Constructing and exploiting an automatically annotated resource of legislative texts

Stefan Höfler, Kyoko Sugisaki
Institute of Computational Linguistics, University of Zurich
Binzmühlestrasse 14 8050 Zürich, Switzerland
{hoefler,sugisaki}@cl.uzh.ch

Abstract

In this paper, we report on the construction of a resource of Swiss legislative texts that is automatically annotated with structural, morphosyntactic and content-related information, and we discuss the exploitation of this resource for the purposes of legislative drafting, legal linguistics and translation and for the evaluation of legislation. Our resource is based on the classified compilation of Swiss federal legislation. All texts contained in the classified compilation exist in German, French and Italian, some of them are also available in Romansh and English. Our resource is currently being exploited (a) as a testing environment for developing methods of automated style checking for legislative drafts, (b) as the basis of a statistical multilingual word concordance, and (c) for the empirical evaluation of legislation. The paper describes the domain- and language-specific procedures that we have implemented to provide the automatic annotations needed for these applications.

Keywords: legal texts, domain-specific annotation, style checking

1. Introduction

In this paper, we report on the construction of a resource of Swiss legislative texts that is automatically annotated with structural, morphosyntactic and content-related information, and we discuss the exploitation of this resource for the purposes of legislative drafting, legal linguistics and translation and for the evaluation of legislation.

We have detailed individual aspects and components of this resource in previous publications mentioned throughout the text. In this paper, we provide a synthesis, report on recent developments and introduce two novel applications of our resource.

The paper is organised as follows. We will first characterise the texts contained in the resource (section 2), then detail its automatic annotation (section 3) and finally outline its multiple areas of application (section 4).

2. Text basis

Our resource is based on the classified compilation of Swiss federal legislation, i.e. the up-to-date collection of statutory law of the Swiss Confederation.[1] It comprises 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.

All texts contained in the classified compilation exist in German, French and Italian. All three language versions are considered equally authentic (Lötscher, 2009).[2] For this reason, each provision in the texts can be referenced unequivocally by indicating its position in the text (article, paragraph, sentence, enumeration item), independent of the language. Even in their non-annotated form, the language versions contained in the collection are thus precisely aligned down to the level of individual sentences and enumeration items. The collection thus amounts to an inherently aligned parallel corpus.

In total, the classified compilation consists of more than 1900 texts 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).[3]

3. Construction

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. These annotations provide information on (a) the boundaries of text segments (implemented for German, French, Italian and Romansh), (b) parts of speech and lemmas (implemented for German, French and Romansh), (c) morphosyntactic features and (d) content types (implemented for German only).

Whether some information has been added to the resource or not is driven by the applications for which it is used; hence the differences between the individual language versions. The German-language version has been used as a testing environment for the development of an automatic style checker for legislative drafting (cf. Section 4.1.) and as a resource for gaining empirical indications on the quality of legislative texts (cf. Section 4.3.). For these applications, all levels of annotation are needed. (The annotation of the boundaries of text segments is more or less language-independent and has thus been implemented for all language versions.) The German, French and Romansh version have further been used as the input to a statistical multilingual word concordance (cf. Section 4.2.). This application only required the annotation of parts of speech and lemmas; only these levels of annotation have thus also been implemented for French and Romansh.

[1] `www.admin.ch` > Federal law > Classified compilation

[2] Some of the texts are also available in Romansh and English; however, these versions do not have legal force.

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

3.1. Text segmentation and POS-tagging

Law texts are heavily structured: they are partitioned into numbered chapters, sections, articles, paragraphs, sentences and enumeration items. We have developed a tool that automatically detects the boundaries of such structural units and marks them in the XML representation. The tool employs a line-based pattern-matching algorithm with look-around (Höfler and Piotrowski, 2011). As it mainly exploits formatting information, it is more or less language-independent and has consequently been implemented for all language versions contained in our resource.

The German and French version have additionally been annotated with part-of-speech and lemma information provided by TreeTager (Schmid, 1994). To this aim, domain-specific expressions had to be pre-tagged in order to avoid part-of-speech tagging errors, and TreeTagger's own list of abbreviations had to be complemented with a list of abbreviations specific to Swiss federal laws.

3.2. Morphosyntactic analysis

The tokens of the German version of the resource have been further annotated with morphological (case, number, person, tense, etc.) and partial syntactic information (grammatical function, topological field).

