

xsl-tokenizer

Schopper, Daniel; daniel.schopper@oeaw.ac.at

Unter Tokenisierung versteht man die Zerlegung eines Fließtextes in Einzelsegmente (Tokens), in aller Regel in Wörter, aber auch kleinere (Zeichen) oder größere Einheiten (*Multi Word Items*). Tokenisierung stellt den ersten Verarbeitungsschritt zur (semi-)automatischen linguistischen (Part-of-Speech-Tagging oder Lemmatisierung) oder semantischen Annotation (*Named Entity Recognition/NER*) dar; gleichzeitig ist sie auch Teil des Indizierungsprozesses für die Volltextsuche.

In Digitalen Editionen häufig notwendige komplexe Markup-Strukturen, insbesondere voneinander unabhängige Textflüsse in einem Dokument (wie z. B. in den Haupttext eingebettete Fußnoten oder ein textkritischer Variantenapparat), stellen eine Herausforderung für die Tokenisierung dar.

Der *xsl-tokenizer* ist eine auf XSLT 2.0 aufbauende Softwarelösung, die es ermöglicht, XML-Instanzen regelbasiert zu tokenisieren und dabei bestehende Dokumentstrukturen zu erhalten. Er ist vollständig parametrierbar und kann somit für unterschiedliche XML-Schemata, Tagsets und Annotationsrichtlinien verwendet werden. Das Ergebnis der Prozessierung wird im Quelldokument mit den TEI-Elementen `<w>` bzw. `<pc>` kodiert. Wo durch die Tokenisierung überlappende XML-Hierarchien (TEI Guidelines, Kapitel 20: Non-hierarchical Structures) entstünden, wird ein Token in mehreren Elemente abgebildet, die mit `@part` markiert und durch `@prev` bzw. `@next` verbunden sind. Weiters besteht die Option, eine verflachte Tokenliste mit vereinfachter Dokumentstruktur auszugeben, die von einem Tagger angereichert und anschließend wieder in das Quelldokument integriert werden kann.

Literatur:

- 16 Linking, Segmentation, and Alignment. URL: <https://tei-c.org/release/doc/tei-p5-doc/en/html/SA.html>

Software:

xsl-tokenizer, acdh-spacytei, No Sketch Engine

Verweise:

NLP, Textannotation, Lemmatisierung, Named Entity Recognition / NER, POS-Tagging, Tagger, Tagsets, XSLT

Projekte:

Corpus Thomasticum, Mittelhochdeutsche Begriffsdatenbank (MHDBDB)

Themen:

Natural Language Processing, Software und Softwareentwicklung

Zitiervorschlag:

Schopper, Daniel. 2021. xsl-tokenizer. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: <https://gams.uni-graz.at/o:konde.216>