Controlled Natural Language for Knowledge-Based Legal Information Systems

Stefan Hoefler and Alexandra Bünzli

University of Zurich, Institute of Computational Linguistics,
Binzmühlestrasse 14, 8050 Zurich, Switzerland
{hoefler,buenzli}@cl.uzh.ch

Abstract. Controlled Legal German (CLG) is a subset of legal German specifically designed to facilitate the automated semantic processing of Swiss statutes and regulations. This paper describes the methods CLG uses to reduce ambiguity and underspecification in order to ensure that statutes and regulations can be deterministically translated into formal logical representations. CLG aims at bridging the gap between legal texts, written in natural language, and knowledge-based legal information systems, operating on the basis of formal logical representations.

Key words: controlled language, AI and law, legislative drafting

1 Introduction

Despite considerable progress in the field of artificial intelligence and law in the last two decades [1], one major obstacle remains as yet unresolved [2]: knowledge-based legal information systems operate on the basis of formal logical representations but legal knowledge is encoded in natural language (statutes, regulations, cases, etc.). While a manual translation of legal texts into formal logic is both costly and error-prone, state-of-the-art natural language processing systems continue to struggle with the notoriously difficult resolution of ambiguity and underspecification.

The Collegis project, a collaboration of computational linguistics at the University of Zurich and legal editors at the Swiss Federal Chancellery, addresses this problem from the perspective of legislative drafting. We develop Controlled Legal German (CLG), a restricted version of Swiss legal German specifically constructed to facilitate the automated semantic processing of statutes and regulations.

Controlled languages restrict the vocabulary, syntax and/or semantics of a natural language in order to reduce its ambiguity and complexity. While early versions of controlled languages were mainly devised to improve the readability and translatability of texts, recently, the method has been used to define subsets of natural languages that can be unambiguously translated into formal logic (see [3] for an overview). Controlled languages have been developed for the domains of technical documentation and requirements engineering and for general-purpose knowledge representation. There have also been first attempts

to apply the method to defining business rules [4] and writing contracts [5]. In this paper, we introduce legislative drafting as another promising area of application. Legislative drafting, by definition, already exerts a certain degree of control on legal language, thereby pursuing aims similar to those of controlled languages: the reduction of ambiguity and sufficient specification of rules.

This paper describes the methods CLG applies to control lexical ambiguity (section 2), syntactic ambiguity (section 3), semantic ambiguity (section 4) and underspecification (section 5) in the language of Swiss statutes and regulations.

2 Controlling Lexical Ambiguity

While early controlled languages primarily focused on controlling the vocabulary, languages such as ACE [6] or PENG [7], which aim at providing interfaces to formal logic, prescribe the semantics of syntactic constructions and function words (articles, conjunctions, prepositions, pronouns, some adverbs) but leave the definition of content words (nouns, verbs, adjectives, some adverbs) to the user. The same policy is applied in CLG: content words are interpreted as logical constants whose meaning needs to be defined in CLG-external terminology databases or ontologies. Thus, CLG does not infringe on the often intended open-texturedness and vagueness of the concepts represented by content words; it only prescribes the syntactic and semantic frames for the respective grammatical categories.

CLG does, however, pre-define certain domain-specific content words. Some of these words are ambiguous in full natural language but have acquired a default interpretation in legal language. In ordinary German, the adverb *grundsätzlich*, modifying an obligation or permission, can have two directly opposed interpretations: if interpreted in the sense of 'strictly' or 'categorically', it denotes that the respective rule does not allow for exceptions; if interpreted as 'generally' or 'in principle', it indicates that the rule is defeasible. By convention, *grundsätzlich* is always used in the latter sense in Swiss legal German. CLG therefore devises an interpretation rule defining that *grundsätzlich* is always interpreted as an explicit defeasibility marker:

(1) Die Veröffentlichung der Entscheide hat grundsätzlich in anonymisierter Form zu erfolgen.(Art. 27 Abs. 2 BGG[1])
'*In principle*, the decisions must be published in anonymized form.'

Note that unlike ordinary adverbs, *grundsätzlich* does not modify the verb but the obligation as a whole. CLG defines a number of words and fixed expressions that are not interpreted like other items of the same grammatical category but obtain domain-specific interpretations. Further examples are: *in der Regel* ('as a rule'), *insbesondere* and *namentlich* ('in particular'), *sinngemäss* ('analoguously'), *gemäss* ('according to') and *im Rahmen von* ('within the scope of').

[1] Bundesgerichtsgesetz (Federal Supreme Court Act), SR 173.110

3 Controlling Syntactic Ambiguity

Syntactic ambiguity occurs if a sentence can be syntactically analyzed in more than one way. Especially attachment ambiguity represents one of the main obstacles to the semantic processing of legal texts [2].

(2) Das Bundesgericht deckt seinen Bedarf an Gütern und Dienstleistungen im Bereich der Logistik selbständig. (Art. 25a Abs. 2 BGG)

'The Federal Supreme Court supplies its need for goods and services in the sector of logistics autonomously.'

