

Towards Data-Driven Style Checking: An Example for Law Texts

Kyoko SUGISAKI ^{a,1}

^a*Institute of Computational Linguistics, University of Zurich, Switzerland*

Abstract. We present a novel approach to detecting syntactic structures that are inadequate for their domain context. We define writing style in terms of the choices between alternatives, and conducted an experiment in the legislative domain on the syntactic choice of nominalization in German, i.e. complex noun phrase vs. relative clause. In order to infer the stylistic choices that are conventional in the domain, we capture the contexts that affect the syntactic choice. Our results showed that a data-driven binary classifier can be a viable method for modelling syntactic choices in a style-checking tool.

Keywords. Natural language processing, style checking, law texts

1. Introduction

Law texts are often criticized as being incomprehensible to non-lawyers (e.g. [1,2,3]). Legal texts are normative, and they describe legal conditions and consequences. Legal conditions are complex because they attempt to include all imaginable cases, while being both general and sufficiently precise. This requirement leads to the excessive use of complex syntactic structures, such as coordination structures, and clausal modifiers, such as relative clauses and subordinate clauses. Legislative language is characterized by long sentences, nominalization, complex morphological derivations, personalization, and archaic words and phrases [4].² Long sentences and nominalization have been regarded as contributing to the complexity of texts (e.g. [4,6,7]).

To improve the comprehensibility of Swiss law texts, legislative drafts are edited by linguistic and legal experts who improve the quality of the language (cf. [8]). Text editing is a time-consuming task, and style checkers have been developed since the 1980s. Examples are UNIX Writers Workbench [9], IBMs CRITIQUE [10], MultiLint [11], FLAG [12], and Check-Point [13]. In these systems, style errors are modeled by anticipating error types, that is, by pre-defining them in form of rules, based on expert knowledge. These rules are then applied to passages in texts. In the UNIX Writers Workbench, for example, the style rules are based on style guidelines, writing standards in rhetorical traditions, and the results of psychological and linguistic research (Frase 1983).

¹The project is funded under SNSF grant 134701.

²In Italian and English law texts, unlike newspapers, prepositional phrases are used extensively, whereas verbal phrases are used less often [5].

However, rule-based methods used to detect style errors present two challenges. First, the tipping point between acceptable and unacceptable choices is typically unknown. As Ravin (1988, p. 109) succinctly noted, the perception of style errors or style weaknesses depends on the writer, the reading audience, and the type of document involved. Therefore, style checking tools are often limited to a certain sub-language, such as technical documentation (e.g. MultiLint). Second, in rule-based methods, style errors are detected uniformly if they belong to error types (e.g. passive sentence). Ideally, however, each instance of a type of error is judged individually based on its context. Nonetheless, this issue has received little attention since the 1980s. An exception is MultiLint, in which the context of the discourse triggers the identification of style errors [11].

Our study extends the MultiLint system by deducing from a domain corpus the contexts that might affect choices of syntax. We investigate ways to model stylistic choices computationally. Our goal is to develop a method that differentiates individual style violations according to the context and then apply it in a domain-specific style checking tool. The study focuses on the domain of German-language law texts from Switzerland and the syntactic choice of complex noun phrase vs. relative clause, that is, the violation of the domain style rule “avoidance of complex noun phrases.”

The paper is organized as follows: In the next section, we propose a methodology used to model syntactic choice. We then experiment with different models by applying them to the domain corpus. We conclude the paper by providing an evaluation of the context-sensitive style error detection.

2. Computational Model of Syntactic Choice

Style guidelines suggest rephrasing complex noun phrases into relative clauses (e.g. The Zürich style guideline [14]; The Bern style guideline [15]). In practice, it can be observed that this rule of the guidelines is only loosely adhered to: complex noun phrases are used frequently in actual Swiss law texts [16].

To detect violations of this style rule, we need to know the tipping point between acceptable and unacceptable syntactic complexity of complex noun phrases. Even if the examples in the guidelines show stylistically incorrect complex noun phrases (1), it is difficult to decide a clear-cut threshold for style violations. We do not know, for example, whether simplified participle phrases such as (2) and (3) would still be acceptable, or if the one would be acceptable while the other one would not.

