Stefan Höfler, Michael Piotrowski*

Building Corpora for the Philological Study of Swiss Legal Texts

We describe the construction of two corpora in the domain of Swiss legal texts: The DS21 corpus is based on the Collection of Swiss Law Sources and contains historical legal texts from the early Middle Ages up to 1798; the Swiss Legislation Corpus (SLC) is based on the Classified Compilation of Swiss Federal Legislation and contains all current Swiss federal laws. The paper summarizes the key properties of both corpora, discusses issues encountered while building them, and outlines some applications.

1 Introduction

Legal texts are a fruitful object of study for the humanities. For historians, they represent a crucial source of information on the distribution of power in societies past and present, the values holding these societies together, their methods of resolving conflicts, and on the ways in which both the distribution of power and the underlying values have changed over time. For linguists, legal texts constitute a special case of highly conventionalized language use that strives, and often struggles, to find an optimal balance between rigorousness and flexibility, formality and understandability, that is, between expressing authority and yet being grounded in the everyday life of those affected by it. In fact, legislative texts do not merely describe, but create law.

The present paper introduces two annotated corpora of Swiss legal texts that have been compiled to provide scholars in the humanities with additional means to study this domain and to support the development of domain-specific natural language processing (NLP) tools. Both corpora are based on preexisting text collections, which were compiled according to criteria that are specific to this particular domain: The DS21 corpus is based on the *Collection of Swiss Law Sources* and the Swiss Legislation Corpus (SLC) is based on the *Classified Compilation of Swiss Federal Legislation*. The DS21 corpus is a corpus of historical legal texts, while the SLC comprises texts that constitute contemporary statutory law. We describe the range of texts contained in either corpus (sections 2.1 and 3.1 respectively), detail the characteristics of these texts (sections 2.2 and 3.2), and discuss some of the implications that such domain-specific features have for the automatic annotation of the texts (sections 2.3 and 3.3). The presentation of either corpus is followed by a brief survey of the range of humanities research it facilitates (sections 2.4 and 3.4). We conclude with a summary of the most important properties of the two corpora (Table 2).

*Authors are given in alphabetical order. MP is responsible for the DS21 corpus (partly funded under SNSF grant no. 124427); SH for the SLC (funded under SNSF grant no. 134701).

2 A Corpus of Historical Legal Texts

2.1 The Collection of Swiss Law Sources

The *Collection of Swiss Law Sources* is an edition of historical *sources of law* created on Swiss territory from the early Middle Ages up to 1798 (the downfall of the *ancien régime* in Switzerland). The Collection employs a broad definition of *law source* and includes not only acts, decrees, and ordinances, but also indentures, administrative documents, court transcripts, and other types of documents. It is edited by the Law Sources Foundation, which was established in 1894 specifically for this task. Since then the Foundation has edited and published over 100 volumes containing more than 60,000 pages of source material and commentary.

The primary sources are manuscripts in various regional historical forms of German, French, Italian, Rhaeto-Romanic, and Latin, which are transcribed, annotated, and commented by the editors. The critical apparatuses are in modern German, French, or Italian. Each volume contains an index of persons and places and a combined subject index and glossary.

The Collection organizes the sources by cantons and then generally subdivides them by areas of jurisdiction, such as towns or bailiwicks. At the time of this writing, it covers 17 of the 26 Swiss cantons to different extents. The edition of the Collection of Swiss Law Sources is an ongoing project and further volumes are in preparation.

Historians and historians of law are currently the primary users of the Collection, but it is also an important source for the Swiss-German Dictionary ("Idiotikon"). See Gschwend (2008) for a description of the Collection from a historian's point of view.

From 2009 to 2011 the Swiss National Science Foundation funded the digitization the Collection. As a result, the complete Collection is now available online in facsimile form[1]. The tables of contents were digitized using optical character recognition (OCR) with extensive post-editing to create an XML registry of the titles and the dates of creation of all texts in the Collection.

