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

# Tokenized contract sentences 2.3.0.

Dieser Korpus enthält Verträge der Hamburger Stadtverwaltung und der Bremer Stadtverwaltung. Auch einige Kooperationsverträge zwischen Hochschulen sind enthalten. Unter diesen Verträgen befinden sich auch einige Verträge, die Hochschulen mit externen Dienstleistern abgeschlossen haben. Alle Verträge sind im PDF-Format verfügbar. Die Verträge in diesem Korpus stammen aus den Jahren 2014 bis 2019. Die Quellen: Stadtverwaltung Hamburg (http://suche.transparenz.hamburg.de/dataset?q=vertrag&esq_title=&check_all_ ), Stadtverwaltung Bremen (Stichwort: "Vertrag", https://www.transparenz.bremen.de ) , Kooperationsverträge zwischen Hochschulen und auch zwischen Hochschulen und Dienstleistern: Wir haben gezielt nach Vertragsdateien auf Webseiten von Hochschulen gesucht und diese dem Vertragskorpus hinzugefügt.

Die Vertragstexte mussten extrahiert, bereinigt und für die Weiterverarbeitung vorbereitet werden. Die Qualität der gescannten Verträge aus den Stadtverwaltungen ist nicht besonders gut. Es gibt Dokumente mit einer geringeren Scan-Auflösung, Seiten wurden in unterschiedlichen Winkeln gescannt usw. Außerdem wurden alle Informationen, die persönliche Daten darstellen, geschwärzt. In diesem Korpus weisen die Texte viele OCR-Fehler auf, die bereinigt werden mussten.

Zusätzlich zur OCR-Korrektur wurden auch Worttrennungen am Zeilenende korrigiert. Anschließend wurde eine Satzsegmentierung durchgeführt. Da in juristischen Texten häufig Abkürzungen verwendet werden, waren viele Satzsegmentierungen fehlerhaft. Diese fehlerhaft segmentierten Sätze wurden neu zusammengesetzt. Dazu wurde zunächst der Satz-Tokenizer aus dem NLTK-Paket (nltk.org/api/nltk.tokenize.html) auf juristische Texte trainiert und dann angewendet. Anschließend wurde eine manuell erstellte Liste mit juristischen Abkürzungen verwendet, um falsch getrennte Sätze wieder zusammenzufügen. Alle Textelemente der Verträge werden in Textklassen wie Überschrift, Fließtext, … klassifiziert. Dieser Vorgang wird beschrieben in:

*Josi, Frieda und Wartena, C.: Structural Analysis of Contract Renewals. In: Proceedings of the ACM CIKM 2018 Workshops. Turin (2018), Link*

Dieser Korpus enthält die tokenisierten Sätze aus dem Korpus 2.2.0. 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 Überschrift und jeder Satz aus den Urteilen entspricht einer Zeile in der Textdatei.

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:** Die Dateien sind unter der Lizenz Data License Germany Attribution 2.0 oder Data License Germany Zero Version 2.0 öffentlich verfügbar.

**Umfang:** 2.126 Dokumente

**Größe:** Ca. 270 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