A Dataset of German Legal Documents for Named Entity Recognition

Elena Leitner, Georg Rehm, Julián Moreno-Schneider
DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
{firstname.lastname}@dfki.de

Abstract

We describe a dataset developed for Named Entity Recognition in German federal court decisions. It consists of approx. 67,000 sentences with over 2 million tokens. The resource contains 54,000 manually annotated entities, mapped to 19 fine-grained semantic classes: *person*, *judge*, *lawyer*, *country*, *city*, *street*, *landscape*, *organization*, *company*, *institution*, *court*, *brand*, *law*, *ordinance*, *European legal norm*, *regulation*, *contract*, *court decision*, and *legal literature*. The legal documents were, furthermore, automatically annotated with more than 35,000 TimeML-based time expressions. The dataset, which is available under a CC-BY 4.0 license in the CoNNL-2002 format, was developed for training an NER service for German legal documents in the EU project Lynx.

Keywords: Named Entity Recognition, NER, Legal Documents, Legal Domain, Corpus Creation, Corpus Annotation

1. Introduction and Motivation

Just like any other field, the legal domain is facing multiple challenges in the era of digitisation. Document collections are growing at an enormous pace and their complete and deep analysis can only be tackled with the help of assisting technologies. This is where content curation technologies based on text analytics come in Bourgonje et al. (2016). Such domain-specific semantic technologies enable the fast and efficient automated processing of heterogeneous document collections, extracting important information units and metadata such as, among others, named entities, numeric expressions, concepts and topics, time expressions, and text structure. One of the fundamental processing tasks is the identification and categorisation of named entities (Named Entity Recognition, NER). Typically, NER is focused upon the identification of semantic categories such as *person*, *location* and *organization* but, especially in domain-specific applications, other typologies have been developed that correspond to task-, language- or domain-specific needs. With regard to the legal domain, the lack of freely available datasets has been a stumbling block for text analytics research. German newspaper datasets from CoNNL 2003 (Sang and Meulder, 2003) or GermEval 2014 (Benikova et al., 2014) are simply not suitable in terms of domain, text type or semantic categories covered.

The work described in this paper was carried out under the umbrella of the project *Lynx: Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe*, a three-year EU-funded project that started in December 2017 (Montiel-Ponsoda et al., 2017).[1] Its objective is the creation of a legal knowledge graph that contains different types of legal and regulatory data (Schneider and Rehm, 2018a; Schneider and Rehm, 2018b; Moreno-Schneider et al., 2020). Lynx aims to help European companies, especially SMEs, that want to become active in new European countries and markets. The project offers compliance-related services that are currently tested and validated in three use cases (UC): (i) UC1 aims to analyse contracts, enriching them with domain-specific semantic information (document structure, entities, temporal expressions, claims, summaries, etc.); (ii) UC2 focuses on compliance services related to geothermal energy operations, where Lynx supports the understanding of regulatory regimes, including norms and standards; (iii) UC3 is a compliance solution in the domain of labour law, where legal provisions, case law, and expert literature are interlinked, analysed, and compared to define legal strategies for legal practice. The Lynx services are developed for several European languages including English, Spanish, and – relevant for this paper – German (Rehm et al., 2019).

Documents in the legal domain contain multiple references to named entities, especially domain-specific named entities, i. e., jurisdictions, legal institutions, etc. Legal documents are unique and differ greatly from newspaper texts. On the one hand, the occurrence of general-domain named entities is relatively rare. On the other hand, in concrete applications, crucial domain-specific entities need to be identified in a reliable way, such as designations of legal norms and references to other legal documents (laws, ordinances, regulations, decisions, etc.). However, most NER solutions operate in the general or news domain, which makes them inapplicable to the analysis of legal documents (Bourgonje et al., 2017; Rehm et al., 2017). Accordingly, there is a great need for an NER-annotated dataset consisting of legal documents, including the corresponding development of a typology of semantic concepts and uniform annotation guidelines. In this paper, we describe the development of a dataset of legal documents, which includes (i) named entities and (ii) temporal expressions.

The remainder of this article is structured as follows. First, Section 2 gives a brief overview of related work. Section 3 describes, in detail, the rationale behind the annotation of the dataset including the different semantic classes annotated. Section 4 describes several characteristics of the dataset, followed by a short evaluation (Section 5) and conclusions as well as future work (Section 6).

