

# Contents

<b>1</b>	<b>Turkish and its Challenges for Language and Speech Processing</b> . . . .	<b>1</b>
	Kemal Oflazer and Murat Saraçlar	
1.1	Introduction . . . . .	1
1.2	Turkish Morphology . . . . .	3
1.3	Constituent Order and Morphology-Syntax Interface . . . . .	7
1.4	Applications . . . . .	10
1.5	State-of-the-art Tools and Resources for Turkish . . . . .	15
	References . . . . .	17
<b>2</b>	<b>Morphological Processing for Turkish</b> . . . . .	<b>21</b>
	Kemal Oflazer	
2.1	Introduction . . . . .	21
2.2	Overview of Turkish Morphology . . . . .	22
2.3	Morphophonology and Morphographemics . . . . .	23
2.4	Root Lexicons and Morphotactics . . . . .	27
2.4.1	Representational Convention . . . . .	28
2.4.2	Nominal Morphotactics . . . . .	29
2.4.3	Verbal Morphotactics . . . . .	29
2.4.4	Derivations . . . . .	30
2.4.5	Examples of Morphological Analyses . . . . .	32
2.5	The Architecture of the Turkish Morphological Processor . . . . .	34
2.6	Processing Real Texts . . . . .	35
2.6.1	Acronyms . . . . .	35
2.6.2	Numbers . . . . .	36
2.6.3	Foreign Words . . . . .	36
2.6.4	Unknown Words . . . . .	36
2.7	Multiword Processing . . . . .	37
2.7.1	Lexicalized Collocations . . . . .	38
2.7.2	Semi-lexicalized Collocations . . . . .	38
2.7.3	Non-lexicalized Collocations . . . . .	40
2.8	Conclusions . . . . .	44

References .....	50
<b>3 Morphological Disambiguation for Turkish</b> .....	<b>53</b>
Dilek Zeynep Hakkani-Tür, Murat Saraçlar, Gökhan Tür, Kemal Oflazer and Deniz Yuret	
3.1 Introduction .....	53
3.2 Challenges .....	55
3.3 Previous Work .....	55
3.3.1 Rule-based Methods .....	56
3.3.2 Learning the Rules .....	57
3.3.3 Models Based on Inflectional Group <i>n</i> -grams .....	59
3.3.4 Discriminative Methods for Disambiguation .....	60
3.4 Discussion .....	63
3.4.1 Data Sets .....	63
3.4.2 Experimental Results .....	64
3.5 Conclusions .....	65
References .....	65
<b>4 Language Modeling for Turkish Text and Speech Processing</b> .....	<b>69</b>
Ebru Arısoy and Murat Saraçlar	
4.1 Introduction .....	69
4.2 Language Modeling .....	70
4.3 Challenges in Statistical Language Modeling for Turkish .....	73
4.4 Sub-lexical Units for Statistical Language Modeling .....	75
4.4.1 Linguistic Sub-lexical Units .....	76
4.4.2 Statistical Sub-lexical Units .....	77
4.5 Statistical Language Modeling for Turkish .....	78
4.5.1 Language Modeling with Linguistic Sub-lexical Units ...	78
4.5.2 Statistical Sub-lexical Units – Morphs .....	81
4.6 Discriminative Language Modeling for Turkish .....	81
4.6.1 Discriminative Language Model .....	82
4.6.2 Feature Sets for Turkish DLM .....	83
4.7 Conclusions .....	89
References .....	89
<b>5 Turkish Speech Recognition</b> .....	<b>95</b>
Ebru Arısoy and Murat Saraçlar	
5.1 Introduction .....	95
5.2 Foundations of Automatic Speech Recognition .....	96
5.3 Turkish Language Resources for ASR .....	100
5.3.1 Turkish Acoustic and Text Data .....	100
5.3.2 Linguistic Tools Used in Turkish ASR .....	105
5.4 Turkish ASR Systems .....	106
5.4.1 Newspaper Content Transcription System .....	106
5.4.2 Turkish Broadcast News Transcription System .....	108
5.4.3 LVCSR System for Call Center Conversations .....	112

