic Journal Research Corpus, academic writing
L2-ARCTIC corpus, 8
lexical bundles, 22, 67, 124, 138, 148–9, 153
Index

lexical growth curves, 140, 141
see also type, linguistic
lexical saturation, 140–1
see also type, linguistic
linguistic distribution, 35, 53–6, 60–3, 122–3, 151, 158, 220, 221
defined, 122
see also distribution considerations
linguistic parameter, 221
defined, 55
estimation, 35, 53, 55, 56, 59, 60–2, 64, 73, 156–60, 216–17
linguistic parameter estimation, 221
linguistic representativeness, 35, 61, 221
see also representativeness, distribution considerations
linguistic research goal. See research questions linguistic variable, 60–1, 122–4, 159–60, 180, 182, 221
linguistics
applied, 1, 92
cognitive, 4
historical, 4
sociolinguistics, 1, 4, 5–6
London-Lund Corpus of Spoken English (LLC), 250, 259, 265, 268
Longman Spoken and Written English (LSWE) Corpus, 233, 263, 265
Louvain Corpus of Native English Essays (LOCNESS), 246, 258
Louvain International Database of Spoken English Intergloss (LINDSEI), 245, 268, 269
mean, 55, 131–8, 139, 164–5, 221
Michigan Corpus of Academic Spoken English (MICASE), 25, 239, 265, 267, 269
Michigan Corpus of Upper-Level Student Papers (MICUSP), 240, 263, 265
news magazines, 92–3
News on the Web (NOW) corpus, 231, 259, 265
newspaper articles, 86–7, 88–9, 91, 101–2, 131, 141, 146
nonlinguistic parameters, 54, 58, 69, 73
see also domain-internal categories, domain considerations, situational representativeness, stratification
nonrandom sampling, 6, 159, 221
and bias, 112, 158, 162
examples, 104, 105, 267
online restaurant reviews, see Corpus of Recent Yelp Restaurant Reviews (CRYRR)
operational domain, 58–60, 70–1, 91–3, 222
boundaries, 94–5
considerations when specifying, 93–7
coverage, 98
evaluation, 97–8
strata, 95–6
versus sampling frame, 96–7
see also domain considerations, domain description
oversampling, 100, 124–5, 128–9, 222
see also sample size, undersampling
Oxford English Dictionary (OED), 8
parameter, 35–6, 53–6, 60–4
see also accuracy, generalization, linguistic parameter estimation, precision
parameter estimation. See linguistic parameter estimation
Philadelphia Neighborhood Corpus (PNC), 254, 261
population, 6–7, 11, 12, 29, 33–6, 37, 38, 39, 59, 68–9, 222
see also domain, operational domain, representativeness, sampling
precision, 35–6, 40, 124–6, 129, 134–7, 158–61, 166–9, 190, 191, 199, 213, 216, 222
deﬁned, 123
for a priori sample size determination, 130–3
for an existing corpus, 134–6
see also accuracy, linguistic parameter estimation, representativeness, evaluation of press brieﬁngs, White House, 73, 78, 83, 94, 95, 96, 97, 98, 105, 117
principled, 2, 3, 6, 7, 10–11, 25, 104
see also corpus
probability level, 130
see also distribution considerations, conﬁdence interval, precision
product manuals, 73, 76, 81, 82, 94, 95, 96, 97, 99
proportionality, 28–9, 33–4, 37–8, 59, 100–3, 222
and linguistic research questions, 101–3
see also domain considerations, balance, sample size, representativeness, stratification
Published Research Article Corpus (PRAC), 22
random sampling, 6, 31–2, 39, 59, 222
and bias, 103, 158
and sampling frames, 103
definition, 103
Index