2.2 The Corpus: DS21

The availability of online facsimiles of all the volumes of the Collection represents a significant advance. It would, nevertheless, be desirable to have the full text of the historical sources and the apparatuses available in digital form.

During the retrodigitization project we discovered that digital typesetting data (FrameMaker 3 and 6 files) still exists for the 22 latest volumes. This provides us with a sizeable collection (about 4 million running word forms) of medieval and early modern texts free from errors introduced by digitization (whether OCR or retyping). We refer to this digital subset of the Collection as DS21.

DS21 contains volumes from ten cantons representing most linguistic and geographic regions of Switzerland. The subset also covers the full period of time documented by

[1] `http://ssrq-sds-fds.ch/online/`

Volume ID	Canton	Primary Language(s) of the Sources	Period Covered	Pages	Texts
SSRQ AG II 9	Aargau	German	1301–1798	735	415
SSRQ AG II 10	Aargau	German	1303–1797	735	530
SSRQ AR/AI 1	Appenzell	German	1409–1632	658	10
SSRQ BE I/13	Berne	German	1230–1796	1143	468
SSRQ BE II/9	Berne	German	1267–1797	992	690
SSRQ BE II/10	Berne	German	1277–1797	1191	808
SSRQ BE II/11	Berne	German	1256–1795	1305	894
SDS FR I/2/6	Fribourg	French, German	1296–1795	582	639
SSRQ GR B II/2	Grisons	German	1289–1832	1403	850
SSRQ LU I/1	Lucerne	German	1178–1501	592	436
SSRQ LU I/2	Lucerne	German	1426–1463	481	476
SSRQ LU I/3	Lucerne	German	1425–1489	731	465
SSRQ LU II/2	Lucerne	German	1301–1799	2428	692
SSRQ SG I/2/3	St. Gallen	German	754–1797	1173	518
SSRQ SG II/1/1	St. Gallen	German	1302–1601	492	16
SSRQ SG II/1/2	St. Gallen	German	1452–1701	538	300
SSRQ SG II/2/1	St. Gallen	German	1229–1799	1184	537
FDS TI A/1	Ticino	Italian	1286–1799	401	623
SDS VD B/2	Vaud	Latin, French	1211–1797	622	448
SDS VD C/1	Vaud	French, German	1530–1797	971	537
SDS VD C/2	Vaud	French	1539–1770	925	21
SSRQ ZH NF II/1	Zurich	German	1307–1794	522	318
				Total 19,804	10,691

Table 1: Composition of the DS21 collection. The languages given in the table refer to historical variants of these languages from the periods indicated.

the Collection: with the oldest text being from 754 and the most recent one being from 1832, it spans 1078 years. We therefore believe DS21 to be a good sample of the legal documents from the relevant period preserved in Swiss archives. Table 1 gives details of the composition of DS21.

We are now working on the conversion of the FrameMaker files of DS21 into a form that is usable for historians, linguists, and other researchers; more specifically, we want to create a corpus marked up according to the TEI P5 guidelines[2]. Several steps are required to reach this goal: First, the FrameMaker files must be converted into an open format; then the markup contained in these files must be regularized; the regularized markup subsequently enables the inference of the semantics of marked-up elements and the upconversion into a format that makes the semantics explicit. After this, links between textual elements—in particular between source text and annotations—can be detected and also made explicit using TEI markup.

To this end, we have developed a multi-stage conversion process for automatically converting FrameMaker files into valid TEI documents. The FrameMaker files contain

[2] http://tei-c.org/

only visually oriented markup, i.e., text is marked up as bold, italics, superscript, etc., not as title, date, or apparatus text. The first step is the conversion of the FrameMaker files into valid XHTML files with CSS stylesheets that closely mirror the layout and formatting of the FrameMaker documents (Piotrowski, 2010a). This conversion is more challenging than it may appear at first sight and requires deep processing of the FrameMaker files, in particular the tracking of style inheritance and font changes. Furthermore, while superficially similar, the conceptual models of FrameMaker and CSS differ significantly in details.