2. Related Work

Until now, NER has not received a lot of attention in the legal domain, developed approaches are fragmented and inconsistent with regard to their respective methods, datasets and typologies used. Among the related work, there is

[1] http://www.lynx-project.eu

no agreement regarding the selection of relevant semantic categories from the legal domain. In addition, corpora or datasets of legal documents with annotated named entities do not appear to exist, which is, obviously, a stumbling block for the development of data-driven NER classifiers.
Dozier et al. (2010) describe five classes for which taggers are developed based on dictionary lookup, pattern-based rules, and statistical models. These are *jurisdiction* (a geographic area with legal authority), *court*, *title* (of a document), *doctype* (category of a document), and *judge*. The taggers were tested with documents such as US case law, depositions, pleadings etc. Cardellino et al. (2017) develop an ontology of legal concepts, making use of NERC (6 classes), LKIF (69 classes) and YAGO (358 classes). On the NERC level, entities were divided in *abstraction*, *act*, *document*, *organization*, *person*, and non-entity. With regard to LKIF, *company*, *corporation*, *contract*, *statute* etc. are used. Unfortunately, the authors do not provide any details regarding the questions how the entities were categorised or if there is any correlations between the different levels. They work with Wikipedia articles and decisions of the European Court of Human Rights. Glaser et al. (2018) use GermaNER (Benikova et al., 2015) and DBpedia Spotlight (Mendes et al., 2011; Daiber et al., 2013) for the recognition of *person*, *location* and *organization* entities. References are identified based on the rules described by Landthaler et al. (2016). The authors created an evaluation dataset of 20 court decisions.

3. Annotation of the Dataset

In the following, we describe the rationale behind the annotation of the dataset including the definition of the various semantic classes and the annotation guidelines.

3.1. Named Entities vs. Legal Entities

Named Entity An entity is an object or set of objects in the real world and can be referenced in a text with a proper name, noun or pronoun (Linguistic Data Consortium, 2008). The examples (1–3) show corresponding sentences that contain the named mention 'John', the nominal mention 'the boy' and the pronominal mention 'he'. This distinction between names on the one hand and pronominal or nominal mentions on the other can also be applied to the broad semantic set of named entities from the legal domain, see (4–6). Thus, (1, 4) contain actual named entities.

(1) John is 8 years old.

(2) The boy is 8 years old.

(3) He is 8 years old.

(4) The BGB regulates the legal relations between private persons.

(5) The law regulates the legal relations [...].

(6) It regulates the legal relations [...].

Legal Entity Basically, legal entities are either designations or references. A designation (or name) is the title of a legal document. In law texts, the title is strictly standardised and consists of a long title, short title and an abbreviation (Bundesministerium der Justiz, 2008, margin nos 321 et seqq.). The title of the Act on the Federal Constitutional Court is: 'Gesetz über das Bundesverfassungsgericht (Bundesverfassungsgerichtsgesetz – BVerfGG)', where 'Gesetz über das Bundesverfassungsgericht' is the long title, 'Bundesverfassungsgerichtsgesetz' is the short title, and 'BVerfGG' is the abbreviation. A reference to a legal norm is also fixed with rules for short and full references (Bundesministerium der Justiz, 2008, margin nos 168 et seqq.). Designations or references of binding individual acts such as regulations or contracts, however, are not uniformly defined.

Personal Data A fundamental characteristic of the published decisions, that are the basis of our dataset, is that all personal information have been anonymised for privacy reasons. This affects the classes *person*, *location* and *organization*. Depending on the respective federal court, different rules were used for this anonymisation process. Named entities were replaced by letters or abbreviated (7), sometimes ellipsis were used (8, 9). Some anonymised *locations* are mentioned with terms such as "street", "place", "avenue", etc. that are part of this named entity (9).

(7) Fernsehmoderator G. **PER**
'television presenter G.'

(8) Firma X... **UN**
'company X...'

(9) in der A-Straße **STR** in ... **ST**
'in the A-Street in ...'

3.2. Semantic Classes

We defined 19 fine-grained semantic classes. The (proto)typical classes are *person*, *location* and *organization*. In addition, we defined more specific semantic classes for the legal domain. These are the coarse-grained classes *legal norm*, *case-by-case regulation*, *court decision* and *legal literature*. The classes *legal norm* and *case-by-case regulation* include designations and references, while *court decision* and *legal literature* include only references.
In the process of developing the typology and annotation guidelines, the fine-grained classes *continent* `KONT` (which belongs to *location*), *university* `UNI`, *institute* `IS` and *museum* `MUS` (which belonged to *organization*) were eliminated due their low frequency in the corpus (less than 50 occurrences). This is why *university*, *institute* and *museum* were subsumed under the fine-grained class *organization*. *Continent* was integrated into *landscape*.
The specification of the 19 fine-grained classes was motivated by the need for distinguishing entities in the legal domain. A first distinction was made between standards and binding acts. Standards, which belong to *legal norm*, are legal rules adopted by a legislative body in a legislative process. We can distinguish further between *law*, *ordinance* (German national standards) and *European legal norm*. Binding acts (circulars, administrative acts, contracts, administrative regulations, directives, etc.) belong to the category of *case-by-case-regulation*. It includes *regulation* (arrangements or instructions on subjects) and *contract* (agreements between subjects). In addition, *court de-*

cision and *legal literature*, which are important in the decision making process, were put into their own categories. Within *person*, we distinguish between *judge* and *lawyer*, key roles mentioned frequently in the decisions. Locations are categorised in terms of their size in *country*, *city* and *street*. Organizations are divided based on their role in the process, into public or social *organization*, state *institution*, (private) economic *company*, mostly as a legal entity, and *court* as an organ of jurisprudence.

