

Index

A

Absolute distance metric, 232–233
Adjective phrase grammar, 182
Adjectives, 48–49
Adverbial phrase reordering (AdvR), 231
Adverbial phrases, 183
Agglutinative morphology, 54
Alternative path model, 227
American National Corpus, 296
Automatic Content Extraction (ACE) program, 116
Automatic speech recognition (ASR), 12, 13, 73, 77–79
 acoustic processor (front-end), 95
 Bayes' formula, 95
 1-best hypothesis, 97
 broadcast news transcription system, 106–110
 communication theory, 94
 composite HMM model, 96
 context-dependency network, 96
 decoder, 96
 deep neural networks, 95
 dictation system, 94, 95
 DNNs, 96
 GMMs, 96
 language resources
 acoustic and text data, 98–103
 linguistic tools, 103–104
 LVCSR system, call center conversations, 110–111
 machine translation system, 94

 newspaper content transcription system, 104–106
 OOV, 95
 oracle error rate, 97
 source-channel model, 94
 spoken document retrieval systems, 94
 3-state HMM model, 95
 voice search, 94
 WER, 97
 WFSTs, 96
Average polarity, 264
Azerbaijani, 239
Azerbaijani-Turkish systems, 238

B

Back-off smoothing, 72
Bag-of-words (BoW), 257
Balkanet Consortium, 318
Balkanet Project, 17
Bayes' formula, 95
Bayesian word alignment, 210
Bigram stem model, 79
BILOU representation, 118
BLEU+ machine translation evaluation metric, 250
BLEU metric, 250
BLEU scores, 222, 226, 228, 229, 232
Booster words, 263
Bootstrapping algorithm, 127
BOUNCorpus, 309–310
British National Corpus (BNC), 296, 327

Brown Corpus, 296–297
Bulgarian, 317

C

Cambridge and Nottingham Corpus of Discourse in English (CANCODE), 297
Case-marked modifiers, 167
Catalan, 318
Center for Spoken Language Research (CSLR), 100
Centering theory, 355
CHILDES database, 155
Chinese, 318
Classifier-based parser, 141
Combinatory Categorical Grammar (CCG), 159–162
Comprehensive sentence, 190
Conditional random fields (CRFs), 63, 82, 128–130
Confusion constraint (CC), 81
CoNLL metric, 119
CoNLL representation, 144
CoNLL sentence, 280–283
CoNLL-U representation, 286–287
CoNLL XI Shared Task, 283
Consonant, morphophonology and morphographemics
 changes, 27
 deletion, 26
 devoicing, 26–27
 gemination, 27
 voicing, 26
Constituent reordering, 232
Constituent Structure (c-structures), 104, 176, 178, 196, 205
Coordination, 193–195
Copula, 30
Copular sentences, 190
Corpus, 294
Corpus linguistics, 294
Corpus of Contemporary American English (COCA), 298
Corpus of Professional Spoken American English (CPSA), 296
Crimean Tatar-Turkish machine translation systems, 238
Crossing alignments metric, 232–233
Cross-validation, 256
Czech, 317
Czech-to-English statistical machine translation, 209

D

Dangling nodes, 322
Deep neural networks (DNNs), 96
Dependency graph, 135–137, 141–143, 145, 163, 339
Dependency parsers, 104
Dependency parsing
 conditional probability model, 135
 data-driven statistical dependency parsing system
 evaluation metrics, 146
 methodology, 141–143
 modeling Turkish, 144–146
 formal languages, 134
 graph-based parsers, 136
 MaltParser, 135
 maximum likelihood estimation, 135
 morphological units, 135
 morphology and dependency relations, 136–141
 parsing algorithm, 135
 pseudo-deterministic approach, 134
 syntactic structure, 135
 transition-based parsers, 136
Dependency treebank, 163
Derivational boundaries (DB), 138
Derivational morphemes, 5, 6
Deterministic parsing algorithms, 141
Diachronic corpora, 297
Discourse Annotation Tool for Turkish (DATT), 343
Discriminative language models (DLMs), 73, 108–110
 ASR and MT systems, 82
 baseline acoustic and in-domain language models, 83
 components of, 82–83
 feature-based sequence modeling approach
 basic *n*-gram features, 84–85
 linguistically motivated features, 85–87
 statistically motivated features, 87–89
 gold-standard hypothesis, 84
 lowest-WER hypothesis, 84
 n-gram features, 83
 perceptron algorithm, 84
Distinctive feature (DF), 81
Dominant polarity, 265
Dutch, 318