The bodies of the books are then converted to TEI, while the indices are converted into an application-specific XML format. At this point, the TEI markup is still similar to the XHTML markup, but the differences between XHTML and TEI require certain structural changes; for example, the `<br/>` element in XHTML marks the end of a line, whereas the `<lb/>` element in TEI marks the beginning of a line.

The upconversion primarily relies on the normalized typographic information produced in the previous conversion steps: it allows the automatic identification of article headings, footnote markers and the corresponding footnote text, and source text and commentary.

2.3 Automatic Annotation

DS21 is based on scholarly editions of historical texts: Each text is accompanied by a modern-language summary or title, the date and place of creation (as far as they are known), a description of the original physical document (archive location, writing material, measurements, etc.), and typically notes and a critical apparatus (describing additions, deletions, alterations, etc.). Figure 1 shows a text from DS21 as it appears in the printed version.

Together, the apparatuses, indices and glossaries form a rich annotation of the source texts. In contrast to summaries and apparatuses, which are located directly before and after each text, indices are stored separately in printed books, and while the index entries point to the relevant points in the text (by giving page and line numbers), there are no pointers *from the text* to the indices. In the TEI-encoded version, however, we want to have the information from the indices integrated into the text.

The FrameMaker files were used for printing the books and thus have the exact same line breaks and pagination. Throughout the conversion process this information is preserved, so that it is possible to identify the locations—with a precision of about one line or 10–15 word forms—to which index entries refer.

We have developed a tool that reads a TEI document and an index of persons and places (in an intermediate XML format), analyzes the index entries and inserts a `<persName>` or `<placeName>` element at the beginning of the corresponding line in the TEI document. For our prototype volume the element is then manually moved to the exact position in the text, so that it encloses the occurrence of the name. As index entries typically list some of the most frequent spelling variants (e.g., "Ausserrhoden: Äussere Rhoden, Außer Rhoden, Ußeren Roden, Ussern Rooden, Usroden, Uß

Roden, Ussroden"), we intend to have the program identify the occurrence in the text automatically.

Glossary entries are linked to the lines to which they refer, as the identification of the exact referent of a glossary entry is often hard, e.g., due to inflectional variation or different word order. All glossary entries referring to a particular text are used to generate keywords for this text, complementing the full-text search that is always possible for electronic texts but difficult for historical texts in different languages. We are currently developing a controlled vocabulary based on the combined glossaries of DS21.

Automatic linguistic annotation of the historical texts in DS21 would be nontrivial; it therefore currently does not contain linguistic annotation beyond what is provided by the glossaries. However, future annotation with further linguistic information is not precluded. Even though not all of existing annotation from the indices and glossaries is directly useful for linguistic purposes, it is nevertheless important to preserve this information, in order not to exclude other uses of DS21. While most historical corpora are based on scholarly editions, they usually do not preserve the critical apparatuses; Boschetti (2007) points out that, for the philologist, even the text of an authoritative edition has "no scientific value without the apparatus."

2.4 Applications

Work on the DS21 corpus is still ongoing, so we cannot report on actual uses of this corpus yet. However, we want to outline some potential uses of the corpus.

First, it will be possible to use the corpus for all types of research for which one would formerly have used the printed volumes or the digital facsimiles of the Collection. Typical research questions come from legal, economic, and social history. Even if a scholar does not use any new research methods, access to the texts will be easier and faster, and digital full text generally offers users more convenience, e.g., text can easily be copied. However, the corpus will also help to investigate research questions which the printed indices were not designed to support, it will facilitate studies that span several of the traditional volumes, and it will make it possible to explore topics orthogonal to the traditional regional organization of the Collection.

