Controlled Language Applications Workshop (CLAW)

Workshop Programme

14:00 – 14:15	Key-Sun Choi, Hitoshi Isahara, Kiyong Lee and Christian Galinski: *Introduction about ISO and CNL*
14:15 – 14:30	Hitoshi Isahara and Tetsuzo Nakamura: *Report from Japan*
14:30 – 14:55	Adam Wyner, Francois Lévy and Adeline Nazarenko: *An Underspecified Approach to a Controlled Language for Legal Texts - a Position Paper -*
14:55 – 15:20	Christian Galinski and Blanca Stella Giraldo Pérez: *Rule-Based Technical Writing: A Meta-Standard on Controlled Language Extended towards Controlled Communication*
15:20 – 15:45	Sylviane Cardey: *A Controlled Language for Sense Mining and Machine Translation for Applications in Mission-Critical Domains*
15:45 – 16:10	Christina Lohr and Robert Herms: *A Corpus of German Clinical Reports for ICD and OPS-based Language Modeling*
16:10 – 16:30	Coffee break
16:30 – 16:55	Xiaofeng Wu, Liangyou Li, Jinhua Du and Andy Way: *ProphetMT: Controlled Language Authoring Aid System Description*
16:55 – 17:30	Rei Miyata, Anthony Hartley, Cécile Paris and Kyo Kageura: *Evaluating and Implementing a Controlled Language Checker*
17:30 – 17:40	Closing

Editors

Key-Sun Choi	KAIST
Sejin Nam	KAIST

Workshop Organizers

Key-Sun Choi	KAIST
Hitoshi Isahara	Toyohashi University of Technology
Christian Galinski	Infoterm
Andy Way	Dublin City University
Teruko Mitamura	Carnegie Mellon University

Workshop Programme Committee

Hitoshi Isahara	Toyohashi University of Technology
Andy Way	Dublin City University
Christian Galinski	Infoterm
Teruko Mitamura	Carnegie Mellon University
Kiyong Lee	ISO/TC37
Key-Sun Choi	KAIST

Table of Contents

Author Index

Preface

Following the highly successful workshops on the Controlled Natural Language Simplifying Language Use at LREC2014, we are pleased to announce the 6th CLAW workshop, embracing an open range both of applications to standardizations, in conjunction with the 10th edition of the Language Resources and Evaluation Conference (LREC2016), 23-28 May 2016, Grand Hotel Bernardin Conference Center, Portorož, Slovenia.

This workshop will focus more on the issues like standardization toward the Controlled Language Applications and their related supporting research and implementation issues in cooperation with the controlled language application, ISO/TC37 standardization, and semantic web communities.

The workshop on the Controlled Language Applications invite papers for the current progress and results toward the standardizations of controlled language. This workshop also would like to encourage submissions on any of (but not limited to) the following topics: human communication protocols, controlled text authoring, conformance checking systems, controlled language authoring aids, memory-based authoring, (re-)authoring combined with translation, issues in Controlled Language design, industrial experience and evolving requirements, models, processing algorithm, terminology aspects, R&D projects, use case, related topics on summarization, question and answering, machine translation, quality and usability evaluations of controlled language.

The workshop will give equal emphasis to the academic, corporate and industrial perspectives, while bringing together researchers, developers, users, and potential users of controlled language systems from around the world. The goal of this workshop will be to bridge the gap between the theory, practice and applications of controlled language and to identify the existing and possible future controlled language applications, and what should be kept in a standard for controlled language application.

An Underspecified Approach to a Controlled Language for Legal Texts - a Position Paper -

Adam Wyner[1], Francois Lévy[2], and Adeline Nazarenko[2]
[1]University of Aberdeen, Aberdeen, Scotland
azwyner@abdn.ac.uk
[2]University of Paris 13, Paris, France
Francois.Levy@lipn.univ-paris13.fr, Adeline.Nazarenko@lipn.univ-paris13.fr

Abstract

The texts of legislation and regulation must be structured and augmented in order to allow for semantic web services (querying, linking, and inference). However, it is difficult to accurately parse and semantically represent such texts due to conventional practices of the legal community, the length and complexity of legal language, and the textual ground of the law. Controlled natural languages have been proposed as an approach to adjust to the difficulties, where the source text is rewritten in some standard form. However, such an approach has not suited legal language due to its requirements and complexities, so standardization has been difficult to achieve. To navigate between the requirements and complexities of legal language, standardization, and a fully controlled natural language, we take a position to propose and exemplify an approach to a high-level controlled language, which is adapted to the legal domain, correlates with the source text, and also facilitates analysis for semantic web applications. The approach can make use of some available NLP processing tools.