In sentence (2), the prepositional phrase *im Bereich der Logistik* ('in the sector of logistics') could theoretically be attached to *deckt* ('supplies'), to *Güter und Dienstleistungen* ('goods and services'), or only to *Dienstleistungen* ('services').

In CLG, constituents are always attached to the closest possible candidate. Thus, if (2) was a CLG sentence, the PP *im Bereich der Logistik* ('in the sector of logistics') would modify *Dienstleistungen* ('services'), but not *Güter* ('goods'). The sentence would have to be rephrased if it was to express one of the other two interpretations. To modify the verb, the PP would have to be moved in front of the coordinated direct object:

(3) Das Bundesgericht deckt im Bereich Logistik seinen Bedarf an Gütern und Dienstleistungen selbständig.

'The Federal Supreme Court supplies, in the sector of logistics, its need for goods and services autonomously.' (German word order)

To modify both *Güter* ('goods') and *Dienstleistungen* ('services'), the PP would have to be repeated after each of these elements. CLG includes a stylistic convention common to legal language that makes such lists easier to read:

(4) Das Bundesgericht deckt selbständig seinen Bedarf an:
   a. Gütern im Bereich der Logistik;
   b. Dienstleistungen im Bereich der Logistik.

'The Federal Supreme Court autonomously supplies its need for:
   a. goods in the sector of logistics;
   b. services in the sector of logistics.'

To avoid the necessity for cumbersome repetitions, CLG also offers the option to relax the aforementioned interpretation rule for constituents attached to coordinations. If the user chooses this option, an interactive authoring tool will, upon the occurrence of an attachment after a coordination, ask the user to indicate whether the respective constituent is meant to be attached to both components of the coordination or only to the last one. In the example above, the user could thus modify both *Güter und Dienstleistungen* ('goods and services') without having to repeat *im Bereich der Logistik* ('in the sector of logistics') for both components. The CLG authoring tool will then record the decisions made by the user in a so-called disambiguation protocol that is to be stored together with the generated logical representation of the respective legal text.

4 Controlling Semantic Ambiguity

Semantic ambiguity occurs if a sentence has only one syntactic structure but can be assigned two or more non-equivalent representations in formal logic. Example (5) contains two common types: plural ambiguity and scope ambiguity.

(5) Die Parteivertreter und -vertreterinnen haben sich durch eine Vollmacht auszuweisen. (Art. 40 Abs. 2 BGG)

'The party representatives have to identify themselves with a letter of attorney. '

Plural noun phrases like *die Parteivertreter* can be interpreted distributively or collectively [8]: the party representatives can identify themselves individually or as a group. In CLG, definite plural NPs are always interpreted distributively; a collective reading has to be expressed with a separately introduced singular term, here e.g. with *die Parteivertetung* ('the party representation') or *die Gesamtheit der Parteiverteter* ('the body of party representatives').

CLG also offers the option to leave plurals underspecified: if this option is chosen by the user, plurals are only interpreted as either distributive or collective if they are accompanied by predefined disambiguation markers such as *einzeln* 'individually' or *gemeinsam* 'together'. Such underspecification is sometimes intended: in sentence (6), for instance, the legislators deliberately avoided specifying whether the judges are to be elected individually or as a body.

(6) Die Bundesversammlung wählt die Richter und Richterinnen. (Art. 5 Abs. 1 BGG)

'The Federal Assembly elects the judges.'

However, if one assumes a distributive interpretation for its subject, sentence (5) comes to exhibit scope ambiguity: either the universally quantified phrase *die Parteivertreter* has wide scope over the existentially quantified phrase *eine Vollmacht*, or vice-versa:

(7) a. $\Box\ \forall x(party\_rep(x) \rightarrow \exists y(letter\_of\_attorney(y) \wedge identified\_by(x, y)))$
b. $\Box\ \exists y(letter\_of\_attorney(y) \wedge \forall x(party\_rep(x) \rightarrow identified\_by(x, y)))$

Like ACE [6], CLG interprets scopes according to the surface order of the quantifiers in the sentence – which is usually also the more intuitive reading. In CLG, the semantics of sentence (5) would thus be analyzed as in (7a). To express meaning (7b), the sentence would have to be rearranged, e.g. as shown in (8).

(8) Es ist eine Vollmacht vorzulegen, die die Parteivertreter und -vertreterinnen ausweist.

'A letter of attorney which identifies the party representatives has to be provided. '

5 Controlling Underspecification

Underspecification can become a problem for automated reasoning if the logical representation of a sentence warrants unintended inferences. An example is (9).

(9) Bei der Geburt eines Kindes hat der Angestellte Anspruch auf eine einmalige Zulage von 530 Franken. (Art. 55 Abs. 1 AngO ETH-Bereich[2])

'Upon the birth of a child, the employee is entitled to a one-time allowance of 530 francs.'