E

Empty synsets, 322
English, SVO constituent order, 207

English tokens, 229
 EUbookshop, 234
 Europe Media Monitor (EMM), 126
 Euro Wordnet project, 320, 323, 332
 Extensible Markup Language for Discourse
 Analysis (EXMARaLDA), 306

F

Factored language models, 60
 Factored representation, 227
 Feasible morpheme pairs, 252
 Feature structures (f-structures), 176–178,
 181–187, 189–191, 194–197, 199,
 200, 202, 205
 Finite state model, 217
 Finite-state morphological analyzers, 177
 Fold-specific language model, 109
 Free word order, 75
 Freiburg-LOB Corpus of British English
 (FLOB), 297
 French, 318
 Full embedding, 345–346

G

Gaussian Mixture Models (GMMs), 96
 GenCor, 310
 General corpora, 296
 GNOME localization documents, 234
 Google Speech Recognition Service, 122
 Google Universal Part-of-Speech Tagset, 286
 Grammar-based approach, 344
 4-gram root language model, 230
 Grapheme-based acoustic models, 13
 Greedy Prepend Algorithm (GPA), 58
 Greek, 317, 318

H

Hand-crafted rules, 57
 Hebrew, 318
 Hidden Markov Models (HMMs), 59, 65, 95,
 96
 Hierarchical lexicon model, 209
 Hindi, 318
 Hyponyms, 322–324

I

IARPA Babel Turkish Language Pack
 (IARPA-babel105b-v0.5), 101
 IMDB internet movie database, 258
 IMST, 282

IMST-UD, 286
 Independent relations, 345
 Inflectional groups (IGs), 8–10, 79, 138, 166,
 167, 175, 275
 Information retrieval (IR), 13
 Intelligence Advanced Research Projects
 Activity (IARPA) Babel program,
 101
 Inter-Lingual Index (ILI), 320
 Internal dependency structure, 277
 International Corpus of English (ICE), 297
 International Corpus of Learner English
 (ICLE), 298
 Interpolation smoothing method, 72
 IOB2 representation, 117, 118
 Italian, 318
 ITU-METU-Sabancı Treebank, 282
 ITU validation set (IVS), 64, 283

J

Japanese, 318

K

Kazakh, 239
 KDE4 localization files, 234
 Korean National Corpus, 296
 Kurdish, 318
 Kyrgyz, 239

L

Lancaster/IBM Spoken English Corpus (SEC),
 297
 Lancaster-Oslo-Bergen Corpus of British
 English (LOB), 297
 Language modeling
 chain rule, 70
 DLMS, 73
 ASR and MT systems, 82
 baseline acoustic and in-domain
 language models, 83
 components of, 82–83
 feature-based sequence modeling
 approach, 83
 gold-standard hypothesis, 84
 lowest-WER hypothesis, 84
n-gram features, 83
 perceptron algorithm, 84
 equivalence classes, 70, 71
 factored language model, 72
 feature-based models, 72
 maximum entropy language model, 72