DS21 contains much information that is relevant for completely different fields of research besides legal history. For example, it contains many historical documents mentioning animals or animal products, such as regulations of hunting, fishery, butchery, and livestock. For archeozoology, these documents may represent valuable evidence complementing archeological findings and documenting, e.g., the presence of certain animals in a certain region. In fact, an archeozoologist has manually analyzed the indices of one printed volume to find texts to animals and animal products and compiled the species of fish mentioned documented for Lake Murten. However, this was tedious work, as the printed indices were not designed for this type of questions. The DS21 corpus will make such new uses of the texts much easier and will allow a multitude of new, as yet unanticipated, uses of the information it contains.

144. Strassenverordnung

1545 September 7* (menntags an unnser liebenn frouwen abend gepurt). *Rapperswil

Witter der straßen halb, so jn der statt verleitt, ist erkentt, wo stein, misthůffen, schiter oder holz, was da schådlich so jn straßen, alles dannen than und niemands sol verschont werden, sol stattknecht versorgen und sagen, darzů ist wachtmeister geordnot.

Rats-, Gerichtsprotokolle: *StadtARap, Bd. B 1, fol. 82r, Pap. 21,5/22 x 32 cm.*

Bemerkung

1566 Mai 14 (zinstags nach Panngratzi). *Rapperswil:* Joner unnd Bußkilcher sonnd einandren den weg zů Bußkilch uff der allmendt machen, damitt jederman tags und nachts wandlen mögen etc. unnd sonnds angends thůn etc. *(Rats-, Gerichtsprotokolle: StadtARap, Bd. B 2, S. 110). – Vgl. auch Ratsprotokolle: StadtARap, Bd. B 28, S. 204; B 31, S. 420; B 32, S. 515; 516; B 33, S. 151; B 39, S. 110; B 51, S. 205; Nr. 203.*

Figure 1: Example of a text contained in DS21: Ordinance on streets by the city of Rapperswil from September 7, 1545 (reproduced from Rechtsquellenstiftung des Schweizerischen Juristenverbandes, 2007, p. 407)

Second, DS21 will enable new modes of access, for example geographic browsing. We have created a prototype system that offers users an interactive map on which all places mentioned in the texts are marked (see Piotrowski, 2010b). By clicking on a place marker, the titles of the sources associated with a place are listed; clicking on a title brings up the corresponding source for reading. Other non-textual access modes could be based on persons or dates.

Third, the annotated electronic text facilitates interlinking with complementary resources, e.g., the *Deutsches Rechtswörterbuch* (DRW), the Swiss-German Dictionary, `e-codices.ch`, or `monasterium.net`.[3]

Finally, the corpus will be invaluable for research into NLP for historical languages, especially for research on normalization of spelling variants, as the glossaries provide lemmas and glosses for the most important words and because historical spelling variation is not confounded with digitization errors.

3 A Corpus of Contemporary Legislative Texts

3.1 The Classified Compilation of Swiss Federal Legislation

The Classified Compilation of the Federal Legislation (abbreviated SR) is a systematic collection of the contemporary statutory law of the Swiss Confederation. It comprises federal acts, ordinances issued by the federal authorities, the federal constitution, all

[3] We are working with these projects to create linkages between the various resources; for example, the DRW already links evidence from the Collection to the digital facsimiles.

cantonal constitutions, federal decrees, and treaties between the confederation and individual cantons or municipalities.

Each text is published in the three official languages of the Swiss Confederation: German, French, Italian. While the German and the French version of a legislative text are usually drafted and edited in parallel, the Italian version is, in most cases, merely a translation. However, all three official language versions are considered authentic, i.e., they all have equal legal force (see Lötscher, 2009).