$\Box \forall x((\mathit{is\_born}(x) \wedge \mathit{child}(x)) \rightarrow \forall y(\mathit{employee}(y) \rightarrow \mathit{entitled}(y)))$

The problem with sentence (9) is that the condition (*at the birth of a child*) does not share any discourse referent with its consequence (*the employee is entitled to a one-time allowance of 530 francs*): the sentence does not specify explicitly that the employee does not receive an allowance on the occasion of the birth of just *any* child but only if he or she is the parent of that child. Human readers will easily infer this missing bit of information from the context and thus reduce the number of warranted inferences. An automated reasoner, on the other hand, may in the worst case combine the logical representation of (9) with the knowledge that approximately 216,000 children are born every day, and deduce that an employee is to receive total allowances of 114,480,000 francs per day.

To avoid this problem, CLG prescribes that the condition of a norm always has to share a discourse referent with its consequence. Sentence (9) would thus only be a correct CLG sentence if this connection were established, e.g. by modifying the noun phrase *eines Kindes* with a relative clause:

(10) Bei der Geburt eines Kindes, *gegenüber dem er elterliche Pflichten hat*, hat der Angestellte Anspruch auf eine einmalige Zulage von 530 Franken.

'Upon the birth of a child *toward whom he or she has parental duties*, the employee is entitled to a one-time allowance of 530 francs.'

The same effect can be achieved by adding another condition to the end of the sentence:

(11) Bei der Geburt eines Kindes hat der Angestellte Anspruch auf eine einmalige Zulage von 530 Franken, *sofern er gegenüber dem Kind elterliche Pflichten hat.*

'Upon the birth of a child, the employee is entitled to a one-time allowance of 530 francs, *provided that he or she has parental duties toward the child.*'

Note that controlling underspecification can be beneficial not only to automated semantic processing but also to legislative drafting. Had they been forced to provide the additional specification required by CLG, legislators would have automatically closed an overlooked regulatory loophole, namely that biological parents who are not liable for support should not be entitled to an allowance while foster parents should.

[2] Angestelltenordnung ETH-Bereich (Employee Regulation ETH), SR 172.221.106.2

6 Conclusion

This paper has introduced CLG, a restricted version of Swiss legal German specifically designed to facilitate the automatic translation of statutes and regulations into formal logical representations. CLG is currently in a stage of development; in this paper, we have explained the methods it applies to control lexical, syntactic and semantic ambiguity as well as underspecification. CLG thus eliminates some major obstacles to successful semantic processing of legal texts.

The employment of CLG – or a similar processing-oriented standard – in legislative drafting contributes to the development of knowledge-based legal information systems as it bridges the gap between natural language legal texts and their representation in formal logic. However, the success of such a standard will depend on its acceptance by professional legal editors: CLG must be easy to learn and close to conventional legal language both in terms of expressiveness and style. We have shown an additional factor that may increase acceptance: the fact that the employment of CLG can be beneficial not only to automatic processing but also to legislative drafting. It can point legal editors to ambiguous passages and regulatory loopholes they might otherwise have overlooked.

References

1. Rissland, E.L., Ashley, K.D., Loui, R.P.: AI and Law: A Fruitful Synergy. Artificial Intelligence 150(1–2), 1–15 (2003)
2. McCarty, L.T.: Deep Semantic Interpretations of Legal Texts. In: Proceedings of the 11th International Conference on Artificial Intelligence and Law, pp. 217–224. ACM Press, New York (2007)
3. Pool, J: Can Controlled Languages Scale to the Web? In: 5th International Workshop on Controlled Language Applications (2006)
4. Spreeuwenberg, S., Anderson Healy, K.: SBVR's Approach to Controlled Natural Language. In: Fuchs, N.E. (ed.): Pre-Proceedings of the Workshop on Controlled Natural Language (CNL 2009). CEUR-WS (2009)
5. Pace, G.J., Rosner, M.: A Controlled Language for the Specification of Contracts. In: Fuchs, N.E. (ed.): Pre-Proceedings of the Workshop on Controlled Natural Language (CNL 2009). CEUR-WS (2009)
6. Fuchs, N.E., Kaljurand, K., Kuhn, T.: Attempto Controlled English for Knowledge Representation. In: Baroglio, C., Bonatti, P.A., Maluszynski, J., Marchiori, M., Polleres, A., Schaffert, S. (eds.): Reasoning Web 2008, pp. 104–124. Springer, Berlin (2008)
7. Schwitter, R., Tilbrook, M.: Let's Talk in Description Logic via Controlled Natural Language. In: Proceedings of the 3rd International Workshop on Logic and Engineering of Natural Language Semantics, pp. 193–207. Tokyo (2006)
8. Schwertel, U.: Controlling Plural Ambiguities in Attempto Controlled English. In: Proceedings of the 3rd International Workshop on Controlled Language Applications. Seattle, Washington (2000)