- MLE, 71
n-gram probabilities, 71, 72
 perplexity, 71
 smoothing technique, 72
 statistical language model, 69, 70
 challenges in, 73–75
 sub-lexical units, 75–78
 for Turkish, 78–81
 structured language models, 72
 Super ARV language models, 72
- Language resources
 acoustic and text data
 BN database, 102
 GlobalPhone project, 98, 99
 Hub4 BN transcription guidelines, 102
 LVCSR system, 101
 newspapers portals, 103
 NIST STM format, 102
 PCM format, 102
 read-speech data collection, 99
 SONIC tool, 100
 speech and text corpora, 101
 syntax, 104
 text-to-speech synthesis, 100
 transcripts and *n*-gram language models, 99
 triphone-balanced sentence set, 100
 linguistic tools, 103–104
- Large vocabulary continuous speech recognition (LVCSR), 12
- Latent Dirichlet Allocation (LDA) approach, 257
- Learner corpora, 298
- Lexemic lexicon extraction, *see* Word-based CCG lexicon extraction
- Lexical-Functional Grammar (LFG)
 formalism, 16, 56–57
 adjective phrases, 182
 adverbial phrases, 183
 causatives, 195–198
 CCG framework, 180
 coordination, 193–195
 c-structures, 178–180
 derivational suffixes, 178
 free-constituent order, 180
 f-structure representation, 177–178
 handling constituent order variations, 190–193
 head-driven phrase structure grammar, 180
 inflectional groups, 175
 lexical integrity tests, 179
 manual test sets, 203
 non-canonical objects, 200–202
 noun phrases, 181–182
 noun phrase test suite, 204
 noun-to-adjective derivational suffix, 179
 ParGram project, 176
 passive construction, 198–200
 postposition analysis, 183–184
 sentences, 190–191
 sentence test suite, 203–204
 sentential derivations (*see* Sentential derivations)
 temporal phrases, 184
 XLE, 176–177
- Lexicalized MWEs, 244
- Lexical morphemes, 36
- Lexical stem+ending model, 108
- Lexicons, 177
- “Lexico-semantic expansion” module, 334
- LibSVM package, 266
- Linguistica software, 77
- Linguistic corpora
 automatic morphological analysis, 300
 balance, 296
 Bank of English, 293
 Birmingham Corpus, 293
 British National Corpus, 293
 Brown Corpus, 293
 corpus linguistics, 294–298
 empirical support, 298
 external and internal criteria, 295
 grammatical studies, 299
 idealizations and abstractions, 291
 large-scale corpora, 300
 large-sized general linguistic corpora, 299
 machine-aided translation, 292
 meta-information, 298
 METU-Turkish Corpus, 301–303
 modern diachronic corpus, 293
 monitor corpus, 293
 natural language processing tasks, 292
 NLP corpora, 299
 preelectronic corpora, 299
 representativeness, 295
 small-sized specialized corpora, 299
 speech recognition, 292
 STC (*see* Spoken Turkish Corpus (STC))
 TNC, 303–305
- Linguistic sub-lexical units, Turkish language
 lexical form stem+ending model, 81
 surface form stem+ending model, 80
- London-Lund Corpus (LLC), 297
- M**
- Machine translation methodology
 Azerbaijani-Turkish systems, 238

