Extraktion von **ju**ristischen **Ver**weisstrukturen: Intertextualität in Gesetzen, Kommentaren und Gerichtsurteilen

Tokenized case law sentences 1.3.0.

Das Korpus enthält anonymisierte deutsche Gerichtsentscheidungen aus deutschen Strafverfahren aus den Jahren 2015 bis Anfang 2020. Die Gerichtsentscheidungen werden vom Bundesgerichtshof (BGH) veröffentlicht. Die Gerichtsentscheidungen wurden direkt von der Website der Rechtsprechungsdatenbank gecrawlt und stehen im HTML-Format für die weitere Arbeit zur Verfügung. Die Quelle für die Dokumente im Rechtsprechungskorpus finden Sie unter: hrr-strafrecht.de.

Es wurde eine Satzsegmentierung durchgeführt. Da juristische Texte häufig Abkürzungen verwenden, waren viele Satzsegmentierungen falsch. Diese falsch segmentierten Sätze wurden wieder zusammengeführt. Dazu wurde zunächst der Sentence-Tokenizer aus dem NLTK-Paket (nltk.org/api/nltk.tokenize.html) auf Rechtstexte trainiert und auf die Texte der Urteile angewendet. Anschließend wurde eine manuell erstellte Liste mit juristischen Abkürzungen verwendet, um falsch getrennte Sätze wieder zusammenzufügen.

Dieser Korpus enthält die tokenisierten Sätze aus dem Korpus 1.2.0. Neben den Token sind auch die Lemma und Part-of-Speech-Tags enthalten. Die Tokenisierung wurde mit dem Wort-Tokenizer aus dem NLTK-Paket (nltk.org/api/nltk.tokenize.html) durchgeführt. Die Lemmatisierung und ein Part-of-Speech-Tagging erfolgte mit dem HanTa-Tagger (github.com/wartaal/HanTa, C. Wartena), ein zusätzliches Part-of-Speech-Tagging erfolgte mit dem TreeTagger (cis.uni-muenchen.de/~schmid/tools/TreeTagger/, H. Schmidt).

Format der Dateien: Die Token, Lemma und Part-of-Speech-Tags sind in einem *CSV* Dateiformat. Die Separatoren sind Taps (\t).

Struktur der Dateien: Die Spalten in den Dateien sind: Das Word, das Lemma_erstellt mit HanTa, der PoS-Tag erstellt mit HanTa und der PoS-Tag erstellt mit dem TreeTagger. Satzanfänge werden mit <s> <s> <s> <s> dargestellt und das Satzende mit </s> </s> </s> </s>.

Lizenz: Siehe hrr-strafrecht.de

Umfang: 4.250 Dokumente

Größe: Ca. 194 MB

Datum: Juli 2021

Projektleiter: Prof. Dr. Christian Wartena

Verantwortlich Rechtswissenschaften: Prof. Dr. Fabian Schmieder

Projektlaufzeit: 01.10.2019 bis 30.09.2022

Projektförderung: Niedersächsischen Ministeriums für Wissenschaft und Kultur, Förderinitiative: Niedersächsisches Vorab, Förderlinie: Geistes- und Kulturwissenschaften – digital: Forschungschancen, Methodenentwicklung und Reflexionspotenziale

Projektwebseite: http://textmining.wp.hs-hannover.de/juver.html



Kontakt:
Hochschule Hannover
Prof. Dr. Christian Wartena
Expo Plaza 12
30539 Hannover
E-Mail: christian.wartena@hs-hannover.de