- (1) Complex noun phrase provided as example of bad style in the Zurich style guideline [14]:
Die Berechnung erfolgt auf Grund *der am 1. Januar des dem Auszahlungsjahr vorangehenden Jahres im Kanton bekannten definitiven Steuerfaktoren*
‘Calculation is based on the definitive taxation factors known in the canton on 1 January of the year before the expenditure took place’
- (2) Simplified version of (1)
Die Berechnung erfolgt auf Grund *der am 1. Januar des dem Auszahlungsjahr vorangehenden Jahres bekannten definitiven Steuerfaktoren*
‘Calculation is based on the definitive taxation factors known on 1 January of the year before the expenditure took place’

- (3) Simplified version of (1)
Die Berechnung erfolgt auf Grund *der im Kanton bekannten definitiven Steuerfaktoren*
'Calculation is based on the definitive taxation factors known in the canton'

To understand the tipping point between acceptable and unacceptable, we employ statistical classification as a method. We define writing style as the choice between alternatives following the stylistic definition [17, pp. 5ff]. Here, the syntactic choice to be investigated is the one between complex noun phrases and relative clauses. Complex noun phrases can be rephrased into relative clauses and vice versa, without fundamentally changing the semantic content. To operationalise the modelling of syntactic choice, decision factors are integrated as features. We hypothesise that (i) the main factor in decisions about syntactic choice is syntactic complexity and (ii) complex noun phrases that are similar to relative clauses in syntactic complexity are more likely to be rephrased by editors. The idea is that the classifiers learn, based on these decision factors, how the two syntactic variants were selected in current Swiss law texts. Therefore, we use current law texts as training data for the classifiers.

To this end, we created a test suite of complex noun phrases and relative clauses extracted from Swiss law texts. The test suite is built automatically using a supertagger [18,19,20]. For the validation of our method, we split the test suite into a training set, cross-validation set and test set. For the test set, two law texts were selected from different periods because writing style evolves over time: One text is the Swiss Civil Code. Written in 1907, it is one of the oldest law texts. The second text is the Animal Protection Act, which was written in 2005 and can thus be considered a current law text. The remaining items of extracted complex noun phrases and relative clauses are separated into a training set (80%, each 17,420 items for complex noun phrases and relative clauses) and a cross-validation set (20%, each 4,355 items for complex noun phrases and relative clauses). The cross-validation set was used in experiments to determine the best classifiers and features for the task.

2.1. Method and Features

Method The task of the classifier is to discriminate complex noun phrases into two classes: (i) stylistically adequate complex noun phrases and (ii) stylistically inadequate complex noun phrases (i.e. relative-clause-like complex noun phrases). For this purpose, we used supervised machine learning methods, particularly Naive Bayes (NB) and Support Vector Machines (SVM). Naive Bayes (NB) is a simple and efficient supervised method based on the naive assumption of Bayes' theorem that features are mutually independent given the context of the class. We used two types of NB: Gaussian NB and multinomial NB. In Gaussian NB, the likelihood of the features is assumed to be Gaussian, that is, normally distributed, whereas it is assumed to be multinomial in multinomial NB. SVM is a hyperplane-based discriminative classifier and is the state-of-the-art method for classification. In corpus linguistics, logistic regression has been widely used to investigate multifactorial language data in syntactic alternation, particularly dative shift [21,22], genitive alternation [23], and heavy NP shift [22]. Compared with other machine learning methods, logistic regression is advantageous in linguistic research because the estimated coefficients (or parameters) are explanatory and help in interpreting how they affect the probabilities of events (cf. [24]). However, we use these three

Features	Corresponding tokens in (4-a) & (4-b)#
(A1) Complex NP: # of tokens in the embedded phrase of complex NP RC: # of corresponding tokens in RC	<i>für die Erfüllung ... notwendigen/notwendig</i> 9
(A2) # of nouns and pronouns in (A1)	<i>Erfüllung, Schutzpflichten, Schweiz</i> 3
(A3) # of prepositions, comparative conjunctions and pronominal adverbs in (A1)	<i>für</i> 1
(A4) # of commas and coordinating conjunctions in (A1)	- 0
(A5) # of predicates in (A1)	<i>notwendigen/notwendig</i> 1
(A6) Complex NP: # of tokens in the head NP of a complex NP RC: # of tokens in an antecedent NP	<i>die, Massnahmen</i> 2
(B1) # of remaining tokens in the projected NP of (A6)	<i>auf ihrem Gebiet</i> 3
(B2) # of nouns and pronouns in (B1)	<i>Gebiet</i> 1
(B3) # of prepositions, comparative conjunctions and pronominal adverbs in (B1)	- 0
(B4) # of commas and coordinating conjunctions in (B1)	- 0
(C1) # of remaining tokens in the sub-field of (B1)	<i>in Absprache mit fedpol</i> 4
(C2) # of nouns and pronouns in (C1)	<i>Absprache, fedpol</i> 2
(C3) # of prepositions, comparative conjunctions and pronominal adverbs in (C1)	<i>in, mit</i> 2
(C4) # of commas and coordinating conjunctions in (C1)	- 0

Table 1. Features for the classification of complex noun phrases and relative clauses

machine-learning methods because of their high performance in NLP tasks. NB methods have been successfully applied to text documentation (cf. [25]). SVM has been successfully applied in NLP tasks, such as in parsing (e.g. [26]) and text documentation (e.g. [27]).