5.5	Conclusions . . . . .	113
	References . . . . .	114
<b>6</b>	<b>Turkish Named Entity Recognition . . . . .</b>	<b>119</b>
	Reyyan Yeniterzi, Gökhan Tür and Kemal Oflazer	
6.1	Introduction . . . . .	119
6.2	NER on Turkish . . . . .	120
6.3	Task Description . . . . .	121
6.3.1	Representation . . . . .	121
6.3.2	Evaluating NER Performance . . . . .	122
6.4	Domain and Datasets . . . . .	124
6.4.1	Formal Texts . . . . .	124
6.4.2	Informal Texts . . . . .	125
6.4.3	Challenges of Informal Texts for NER . . . . .	126
6.5	Preprocessing for NER . . . . .	126
6.5.1	Tokenization . . . . .	127
6.5.2	Morphological Analysis . . . . .	127
6.5.3	Normalization . . . . .	127
6.6	Approaches used in Turkish NER . . . . .	128
6.6.1	Rule-based Approaches . . . . .	129
6.6.2	Hybrid Approaches . . . . .	130
6.6.3	Machine Learning Approaches . . . . .	131
6.7	Conclusions . . . . .	134
	References . . . . .	134
<b>7</b>	<b>Dependency Parsing of Turkish . . . . .</b>	<b>137</b>
	Gülşen Eryiğit, Joakim Nivre and Kemal Oflazer	
7.1	Introduction . . . . .	137
7.2	Dependency Parsing . . . . .	139
7.3	Morphology and Dependency Relations in Turkish . . . . .	140
7.3.1	Dependency Relations in Turkish . . . . .	143
7.4	An Incremental Data-driven Statistical Dependency Parsing System . . . . .	144
7.4.1	Methodology . . . . .	145
7.4.2	Modeling Turkish . . . . .	147
7.4.3	Evaluation Metrics . . . . .	150
7.5	Related Work . . . . .	151
7.6	Conclusions . . . . .	151
	References . . . . .	152
<b>8</b>	<b>Wide-coverage parsing, semantics and morphology . . . . .</b>	<b>157</b>
	Ruket Çakıcı, Mark Steedman and Cem Bozşahin	
8.1	Introduction . . . . .	157
8.2	Morphology and Semantics . . . . .	160
8.3	Radical Lexicalization and Predicate-Argument Structure of sub-lexical Elements . . . . .	161

8.4	Combinatory Categorical Grammar: CCG .....	162
8.5	The Turkish Categorical Lexicon .....	166
8.5.1	The Lexemic Model .....	168
8.5.2	The Morphemic Model .....	170
8.6	Parsing with Automatically Induced CCG Lexicons .....	172
8.7	Conclusion .....	174
	References .....	175
<b>9</b>	<b>Deep Parsing of Turkish with Lexical-Functional Grammar .....</b>	<b>179</b>
	Özlem Çetinoğlu and Kemal Oflazer	
9.1	Introduction .....	179
9.2	Lexical-Functional Grammar and Xerox Linguistic Environment ..	180
9.3	Inflectional Groups as First-class Syntactic Citizens .....	181
9.4	Previous Work .....	184
9.5	LFG Analyses of Various Linguistic Phenomena .....	185
9.5.1	Noun Phrases .....	185
9.5.2	Adjective Phrases .....	186
9.5.3	Adverbial Phrases .....	187
9.5.4	Postpositional Phrases .....	187
9.5.5	Temporal Phrases .....	188
9.6	Sentential Derivations, Sentences and Free Constituent Order .....	189
9.6.1	Sentential Derivations .....	189
9.6.2	Sentences .....	194
9.6.3	Handling Constituent Order Variations .....	195
9.7	Coordination .....	198
9.8	Valency Alternations .....	199
9.8.1	Causatives .....	199
9.8.2	Passives .....	202
9.9	Non-canonical Objects .....	204
9.10	Evaluation .....	206
9.10.1	Manual Test Sets .....	207
9.10.2	Sentence Test Suite .....	207
9.10.3	Noun Phrase Test Suite .....	208
9.11	Conclusions .....	208
	References .....	209
<b>10</b>	<b>Statistical Machine Translation and Turkish .....</b>	<b>213</b>
	Kemal Oflazer, Reyhan Yeniterzi, and İlknur Durgar-El Kahlout	
10.1	Introduction .....	213
10.2	Handling Morphology in Statistical Machine Translation .....	215
10.3	The Morpheme Segmentation Approach .....	216
10.3.1	Experiments and Results .....	219
10.3.2	Word Repair .....	222
10.3.3	Sample Translations .....	223
10.3.4	Observations on the Morpheme Segmentation Approach ..	224
10.4	The Syntax-to-Morphology Mapping Approach .....	225