examples, 34, 45, 48, 90, 105, 109, 125–7, 148, 162, 166–74, 176, 188–9, 195–6, 208, 210, 267
see also bias, sampling
see also domain considerations, domain-internal categories, situational variables, stratification
relative standard error, 136, 154, 213–15
sampling frame, 68, 98, 103, 117, 162, 166, 170, 176, 187, 195, 222
defined, 59
versus operational domain, 96–7
see also domain considerations, operational domain, sample, random/probability
formula, 134
see also confidence interval, precision, standard deviation, standard error
representativeness, 28–51
absence of selective forces, 31–2
and balance, 37–9
as a continuous construct, 62–3
conceptualizing, 42
coverage of the population’s heterogeneity, 34–5
defined, 53, 222
designed for a particular purpose, 36
evaluation of, 97–8, 104–5, 130–3
framework for, 53–63
general acclaim for data, 30–1
miniature of the population, 33
permitting good estimation, 35–6
pillars of, 58–62
possibility of, 39–40
statistical meaning of, 28–30
steps as achieving and evaluating, 178–82
typical or ideal cases, 32
very large corpus, 36–7
see also corpus, distribution considerations, documentation, domain considerations, evaluation
research questions, 3, 12, 36, 55, 57–8, 101, 178–9, 180, 221
sample, 3, 5–8, 222
nonrandom, 103–4
planning, 98–104
random, 103
simple, 103, 222
stratified, 103, 223
see also corpus, domain, representative-ness, sample size, stratification
sample size, 124–9, 222
estimating required (a priori), 130–3
evaluating adequacy (post hoc), 134–6
for linguistic tokens, 129–36
for linguistic types, 138–51
misconceptions about, 36–7, 136–8
see also distribution considerations, sample, oversampling, undersampling
sampling methods, 98–104, 222
selection bias. See bias, selection theory, 28–30, 68–9
see also representativeness, sample
sampling frame, 68, 98, 103, 117, 162, 166
defined, 59
versus operational domain, 96–7
see also domain considerations, operational domain, sample, random/probability
sampling unit, 210, 212, 222
full texts versus text excerpts, 98–9, 216, 263
texts, 7–8
see also sample, sampling, texts
Santa Barbara Corpus of Spoken American English, 241, 258, 265, 268
selection bias. See bias, selection
simple random sample. See sample, random, simple
situational representativeness, 222
see also domain considerations, representativeness
situational variables, 88, 223
see also demographic variables, domain considerations, stratification
standard deviation, 123, 130, 131–4, 136, 138, 153–4, 158, 164, 223
standard error, 134–5, 213, 223
formula, 134
see also relative standard error
stratification, 100–3, 223
and proportionality, 100–3
see also balance, domain considerations, proportionality
stratified random sample. See sample, random, stratified
stratified sampling. See stratification
stratum (pl. strata).
See stratification
Switchboard Corpus, 242, 258, 265, 268
text, 7–8, 223
defined, 7–8
see also corpus, domain considerations, sampling unit
text type, 34, 54, 55, 59, 62–3, 70–1, 81, 86, 88, 100–1, 223
textbooks, 206, 209
psychology, 86, 117, 140
see also academic writing
Timestamped JSI web corpus, 9
TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) corpus, 234, 265, 269
Index

token, linguistic, 129–36, 223
  see also sample size, for linguistic tokens
tolerable error, 35
  see also distribution considerations, confidence interval
t-value, 130, 136, 189, 197, 213
type, linguistic, 138–51, 223
  and vocabulary lists, 142–51
  purposes for analyzing, 138–9
  see also sample size, for linguistic types
undersampling, 124, 125–6, 190, 223
  see also sample size, oversampling
variables. See demographic variables, linguistic variables, situational variables
variance, 130, 134, 138, 215
  see also confidence interval, standard deviation, standard error

Web as corpus, 10–11, 39
WebCorp, 10
Wikipedia, 82–3
representativeness experiments, 161–73
word frequency bands, 142, 146, 150
  see also type, linguistic
Wordbanks Online, 248, 262
Yahoo! Question and Answer (Q+A) corpus, 85
Yelp Academic Challenge Dataset (YAD), 186–8, 189
  see also Corpus of Recent Yelp Restaurant Reviews (CRYRR)
YouTube, 191–201
  see also Corpus of YouTube Vlogs (CYTV)
Zipfian distribution (Zipf’s Law), 142, 223