As opposed to historical legal texts, contemporary laws are relatively easy to obtain—which greatly facilitates the process of corpus building. Nowadays, most legislative texts are published online, and the texts are usually not subject to copyright provisions that would prevent their use and distribution for research purposes. The Classified Compilation of Swiss Federal Legislation is no different in this regard: The collection can be accessed online at the website of the Swiss federal authorities, where all texts are available in HTML and in PDF format.[4]

3.2 The Corpus: The SLC

We have exploited the Classified Compilation of Swiss Federal Legislation to build an annotated corpus of contemporary legislative texts: the Swiss Legislation Corpus (SLC). This corpus has the following characteristics.

First, the SCL is *domain-complete*. It contains all texts published in the Classified Collection and thus comprises the whole body of contemporary legislative writing of the Swiss Confederation. In total, the SLC consists of 1915 texts per language. The sizes of the individual texts range from roughly 800 words (Federal Decree on the Coat of Arms, SR 111) to over 1.3 million words (Code of Obligations, SR 220).[5]

Second, the SLC is a *parallel corpus*. All texts are available in German, in French and in Italian. The conventions of Swiss legislative drafting ensure that even in their raw form, the texts exhibit a precise alignment of all language versions down to the level of individual sentences and enumeration items. Legal technicalities make it mandatory that each sentence and enumeration item of a legislative text can be identified unequivocally by naming the respective law and the number of the article, the paragraph and, where applicable, the sentence or enumeration item (e.g., *Art. 6 Par. 2 Ltr. b Federal Act on Professional Education, SR 412.10*). This identifier is language independent and thus ensures alignment between the text versions (see Figure 2). Occasionally, translation issues make it necessary that a statement that is expressed in a single sentence in one language has to be rendered as two sentences in another language. In these cases, the two sentences in the latter version are separated by a semicolon rather than a full stop. Thus, it can be guaranteed that the respective passage can still be referred to by one and the same sentence identifier in all language versions of the text.

Third, the SCL exhibits both *inter- and intra-textual time depth*. Despite the fact that all material found in the SLC constitutes contemporary Swiss federal law, there is

[4] `http://www.admin.ch/ch/d/sr/`

[5] The sizes refer to the German versions of the texts.

Art. 6 Verständigung und Austausch zwischen den Sprachgemeinschaften

[1] Der Bund kann Massnahmen im Bereich der Berufsbildung fördern, welche die Verständigung und den Austausch zwischen den Sprachgemeinschaften verbessern.

[2] Er kann insbesondere fördern:

a. die individuelle Mehrsprachigkeit, namentlich durch entsprechende Anforderungen an die Unterrichtssprachen und die sprachliche Bildung der Lehrkräfte;

b. den durch die Kantone, die Organisationen der Arbeitswelt oder die Unternehmen unterstützten Austausch von Lehrenden und Lernenden zwischen den Sprachregionen.

Art. 6 Compréhension et échanges entre les communautés linguistiques

[1] Dans le secteur de la formation professionnelle, la Confédération peut encourager les mesures qui favorisent la compréhension et les échanges entre les communautés linguistiques.

[2] Elle peut notamment encourager:

a. le plurilinguisme individuel, en veillant en particulier à la diversité des langues d'enseignement ainsi qu'à la formation des enseignants sur le plan linguistique;

b. les échanges d'enseignants et de personnes en formation entre les régions linguistiques, s'ils sont soutenus par les cantons, les organisations du monde du travail ou les entreprises.

Figure 2: Example of the alignment between the language versions of contemporary Swiss laws (excerpts from the German and the French versions of the Federal Act on Professional Education, SR 412.10)

a considerable diachronicity both between and within individual texts. As a whole, the corpus exhibits a time depth of 136 years: Its oldest text dates from June 22, 1875, its most recent text from March 30, 2011. Likewise, a time span of up to 122 years can be found within individual texts. Laws are subject to continuous alterations: articles, paragraphs, sentences or enumeration items may be added, changed or removed by the legislator. The Federal Act on Debt Enforcement and Bankruptcy (SR 281.1), for instance, originates from April 11, 1889, but its most recent update—in which an article and an enumeration item were added—only dates from September 1, 2011.