10.4.1	Mapping Source-side Syntax to Target-side Morphology . . . . .	226
10.4.2	Experimental Setup and Results . . . . .	230
10.4.3	Experiments with Constituent Reordering . . . . .	237
10.5	Conclusions . . . . .	239
	References . . . . .	241
<b>11</b>	<b>Machine Translation Between Turkic Languages . . . . .</b>	<b>245</b>
	A. Cüneyd Tantuğ and Eşref Adalı	
11.1	Introduction . . . . .	245
11.2	Turkic Languages . . . . .	246
11.2.1	Similarities and Differences of Turkic Languages . . . . .	246
11.3	Machine Translation between Turkic Languages . . . . .	250
11.3.1	Preprocessing . . . . .	252
11.3.2	Morphological Disambiguation . . . . .	254
11.3.3	Morphological Feature Transfer . . . . .	254
11.3.4	Lexical Transfer . . . . .	254
11.3.5	Statistical Disambiguation Module . . . . .	255
11.3.6	Sentence Level Rules . . . . .	257
11.3.7	Morphological Generation . . . . .	258
11.4	Machine Translation Evaluation on Turkic Languages . . . . .	258
11.4.1	Root Matching . . . . .	259
11.4.2	Feasible Suffix Pairs . . . . .	260
11.5	Conclusions . . . . .	261
	References . . . . .	261
<b>12</b>	<b>Sentiment Analysis in Turkish . . . . .</b>	<b>265</b>
	Gizem Gezici and Berrin Yanıkoğlu	
12.1	Introduction . . . . .	265
12.2	Related Work . . . . .	268
12.3	Main Difficulties for Turkish Sentiment Analysis . . . . .	270
12.4	Practical Sentiment Analysis for Turkish . . . . .	271
12.4.1	Resources . . . . .	271
12.4.2	Methodology . . . . .	273
12.5	Experimental Evaluation . . . . .	276
12.5.1	Data . . . . .	276
12.5.2	Results . . . . .	277
12.6	Conclusions . . . . .	278
	References . . . . .	279
<b>13</b>	<b>The Turkish Treebank . . . . .</b>	<b>283</b>
	Gülşen Eryiğit, Kemal Oflazer, and Umut Sulubacak	
13.1	Introduction . . . . .	283
13.2	What information needs to be represented? . . . . .	284
13.2.1	Representing Morphological Information . . . . .	284
13.2.2	Representing Syntactic Relations . . . . .	286
13.2.3	Example of a Treebank Sentence . . . . .	288

13.3	Evolution of the Turkish Treebank .....	290
13.3.1	The CoNLL Format .....	290
13.3.2	Branches of the Turkish Treebank .....	292
13.4	The ITU Web Treebank .....	293
13.5	The Annotation Tool .....	294
13.6	The Turkish Universal Dependencies Treebank .....	296
13.7	Conclusions .....	297
	References .....	298
<b>14</b>	<b>Linguistic corpora: A view from Turkish</b> .....	<b>301</b>
	Mustafa Aksan and Yeşim Aksan	
14.1	Introduction .....	301
14.2	Brief History of Corpus Linguistics .....	302
14.3	Linguistic Corpora and Corpus Linguistics .....	304
14.4	Use of Corpora in Linguistics .....	308
14.5	Turkish Linguistic Corpora .....	309
14.5.1	METU-Turkish Corpus .....	310
14.5.2	Turkish National Corpus (TNC) .....	313
14.5.3	Spoken Turkish Corpus (STC) .....	315
14.6	Conclusions .....	320
	References .....	321
<b>15</b>	<b>Turkish Wordnet</b> .....	<b>327</b>
	Özlem Çetinoğlu, Orhan Bilgin and Kemal Ofazer	
15.1	Introduction .....	327
15.2	Basic Structure of Turkish Wordnet .....	328
15.2.1	Semantic Relations .....	328
15.2.2	Linking Wordnets to Each Other .....	329
15.3	Design Decisions .....	330
15.3.1	Merge vs. Expand .....	331
15.3.2	Parts-of-Speech, Definitions and Sense Numbers .....	331
15.3.3	Lexical Gaps .....	332
15.3.4	No Dangling Nodes or Relations .....	332
15.3.5	Validating Semantic Relations .....	333
15.4	The Development Process .....	333
15.4.1	First Set of Concepts (Subset I) .....	333
15.4.2	Extracting Semantic Relations from Monolingual Resources .....	334
15.4.3	Second Set of Concepts (Subset II) .....	336
15.4.4	Shifting to Princeton Wordnet 1.7.1 .....	337
15.4.5	Third Set of Concepts (Subset III) .....	338
15.4.6	Shifting to Princeton Wordnet 2.0 .....	338
15.4.7	Adding Balkanet-specific Concepts .....	338
15.4.8	Final Expansion .....	339
15.5	Current Status of Turkish Wordnet .....	339
15.6	Quality Validation and Coverage Tests .....	340

15.7	Applications of Turkish Wordnet .....	342
15.7.1	Capturing Semantic Relations through Morphology .....	342
15.7.2	Turkish Wordnet in Use .....	344
15.8	Conclusion and Directions for Future Work .....	345
	References .....	345
<b>16</b>	<b>Turkish Discourse Bank: Connectives and Their Configurations</b> .....	<b>349</b>
	Deniz Zeyrek, Işın Demirşahin, Cem Bozşahin	
16.1	Introduction .....	349
16.2	The TDB Annotation Cycle .....	351
16.2.1	Major Sources of Disagreements among Annotators .....	353
16.2.2	The Discourse Annotation Tool for Turkish .....	355
16.3	Connectives and Discourse Structure .....	355
16.4	Discourse relation configurations in the TDB .....	356
16.4.1	Independent Relations .....	357
16.4.2	Full Embedding .....	358
16.4.3	Nested Relations .....	358
16.4.4	Shared Argument .....	360
16.4.5	Properly Contained Argument .....	360
16.4.6	Properly Contained Relation .....	362
16.4.7	Partially Overlapping Arguments .....	362
16.4.8	Pure Crossing .....	364
16.5	Results and Conclusions .....	366
	References .....	368