- Catalan-Aranese Occitan, 237
- Crimean Tatar-Turkish machine translation systems, 238
- Czech-Lower Serbian and Macedonian, 237
- Czech-Russian, 237
- Irish-Scottish Gaelic, 237
- Spanish-Portuguese, 237
- Tatar-Bashkir machine translation, 238
- Turkic languages (*see* Turkic languages)
- Turkmen-Turkish machine translation system, 238
- MaltParser, 135
 - framework, 16
- Manual test sets, 203
- Maximum entropy, 260
- Maximum likelihood estimation (MLE), 71
- Maximum mutual information (MMI), 82
- Mel frequency cepstral coefficients (MFCC), 95
- Message understanding conferences (MUC), 116
- METEOR metric, 250, 251
- METEOR scores, 219
- METU-Sabancı Turkish Treebank (Insert Symbol), 64, 65, 282, 286, 301
- Michigan Corpus of Academic Spoken English (MICASE), 296
- Middle East Technical University (METU) Turkish Corpus, 301–303
- Minimum description length (MDL) principle, 77, 78
- Minimum edit distance (MED), 88
- Minimum phone error (MPE), 82
- 240 missing hypernyms, 326–327
- Monitor corpora, 298
- Morfessor algorithm, 78, 81
- Morph-based language model, 105
- Morpheme-based language models, 76–78
- Morphemes, 4, 5
- Morpheme segmentation approach, 210
 - baseline, 214
 - BLEU results, 215
 - full morphological segmentation, 214
 - GIZA++ tool, 216
 - lexical morphemes, 211
 - Moses toolkit, 213
 - observations on, 219–220
 - root+morphemes segmentation, 214
 - sample translations, 218
 - selective morphological segmentation, 214
 - SRILM language modelling toolkit, 213
 - Turkish Ministry of Foreign Affairs data, 212–213
 - word repair, 216–218
- Morpholexical language models, 81
- Morphological disambiguation (MD), 14
 - challenges, 55
 - data sets, 64
 - discriminative methods, 61–63
 - experimental results, 64–65
 - feature templates, 63
 - lexical morphemes, 54
 - POS, 53
 - rule-based methods
 - constraint-based morphological disambiguation, 56–57
 - constraints with voting, 57
 - inflectional group *n*-grams, 59–61
 - learning process, 57–59
- Morphological features and pronunciation, 36
 - adjectives, 48–49
 - major root parts of speech, 46
 - minor parts of speech, 46–47
 - morphological disambiguation, 51
 - nominal forms, 48
 - semantic markers, derivations, 50–51
 - verbs, 49–50
- Morphological representation, 250
- Morphology
 - ambiguity, 21
 - feature symbols, 22
 - lexical morphemes, 22
 - morphemes, 22
 - morphophonology and morphographemics, 23–27
 - morphotactics, 22
 - multiword processing
 - lexicalized collocations, 39–40
 - non-lexicalized collocations, 42–45
 - semi-lexicalized collocations, 40–42
 - orthography, 23
 - processing real texts
 - acronyms, 36–37
 - foreign words, 37–38
 - numbers, 37
 - unknown words, 38
 - root lexicons and morphotactics
 - derivations, 30–32
 - morphological analyses, 32–34
 - Morphological Processor, 34–36
 - nominal morphotactics, 29
 - representational convention, 28–29
 - verbal morphotactics, 30
 - segmented lexical morphographic representation, 23
 - surface morphemes, 22
 - wide-coverage parsing, 156–157

Morphophonology and morphographemics

- consonant
 - changes, 27
 - deletion, 26
 - devoicing, 26–27
 - gemination, 27
 - voicing, 26
- orthography, 24
- root-final plosives, 24
- velar consonants, 24
- vowel deletion, 26
- vowel harmony, 25–26

Moses toolkit, 15, 213

MTC, 301–303

MUC metric, 119

Multi-level alignment scheme, 210

Multi-word expressions (MWEs), 241, 244

Multiword processing

- lexicalized collocations, 39–40
- Non-lexicalized collocations, 42–45
- semi-lexicalized collocations, 40–42

MWE translation, 245

N

Naive Bayes approach, 260, 266

Named-entity recognition (NER), 14

- CRF-based systems, 117
- domain and datasets
 - formal texts, 120–121
 - informal texts, 121–122
- ENAMEX, 116, 120, 121
- evaluation of, 119–120
- hand-crafted rule-based systems, 125–126
- HMM-based NER system, 116
- hybrid approaches, 126–127
- language independent bootstrapping
 - algorithm, 116
- machine learning, 115, 117
- machine learning approaches, 127–130
- natural language processing tasks, 115
- NUMEX, 116, 120, 121
- preprocessing steps
 - morphological analysis, 123–124
 - normalization, 124
 - tokenization, 123
- representation scheme, 117–118
- Scratch approach, 117
- TIMEX, 116, 120, 121

