Constructing and exploiting an automatically annotated resource of legislative texts

Stefan Höfler / Kyoko Sugisaki

University of Zurich UZH

Summary

We build an automatically annotated resource of legislative texts with a **particularly broad range of applications in legislative drafting, legal linguistics and the evaluation of legislation**.

We have developed the **domain- and language-specific NLP tools** required to provide the automatic annotations needed for the applications.

Our **Swiss Legislation Corpus** is based on the up-to-date collection of statutory law of the Swiss Confederation. All texts contained in the Swiss Legislation Corpus exist in German, French and Italian.

Swiss Legislation Corpus

The **Swiss Legislation Corpus** is based on the classified compilation of Swiss federal legislation. The Swiss Legislation Corpus:

– comprises the **up-to-date collection of statutory law of the Swiss Confederation**, i.e. the federal and all cantonal constitutions, federal acts, ordinances issued by the federal authorities, federal decrees and treaties between the Confederation and individual cantons or municipalities.

– contains parallel texts in Switzerland's three official languages, **German, French and Italian**. In particular, the German-language version is annotated syntactically by exploiting domain-specific NLP tools.

– consists of **more than 1900 texts** in total per language. The sizes of the individual texts range from roughly 800 words (Federal Decree on the Coat of Arms) to over 1.3 million words (Code of Obligations).

Domain-Specific NLP Tools

The texts contained in the classified compilation are available online in HTML and PDF format. We have converted the HTML files into a simple XML representation, to which we have added our automatic annotations (cf. Fig. 2). These annotations provide information on:

1. the boundaries of **text segments** (implemented for German, French, Italian and Romansh)

2. **parts of speech** and **lemmas** (implemented for German, French and Romansh)

3. **morphosyntactic features** and **grammatical functions** (implemented for German only)

4. **content types**, e.g. the goal and the scope of the respective law, definitions of terms (implemented for German only)

Applications

– We use the German-language part of the Swiss Legislation Corpus as a testing environment for the development of an **automatic style checker for legislative drafting**. This tool is aimed at detecting potential violations of domain-specific style guidelines in drafts of new legislation (cf. Fig. 1).

– The Swiss Legislation Corpus has also been used as the input to Bilingwis, a **statistical multilingual word concordance** (Volk et al., 2011). Bilingwis allows for translators of legislative texts to search for specific terms.

– The most recent strand of research exploiting the Swiss Legislation Corpus is concerned with gaining **empirical indications on the quality of legislative texts** (Uhlmann, 2014).