Features To capture the decision factors of the syntactic choice as features, we measured the syntactic complexity of the following three context zones by counting the occurrences of syntactic categories, according to the methods of readability assessment (e.g. [28,29]).

- (A) A complex noun phrase,
- (B) The maximally projected noun phrase of a complex noun phrase,
- (C) The sub-fields of a complex noun phrase (i.e. vorfeld, mittelfeld, nachfeld).

To operationalise the syntactic choice, we created the syntactic features so that the features of relative clauses (RC) correspond to those of complex noun phrases (complex NP). Table 1 provides an overview of the syntactic features in our task. We illustrate the features using the sentence pair (4-a) and (4-b).³ For the feature extraction, the sentences in the test suite were parsed with the supertagger and the coreference resolution system CorZu [30].

- (4) a. Complex noun phrase:
Die Kantone treffen in Absprache mit fedpol *die für die Erfüllung der völkerrechtlichen Schutzpflichten der Schweiz notwendigen Massnahmen* **auf ihrem Gebiet**; ...
- b. Relative clause:
Die Kantone treffen in Absprache mit fedpol *die Massnahmen* **auf ihrem**

³Feature A = *italic*, Feature B = **bold**, Feature C = underlined

	Local context (feature A)		Medium context (feature A+B)		Global context (feature A+B+C)	
	Complex NP	RC	Complex NP	RC	Complex NP	RC
GNB	71.80 (59.84/89.74)	53.02 (79.49/39.77)	68.01 (86.33/56.10)	77.53 (67.48/91.11)	80.08 (83.65/76.81)	81.64 (78.56/84.98)
MNB	74.61 (66.56/84.87)	66.51 (79.13/57.36)	67.03 (81.52/56.92)	75.68 (66.91/87.10)	68.61 (84.15/57.91)	77.08 (67.92/89.09)
SVM	84.82 (82.34/87.46)	83.85 (86.63/81.24)	87.13 (85.35/88.98)	86.57 (88.49/84.73)	88.84 (87.52/90.20)	88.49 (89.89/87.14)

Table 2. F1 score (precision/recall) of the prediction of the class *complex noun phrase* and *relative clause*

Gebiet, die für die Erfüllung der völkerrechtlichen Schutzpflichten der Schweiz notwendig sind; ...

‘The cantons take, in alignment with fedpol, measures in those areas that are necessary to fulfil the international obligations of Switzerland with regard to protection.’

2.2. Experiments

We trained all three classifiers,⁴ with three types of features: syntactic complexity of the local context (A), medium context (A) and (B) and global context (A), (B) and (C) (cf. Table 1).

We tested the trained models on the cross-validation data, and measured the performance of the classification for each class, computing the F1 score, precision, and recall. The results are shown in Table 2. All three features of the SVM model achieved the best F1 score, precision, and recall in the prediction of the class *complex noun phrase* and F1 score and precision in the prediction of class *relative clause*. The Gaussian NB model with features A and B outperformed the SVM model in the score for recall in the prediction of the class *relative clause*. The SVM model constantly increased the F1 score, precision, and recall by increasing the number of features. The Gaussian and multinomial NB models also tended to do the same, which indicates that not only the local contexts but also wider contexts affected the syntactic choice.

Because the objective of the error detection task in a style checking tool is to detect style errors - in our case, relative-clause-like complex noun phrases - the classifier is required to be optimized for the class *relative clause*. We assumed that there are fewer stylistically inadequate complex noun phrases than stylistically adequate ones. Swiss law texts are edited thoroughly by linguistic experts, which supports our assumption. To cope with this problem of class imbalance [31], the class *relative clause* was weighted so that the classifier was biased toward the minority class. For class weighting, we used the best model, that is, the SVM model with all three features. The results are shown in Table 3. The weighting of the class improved the score of recall with an increase in the weight and a decrease in precision and F1 score. In the class of *complex noun phrases*, the class weighting caused the opposite effects. With the increase in class weighting, the score of precision increased and the score of recall decreased.