Natural language processing (NLP), 261

neo-Firthians, 295

Neural network language models, 210

NewsCor, 310

NIST metric, 250

Nominal forms, 48

Non-canonical objects, 200–202

NooJ_TR module, 304

Normalization and morphology layer, 285

Noun+Adj group of transformations, 229

Noun+Adj+Verb+Adv+PostP transformations, 230, 231

Noun phrases, 181–182

Noun phrase test suite, 204

O

Object reordering (ObjR), 231

OpenSubtitles, 234

Opinion strength, 255

OrienTel Turkish database, 101

Orthography, 13

Out-of-vocabulary (OOV), 74–76, 79, 95, 104–107

P

Pair annotation, 341

Parallel Grammars Project, 17

ParGram grammars, 203

Partially overlapping arguments, 350

Part-of-speech (POS) tagging, 53

Passive sentence agent reordering (PassAgR), 231

Penn Discourse Treebank, 17

Perceptual Linear Prediction (PLP), 95

Persian, 318

Person-Number Agreement, 30

PHP manual, 234

Phrase-based back-off models, 209

Polarity, 30

Polarity lexicon, 256, 261–262

Polish National Corpus, 296

Postposition analysis, 183–184

Precision, 119, 170, 171, 230, 250

Princeton Wordnet, 318, 326, 327, 332, 334

Princeton Wordnet 1.5, 327

Princeton Wordnet 1.7.1, 327

Princeton Wordnet 2.0., 328, 333

Pronunciation modeling, 13

Properly contained argument, 348–349

Properly contained relation, 349–350

Pure crossing, 351–353

Q

Question sentences, 193

R

Raw representation, 118
 Real texts processing
 acronyms, 36–37
 foreign words, 37–38
 numbers, 37
 unknown words, 38
 Recall, 119, 170, 171, 233, 354
 Reference corpus, 296
 Relative clauses, 188–190
 Romanian, 317
 Root BLEU scores, 219
 Root lexicons and morphotactics
 derivations, 30–32
 morphological analyses, 32–34
 Morphological Processor, 34–36
 nominal morphotactics, 29
 representational convention, 28–29
 verbal morphotactics, 30
 Root matching strategy, 251
 Rule-cased approach, 57
 Russian, 318

S

Santa Barbara Corpus of Spoken American
 English (SBCSAE), 297
 Seed words, 262–263, 265–266
 Segmentation ambiguity, 5
 Semantics
 markers, derivations, 50–51
 wide-coverage parsing, 156–157
 Semi-Lexicalized MWEs, 244
 Sentence level rules, 249
 Sentence splitting, 242
 Sentence subjectivity, 259
 Sentence test suite, 203–204
 Sentential adjuncts, 187–188
 Sentential complement, 185–187
 Sentential derivations
 relative clauses, 188–190
 sentential adjuncts, 187–188
 sentential complements, 185–187
 SenticNet, 258
 Sentiment analysis
 automatic extraction, 255
 bag-of-words approach, 257
 BeyazPerde, 259
 classification accuracy, 268
 classification problem, 256
 cross-validation, 256
 data type, 257–258
 deep learning approaches, 257
 domain dependence, 258
 emotion analysis, 260
 feature efficacy, 268
 fuzzy-logic representation, 260
 LDA approach, 257
 learn from data, 256
 lexicon-based framework, 260
 lexicon effect, 268–269
 linguistic/lexicon-based, 256
 main difficulties, 260–261
 methodology, 263–266
 Naive Bayes, 267
 polarity lexicon, 256
 positive and negative movie reviews, 267
 regression problem, 256
 resources, 261–263
 rule-based methods, 258
 semantic orientation, 255
 SemEval, 259
 SentiWordNet, 259
 subjectivity, 259
 SVM classifiers, 258, 267
 train and validation, 256
 training data, 256
 Sentiment polarity, 255
 SentiTurkNet, 262, 265
 SentiWordNet, 258
 Serbian, 317
 SETIMES, 234
 Spanish, 318
 Specialized corpora, 296
 Spoken corpora, 297
 Spoken Turkish Corpus (STC), 297
 BOUNCORPUS, 309–310
 corpus-driven approach, 307
 EXMARaLDA, 306
 EXMARaLDA's Partitur Editor, 306
 metadata annotation, 306–307
 NLP corpora, 309
 text type, 306
 Turkish Discourse Bank's style of
 annotation, 308
 TurkishWaC, 309–310
 WebCorp, 310
 SRILM language modelling toolkit, 213
 SRILM toolkit, 105
 Standard coordination, 193, 194
 Statistical language model (SLM), 69, 70,
 248–249
 challenges in, 73–75
 sub-lexical units
 linguistic units, 76–77
 statistical units, 77–78
 for Turkish
 linguistic sub-lexical units, 79–81