For an initial morphological analysis, we use Gertwol (Haapalainen and Majorin, 1994), a classical two-level rule-based morphological analyser that provides fine-grained morphological features. However, the fact that German is a morphologically rich language made it necessary to develop our own post-processing routines to further disambiguate the output delivered by Gertwol: we have developed a rule-based disambiguation system in the framework of Constraint Grammar (Karlsson et al., 1995)[4], a grammar formalism that has been successfully employed for morphological disambiguation in English (Voutilainen, 1995) as well as in morphologically rich languages such as Irish (Uí Dhonnchadha, 2006) and Icelandic (Loftsson, 2008).

For the disambiguation of verbs, we exploit the theory of topological fields (vorfeld, mittelfeld, nachfeld) developed in traditional German grammar. This theory categorises German clauses into three types depending on the positioning of their verbal elements: verb-first, verb-second and verb-final clauses. From the constraints that apply to each of these types, we have derived a set of heuristics that allow us at the same time to (a) further disambiguate the verbal elements and (b) identify the boundaries of the topological fields (Sugisaki and Höfler, 2013b). As an example, a verb form that could be 1st person plural, 3rd person plural or infinitive (e.g. *schreiben* 'write') must be an infinitive if it occurs in a verb-second clause and its left-most verbal neighbour is a modal. At the same time, the modal marks the boundary between the vorfeld and the mittelfeld of that clause and the infinitive marks the boundary between mittelfeld and nachfeld.

Like this, we are able to reduce the rate of POS-tagging mistakes from 10.2% to 1.6%. Our evaluation has shown that the largest part of this reduction is achieved by heuristics that check the compatibility of morphological features within the long-distance relationships of discontinuous verbal elements. Since in law texts, the average distance between the left and right brackets of clauses is relatively large (9.5 tokens in our test data), this domain also makes it necessary that a wide context window is used for the morphosyntactic disambiguation of verbs.

As German is a dependent-marking language and exhibits relatively free word order, disambiguating the morphology of nouns is essential for the recognition of grammatical functions. We have developed a heuristics-based disambiguation strategy that exploits the fact that nominal elements must exhibit agreement with other elements within (a) the noun phrase, (b) potential superordinate noun phrases and (c) the clause. Agreement within each of these three contexts is checked successively, and after each check only those morphological analyses remain that fulfill the agreement requirements for the respective context. If, for instance, a noun could be either nominative or accusative case and it appears in a clause with no other nominal elements that could be nominative case, then it must be nominative as each clause must have a subject. Like this, we are able to reduce the rate of morphological ambiguity in nouns from 91.12% to 32.31% (Sugisaki and Höfler, 2013a).

With regard to the syntactic analysis of the texts, our approach thus amounts to supertagging (Bangalore and Joshi, 1999) in the sense that we annotate rich syntactic information such as grammatical functions and typological fields, which could then be combined to obtain a coherent syntactic parse. Similar approaches have been proposed for dependency grammar (Foth et al., 2010; Harper and Wang, 2010), Tree Adjoining Grammar (Bangalore and Joshi, 1999), Head-driven Phrase Structure Grammar (Zhang et al., 2009) and Categorical Grammar (Clark, 2011). What is new about our approach is that we combine supertagging with heuristics derived from the theory of topological fields to disambiguate verbal elements.

3.3. Recognition of content types

We also annotate individual text segments with information on the content they express. While most articles in a legislative text consist of ordinary norms, some serve special functions. Among these are articles containing transitional provisions, repeals and amendments of current legislation, definitions of the subject matter, the goal and the scope of the respective law, definitions of terms, as well as preambles and commencement clauses. We use a range of features to automatically identify such contents: e.g. the position in the text, certain keywords and typical sentence patterns. The article defining the goal of a law, for instance, usually appears at the beginning of the text and its header contains the words *Zweck* ('purpose') or *Ziel* ('aim').

The content type most difficult to detect automatically are definitions of terms. Three general forms of definitions of terms can be distinguished: bracketed definitions, enumerated definitions and sentential definitions (Höfler et al., 2011). In bracketed definitions, the defined term or abbreviation occurs in parentheses after its definition:

(1) *Der Bundessicherheitsdienst (Dienst) übt die*

[4]We employ VISLCG21 (`http://beta.visl.sdu.dk`; last visited on 12/10/2013) to compile hand-crafted Constraint Grammar rules.

