Index

analysis tools: see also methodological issues, tagging programs
advantages of developing your own computer programs, 16, 59–60, 107–8, 112–13
concordancing software, 13, 15, 254: see also concordancing interactive software, 73, 112–16
association patterns, 5–8
discourse, see discourse analysis
grammatical, 56: see also grammatical analysis
lexico-grammatical, 84–5: see also lexico-grammatical analysis
lexical, 21–2: see also lexical analysis
linguistic, 6–7
non-linguistic, 5, 7
author style, 222–8

concordancing, 239–40, 254–5 for lexical analysis, 26–31, 36–7 for morphological analysis, 58–9
KWIC concordance, 26–8, 87, 96 limitations, 107, 254–6
software, 15, 28–9, 37, 285–6
c-o-ccurrence of linguistic features, 144–6: see also multidimensional analysis, register variation
corpora: see also tagging programs
Arizona Corpus of elementary student writing, 175–7
ARCHER, 124, 206–8, 210, 211, 216, 223, 251, 252–3, 283
British National Corpus (BNC), 13, 14, 73, 86, 101, 206, 207, 282
CHILDES, 175, 177, 282
commercially available, 281–7
Corpus of elementary student language, 175–7; creating the corpus, 175–77
Corpus of writing in the disciplines, 158–66; creating the corpus, 158–9
Helsinki Corpus, 205, 251, 252, 283
Lancaster–Oslo/Bergen (LOB), 13, 14, 26, 28, 29, 31, 32, 110, 112, 140, 145, 258, 282, 284
London–Lund, 13, 14, 60, 71, 110, 112, 140, 145, 282, 284
Longman–Lancaster, 13, 15, 26, 28, 34, 36, 41, 52, 60, 73, 78, 80, 86, 96, 99, 101, 206, 207, 266, 267, 282

297
Index

298

major corpora used in book, 12–13, 14–15
corpus design issues, 246–53
diachronic, 251–3
diversity, 34–5, 137, 248: see also
register variation
size, 30, 41, 80, 136, 142–3, 248–9
representativeness, 34–5, 158–9, 173, 177, 215, 246–50
tagging, 257–9
corpus-based approach
advantages of, 9–12, 173, 233–4
characteristics of, 3–5, 11–12, 18
compared to other approaches, 9–11
other research areas, 234–6
used for language education, 236–8
dialect(s)
defined, 135
female/male, 216–22
discourse analysis: see also referring
expressions
difficulty of, 106–8
discourse maps, 126–30
text subsections, 122–30, 166–9

English for Specific Purposes (ESP), 157–69
error analysis, 197–200
ESL textbooks, comparison of, 80–2
frequency counts
issues in computing counts, 66–8, 91–3
normalized, 32–3, 60, 263–4
raw, 263–4
gender differences, 215–22
grammatical analysis: see also
grammatical categories and
structures
compared to ESL textbook
descriptions, 80–2
descriptive vs. prescriptive
grammar, 55–6
diachronic, 205–9
lexical associations: see also lexicogrammatical analysis
register distribution, 60, 63, 68–9, 74–5
synonymous structures, 58, 76–80, 101, 105, 100–4
frequency of features, 60, 66–7, 91–3, 206
grammatical categories and structures
active vs. passive, 76, 124–30
adjectives, attributive and
predicative, 85–6
complement clauses (that-clauses
and to-clauses), 3, 7, 10, 70–81, 85, 96, 98–100, 101–5, 269–73
dependent clauses, 139–43
grammaticalization, 210
modals and semi-modals, 205–10
morphological features, 58–9
nominalizations, 59–78
noun/verb ratios, 57, 67–9, 75
transitive vs. intransitive verb
patterns, 85, 95–100
t-units, 178–80
verbs, 102–4
verb tense and voice, 124–5, 126–30
historical studies
corpus design issues, 251–3
historical registers, 210–15
language acquisition and
development
approaches to, 172–3
elementary student vs. adult
register variation, 192–7
elementary student writing
development, 188–92
error patterns, 197–200
individual linguistic features, 178–80
lemma, 29
lexical analysis, 21ff
collocations, 35–9, 41–2, 45–50:
at a distance, 51–3
compared to dictionary descriptions, 39–41
diachronic, 208–9
grammatical functions of a word, 30−1, 33−4, 85−7; see also
lexico-grammatical analysis
register distribution of a word, 32−5, 35−41, 44
statistical measures of lexical associations, 265−8
synonymous words, 43−51, 85−94, 95−100
word meanings, 26−8, 35, 36−41, 43−51
word frequency, 28−31, 43−4, 208
lexico-grammatical analysis
grammatical associations of words, 30−1, 33−4, 85−7, 93−4, 95−100
lexical associations of grammatical constructions, 100−4
verb valency patterns, 85, 95−100
methodological issues: see also
analysis tools, corpus design
issues, frequency counts, multi-dimensional analysis
automatic vs. “human” analytical techniques, 66−7, 71−3, 87−93, 97
determining the unit of analysis, 269−74
functional interpretation, 8−9, 139, 149−51
interactive analysis, 112−16
quantitative analysis, 8−9, 139
multi-dimensional analysis
adult vs. elementary student
dimensions of variation, 192−7
dimensions in 1988 model of
register variation, 147−55
dimensions in model of elementary
student registers, 181−8
author style, 222−8
biology vs. history academic texts, 158−66
ESP registers, 157−69
historical evolution of registers, 210−15
female/male dialect differences, 215−22
methodological procedures:
dimension scores, 151, 278−80;
factor analysis, 147−8, 278−80;
functional interpretations,
149−51; identifying and
interpreting dimensions, 145−51
research article subsections, 166−9
student writing development, 188−92
mutual information score, 265−7
normalizing counts, see frequency
counts
on-line dictionary, 71, 261
referring expressions, 108, 110,
116−22
classifying referents, 109−11
given/new information, 107−9, 110
reference distance, 110−11
register distribution of, 116−22
register descriptions
academic writing, 141−3, 144,
149, 158−69
biology vs. history research articles, 158−66
conversation, 141−2, 144, 150
drama, 210−15
elementary student fiction and
textbooks, 185, 186
elementary student writing, 185,
188−92
ESP, 157−8
fiction, 155−6, 224−7
historical, 210−15; see also
diachronic studies
medical writing, 210−15, 252
personal letters, 215−22
research article subsections, 166−9
register variation: see also multi-dimensional analysis.
register descriptions
comprehensive analysis of, 136−7
co-occurrence patterns, 143−5, 146−8
diachronic, 205−15
elementary student speech and
writing, 181−97
ESP, 157−8
individual linguistic features,
139−43, 178−80
Index

"register" defined, 135
spoken vs. written, 143–56
sampling, see corpus design issues
second language writing, 197–200
statistical measures: see also frequency counts
ANOVA, 273–4, 276
Chi-squared, 272–3, 276–7
determining the unit of analysis, 269–74
factor analysis, 278–80
F-score, 124, 152, 154, 275–6
mutual information score, 265–7
Pearson correlation, 124, 152, 154, 276
significance tests, 124, 276
standardized scores, 280
strength of association, 273, 275
t-scores, 267–8
T-test, 275–6
Varbrul, 273
z-scores, 268
tagging programs, 86, 261–2; see also analysis tools
advantages of, 254–7
accuracy of, 87–93, 262
interactive and hand editing, 71, 73, 87–93, 97, 262
vs. parsers, 260
probabilistic tagging, 261
tagged corpora, 257–60
writing development, 178–92