- statistical sub-lexical units, 81
- Statistical machine translation (SMT), 14–16, 73
 - building machine translation systems, 208
 - handling morphology, 209–210
 - inflected forms, 208
 - morpheme segmentation approach (*see* Morpheme segmentation approach)
 - syntax-to-morphology mapping approach (*see* Syntax-to-morphology mapping approach)
- Stem+ending model, 79, 80
- Stem error rate (SER), 108
- Subject object verb (SOV), 240
- Sub-lexical language model, 13
- Subordinate clause reordering (SubCR), 231
- Support vector machines (SVM), 111, 143, 258, 260
 - classifier, 266
- Surface morphemes, 34
- Synchronic corpora, 297
- Synonym sets, 318
- Syntactic layer, 285
- Syntactic transformations, 210
- Syntax, 86, 104
- Syntax-to-morphology mapping approach, 210
 - applying transformations, 228–230
 - baseline system, 226–228
 - BLEU scores, 226
 - constituent reordering, 231–233
 - English prepositional phrase transformation, 221
 - higher-order language models, 230
 - MaltParser, 224
 - mapping source-side syntax to target-side morphology, 221–224
 - Moses toolkit, 226
 - multiple transformations, 224, 225
 - Penn Treebank Tagset, 224
 - phrase extraction algorithms, 222
 - TreeTagger, 224
 - Turkish inflectional morphology, 220
- Syntax-to-morphology scheme, 15

- T**
- Tanzil Project, 234
- Tarama Sözlüğü*, 300
- Tatar-Bashkir machine translation, 238
- Tatoeba, 234
- TDB 1.0, 341
- Temporal phrase subgrammar, 184
- Tense-Aspect-Mood, 30
- TER metric, 250
- Text-to-Speech (TTS) systems, 14
 - 1228 additional synsets, 327
- TNC v3.0, 305
- Tokenization tools, 242
- Tokenizers, 177
- Traditional *n*-gram language models, 11
- Training data, 256
- Treebank sentence, 278–280
- Treebank tag set, 275
- Trigram tag model, 59
- TripAdvisor website, 257
- Turkic languages
 - Azerbaijani, 239
 - example sentences in, 241
 - family tree of, 238–239
 - geographical and political reasons, 239
 - geographical map, 240
 - intense Russian influence, 238
 - Kazakh, 239
 - Kyrgyz, 239
 - machine translation
 - direct translation, 242
 - evaluation, 250–253
 - lexical transfer, 246–248
 - morphological disambiguation module, 245–246
 - morphological feature transfer, 246
 - morphological representation, 250
 - preprocessing, 242, 244–245
 - sentence level rules, 249
 - statistical disambiguation module, 248–249
 - system architecture, 243, 244
 - past definitive tense, 240
 - Russian and Arabic influence, 242
 - SL MWE, 242
 - syntactical structure, 241
 - Turkmen, 239
 - Uyghur, 239
 - Uzbek, 239
- TurkIE dataset, 121
- Turkish Broadcast News (BN), 102, 107
- Turkish Discourse Bank (TDB)
 - annotation scheme
 - abstract objects, 340
 - allomorph inflections, 343
 - average *K* values, 339
 - DATT, 343
 - discourse connectives, 339
 - inconsistencies, 341
 - inter-coder agreement, 339
 - minimality principle, 339
 - pair annotation, 341
 - phrasal expressions, 340