Person The coarse-grained class *person* PER contains the fine-grained classes *judge* RR, *lawyer* AN and *person* PER (such as accused, plaintiff, defendant, witness, appraiser, expert, etc.), who are involved in a court process and mentioned in a decision. In example (10), the same surname occurs twice in a sentence, one as *judge* and one as *person*.

(10) Zwar ist Paul Kirchhof **RR** mit dem Vizepräsidenten Kirchhof **PER** als dessen Bruder in der Seitenlinie im zweiten Grade verwandt…
'Although Paul Kirchhof is related to the Vice President Kirchhof as his brother in the second-degree sidelines…'

Location The coarse-grained class *location* LOC contains names of topographic objects, divided into *country* LD, *city* ST, *street* STR and *landscape* LDS. *Country* (11) includes countries, states or city-states and *city* (12) includes to cities, villages or communities. *Street* (13) refers to avenues, squares, municipalities, attractions etc., i. e., named entities within a city or a village. *Landscape* (14) includes continents, lakes, rivers and other geographical objects.

(11) … hat bislang nur das Land Mecklenburg-Vorpommern **LD** Gebrauch gemacht.
'So far, only the state of Mecklenburg-Vorpommern has made use of it.'

(12) Dem Haftbefehl liegt eine Entscheidung des Berufungsgerichts in Bukarest **ST** vom 18. Februar 2016 zugrunde …
'The arrest warrant is based on a decision of the Appeal Court in Bucharest of 18 February 2016 …'

(13) Zwar legt der Bezug auf die Grenzwertüberschreitung 2015 insbesondere in der Corneliusstraße **STR** …
'Admittedly, the reference to the exceedance of the 2015 threshold applies in particular to Corneliusstraße …'

(14) … aus der Region um den Fluss Main **LDS** stammen bzw. dort angeboten werden …
'… come from the region around the river Main or are offered there…'

Organization The coarse-grained class *organization* ORG is divided into public/social, state and economic institutions. Social and public institutions such as parties, associations, centres, communities, unions, educational institutions or research institutions are grouped into the fine-grained class *organization* ORG (15). *Institution* INN (16) contain public administrations, including federal and state ministries and the constitutional bodies of the Federal Republic of Germany at the federal and state level, i. e., the Federal Government, the Federal Council, the Bundestag, the state parliaments and governments. *Company* UN (17) includes commercial legal entities.

(15) Der FC Bayern München **ORG** schloss den Beschwerdeführer … aus dem Verein aus …
'Bayern Munich closed the complainant … from the club …'

(16) Die Landesregierung Rheinland-Pfalz **INN** hat von einer Stellungnahme abgesehen.
'The state government of Rhineland-Palatinate refrained from commenting.'

(17) … eingeführte Smartphone-Modellreihe des US-amerikanischen Unternehmens Apple **UN** …
'… introduced smartphone model series of the US company Apple …'

Court designations play a central role in decisions, which is why they are collected in their own class *court* GRT. These are designations of federal, supreme, provincial and local courts. The designations of the courts at the country level are composed of the names of the ordinary jurisdiction and their location (18). Furthermore, brands are often discussed in decisions of the Federal Patent Court. They are subsumed under *brand* MRK, which can be contextual and semantically ambiguous, such as 'Becker' from (19), which has evolved from a personal name.

(18) Diesen Anspruch hat das LSG Mecklenburg-Vorpommern **GRT** mit Urteil vom 22.2.2017 verneint …
'This claim was rejected by the LSG Mecklenburg-Vorpommern by judgment of 22.2.2017 …'

(19) Vorliegend stehen sich die Widerspruchsmarke Becker Mining **MRK** und die angegriffene Marke Becker **MRK** gegenüber.
'In the present case, the opposing brand Becker Mining and the challenged brand Becker face each other.'

Legal Norms Norms are divided according to their legal status into the fine-grained classes of *law* GS, *ordinance* VO and *European legal norm* EUN. *Law* is composed of the standards adopted and designated by the legislature (Bundestag, Bundesrat, Landtag). *Ordinance* includes standards adopted by a federal or provincial government or by a ministry. *European legal norm* includes norms of European primary or secondary legislation, European organizations and other conventions and agreements.

Example (20) includes a reference to the 'Part-Time and Limited Term Employment Act' and the designation 'Basic Law'. The complex reference consists of the reference to the particular section of the law, its name and abbreviation (in brackets), date of issue, the reference in parenthesis and the details of the most recent change. Cases such as this one are a full reference. Example (21), on the other hand, shows a short reference consisting of information on the corresponding section of the law and the abbreviated name of the statutory order.