Fourth, the SCL is an *annotated corpus*. The texts have been converted into XML. At present, they are enriched with tags providing meta information (dates of issue and last update, title and number of the text, issuing authority, legal basis of the law), delineate structural units (chapter, section, article, paragraph, sentence and enumeration item boundaries) and indicate parts of speech. The annotation of syntactic structures is in preparation. To facilitate querying, the SCL has been imported into the ILM Corpus Workbench (Christ, 1994).

3.3 Automatic Annotation

The annotation of the SLC is largely determined by the characteristics of the domain of legislative texts. One task that plays a much more central role in the processing of laws than it does in other domains is text segmentation. As we have illustrated in the previous section, laws are heavily structured: They are partitioned into numbered chapters, sections, articles, paragraphs, sentences and enumeration items. Marking the boundaries of these structural units is crucial if one wants to preserve of the alignment between the individual language versions of the texts. Furthermore, the availability of such an annotation is a prerequisite for corpus-based studies into the discourse structure of legislative texts.

We have developed a tool that automatically marks the boundaries of textual units. The method that we employ combines line-based pattern matching with a look-behind strategy. For example, a line is recognized as an enumeration item if (a) it begins with a lowercase character (optionally accompanied by a Latin ordinal such as *bis, ter, quater*, etc.), followed by a full stop and one or more words, and if (b) the previous line has already been tagged either as an enumeration item or as the introductory sentence of an enumeration. The second line of the following excerpt (Article 26 of the Federal Act on Forest, SR 921.0), for instance, will thus be annotated as an enumeration item:

```
1  <paragraph issue_date="04/10/1991"><par_nr>1</par_nr>
   <enum_intro_sentence>Der Bundesrat erlässt Vorschriften über forstliche
   Massnahmen:</enum_intro_sentence>
2  a.  zur Verhütung und Behebung von Waldschäden;
```

Another feature of legislative texts is that each structural unit can be associated with a number of dates: the date of its first publication as part of a decree, the date of its official approval by the parliament or the people, and finally the date of its commencement. If the dates for a specific textual unit differ from those of the text as a whole, they are listed in a footnote attached to that unit (see Figure 3). Text segmentation must therefore also include date stamping.

We use pattern-matching methods to extract the dates from the footnotes and make them explicit in the markup of the corresponding text unit. In the markup, all text segmentation tags are augmented with attributes denoting the dates of the respective unit (e.g. `<paragraph issue_date="04/10/1991">`). By default, they are assigned the dates associated with the whole text. If, however, the respective unit is accompanied by a footnote mentioning a specific date (as it is the case with paragraphs 3 and 4 in Figure 3), that date is extracted from the footnote (e.g. by matching the string `'Angenommen in der Volksabstimmtung vom DATE'`) and inserted in the paragraph tag (thus replacing the default date).

The provision of precise date stamping for each textual unit allows for diachronic analyses of the linguistic material found in a text: We can, for instance, study if the language of earlier passages deviates from the language of passages that were added

Art. 175 Zusammensetzung und Wahl

[1] Der Bundesrat besteht aus sieben Mitgliedern.

[2] Die Mitglieder des Bundesrates werden von der Bundesversammlung nach jeder Gesamterneuerung des Nationalrates gewählt.

[3] Sie werden aus allen Schweizerbürgerinnen und Schweizerbürgern, welche als Mitglieder des Nationalrates wählbar sind, auf die Dauer von vier Jahren gewählt.[89]

[4] Dabei ist darauf Rücksicht zu nehmen, dass die Landesgegenden und Sprachregionen angemessen vertreten sind.[90]

89 Angenommen in der Volksabstimmung vom 7. Febr. 1999 (BB vom 9. Okt. 1998, BRB vom 2. März 1999 – AS **1999** 1239; BBl **1993** IV 554, **1994** III 1370, **1998** 4800, **1999** 2475 8768).

