GeLeCo: A large **G**erman **Le**gal **Co**rpus of laws, court decisions and administrative regulations issued in Germany at federal level

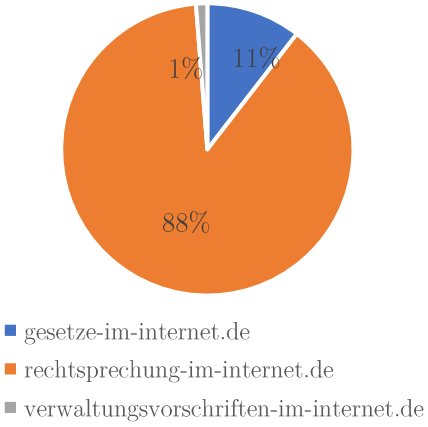
1. Introduction

The GeLeCo corpus is a complete collection of federal laws, administrative regulations and court decisions which were published on three online databases by the German Federal Ministry of Justice and Consumer Protection and the Federal Office of Justice.[1]

2. Corpus design

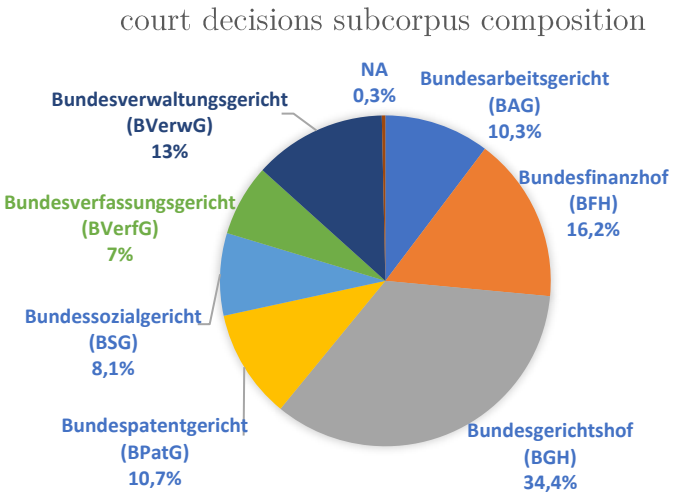
2.1. Composition

subcorpora	period	text count	token count
laws (gesetze-im-internet.de)	1860-2020	6,577	22,502,937
court decisions (rechtsprechung-im-internet.de)	2010-2020	55,361	167,210,730
administrative regulations (verwaltungsvorschriften-im-internet.de)	1970-2020	787	3,469,166
total count		**62,725**	**193,182,833**



The largest subcorpus (the corpus of court decisions published on rechtsprechung-im-internet.de) has the following composition:

issuing court	text count	%
Bundesarbeitsgericht (BAG)	5,697	10,3%
Bundesfinanzhof (BFH)	8,964	16,2%
Bundesgerichtshof (BGH)	19,069	34,4%
Bundespatentgericht (BPatG)	5,913	10,7%
Bundessozialgericht (BSG)	4,460	8,1%
Bundesverfassungsgericht (BVerfG)	3,878	7,0%
Bundesverwaltungsgericht (BVerwG)	7,189	13,0%
NA	191	0,3%
total	**55,361**	**100,0%**



[1] www.gesetze-im-internet.de, www.rechtsprechung-im-internet.de, www.verwaltungsvorschriften-im-internet.de.

2.2. Annotation scheme

The corpus was built in vertical or word-per-line (WPL) format, as required by SketchEngine and NoSketchEngine and marked-up with contextual (metadata), structural (text and sentence boundaries) and linguistic (POS tagging, lemmatisation) annotation (s. below). The complete POS tagset is available on spacy.io.[2]

<corpus>
<text type="Gerichtsentscheidung" level="Bund" title="GmbH: Beschränkung der Stimmrechtsausübungsfreiheit eines Gesellschafters aufgrund der Treuepflicht" title_abbreviation="NA" drafting_date="12.04.2016" decade="2010" database_URL="rechtsprechung-im-internet.de" court="BGH" court_detail="BGH 2. Zivilsenat" reference="II ZR 275/14" year="2016" decision_type="Urteil" ECLI="ECLI:DE:BGH:2016:120416UIIZR275.14.0">
<s>

Nach	ADP	Nach
dem	DET	der
1.	ADJ	1.
April	NOUN	April
1941	NUM	1941
werden	AUX	werden
Gewinnausschüttungen	NOUN	Gewinnausschüttungen
von	ADP	von
Versorgungsunternehmen	NOUN	Versorgungsunternehmen
an	ADP	an
Gemeinden	NOUN	Gemeinde
anerkannt	VERB	anerkennen
.	PUNCT	.

</s>
</text>
</corpus>

2.3. Metadata

Contextual information marked-up for each text includes:

- title
- title_abbreviation
- type: can take one of the following values: *Gesetz* (law), *Gerichtsentscheidung* (court decision), *Verwaltungsvorschrift* (administrative regulation)
- level: indicates whether the law, regulation or court decision was published at federal or *Länder* level. This metadatum was included with sight to a possible extension of the corpus to laws, regulations and court decisions published at *Länder* level. It can take the following values: *Bund* (federal level), *Land* (*Länder* level, not present in this corpus so far)
- drafting_date: this corresponds to the *Ausfertigungsdatum* of laws and the *Entscheidungsdatum* of court decisions.
- year
- decade

[2] https://spacy.io/api/annotation#pos-de

- database_URL: can take the following values: *gesetze-im-internet.de, rechtsprechung-im-internet.de, verwaltungsvorschriften-im-internet.de*
- court: e.g. *"BVerfG"*
- court_detail: e.g. *"BVerfG 1. Senat 3. Kammer"*
- reference: a reference code for court decisions *(Aktenzeichen)*
- decision_type: the type of document for court decisions *(Dokumenttyp)*
- ECLI: the European Case Law Identifier code for court decisions.

3. Corpus building steps[3]

3.1. URL collection

All URLs were collected by means of website-specific web scrapers written in Python. Three lists of URLs were exported in newline-separated .txt files for subsequent text scraping.

3.2. Text scraping and XML tagging

Based on the previously collected URL lists, single legal texts were scraped by means of ad hoc web scrapers written in Python. Text and metadata collection was carried out using the BeautifulSoup Python library. Text contained in different HTML tags was newline-separated, making the subsequent sentence splitting stage easier and faster to carry out. After scraping, texts were merged and a first raw corpus version was exported as a single .txt file for each subcorpus.

3.3. Boilerplate cleaning, deduplication, text filtering

Boilerplate text was eliminated by means of regular expressions. Texts extracted from incorrectly visualized webpages (not containing any law, regulation or court decision) were discarded. Texts also underwent a deduplication process based on metadata equivalence.

3.4. Sentence splitting

After scraping and cleaning, the subcorpora were sentence-splitted. In particular, only lines containing two or more period characters underwent sentence splitting. For this task, Kahn's and Schroeder's sentence-splitter[4] was used, and a list of non-breaking prefixes with legal abbreviations taken from the corpus and from online sources was supplied in order to improve sentence splitting accuracy. After splitting, lines were added opening and closing sentence delimiting tags (<s>).

3.5. POS tagging and lemmatization

The corpus was tagged with Part-Of-Speech tags and lemmas using the SpaCy tagger.[5] The output did not undergo any systematic revision or correction operation; therefore, the corpus may contain minor sentence splitting or metadata errors.

[3] The scripts written for the building of the GeLeCo corpus can be found on https://github.com/antcont/GeLeCo

[4] https://github.com/mediacloud/sentence-splitter

[5] https://spacy.io/models/de