(20) ... § 14 Absatz 2 Satz 2 des Gesetzes über Teilzeitarbeit und I befristete Arbeitsverträge (TzBfG) vom 21. Dezember 2000 (Bundesgesetzblatt Seite 1966), zuletzt geändert durch Gesetz vom 20. Dezember 2011 (Bundesgesetzblatt I Seite 2854) **GS** , ist nach Maßgabe der Gründe mit dem Grundgesetz **GS** vereinbar.

'... section 14 paragraph 2 sentence 2 of the Law on Part-Time and Limited Term Employment Act (TzBfG) of 21 December 2000 (Federal Law Gazette I, page 1966), as last amended by the Law of 20 December 2011 (Federal Law Gazette I, page 2854), shall be published in accordance with the reasons compatible with the Basic Law.'

(21) ... Neuregelung in § 35 Abs. 6 StVO **VO** ...

'... new regulation in sec. 35 para. 6 StVO...'

Case-by-case Regulation The class *case-by-case regulation* `REG` contains individual binding acts. These include *regulation* `VS` and *contract* `VT`. *Regulation* is an internal order or instruction from a superordinate authority to a subordinate, regulating their activities. In addition to administrative regulations, these include guidelines, circulars and decrees. In contrast to *legal norm*, these rules have no direct effect on the citizen. The class *contract* includes public contracts, international treaties and collective agreements. Some designations and references from these classes are similar to *legal norm* (22, 23).

(22) ... insbesondere durch die Richtlinien zur Bewertung des Grundvermögens –BewRGr– vom 19. September 1966 (BStBl I, S. 890) **VS** .

'... in particular by the Guidelines for the Valuation of Real Estate – BewRGr – of 19 September 1966 (BStBl I, p. 890).'

(23) ... fand der Manteltarifvertrag für die Beschäftigten der Mitglieder der TGAOK **VT** (BAT/AOK-Neu **VT**) vom 7. August 2003 Anwendung.

'... the Collective Agreement for the Employees of Members of TGAOK (BAT/AOK-New) was applied of 7 August 2003 ...'

Court Decision The class *court decision* `RS` includes references to decisions. It does not have any subclasses, the coarsed and fine-grained versions are identical. In *court decision*, the name of the official decision-making collection, the volume and the numbered article are cited. Often mentioned are also the court, if necessary the decision type, date and file number. Example (24) cites decisions of the Federal Constitutional Court (BVerfG) and the Federal Social Court (BSG). Decisions of the BVerfG are referenced with regard to pages, while decisions of the BSG are sorted according to paragraphs, numbers and marginal numbers.

Legal Literature *Legal literature* `LIT` also contains references, but they refer to legal commentaries, legislative material, legal textbooks and monographs. The commentary in example (24) includes the details of an author's and/or publisher's name, the name of a legal norm, a paragraph and a paragraph number. Multiple authors are separated by a slash. Textbooks and monographs are cited as usual (author's name, title, edition, year of publication, page number). References of legislative materials consist of a title and reference marked with numbers.

(24) ... vgl zB BVerfGE 62, 1, 45 **RS** ; BVerfGE 119, 96, 179 **RS** ; BSG SozR 4–2500 § 62 Nr 8 RdNr 20 f **RS** ; Hauck/Wiegand, KrV 2016, 1, 4 **LIT** ...

'... cf. i.e. BVerfGE 62, 1, 45; BVerfGE 119, 96, 179; BSG SozR 4–2500 § 62 Nr 8 RdNr 20 f; Hauck/Wiegand, KrV 2016, 1, 4 ...'

4. Description of the Dataset

The dataset[2], which also includes annotation guidelines, is freely available under a CC-BY 4.0 license.[3] The named entity annotations adhere to the CoNLL-2002 format (Sang, 2002), while time expressions were annotated using TimeML (Pustejovsky et al., 2003).

4.1. Original Source Documents

Legal documents are a rather heterogeneous class, which also manifests in their linguistic properties, including the use of named entities and references. Their type and frequency varies significantly, depending on the text type. Texts belonging to specific text type, which are to be selected for inclusion in a corpus must contain enough different named entities and references and they need to be freely available. When comparing legal documents such as laws, court decisions or administrative regulations, decisions are the best option. In laws and administrative regulations, the frequencies of *person*, *location* and *organization* are not high enough for NER experiments. Court decisions,

[2]https://github.com/elenanereiss/Legal-Entity-Recognition
[3]https://creativecommons.org/licenses/by/4.0/deed.en

on the other hand, include *person*, *location*, *organization*, references to *law*, other *decision* and *regulation*.

Court decisions from 2017 and 2018 were selected for the dataset, published online by the Federal Ministry of Justice and Consumer Protection.[4] The documents originate from seven federal courts: Federal Labour Court (BAG), Federal Fiscal Court (BFH), Federal Court of Justice (BGH), Federal Patent Court (BPatG), Federal Social Court (BSG), Federal Constitutional Court (BVerfG) and Federal Administrative Court (BVerwG).