90 Angenommen in der Volksabstimmung vom 7. Febr. 1999 (BB vom 9. Okt. 1998, BRB vom 2. März 1999 – AS **1999** 1239; BBl **1993** IV 554, **1994** III 1370, **1998** 4800, **1999** 2475 8768).

Figure 3: Example of the date stamping of textual units in contemporary legislative texts (excerpt from the Federal Constitution, SR 101)

more recently, or we can investigate how the continuous insertion of additional material has affected the overall structure of a text.

In addition to text segmentation (and the annotation of textual meta information), we have annotated the words in all three language versions of the SLC with their part of speech and their lemma. We used TreeTager (Schmid, 1994) for this task. Domain-specific words unknown to TreeTagger constituted the main problem. However, except for archaisms like *bejahendenfalls* 'in case of affirmation,' most unknown words turned out to be nouns (including proper names and abbreviations) or adjectives. In most cases, TreeTager was able to guess the part of speech of these words correctly; only their lemmas could not be inferred.[6] We are confident that this situation can be remedied by equipping TreeTagger with a hand-made list of domain-specific expressions and their lemmas.

3.4 Applications

The need for annotated corpora of legislative texts has grown with the recent advance of legal linguistics as a theoretical and applied academic discipline (see Grewendorf and Rathert, 2009). The SLC is meant to make a contribution to filling this gap.

We currently use the SLC to study the stylistic properties of legislative texts. We are interested in investigating to what extent present-day Swiss laws comply with established stylistic guidelines for legislative drafting. To this aim, we use the method of error modeling employed in automated language checkers for technical writing: We

[6] The evaluation refers to the German part of the corpus. We have manually assessed the tagging of 1,000 randomly selected tokens. 85 (8.5%) of these tokens were unknown to TreeTagger; in total, 399,872 (6.7%) of the 5,896,451 tokens contained in the corpus were unknown to TreeTagger. For 67 (79%) of the manually evaluated unknown tokens, TreeTagger was able to guess the part of speech correctly; only 18 (21%) were assigned a wrong part of speech.

	DS21	SLC
Source	Collection of Swiss Law Sources	Classified Collection of Swiss Federal Legislation
Temporal classification	Historical	Contemporary
Text types	Statutes, decrees, regulations, indentures, treaties, administrative documents, court transcripts, letters, and others	Federal acts, ordinances, federal and state constitutions, federal decrees, non-international treaties
Languages	German, French, Italian (historical regional variants)	German, French, Italian (parallel)
Units of alignment	N/A	Sentences, enumeration items
Number of texts	10,691 (total)	1915 (per language)
Time depth	1078 years (754–1832)	136 years (1875–2011)
Intra-textual time depth	Unknown	Up to 122 years
Annotated information	Meta information, structural units, persons and place names	Meta information, structural units, parts of speech; syntax in preparation
Format	XML: TEI P5	XML, IMS Corpus Workbench (CWB)
Current applications	Historical research	Stylistic analysis, definition extraction

Table 2: Key properties of the two corpora

first specify linguistic features and textual patterns that indicate the violation of a specific style guideline and then search the SLC for occurrences of these indicators.[7]

The oft-cited rule that, in a good legislative text, an article should not contain more than three paragraphs, a paragraph should only contain one sentence, and a sentence should not make more than one statement may serve as an example (Federal Office of Justice, 2007, p. 358). An evaluation of the first two parts of the rule can be done by accessing the annotated structural units: The rule is violated in articles with more than three paragraphs and in paragraphs with more than one sentence. Violations of the third part of the rule can be found by searching for specific keywords and syntactic structures. Höfler (2011), for instance, points out that, among other things, sentence coordination, relative clauses introduced by the adverb *wobei* 'whereby,' and certain prepositions (e.g., *vorbehältlich* 'subject to' or *mit Ausnahme von* 'with the exception of') indicate that a sentence makes more than one statement.