Illustration: Output of the Style Checker

Figure 1: Example of the output returned by the style checker.
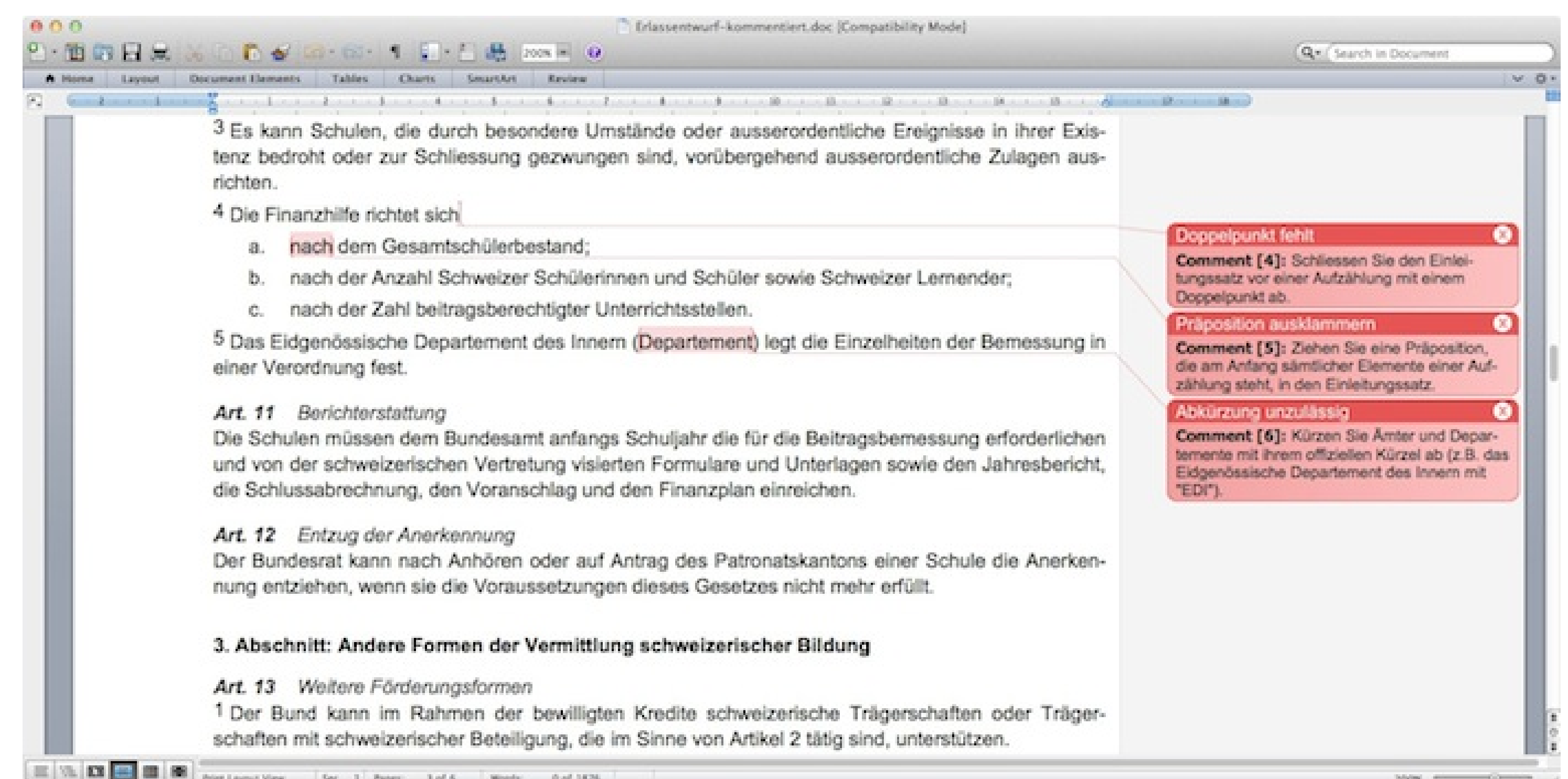
Example: Swiss Legislation Corpus

```
<p type="paragraph" subtype="full" nr="4">
 <s type="other">
  <t id="1456" word="4" pos="CARD" lemma="4"/>
 </s>

 <s type="full" content="legal definition">
  <t id="1457" word="Gewürzmischungen" pos="NN" lemma="Gewü
     rzmischung" dep="SUBJ"/>
  <t id="1458" word="sind" pos="VAFIN" lemma="sein" dep="ROOT"/>
  <t id="1459" word="Mischungen" pos="NN" lemma="Mischung" dep="
     PRED"/>
  <t id="1460" word="," pos="$," lemma=","/>
  <t id="1461" word="die" pos="PRELS" lemma="d" dep="SUBJ"/>
  <t id="1462" word="ausschliesslich" pos="ADJD" lemma="
     ausschliesslich" dep="ADV"/>
  <t id="1463" word="aus" pos="APPR" lemma="aus" dep="OBJP"/>
  <t id="1464" word="Gewürzen" pos="NN" lemma="Gewürz" dep ="PN"/>
  <t id="1465" word="bestehen" pos="VVFIN" lemma="bestehen" dep="
     REL"/>
  <t id="1466" word="." pos="$." lemma="."/>
 </s>
</p >
```

Figure 2: Excerpt from the Swiss Legislation Corpus:
[4] Gewürzmischungen sind Mischungen, die ausschliesslich aus Gewürzen bestehen.

[4] *Spice blends are blends that only consist of spices.*

References

Höfler, S. and Sugisaki, K. (2012). From drafting guideline to error detection: Automating style checking in legislative texts. In Proceedings of the EACL 2012 Workshop on Computational Linguistics and Writing, Avignon, France. Association for Computational Linguistics.

Höfler, S. and Piotrowski, M. (2011). Building corpora for the philological study of Swiss legal texts. Journal for Language Technology and Computational Linguistics (JLCL), 26(2):7789.

Uhlmann, F. (2014). Qualität der Gesetzgebung: Wüsche an die Empirie. In Griffel, A., editor, Vom Wert einer guten Gesetzgebung, pages 171181. Stäpfli, Bern.

Volk, M., Göring, A., Lehner, S., Rios, A., Sennrich, R., and Uibo, H. (2011). Worldaligned parallel text: A new resource for contrastive language studies. In Proceedings of the Conference on Supporting Digital Humanities, Copenhagen, Denmark.

Contact

Address
Institute of Computational Linguistics
Binzmühlestrasse 14, CH-8050 Zürich

This project is funded under Swiss National Science Foundation grant no. 134701.

Email
Stefan Höfler (project leader)
hoefler@cl.uzh.ch