From the table of contents[5], 107 documents from each court were selected (see Table 1). The data was collected from the XML documents, i. e., it was extracted from the XML elements `Mitwirkung`, `Titelzeile`, `Leitsatz`, `Tenor`, `Tatbestand`, `Entscheidungsgründe`, `Gründen`, `abweichende Meinung`, and `sonstiger Titel`. The metadata at the beginning of the documents (name of court, date of decision, file number, European Case Law Identifier, document type, laws) and those that belonged to previous legal proceedings was deleted. Paragraph numbers were removed. The extracted data was split into sentences, tokenised using SoMaJo[6] (Proisl and Uhrig, 2016) and manually annotated in WebAnno[7] (Eckart de Castilho et al., 2016).

The annotated documents are available in CoNNL-2002. The information originally represented by and through the XML markup was lost in the conversion process. We decided to use CoNNL-2002 because our primary focus was on the NER task and experiments. CoNNL is one of the best practice formats for NER datasets. All relevant tools support CoNNL, including WebAnno for manual annotation. Nevertheless, it is possible, of course, to re-insert the annotated information back into the XML documents.

4.2. Annotation of Named Entities

The dataset consists of 66,723 sentences with 2,157,048 tokens (incl. punctuation), see Table 1. The sizes of the seven court-specific datasets varies between 5,858 and 12,791 sentences, and 177,835 to 404,041 tokens. The distribution of annotations on a per-token basis corresponds to approx. 19–23 %. The Federal Patent Court (BPatG) dataset contains the lowest number of annotated entities (10.41 %).

The dataset includes two different versions of annotations, one with a set of 19 fine-grained semantic classes and another one with a set of 7 coarse-grained classes (Table 2). There are 53,632 annotated entities in total, the majority of which (74.34 %) are legal entities, the others are *person*, *location* and *organization* (25.66 %). Overall, the most frequent entities are *law* `GS` (34.53 %) and *court decision* `RS` (23.46 %). The other legal classes (*ordinance* `VO`, *European legal norm* `EUN`, *regulation* `VS`, *contract* `VT`, and *legal literature* `LIT`) are much less frequent (1–6 % each). Even less frequent (less than 1 %) are *lawyer* `AN`, *street* `STR`, *landscape* `LDS`, and *brand* `MRK`.

The classes *person*, *lawyer* and *company* are heavily affected by the anonymisation process (80 %, 95 % and 70 % respectively). More than half of *city* and *street*, about 55 %, have also been modified. *Landscape* and *organization* are affected as well, with 40 % and 15 % of the occurrences edited accordingly. However, anonymisation is typically not applied to *judge*, *country*, *institution* and *court* (1–5 %).

The dataset was originally annotated by the first author. To evaluate and potentially improve the quality of the anno-

Court	Documents	Tokens	Sentences	Annotated tokens
BAG	107	343,065	12,791	19.23%
BFH	107	276,233	8,522	22.43%
BGH	108	177,835	5,858	19.23%
BPatG	107	404,041	12,016	10.41%
BSG	107	302,161	8,083	22.76%
BVerfG	107	305,889	9,237	22.09%
BVerwG	107	347,824	10,216	20.84%
Total	**750**	**2,157,048**	**66,723**	**19.15%**

Table 1: Dataset size (tokens, sentences, annotated tokens)

Classes				#	%
f	1	`PER`	Person	1,747	3.26
f	2	`RR`	Judge	1,519	2.83
f	3	`AN`	Lawyer	111	0.21
c	**1**	**`PER`**	**Person**	**3,377**	**6.30**
f	4	`LD`	Country	1,429	2.66
f	5	`ST`	City	705	1.31
f	6	`STR`	Street	136	0.25
f	7	`LDS`	Landscape	198	0.37
c	**2**	**`LOC`**	**Location**	**2,468**	**4.60**
f	8	`ORG`	Organization	1,166	2.17
f	9	`UN`	Company	1,058	1.97
f	10	`INN`	Institution	2,196	4.09
f	11	`GRT`	Court	3,212	5.99
f	12	`MRK`	Brand	283	0.53
c	**3**	**`ORG`**	**Organization**	**7,915**	**14.76**
f	13	`GS`	Law	18,520	34.53
f	14	`VO`	Ordinance	797	1.49
f	15	`EUN`	EU legal norm	1,499	2.79
c	**4**	**`NRM`**	**Legal norm**	**20,816**	**38.81**
f	16	`VS`	Regulation	607	1.13
f	17	`VT`	Contract	2,863	5.34
c	**5**	**`REG`**	**Case-by-c. regul.**	**3,470**	**6.47**
f	18				
c	**6**	**`RS`**	**Court decision**	**12,580**	**23.46**
f	19				
c	**7**	**`LIT`**	**Legal literature**	**3,006**	**5.60**
Total				**53,632**	**100**