Table 1: Precisions of the individual search patterns. For each pattern, 150 randomly chosen positives were evaluated (or fewer if a smaller total number of positives were returned by the system).

Type (Pattern)	Total Returned	Total Evaluated	True Positives	False Positives	Precision
Bracketed Definitions	7691	150	141	9	0.94
Enumerated Definitions	1072	150	149	1	0.99
Sentential Definitions:					
– *Als **X** gilt/gelten **Y***	1498	150	144	6	0.96
– ***X** umfasst/umfassen **Y***	713	150	121	29	0.81
– ***X** liegt/liegen vor, wenn **Y***	116	116	116	0	1.00
– *Unter **X** ist/sind **Y** zu verstehen*	23	23	23	0	1.00
– ***X** ist/sind **Y***	1727	150	138	12	0.92

Aufgaben im Sinn von Artikel 1 aus.
'The Federal Security Service (Service) performs the tasks according to Article 1.'

Enumerated definitions occur as a list of numbered items:

(2) *In diesem Gesetz bedeuten:*

a. Museum des Bundes: Museum, das organisatorisch zur zentralen oder dezentralen Bundesverwaltung gehört;

b. Sammlung des Bundes: Bestand an beweglichen Kulturgütern, der im Eigentum des Bundes oder einer Einheit der dezentralen Bundesverwaltung steht.

'In this act shall mean:

a. museum of the Confederation: a museum affiliated to the central or decentralised federal administration;

b. collection of the Confederation: a stock of mobile cultural goods in the possession of the Confederation or of a unit of the decentralised federal administration.'

Sentential definitions come in the form of a full sentence:

(3) *Als Rodung gilt die dauernde oder vorübergehende Zweckentfremdung von Waldboden.*
'Clearing shall be deemed to be the permanent or temporary misuse of forest soil.'

We have identified five general patterns that sentential definitions typically follow:

(4) ***Als X gilt/gelten Y***
'**X** is/are deemed to be **Y**'

(5) ***X umfasst/umfassen Y***
'**X** comprises/comprise **Y**'

(6) ***X liegt/liegen vor, wenn Y***
'**X** is/are present if **Y**'

(7) ***Unter X ist/sind Y_1 zu verstehen(, Y_2)***
'**X** is/are to be understood as **Y**'

(8) ***X ist/sind Y***
'**X** is/are **Y**'

We found that bracketed definitions, enumerated definitions and sentential definitions, with the exception of the pattern indicated in (8), can be detected by employing regular expressions operating on the surface of the text alone (Höfler et al., 2011). For the detection of sentential definitions that follow pattern (8), it was necessary that we resorted to additional morphosyntactic information. Clauses matching pattern (8) need to be further filtered in order for the system to only return those copula clauses that constitute definitions of terms. To this aim, we have developed the following filtering rules:

(8') a. The copula is the main verb, in indicative mood and not accompanied by a modal verb.

b. The subject or predicate of the copula clause is not an organisation and does not contain words such as *Zweck* ('purpose'), *Ziel* ('aim'), *Voraussetzung* ('precondition') or *Ausnahme* ('exception').

The following copula clause is, for example, filtered out by rule (8'b):

(9) *Zuständige Behörde ist das Bundesamt.*
'The responsable authority is the Federal Office.'

To determine the *recall* that our search patterns exhibit we had 27 legislative texts manually annotated for legal definitions. The texts were selected from across all domains of law: 2 texts were selected from constitutional law, 2 from private law, 2 from criminal law, 2 from education, science and culture law, 2 from national defence law, 2 from finance law, 3 from energy and transport law, 10 from health, employment and social security law, and 2 from economy law. The annotators were told to mark whatever statement they deemed a legal definition. Of the 225 paragraphs that the annotators had marked as containing legal definitions, our system recognised 210, which amounts to a recall of 91%. *Precision* was evaluated for each pattern individually. The developed search strategies were applied to all texts contained in our corpus. For each pattern, we evaluated a set of 150 randomly chosen instances returned by the system – or the total number of instances returned if it was less than 150. The results are detailed in Table 1. Precision was at 92%