In a related strand of research, we use the SLC to extract legal definitions (Höfler et al., 2011). We exploit the fact that, by convention, legal definitions follow a relatively small inventory of sentence patterns. Searching the SLC for these patterns allows us to identify these definitions. An automatic extraction of the concepts and terms defined in the present legislation can be of value to legal practitioners, to scholars of law, and to professionals involved in the drafting and editing of new acts and ordinances.

[7] We also work on employing the same method to check draft laws for style guideline violations.

4 Summary

In this paper, we have described the construction of two corpora of Swiss legal texts: the DS21 corpus, a corpus of historical legal texts, and the SLC, a collection of contemporary laws. The two corpora are complementary: Together, they reflect the historical development of Swiss legal language almost in its entirety (except for the period from 1798 until the formation of the federal state in 1848).

We have illustrated that the availability of such corpora facilitates a plethora of humanities research, particularly in the fields of history, linguistics, and law. We have also shown that the peculiarities of the legal texts represented in the two corpora had a strong impact on the tasks that had to be solved in order to build them. The work presented in this paper emphasizes that the construction of domain-specific corpora also involves putting work and effort into developing domain-specific annotation tools.

References

Boschetti, F. (2007). Methods to extend Greek and Latin corpora with variants and conjectures: Mapping critical apparatuses onto reference text. In Davies, M., Rayson, P., Hunston, S., and Danielsson, P., editors, *Proceedings of the Corpus Linguistics Conference CL2007*. University of Birmingham.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX 1994, 3rd Conference on Computational Lexicography and Text Research*, pages 23–32.

Federal Office of Justice, editor (2007). *Gesetzgebungsleitfaden: Leitfaden für die Ausarbeitung von Erlassen des Bundes*. Berne, Switzerland.

Grewendorf, G. and Rathert, M., editors (2009). *Formal Linguistics and Law*, volume 12 of *Trends in Linguistics*. Mouton de Gruyter, Berlin, Germany.

Gschwend, L. (2008). Rechtshistorische Grundlagenforschung: Die Sammlung Schweizerischer Rechtsquellen. *Schweizerische Zeitschrift für Geschichte*, 58(1):4–19.

Höfler, S. (2011). "Ein Satz – eine Aussage." Multipropositionale Rechtssätze an der Sprache erkennen. *LeGes: Gesetzgebung und Evaluation*, 22(2):275–295.

Höfler, S., Bünzli, A., and Sugisaki, K. (2011). Detecting legal definitions for automated style checking in draft laws. Technical report, Department of Informatics, University of Zurich.

Lötscher, A. (2009). Multilingual law drafting in Switzerland. In Grewendorf, G. and Rathert, M., editors, *Formal Linguistics and Law*, volume 12 of *Trends in Linguistics*, pages 371–400. Mouton de Gruyter, Berlin, Germany.

Piotrowski, M. (2010a). Document conversion for cultural heritage texts: FrameMaker to HTML revisited. In Antonacopoulos, A., Gormish, M., and Ingold, R., editors, *DocEng 2010: Proceedings of the 10th ACM Symposium on Document Engineering*, pages 223–226, New York, NY, USA. ACM.

Piotrowski, M. (2010b). Leveraging back-of-the-book indices to enable spatial browsing of a historical document collection. In Purves, R., Clough, P., and Jones, C., editors, *Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR'10)*, pages A17/1–2, New York, NY, USA. ACM.

Rechtsquellenstiftung des Schweizerischen Juristenverbandes, editor (2007). *Rechtsquellen der Stadt und Herrschaft Rapperswil*, volume SSRQ SG II/2/1 of *Sammlung Schweizerischer Rechtsquellen*. Schwabe, Basel, Switzerland. Prepared by Pascale Sutter.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.