Table 2: Distribution of fine-grained (f) and coarse-grained (c) classes in the dataset

[4] https://www.rechtsprechung-im-internet.de

[5] http://www.rechtsprechung-im-internet.de/rii-toc.xml

[6] https://github.com/tsproisl/SoMaJo

[7] https://webanno.github.io/webanno/

tations, part of the dataset was annotated by a second linguist (using the annotation guidelines specifically prepared for its construction). We selected a small part that could be annotated in approx. two weeks. For the sentence extraction we paid special attention to the anonymised mentions of *person*, *location* or *organization* entities, because these are usually explained at their first mention. The resulting sample consisted of 2005 sentences with a broad variety of different entities (3 % of all sentences from each federal court). The agreement between the two annotators was measured using Kappa on a token basis. All class labels were taken into account in accordance with the IOB2 scheme (Sang and Veenstra, 1999). The inter-annotator agreement is 0.89, i. e., there is mostly very good agreement between the two annotators. Differences were in the identification of *court decision* and *legal literature*. Some unusual references of *court decision* (consisting only of decision type, court, date, file number) were not annotated such as 'Urteil des Landgerichts Darmstadt vom 16. April 2014 – 7 S 8/13 –'. Apart from missing *legal literature* annotations, author names and law designations were annotated according to their categories (i. e., 'Schoch, in: Schoch/Schneider/Bier, VwGO § 123 Rn. 35', 'Bekanntmachung des BMG gemäß §§ 295 und 301 SGB V zur Anwendung des OPS vom 21.10.2010').
The second annotator had difficulties annotating the class *law*, not all instances were identified ('§ 272 Abs. 1a und 1b HGB', '§ 3c Abs. 2 Satz 1 EStG'), others only partially ('§ 716 in Verbindung mit' in '§ 716 in Verbindung mit §§ 321 , 711 ZPO'). Some titles of *contract* were not recognised and annotated ('BAT', 'TV-L', 'TVÜ-Länder' etc.).
This evaluation has revealed deficiencies in the annotation guidelines, especially regarding *court decision* and *legal literature* as well as non-entities. It would also be helpful for the identification and classification to list well-known sources of *law*, *court decision*, *legal literature* etc.

4.3. Annotation of Time Expressions

All court decisions were annotated automatically for time expressions using a customised version of HeidelTime (Strötgen and Gertz, 2013), which was adapted to the legal domain (Weißenhorn, 2018). This version of Heideltime achieves an F_1 value of 89.1 for partial identification and normalization. It recognizes four TIMEX3-types of time expressions (Verhagen et al., 2010): DATE, DURATION, SET, TIME. DATE describe a calendar date ('23. July 1994', 'November 2019', 'winter 2001' etc). It also includes expressions such as 'present', 'former' or 'future'. DURATION describes time periods such as 'two hours' or 'six years'. SET describes a set of times/periods ('every day', 'twice a week'). TIME describes a time expression ('13:12', 'tomorrow afternoon'). Expressions with a granularity less than 24 hours are of type TIME, all others are of type DATE. The distribution of TIMEX3 types in the legal dataset is shown in Table 3 with a total number of 35,119 time expressions, approx. 94% of which are of type DATE.

(25) . . . vgl. BGH, Beschluss vom <TIMEX3 tid="t14" type="DATE" value="1999-02-03">3. Februar 1999</TIMEX3> – 5 StR 705/98, juris Rn. 2 . . .

	DATE	DURATION	SET	TIME
BAG	6,463	491	99	34
BFH	6,156	189	37	9
BGH	2,819	254	7	22
BPatG	4,576	84	4	12
BSG	4,634	215	64	14
BVerfG	3,595	207	12	20
BVerwG	4,879	178	36	9
Total	**33,122**	**1,618**	**259**	**120**

Table 3: Distribution of time expressions in the dataset

5. Evaluation

The dataset was thoroughly evaluated, see Leitner et al. (2019) for more details. As state of the art models, Conditional Random Fields (CRFs) and bidirectional Long-Short Term Memory Networks (BiLSTMs) were tested with the two variants of annotation. For CRFs, these are: CRF-F (with features), CRF-FG (with features and gazetteers), CRF-FGL (with features, gazetteers and lookup). For BiLSTM, we used models with pre-trained word embeddings (Reimers et al., 2014): BiLSTM-CRF (Huang et al., 2015), BiLSTM-CRF+ with character embeddings from BiLSTM (Lample et al., 2016), and BiLSTM-CNN-CRF with character embeddings from CNN (Ma and Hovy, 2016). To evaluate the performance we used stratified 10-fold cross-validation. As expected, BiLSTMs perform best (see Table 4). The F_1 score for the fine-grained classification reaches 95.46 and 95.95 for the coarse-grained one. CRFs reach up to 93.23 F_1 for the fine-grained classes and 93.22 F_1 for the coarse-grained ones. Both models perform best for *judge*, *court* and *law*.