⁴We used sci-kit learn: <http://scikit-learn.org/stable/>

Weight	Complex NP	RC
60	88.20 (88.73 , 87.67)	88.33 (87.81, 88.86)
70	86.44 (91.16 , 82.18)	87.71 (83.78, 92.03)
80	84.40 (92.72 , 77.45)	86.77 (80.64, 93.92)
90	75.92 (94.59 , 63.40)	82.73 (72.47, 96.37)
95	68.06 (96.07 , 52.70)	79.82 (67.41, 97.84)

Table 3. Weighting of the class *relative clause* for SVM with global feature (A+B+C): F1 (precision, recall)

Classifier	tp, fp, fn	F1	(prec, rec)
SVM A+B+C No_W	6, 2, 14	42.86	(75.00 , 30.00)
SVM A+B+C W60	7, 3, 13	46.67	(70.00, 35.00)
SVM A+B+C W70	8, 5, 12	48.49	(61.54, 40.00)
SVM A+B+C W80	9, 7, 11	50.00	(56.25, 45.00)
SVM A+B+C W90	12, 9, 8	58.54	(57.14, 60.00)
SVM A+B+C W95	14, 13, 6	42.43	(53.85, 35.00)
GNB A+B	10, 9, 10	51.28	(52.63, 50.00)

Table 4. Prediction of the class *relative clause* for complex noun phrases (tp = true positive, fp = false positive, fn = false negative, prec=precision, rec=recall, W=weight)

2.3. Evaluation

We tested two hypotheses on the test set: 1) the stylistic choice is affected by syntactic complexity in contexts; and 2) complex noun phrases that are similar to relative clauses with regard to syntactic complexity are more likely to be rephrased into relative clauses by editors.

To examine the first hypothesis, we tested the SVM classifier (global feature A+B+C and no weight) on 269 complex noun phrases in the two law texts selected for the evaluation. As expected, the distinction between two syntactic alternatives was highly accurate. The accuracy of the classification was 90.33%, that is, 243 instances of 269 were classified as *complex noun phrase*. This result supports our first hypothesis that the syntactic choice between complex noun phrases and relative clauses is affected by syntactic complexity in contexts.

To test the second hypothesis, the classifiers were tested on complex noun phrases that were judged by a legal editor. For the manual annotation, 50 complex noun phrases were randomly selected from the two law texts for evaluation. The editor was asked whether he would rephrase these complex noun phrases. Twenty of the 50 phrases were qualified for rephrasing, and 20 were left as they were. In addition, in 10 cases, the editor was not able to make a clear decision for or against re-phrasing.

Table 4 shows the results of the class *relative clause* prediction in precision, recall, and F1 score. The best precision score was 75%, and 6 of 8 items were correctly classified as style violations by the SVM classifier without weighting. The best score for recall was 60%, and 12 out of 20 instances were correctly classified as class *relative clause*, that is, style violations classified by the SVM with a class weight of 90% biased towards class *relative clause*. As expected, weighting the class *relative clause* retrieved more *relative clause* class instances and improved the recall score at the cost of precision. The Gaussian

NB with features A and B achieved a better score in recall than the SVM without weight. Although the model is simple, the performance of the Gaussian NB was surprisingly high. However, in total, the classifiers were not high in F1 scores, which indicates that the syntactic choice did not correspond to the criteria that the legal editor used to rephrase complex noun phrases.

3. Conclusion

We presented a statistical method for detecting style errors according to the degree to which they violated syntax. To the best of our knowledge, this is the first study to use a classifier to distinguish individual instances of a type of style error. In a test case, we investigated the syntactic choice of complex noun phrase vs. relative clause in German. The SVM classifier accurately distinguished the two syntactic alternatives based on the complexity of the syntax in a particular context. This result indicated that syntactic choices could be predicted by inferring the contexts in a domain corpus. This finding is the first step toward developing a context-sensitive style-checking tool that uses a statistical method. In future research, we will apply the proposed method to active/passive alternation.

