INHALT

1. Einleitung ... 13
1.1 Was ist Part-of-Speech-Tagging? ... 13
1.2 Inhalt und Aufbau der Untersuchung ... 16
1.3 Datengrundlage FOLK ... 18
1.4 Related Work ... 20
1.4.1 Deutsche getaggte Korpora und ihre Unterschiede zum FOLK-Korpus ... 21
1.4.2 Referenzkorpora – Korpora für gesprochene und geschriebene Sprache ... 23
1.4.3 Korpora für gesprochene Sprache ... 28
1.4.4 Zwischenfazit ... 32
1.5 Pilotstudie ... 33
1.6 Zielsetzung der Untersuchung ... 35

2. Theoretische Grundlagen ... 37
2.1 Grammatik der gesprochenen Sprache ... 37
2.2 Segmentierung von Transkripten gesprochener Sprache ... 50
2.3 Kontext und Multimodalität in der Face-to-Face-Interaktion ... 57
2.4 Wortarten in der gesprochenen Sprache ... 68
2.4.1 Gesprächspartikeln ... 68
2.4.2 Exkurs: topologisches Satzmodell ... 70
2.4.3 Exkurs: Umgang mit Mehrworteinheiten ... 71
2.4.4 Problematik verschiedener Definitionen von Partikeln ... 71
2.4.5 Gesprächswörter – Definitionen in der Literatur ... 72
2.4.6 Satz-interne Partikeln ... 83
2.4.7 Satz-unabhängige Partikeln ... 100
2.4.8 Satz-externe Elemente ... 127
2.5 Weitere Klassifikationsprobleme ... 177
2.5.1 Medialitätsübergreifende Abgrenzungsprobleme ... 180
2.5.2 Probleme beim Taggen spezifisch gesprochensprachlicher Phänomene ... 226
2.5.3 Zwischenfazit ... 243

3. Empirischer Teil ... 245
3.1 Zielsetzung der empirischen Arbeit ... 245

3.2 Überblick über die empirische Vorgehensweise ... 246
3.3 Auswahl des Taggers und des Tagsets ... 247
3.4 Funktionsweise des Taggers ... 250
3.5 Möglichkeiten der Anpassung des Taggers und des Tagsets ... 252
3.6 Erstellen des Goldstandards ... 257
3.6.1 Kriterien für die Auswahl der Transkripte des Goldstandards ... 258
3.6.2 Darstellung der Transkripte des Goldstandards ... 261
3.6.3 Sub-Sets des Goldstandards ... 276
3.7 Erste Anpassung des Tagsets und der Guidelines ... 278
3.7.1 Das STTS – Aufbau des Tagsets und der Guidelines ... 279
3.7.2 Die Anwendung des STTS für Transkripte gesprochener Sprache – eine Problemanalyse ... 282
3.7.3 Grundsätze für eine Anpassung des STTS und der Guidelines ... 287
3.7.4 Erste Änderungen am Tagset und an den Guidelines ... 289
3.7.5 Das STTS 2.0 ... 301
3.8 Manuelle Korrektur des Development-Sets ... 303
3.8.1 OrthoNormal, das Tool zur manuellen Korrektur des POS-Taggings ... 303
3.8.2 Annotator*innen und Annotationsprozess ... 306
3.9 Inter-Annotator-Agreement zur manuellen Korrektur des Development-Sets ... 307
3.9.1 Voraussetzungen und Vorgehen ... 310
3.9.2 Ergebnisse des ersten Inter-Annotator-Agreements ... 313
3.10 Einführung eines Post-Processings ... 319
3.11 Zweite Anpassung der Guidelines ... 321
3.12 Entwicklung eines automatisierten Taggings anhand des Development-Sets ... 323
3.12.1 Erstellen des Lexikons ... 323
3.12.2 Neutraining mit Development-Set und Lexikon ... 325
3.12.3 Auswertung ... 327
3.13 Manuelle Korrektur des Goldstandard-Sets ... 329
3.13.1 Inter-Annotator-Agreement zur manuellen Korrektur des Goldstandard-Sets ... 330
3.13.2 Endkorrektur des Goldstandards ... 331
3.14 Entwicklung eines automatisierten Taggings anhand des Goldstandards ... 336
3.14.1 Trainings-Set und Evaluations-Sets ... 336

3.14.2 Segmentierung der Daten anhand von Pausenlänge und Ausschluss von mit Dummys markierten Wortformen 340
3.15 Evaluation des POS-Taggings für spontansprachliche Daten 342
3.15.1 Ergebnisse der automatisierten Annotationen 342
3.15.2 Analyse der Annotationsdifferenzen 345
3.15.3 Ambiguitäten 356
3.16 Fazit 359

4. Abschließende Diskussion und Ausblick 361

5. Literatur 367

6. Anhang 391
6.1 Transkriptionskonventionen 391
6.1.1 Transkriptionskonventionen nach GAT 2 (Selting et al. 2009) 391
6.1.2 Multimodale Konventionen (Kurzversion) 392
6.2 STTS Tag table (1995/1998) 393
6.3 Transkripte des Goldstandards 395
6.4 Heatmap-Plots der Annotationsdifferenzen 406
6.5 Plots für Annotationsunterschiede einzelner Tags 414