	Prec %	**Rec %**	**F_1**
Annotation with fine-grained semantic classes			
CRF-F	94.28	91.85	93.05
CRF-FG	94.31	91.96	93.12
CRF-FGL	94.37	92.12	**93.23**
Annotation with coarse-grained semantic classes			
CRF-F	94.17	92.07	93.11
CRF-FG	94.26	92.20	**93.22**
CRF-FGL	94.22	92.25	**93.22**
Annotation with fine-grained semantic classes			
BiLSTM-CRF	93.80	93.70	93.75
BiLSTM-CRF+	95.36	95.57	**95.46**
BiLSTM-CNN-CRF	95.34	95.58	**95.46**
Annotation with coarse-grained semantic classes			
BiLSTM-CRF	94.86	94.49	94.68
BiLSTM-CRF+	95.84	96.07	**95.95**
BiLSTM-CNN-CRF	95.71	95.87	95.79

Table 4: Precision, recall and F_1 values of the CRF and BiLSTM models for the fine- and coarse-grained classes

6. Conclusions and Future Work

We describe a dataset that consists of German legal documents. For the annotation, we specified a typology of characteristic semantic categories that are relevant for court decisions (i. e., *court*, *institution*, *law*, *court decision*, and *legal literature*) with corresponding annotation guidelines. A functional service based on the work described in this paper will be made available through the European Language Grid (Rehm et al., 2020).

In terms of future work, we will look into approaches for extending and further optimizing the dataset. We will also perform additional experiments with more recent state of the art approaches (i. e., with language models); preliminary experiments using BERT failed to yield an improvement. We also plan to replicate the dataset in one or more other languages, such as English, Spanish, or Dutch, to cover at least one more of the relevant languages in the Lynx project. We also plan to produce an XML version of the dataset that also includes the original XML annotations.

Acknowledgements

This work has been partially funded by the project Lynx, which has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement no. 780602, see http://www.lynx-project.eu.

7. References

Benikova, D., Biemann, C., and Reznicek, M. (2014). NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 2524–2531. European Language Resources Association (ELRA).

Benikova, D., Yimam, S. M., Santhanam, P., and Biemann, C. (2015). GermaNER: Free Open German Named Entity Recognition Tool. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2015, University of Duisburg-Essen, Germany, 30th September - 2nd October 2015*, pages 31–38.

Bourgonje, P., Moreno-Schneider, J., Nehring, J., Rehm, G., Sasaki, F., and Srivastava, A. (2016). Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer. In Harald Sack, et al., editors, *The Semantic Web*, number 9989 in Lecture Notes in Computer Science, pages 65–68. Springer, 6. ESWC 2016 Satellite Events. Heraklion, Crete, Greece, May 29 – June 2, 2016 Revised Selected Papers.

Bourgonje, P., Moreno-Schneider, J., and Rehm, G. (2017). Domain-specific Entity Spotting: Curation Technologies for Digital Humanities and Text Analytics. In Nils Reiter et al., editors, *CUTE Workshop 2017 – CRETA Unshared Task zu Entitätenreferenzen. Workshop bei DHd2017*, Berne, Switzerland, February.

Bundesministerium der Justiz. (2008). Bekanntmachung des Handbuchs der Rechtsförmlichkeit. *Bundesanzeiger*, Jahrgang 60(160a):296, September.

Cardellino, C., Teruel, M., Alemany, L. A., and Villata, S. (2017). A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, pages 9–18, New York, NY, USA. ACM.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pages 121–124.

Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., and Wudali, R. (2010). Named Entity Recognition and Resolution in Legal Text. In Enrico Francesconi, et al., editors, *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, volume 6036 of *Lecture Notes in Computer Science*, pages 27–43. Springer.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In Erhard W. Hinrichs, et al., editors, *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities, LT4DH@COLING, Osaka, Japan, December 2016*, pages 76–84. The COLING 2016 Organizing Committee.

Glaser, I., Waltl, B., and Matthes, F. (2018). Named Entity Recognition, Extraction, and Linking in German Legal Contracts. *IRIS: Internationales Rechtsinformatik Symposium*, pages 325–334.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR*, abs/1508.01991.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270.

Landthaler, J., Waltl, B., and Matthes, F. (2016). Unveiling References in Legal Texts – Implicit versus Explicit Network Structures. *IRIS: Internationales Rechtsinformatik Symposium*, pages 71–78.

Leitner, E., Rehm, G., and Moreno-Schneider, J. (2019). Fine-grained Named Entity Recognition in Legal Documents. In Maribel Acosta, et al., editors, *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*, number 11702 in Lecture Notes in Computer Science, pages 272–287, Karlsruhe, Germany, 9. Springer. 10/11 September 2019.

Linguistic Data Consortium. (2008). ACE (Automatic Content Extraction) English Annotation Guidelines for Entities.