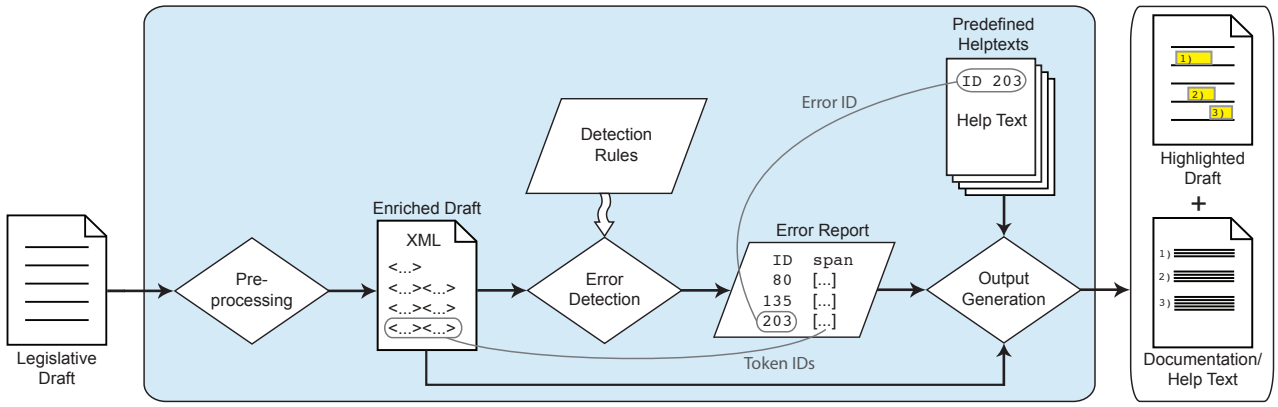

Figure 1: Architecture of the style checking tool.

or above for all but one of the evaluated patterns: sentential definitions with *umfassen* ('comprise') ranged slightly below at only 81% precision.

Initially, our system recognized a total of 4099 copula clauses matching pattern (8). After applying the filtering rules in (8'), the system had identified 1727 of these clauses as definitions of terms. 138 of 150 randomly chosen positives identified by the system were indeed definitions of terms, which amounts to a precision of 92%.

Most of the patterns we devised thus proved to be fairly reliable indicators for the presence of a legal definition.

4. Exploitation

Our resource is currently being exploited (a) as a testing environment for developing methods of automated style checking for legislative drafts, (b) as the basis of a statistical multilingual word concordance, and (c) for the empirical evaluation of legislation.

4.1. Automated style checking

We use the German-language part of our resource 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.

Figure 1 provides an overview of the architecture of the tool. The input document is a legislative draft in Word format. We exploit the XML structure underlying this format. In a first step, the input text is enriched with the various levels of annotation detailed in Section 3. In Figure 1, this step is labelled as "Pre-processing." In a second step, specific detection rules are then applied to the enriched text to identify violations of style guidelines. In Figure 1, this step is labelled as "Error Detection." Finally, the output document is generated by highlighting, in the original document, the passages that have been detected as containing a potential style guide violation and by inserting word comments that provide documentation with regard to the type of error that has been detected. Figure 2 provides an illustration of what the output of the style checking tool looks like.

The main method employed by our tool is that of error modelling. The texts to be assessed are automatically searched for specific features that indicate a style guideline violation. For this to be possible, the specifics of "errors" first have to be anticipated and modelled (Höfler and Sugisaki, 2012). As even laws that are currently in force may contain style guideline violations, our resource provides an ideal environment to test whether particular errors have been modelled correctly or whether the detection strategy grossly over- or undergenerates.

Depending on what type of styleguide violation is to be modelled, different parts of annotated information needs to be accessed. Violations of some stylistic rules can be detected, for instance, purely on the basis of the information on the beginning and end of text segments (e.g. "sections should not contain more than twelve articles, articles should not contain more than three paragraphs and paragraphs should not contain more than one sentence"). For other style guideline violations, the information on the extent of particular text segments has to be combined with pattern matching (e.g. "the sentence introducing an enumeration must end in a colon") or with more complex morphosyntactic features (e.g. "the antecedent of a pronoun must be within the same article as the pronoun"). Morphosyntactic annotations also have to be accessed when checking for violations of rules that pertain to the use of specific terms (e.g. "the modal *sollen* 'should' is to be avoided"), syntactic constructions (e.g. "complex participial constructions preceding a noun should be avoided") or combinations thereof (e.g. "obligations where the subject is an authority must be put as assertions and not contain a modal verb"). Some of these rules only apply to specific contents: the modal *sollen* 'should', for instance, must be avoided in ordinary norms but is acceptable where the goal of a law is defined. To determine whether a particular occurrence of it violates the style guidelines for legislative texts, the style checker thus also needs to resort to the annotations indicating the content that the respective text segment expresses.

