

Erweiterte Standoff-Annotationen zur Verwaltung und Analyse mehr-dimensional annotierter Korpora

Maik Stührenberg

18.11.2009

Linguistische Korpora sind ein unentbehrliches Werkzeug für die analytische Arbeit in einer Reihe sprachwissenschaftlicher Untersuchungen. Ihre Erstellung ist zeitaufwändig und kostenintensiv. Dabei ist zu beobachten, dass die nachhaltige Nutzung und Verwendung von Korpora stark abhängig ist von technischen Errungenschaften, die in den letzten Jahrzehnten gemacht wurden. Waren linguistische Korpora anfänglich sehr stark zugeschnitten auf eine spezielle Forschungsfrage und in proprietären Formaten kodiert, wurde mit der größeren Verbreitung der Metasprachen SGML und XML verstärkt Wert auf Austauschbarkeit und Erweiterbarkeit gelegt. Daran knüpfte sich die Etablierung standardisierter Auszeichnungssprachen (u. a. die TEI, XCES) an, die zur Beschreibung einer Vielzahl von linguistischen Merkmalen genutzt werden können.

Die aktuelle Herausforderung der letzten Jahre liegt darin nicht mehr nur in Bezug auf *eine* Forschungsfrage hin mit textuellen Auszeichnungen angereicherte sprachliche Daten – ein-dimensional annotierte Korpora – sondern mehr-dimensional annotierte Daten verarbeiten und analysieren zu können. Problematisch ist dabei das XML-inhärente formale Modell des einwurzeligen Baumes (Single Rooted Tree) und der syntaktische Constraint, keine überlappenden Strukturen zuzulassen.

Der Vortrag beschäftigt sich mit der Frage, wie mehr-dimensionale Annotationen auf linguistischen Korpora trotz der genannten Einschränkungen angelegt, verwaltet und analysiert werden können. Dabei stehen Aspekte wie Verwendung vorhandener Annotationen und Erweiterbarkeit im Vordergrund.