Ma, X. and Hovy, E. H. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association*

for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In Chiara Ghidini, et al., editors, *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*, ACM International Conference Proceeding Series, pages 1–8. ACM.

Montiel-Ponsoda, E., Rodríguez-Doncel, V., and Gracia, J. (2017). Building the legal knowledge graph for smart compliance services in multilingual europe. In *Proceedings of the 1st Workshop on Technologies for Regulatory Compliance co-located with the 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017), Luxembourg, December 13, 2017*, pages 15–17.

Moreno-Schneider, J., Rehm, G., Montiel-Ponsoda, E., Rodriguez-Doncel, V., Revenko, A., Karampatakis, S., Khvalchik, M., Sageder, C., Gracia, J., and Maganza, F. (2020). Orchestrating NLP Services for the Legal Domain. In Nicoletta Calzolari, et al., editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France. European Language Resources Association (ELRA). Accepted for publication.

Proisl, T. and Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In Paul Cook, et al., editors, *Proceedings of the 10th Web as Corpus Workshop, WAC@ACL 2016, Berlin, August 12, 2016*, pages 57–62. Association for Computational Linguistics.

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Rehm, G., Moreno-Schneider, J., Bourgonje, P., Srivastava, A., Nehring, J., Berger, A., König, L., Räuchle, S., and Gerth, J. (2017). Event Detection and Semantic Storytelling: Generating a Travelogue from a large Collection of Personal Letters. In Tommaso Caselli, et al., editors, *Proceedings of the Events and Stories in the News Workshop*, pages 42–51, Vancouver, Canada, August. Association for Computational Linguistics. Co-located with ACL 2017.

Rehm, G., Moreno-Schneider, J., Gracia, J., Revenko, A., Mireles, V., Khvalchik, M., Kernerman, I., Lagzdins, A., Pinnis, M., Vasilevskis, A., Leitner, E., Milde, J., and Weißenhorn, P. (2019). Developing and Orchestrating a Portfolio of Natural Legal Language Processing and Document Curation Services. In Nikolaos Aletras, et al., editors, *Proceedings of Workshop on Natural Legal Language Processing (NLLP 2019)*, pages 55–66, Minneapolis, USA, June. Co-located with NAACL 2019. 7 June 2019.

Rehm, G., Berger, M., Elsholz, E., Hegele, S., Kintzel, F., Marheinecke, K., Piperidis, S., Deligiannis, M., Galanis, D., Gkirtzou, K., Labropoulou, P., Bontcheva, K., Jones, D., Roberts, I., Hajic, J., Hamrlová, J., Kacena, L., Choukri, K., Arranz, V., Vasiljevs, A., Anvari, O., Lagzdins, A., Melnika, J., Backfried, G., Dikici, E., Janosik, M., Prinz, K., Prinz, C., Stampler, S., Thomas-Aniola, D., Pérez, J. M. G., Silva, A. G., Berrío, C., Germann, U., Renals, S., and Klejch, O. (2020). European Language Grid: An Overview. In Nicoletta Calzolari, et al., editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France, 5. European Language Resources Association (ELRA). Accepted for publication.

Reimers, N., Eckle-Kohler, J., Schnober, C., Kim, J., and Gurevych, I. (2014). GermEval-2014: Nested Named Entity Recognition with Neural Networks. In Gertrud Faaß et al., editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 117–120. Universitätsverlag Hildesheim, Oktober.

Sang, E. F. T. K. and Meulder, F. D. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Walter Daelemans et al., editors, *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

Sang, E. F. T. K. and Veenstra, J. (1999). Representing text chunks. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 173–179. The Association for Computer Linguistics.

Sang, E. F. T. K. (2002). Introduction to the conll-2002 shared task: Language-independent named entity recognition. *CoRR*, cs.CL/0209010.

Schneider, J. M. and Rehm, G. (2018a). Curation Technologies for the Construction and Utilisation of Legal Knowledge Graphs. In Georg Rehm, et al., editors, *Proceedings of the LREC 2018 Workshop on Language Resources and Technologies for the Legal Knowledge Graph*, pages 23–29, Miyazaki, Japan, 5. 12 May 2018.

Schneider, J. M. and Rehm, G. (2018b). Towards a Workflow Manager for Curation Technologies in the Legal Domain. In Georg Rehm, et al., editors, *Proceedings of the LREC 2018 Workshop on Language Resources and Technologies for the Legal Knowledge Graph*, pages 30–35, Miyazaki, Japan, 5. 12 May 2018.

Strötgen, J. and Gertz, M. (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Verhagen, M., Saurí, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, July. Association for Computational Linguistics.

Weißenhorn, P. (2018). *Automatische Identifikation und Normalisierung von Zeitausdrücken in deutschsprachigen rechtlichen Texten*. Bachelor's thesis, Universität Potsdam, Potsdam.