4.2. Multilingual concordance

Our resource has also been used as the input to Bilingwis (formerly known as "Align+Search"), a statistical multilingual word concordance (Volk et al., 2011). Bilingwis allows translators of legislative texts to search for specific terms

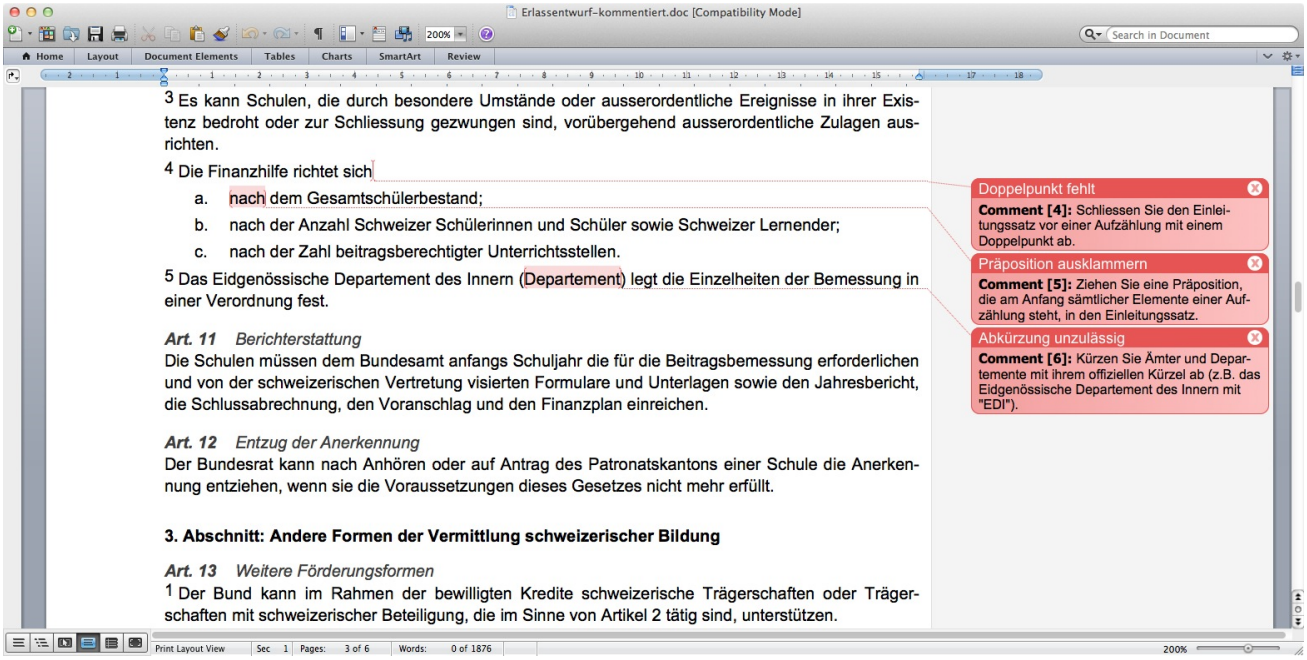

Figure 2: Sample output of the style checking tool.

in existing texts and to inspect the various translations of these terms and the contexts in which they are used. The word alignment provided by Bilingwis is based purely on statistics, which makes it more flexible than systems based on manually compiled dictionaries. Furthermore, the search results can be sorted by frequency and thus conclusions can be drawn on the way individual words are used in the domain.

