Maik Stührenberg

Standards in der linguistischen Annotation
– Abstract –

In der letzten Dekade haben im Bereich der Annotation linguistischer Daten mehrere Umwälzungen stattgefunden: XML hat sich als Grundlage und Metaformat für die Auszeichnung linguistischer Phänomene etabliert und seit einigen Jahren werden standardisierte Formate zur strukturierten Speicherung linguistischer Korpora entwickelt. Diese lassen sich unterteilen in De-facto-Standards, die beispielsweise auf Grund einer starken Verbreitung zum Einsatz kommen, als auch De-jure-Standards, also Spezifikationen, die den Rang einer (inter-)nationalen Norm haben. Während Letztere prinzipiell eine hohe Nachhaltigkeit garantieren sollen, können Erstere theoretisch schneller auf Veränderungen reagieren, da sie im Gegensatz zu Normen im geringeren Umfang restriktiven zeitlichen Abläufen unterworfen sind.

Verbunden mit der voranschreitenden Normierung lassen sich Veränderungen bzgl. der Notation und des formalen Modells in XML-basierten Auszeichnungssprachen feststellen: Inline-Annotation wird mehr und mehr durch Standoff-Notation abgelöst und damit einher geht der Wechsel von Baum-basierten hin zu Graphen-basierten formalen Modellen. Inline-Annotation reichert die Auszeichnung mit Hilfe von Elementen (bzw. deren Start- und End-Tag) um den auszuzeichnenden Inhalt (die Primärdaten) herum an, während Standoff-Notation Primärdaten und Auszeichnung trennt (teilweise auch in separaten Dateien) und über zusätzliche Mechanismen verknüpft. War Erstere über Jahre hinweg das prototypische Verfahren, so können vor allem multiple Annotationen auf Grund der XML-inhärenten Baumstruktur oftmals nur im Standoff-Verfahren kodiert werden, was wiederum den Einsatz formal ausdrucksstärkerer Datenmodelle (bis hin zum Graphen) erlaubt.

Der Vortrag soll aktuelle relevante Standards und Spezifikationen im Bereich der linguistischen Auszeichnung vorstellen und Hilfestellungen für deren Einsatz in konkreten Projekten bieten.