

Annotation von Texten des gesprochenen Italienisch nach dem Standard der Text Encoding Initiative (TEI)

ROSSMANN Barbara
9912246

Zusammenfassung

Die Aufgabenstellung dieser Diplomarbeit bestand darin, Texte aus dem gesprochenen Italienisch nach den Richtlinien der *Text Encoding Initiative* (TEI) zu annotieren. Diese so genannten *TEI-Guidelines* basieren auf der Markupsprache *Extensible Markup Language* (XML) und bilden eine internationale Basis zur Annotation von linguistischen Daten. Die TEI hat sich bereits in den Neunziger Jahren zum Ziel gemacht, den internationalen Datenaustausch zu fördern, um die Forschung in der Sprachwissenschaft zu unterstützen. Die Texte, die dieser Arbeit zu Grunde liegen, stammen aus der Datenbank *Banca dati dell'italiano parlato* (BADIP), die über den Languageserver der Karl-Franzens-Universität Graz abgerufen werden kann (<http://languageserver.uni-graz.at/badip>).

Die beiden Begriffe *Annotation* und *Markup* werden in dieser Arbeit synonym verwendet und bezeichnen das Explizitmachen von Informationen, die den Inhalt eines Textes strukturieren. Für den Menschen sind diese Informationen implizit klar. Erst durch so genannte *Tags* werden sie auch für einen Computer les- und verarbeitbar gemacht. Um das Prinzip der Textannotation besser verstehen zu können, werden einleitend Begriffe wie *Elektronischer Text* und *Texttechnologie* näher erläutert. Sie bilden die Grundlage für das computergestützte Arbeiten mit linguistischen Korpora, im Falle der vorliegenden Diplomarbeit besonders Korpora gesprochener Sprache.

Einen der Schwerpunkte dieser Arbeit stellt die Einführung in die Grammatik von XML dar. Auf diese aufbauend wird das Konzept der TEI Guidelines, in der sich aktuell in Entwicklung befindlichen Version P5, vorgestellt. Anhand von Textausschnitten der Datenbank BADIP wurde letztendlich ein TEI-konformes Annotationsmuster erstellt, das einer Konvertierung der derzeitigen XML-Annotation dienen soll. Der mit Hilfe des XML-Editors Oxygen erstellte Strukturbaum wird im Text dargestellt und Schritt für Schritt erläutert.

Der Diplomarbeit ist eine CD-ROM beigelegt, die die erstellte xml-Datei, ein dazugehöriges im Schema-Editor ROMA generiertes Schema, XSLT-Stylesheets und digitale Literaturquellen, sowie die Arbeit selbst in pdf-Format beinhaltet.