

Einführung in die Computerlinguistik WS 03/04

S. Pado, M. Pinkal

Übungsblatt 7

7.1 Tokenisierung

Lesen Sie Kap. 4.2 und 4.3 aus dem Buch von Carstensen. Tokenisieren Sie folgenden Satz und beschreiben Sie die auftretenden Schwierigkeiten:

Peter Mengel-Kaiser, der 25jährige Gewinner der Rallye Paris-Dakar, sagte bei der Überreichung des Preises in Höhe von 25 000 Euro:
“Mir hat's einfach Spaß gemacht.”

7.2 Markup und Korpora

Sie finden unter <http://www.coli.uni-sb.de/~pado/I2CL/negra1.txt> den ersten Satz des NEGRA-Korpus im Penn-Treebank-Format.

Unter <http://www.coli.uni-sb.de/~pado/I2CL/tiger1.xml> finden Sie einen der ersten Sätze des TIGER-Korpus im TIGER-XML-Format.

Stellen Sie die Struktur dieser beiden Sätze jeweils grafisch dar.
Falls Sie Schwierigkeiten haben, die Bäume zu verstehen, finden Sie Hilfe z.B. unter

<http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html>

bzw.

<http://www.cis.upenn.edu/~treebank/home.html>

7.3 Semantische Annotation

Unter <http://www.ps.uni-sb.de/~pado/moin.cgi> finden Sie das Wiki des SALSA-Projekts und darin 20 hand-annotierte Sätze. Unter <http://www.icsi.berkeley.edu/~framenet/> finden Sie Beschreibungen der verwendeten Frames (Menüpunkt “FN Data”).

Umschreiben Sie mithilfe der Frame-Definitionen, welche “Bedeutung” Satz 53 hat. (“UNKNOWN”-Frames können Sie ignorieren.)

(Beispiel für Satz 52: “Rao nimmt eine Führungsrolle ein (LEADERSHIP, und zwar als Premierminister). Er unternimmt gleichzeitig eine gefährliche Handlung (RISK_ACTION), und zwar eine revolutionäre Wirtschaftreform.”)