

Inhalt

Vorwort	9
1 Einleitung	11
1.1 Über dieses Buch	11
1.2 Korpus- und Computerlinguistik	14
1.3 Grundbegriffe	16
Teil 1: Linguistische Ausgangspunkte	21
2 Lexik	23
2.1 Das Wort	23
2.2 Tokenisierung	25
2.3 Lemmatisierung	27
2.4 Der Wortschatz von Korpora	28
2.5 Kollokationen	32
2.6 Keywords	36
2.7 Beispielstudien	37
2.8 Übungen	39
3 Wortarten	41
3.1 Wortarten in der Linguistik	41
3.2 Wortarten annotieren	44
3.3 Automatisches POS-Tagging	47
3.4 Beispielstudien	51
3.5 Übungen	54
4 Syntax	55
4.1 Konstituentengrammatik	55
4.2 Dependenzgrammatik	59
4.3 Computerbasierte Syntaxanalyse	63
4.4 Beispielstudien	68
4.5 Übungen	69
5 Semantik: Wortfelder	71
5.1 Semantik: Linguistische Grundlagen	71
5.2 Wortfelder	74

5.3	Beispielstudien	77
5.4	Übungen	79
6	Semantik: Sentimentanalyse	81
6.1	Bewertungen in Texten	81
6.2	Lexikonbasierte Sentimentanalyse	82
6.3	Sentimentanalyse mit maschinellem Lernen	87
6.4	Emotionsanalyse	88
6.5	Beispielstudien	90
6.6	Übungen	92
7	Semantik: Distributionelle Semantik	93
7.1	Grundlagen	93
7.2	Ähnlichkeiten berechnen	97
7.3	Word Embeddings	99
7.3.1	Spärliche vs. dichte Repräsentation	99
7.3.2	Word Embeddings berechnen	101
7.3.3	Statische und dynamische Embeddings	103
7.3.4	Mit Word Embeddings arbeiten	104
7.3.5	Evaluation	105
7.4	Beispielstudien	106
7.5	Übungen	108
8	Pragmatik: Referenz	109
8.1	Entitäten und Referenz	109
8.2	Named Entity Recognition	110
8.3	Koreferenz	113
8.4	Beispielstudien	119
8.5	Übungen	122
	Teil 2: Methoden	123
9	Korpussuche und -statistik	125
9.1	Reguläre Ausdrücke	125
9.2	Absolute und relative Frequenzen	128
9.3	Deskriptive Statistik	130
9.4	Visualisierung	134
9.5	Inferenzstatistik	138
9.6	Übungen	141
10	Manuelle Annotation	143
10.1	Manuelle und automatische Annotation	143

10.2	Annotationsrichtlinien	144
10.3	Qualität manueller Annotationen prüfen	147
10.4	Tools zur manuellen Annotation	152
10.5	Übungen	154
11	Maschinelles Lernen	157
11.1	Maschinelles Lernen, künstliche Intelligenz & Co.	157
11.2	Überwachtes und unüberwachtes Lernen	159
11.3	Musterablauf einer Klassifikation	163
11.3.1	Trainingsdaten	163
11.3.2	Merkmale	165
11.3.3	Lernverfahren	166
11.3.4	Evaluation	171
11.4	Übungen	176
12	Deep Learning	179
12.1	Grundlagen	179
12.2	Aufbau eines Deep-Learning-Modells	180
12.3	Training eines Deep-Learning-Modells	183
12.4	Word Embeddings	186
12.5	Recurrent Neural Networks	187
12.6	Transformer	188
12.7	Mit Deep Learning arbeiten	192
12.8	Übungen	192
Teil 3: Gesellschaft		195
13	Computerlinguistik und Ethik	197
13.1	Einführung	197
13.2	Dual Use	198
13.3	Bias und Diskriminierung	200
13.3.1	Beispiele für Bias	201
13.3.2	Ursachen von Bias	203
13.4	Ressourcenverbrauch	206
13.5	Repräsentation	207
Ressourcenverzeichnis		209
Literaturverzeichnis		219
Sachregister		237