References

- [1] Georg Müller. *Elemente einer Rechtssetzungslehre*. Schulthess Verlag, 2 edition, 2006.
- [2] Andreas Lötscher. Der Stellenwert des Verständlichkeit in einer Hierarchie der kommunikativen Werte von Gesetzen. *Bulletin Suisse de Linguistique Appliquée*, 2:109–127, 1995.
- [3] Werner Hauk and Andreas Lötscher. Verständlichkeit von Gesetzen als Problem der Gesetzgebung. *LeGes: Gesetzgebung & Evaluation*, 5(2):91–99, 1994.
- [4] S. Hansen-Schirra and Stella Neumann. Linguistische Verständlichmachung in der juristischen Realität. In Kent D. Lerch, editor, *Die Sprache des Rechts: Recht verstehen: Verständlichkeit, Missverständlichkeit und Unverständlichkeit von Recht*, volume 1. Walter de Gruyter, Berlin, 2004.
- [5] Giulia Venturi. Parsing legal texts. A contrastive study with a view to knowledge management applications. In *Workshop 'Semantic Processing of Legal Texts' (LREC)*, 2008.
- [6] Edmund Benedict Coleman and J. P. Blumenfeld. Cloze scores of nominalizations and their grammatical transformations using active verbs. *Psychological Reports*, 13(3):651–654, 1963.
- [7] Edmund Benedict Coleman. Improving comprehensibility by shortening sentences. *Journal of Applied Psychology*, 46(2):131–134, 1962.
- [8] Markus Nussbaumer. Es gibt nichts Gutes, außer man tut es – Arbeit an der Verständlichkeit von Gesetztexten in der Schweizerischen Bundeskanzlei. *Hermes - Journal of Linguistics*, 29:111–123, 2002.
- [9] N. H. MacDonald. The UNIX Writer's Workbench software: Rationale and design. *The Bell System Technical Journal*, 62(6):1891–1908, 1983.
- [10] Yael Ravin. Grammar errors and style weaknesses in a text-critiquing system. *IEEE Transactions on Professional Communication*, 31(3):108–115, 1988.
- [11] Antje Schmidt-Wigger. Grammar and style checking for German. In *Proceedings of the Second International Workshop on Control Language Applications (CLAW)*, pages 76–86, 1998.
- [12] Andrew Bredenkamp, Berthold Crysmann, and Mirela Petrea. Looking for errors: A declarative formalism for resource-adaptive language checking. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, 2000.
- [13] Berthold Crysmann, Nuria Bertomeu, Peter Adolphs, Dan Flickinger, and Tina Klüwer. Hybrid processing for grammar and style checking. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, pages 153–160, 2008.
- [14] Regierungsrat des Kantons Zürich, editor. *Richtlinien der Rechtssetzung*. 2005.

- [15] Regierungsrat des Kantons Bern, editor. *Rechtsetzungsrichtlinien des Kantons Bern (RSR). Modul 4: Sprache*. Bern, 2000.
- [16] Markus Nussbaumer. 128. *Rhetorisch-stilistische Eigenschaften der Sprache des Rechtswesens*, volume Part 2. Mouton de Gruyter, 2009.
- [17] Hans-Werner Eroms. *Stil und Stilistik*. Erich Schmidt Verlag, Berlin, 2008.
- [18] Kyoko Sugisaki and Stefan Höfler. Incremental morphosyntactic disambiguation of nouns in German-language law texts. In *ESSLLI-13 Workshop on Extrinsic Parse Improvement (EPI)*, Düsseldorf, 2013.
- [19] Kyoko Sugisaki and Stefan Höfler. Verbal morphosyntactic disambiguation through topological field recognition in German-language law texts. In Cerstin Mahlow and Michael Piotrowski, editors, *Systems and frameworks for computational morphology*. Springer, Berlin, Heidelberg, 2013.
- [20] Kyoko Sugisaki. *Automatic Annotation and Assessment of Syntactic Structures in Law Texts Combining Rule-Based and Statistical Methods*. PhD thesis, University of Zurich, 2016.
- [21] Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. Predicting the dative alternation. In Gerlof Bouma, Irene Kramer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam, 2007.
- [22] Jennifer E. Arnold, Thomas Wasow, Anthony Losongco, and Ryan Ginstrom. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55, 2000.
- [23] Stefan Th. Gries and Stefanie Wulff. The genitive alternation in Chinese and German ESL learners: towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics*, 18(3):327–356, 2013.
- [24] Herald Baayen. Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, 11(2):295–328, 2011.
- [25] Andrew McCallum and Kamal Nigam. A comparison of event models for Naive Bayes text classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pages 41–48, 2001.
- [26] Joakim Nivre, Johan Hall, and Jens Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC)*, pages 2216–2219, 2006.
- [27] Torsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML)*, pages 137–142, 1998.
- [28] Lijun Feng, Martin Jansche, and Matt Huenerfauth. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 276–284, Beijing, China, 2010.
- [29] Sarah E Petersen and Mari Ostendorf. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106, 2009.
- [30] Manfred Klenner and Don Tuggener. An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In *Recent Advances in Natural Language Processing (RANLP 2011)*, pages 178–185, 2011.
- [31] Rushi Longadge and Snehalata Dongre. Class imbalance problem in data mining review. *International Journal of Computer Science and Network (IJCSN)*, 2(1), 2013.