- TDB 1.0., 340
 - centering theory, 355
 - connectives and discourse structure, 343–344
 - discourse relations, 337–338
 - full embedding, 345–346
 - independent relations, 345
 - nested relations, 346–347
 - non-discourse-level phenomena, 354
 - partially overlapping arguments, 350
 - properly contained argument, 348–349
 - properly contained relation, 349–350
 - pure crossing, 351–353
 - rhetorical structure trees, 339
 - shared arguments, 347–348
 - Turkish–English parallel corpus, 208
 - Turkish language
 - applications
 - ASR/STT systems, 12–13
 - pronunciation modeling, 12
 - speech retrieval systems, 13–14
 - speech synthesis/TTS, 14
 - spelling checking and correction, 11
 - statistical language modeling, 11–12
 - statistical machine translation, 14–16
 - syntactic modeling, 11
 - tagset design, 11
 - constituent order and morphology-syntax interface, 7–10
 - geography of, 2
 - morphology, 3–7
 - morphophonology, 1, 4
 - orthography and pronunciation, 3
 - sentiment analysis (*see* Sentiment analysis)
 - speakers of, 2
 - state-of-the-art tools and resources for
 - discourse bank, 17
 - LFG-based parser, 16–17
 - miscellaneous corpora and resources, 17
 - morphological analysis, 16
 - morphological disambiguation, 16
 - statistical dependency parsers, 16
 - treebank, 17
 - WordNet, 17
 - syntax, 1
 - Turkish Language Institute (TDK), 300
 - Turkish National Corpus (TNC), 296, 297, 303–305
 - Turkish Radio and Television (TRT), 122
 - Turkish treebank
 - annotation tool, 284–286
 - evolution of
 - branches of, 282–283
 - CoNLL format, 280–282
 - ITUWeb treebank, 283–284
 - lexical items, 274
 - morphological information, 274–276
 - parse and IG statistics, 276
 - Penn Treebank, 273
 - syntactic relations, 276–278
 - treebank sentence, 278–280
 - Turkish Universal Dependencies Treebank, 286–287
 - TurkishWaC, 309–310
 - TurkishWordnet
 - applications of, 331–334
 - Balkanet project, 320
 - CATEGORY_DOMAIN, 319
 - coverage tests, 330–331
 - current status of, 329–330
 - design decisions
 - dangling nodes, 322
 - definitions, 321–322
 - lexical gaps, 322
 - merge *vs.* expand, 321
 - parts-of-speech, 321–322
 - sense numbers, 321–322
 - validating semantic relations, 323
 - development process
 - Balkanet-specific concepts, 328–329
 - first set of concepts, 323
 - hypernyms, 325–326
 - near-antonyms, 326
 - Princeton Wordnet 1.5, 327
 - Princeton Wordnet 1.7.1, 327
 - Princeton Wordnet 2.0, 328
 - second set of concepts, 326–327
 - synonyms, 324
 - third set of concepts, 328
 - Euro Wordnet project, 320
 - HOLO_PART, 319
 - HYPERNYM relation, 319
 - languages, 317
 - NEAR_ANTONYM, 319
 - Princeton Wordnet, 318
 - quality validation, 330–331
 - semantic relations, 318–319
 - synonym sets, 318
 - Turkmen, 239
 - Turkmen-Turkish machine translation system, 238
 - Two-level morphology approach, 21, 154
 - Type raising, 160
- U**
- Ubuntu localization documents, 234

Unification-based grammar formalism, 180
 Universal Dependencies (UD), 286
 Unlabeled attachment score (ASU), 146
 Unsupervised learning process, 57
 Uyghur, 239
 Uzbek, 239

V

Valency alternations
 causatives, 195–198
 passive construction, 198–200
 Verb complex transformation, 223
 Verbs, 49–50
 Voice of America (VOA), 102
 Vowel harmony, 4, 25–26, 211

W

Web 2.0., 283
 WebCorp, 310
 Web Inventory of Transcribed and Translated
 Talks (WIT), 234
 Weighted finite state transducer (WFST), 96,
 108
 Wellington Corpus of Spoken New Zealand
 English (WSC), 297
 WER metric, 250
 wh-question sentences, 192
 Wide-coverage parsing
 automatically induced CCG lexicons,
 169–171

BOUN corpus, 155
 CCG, 159–162
 CHILDES database, 155
 morphemes, 154
 morphology and semantics, 156–157
 radical lexicalization and predicate-
 argument structure, sub-lexical
 elements, 157–158
 Turkish categorial lexicon
 CCG categories, 163
 dependency treebank, 163
 lexemic model, 164–166
 morphemic model, 166–169
 Word-based CCG lexicon extraction, 164
 Word-based language model, 79, 217
 Word-based lexicon induction algorithm, 166
 Word-based sentence model, 139
 Word error rate (WER), 97, 106, 108, 110, 111
 WordNET, 17, 251, 253
 Word sense disambiguation (WSD), 14
 Written corpora, 296–297

X

XCES style annotation, 302
 Xerox Finite State Tools, 34
 Xerox Linguistic Environment (XLE),
 176–177

Y

Yes/no questions, 193