The Bilingwis interface to our resource is currently available for German and French and for German and Romansh.[5]

4.3. Empirical evaluation of legislation

The most recent strand of research exploiting the present resource is concerned with gaining empirical indications on the quality of legislative texts (Uhlmann, 2014). Using similar or even the same procedures that we also employ for domain-specific style checking, we calculate how the individual texts compare with regard to specific features: Which laws exhibit particularly "heavy" articles, i.e. articles consisting of more than three paragraphs? Which laws exhibit particularly long and complex sentences? Which laws are particularly prone to remaining at the relatively vague level of "soft" obligations expressed by the modal *sollen* ('should')? Which laws leave a lot of room for interpretation and discretionary decisions by encompassing particularly high numbers of provisions with the modal verb *können* ('can')? The output of these evaluations serves as the input to research, carried out by law scholars, into the quality of particular pieces of legislation.

[5]The German-French Bilingwis implementation of our resource can be accessed at `http://kitt.cl.uzh.ch/kitt/bilingwis\_scl/slc2` (last visited on 14/10/2013); it has been set up by Roger Wechsler. The German-Romansh implementation can be accessed at `http://kitt.cl.uzh.ch/kitt/bilingwis_derm/` (last visited on 11/03/2014) and has been developed by Manuela Weibel (Weibel, 2014).

5. Conclusion

The present paper introduces an automatically annotated resource of legislative texts with a particularly broad range of applications in legislative drafting, legal linguistics and the evaluation of legislation. It shows that domain- and language-specific procedures are required to provide the automatic annotations needed for these applications.

Acknowledgments

This work has been funded under SNSF grant 134701.

6. References

Bangalore, S. and Joshi, A. K. (1999). Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Clark, S. (2011). Supertagging for Combinatory Categorial Grammar. In *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6)*, pages 19–24.

Foth, K., By, T., and Menzel, W. (2010). Guiding a constraint dependency parser with supertags. In Bangalore, S. and Joshi, A. K., editors, *Supertagging: Using Complex Lexical Descriptions in Natural Language Processing*. MIT Press, Cambridge, Massachusetts and London, England.

Haapalainen, M. and Majorin, A. (1994). GERTWOL: ein System zur automatischen Wortformerkennung deutscher Wörter. Technical report, Lingsoft, Inc.

Harper, M. P. and Wang, W. (2010). Constraint dependency grammars: Superarvs, language modeling, and parsing. In Bangalore, S. and Joshi, A. K., editors, *Supertagging: Using Complex Lexical Descriptions in Natural Language*

Processing. MIT Press, Cambridge, Massachusetts and London, England.

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):77–89.

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*, pages 9–18, Avignon, France. Association for Computational Linguistics.

Höfler, S., Bünzli, A., and Sugisaki, K. (2011). Detecting legal definitions for automated style checking in draft laws. Technical Report CL-2011.01, University of Zurich, Institute of Computational Linguistics, Zürich.

Karlsson, F., Voutilainen, A., Heikkila, J., and Anttila, A. (1995). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

Loftsson, H. (2008). Tagging icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31:47–72, 5.

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.

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.

Sugisaki, K. and Höfler, S. (2013a). Incremental morphosyntactic disambiguation of nouns in german-language law texts. In *ESSLLI-13 Workshop on Extrinsic Parse Improvement (EPI)*.

Sugisaki, K. and Höfler, S. (2013b). Verbal morphosyntactic disambiguation through topological field recognition in german-language law texts. In *Third International Workshop on Systems and Frameworks for Computational Morphology (SFCM 2013)*, Berlin, Germany.

Uhlmann, F. (2014). Qualität der Gesetzgebung: Wünsche an die Empirie. In Griffel, A., editor, *Vom Wert einer guten Gesetzgebung*, pages 171–181. Stämpfli, Bern.

Uí Dhonnchadha, E. (2006). *Part-of-speech tagging and partial parsing for Irish using finite-state transducers and constraint grammar*. Ph.D. thesis, Dublin City University.

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

Voutilainen, A. (1995). A syntax-based part-of-speech analyser. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics (EACL '95)*, pages 157–164, San Francisco, CA, USA. Morgan Kaufmann.

Weibel, M. (2014). Aufbau paralleler Korpora und Implementierung eines wortalignierten Suchsystems für Deutsch – Rumantsch Grischun. Master's thesis, University of Zurich, Zurich, Switzerland. `http://www.cl.uzh.ch/studies/theses/lic-master-theses/MLTA_Masterarbeit_Manuela_Weibel.pdf`.

Zhang, Y.-z., Matsuzaki, T., and Tsujii, J. (2009). HPSG supertagging: A sequence labeling view. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT '09)*.