Keywords: natural language standardization, semantic annotation, legal rules, controlled languages, semantic web

1. Introduction

The increasing complexity and integration of legislation and regulations calls for rich legal content management. The legal semantic web aims at giving a uniform access to legal sources, whatever form they may take or the institution that publishes them. This is traditionally supported by the definition of a meta-data vocabulary and the semantic annotation of the sources. Beyond documents and topic-based annotations, however, legal experts must have direct access to the rules contained in documents and their interpretations. This calls for a rich and structured representation of the rule text fragments.

However, as discussed later, legal texts have proven to be difficult to accurately parse and semantically represent. Controlled natural languages (CNLs) have been proposed as an approach to adapt to the difficulties, where the source text is rewritten in some normative form. However, such an approach does not suit legal professionals, who work and reason with the language of the law strictly as it is, and whose linguistic practice is not fully reflected by semantics. In addition, legal language itself introduces issues that are not straightforward to address in a CNL given sentence length, construction complexity, semantic ambiguity, and domain terminology. The difficulties raise significant issues about standardizing legal language to suit CNLs (though this is a matter relative to what is standardized and to what degree). Nonetheless, some degree of machine-readability would be very valuable. To navigate between the requirements of legal professionals, the complexities of legal language, and a fully controlled natural language, we take a position on and exemplify an approach to a high-level controlled language (hCL), which is adapted to the legal domain, maintains the source text, and facilitates analysis for semantic web applications (querying, linking, and inference). It is a hCL in that we propose units which, when in construction with other units, provide well-formed rules. In this sense, our proposal provides a controlled language along the lines of the controlled language of the syntax of predicate logic.

The novel, significant contribution of this paper is an approach to the analysis and representation of legal text which leaves the original text in place and yet adds a layer of semantic representation. In other words, the original is not transformed into a controlled language, but is 'covered by' a higher-level of representation. In particular, we focus on the analysis of legal rule statements in the source text, connecting portions of the source text with a high-level controlled language for rules.

To ground our discussion and provide a running example, we use a corpus that was previously reported in Wyner and Peters (2011), which is a passage from the US Code of Federal Regulations, US Food and Drug Administration, Department of Health and Human Services regulation for blood banks on testing requirements for communicable disease agents in human blood, Title 21 part 610 section 40.

In the remainder of the paper, we outline existing research to contrast with our proposal (Section 2). We sketch our annotation approach based on hCL in Section 3 and a sample example Section 4. In Section 5, we outline some of the available tools that can be used to support the approach. The paper closes with some discussion.

2. State of the Art

There are a variety of sorts of properties that controlled languages have and purposes that they serve (Wyner et al., 2010a; Kuhn, 2014), allowing for a range of approaches. The fundamental idea about a CNL is that controlled statements would be easier to automatically parse and semantically represent, but still be meaningful and manageable for domain experts. For instance, *Attempto Controlled English*

(ACE) defines unambiguous readings of quantifier scopes and anaphora, and prohibits ambiguous attachments, so that it can be parsed into logical formulae (Fuchs et al., 2008).

The complexity of legal language and regulations has long been an obstacle to the development of legal content management tools. Attempts have been made to parse and automatically formalize legal texts. For instance, C&C/Boxer (Bos, 2008) has been applied to fragments of regulations (Wyner et al., 2012). C&C/Boxer is a wide coverage parser that feeds a tool which generates semantic representations (essentially in First-order Logic). However, as discussed in Wyner et al. (2012), the complexity and ambiguity of the resulting parses and semantic representations make them difficult to evaluate for correctness as well as to exploit for experts in formal languages, *a fortiori* for legal analysts.

Controlling the legal sources has been proposed as an alternative approach. Efforts are made to clarify and simplify the legal language when drafting (*e.g.* in favor of "Plain English" (U.S. Government, 2015), to ease translation (Meunier-Crespo and Damette, 2011; Meunier et al., 2013) or to avoid ambiguity and provide uniformity (Hoefler, 2012)). *Oracle Policy Modeling* (OPM) system (Dayal and Johnson, 2000) is designed to parse structured sets of controlled sentences and make rule-bases available online. *Semantics of Business Vocabulary and Business Rules* (SBVR) has been specifically designed to model business rules (OMG, 2008): it provides elements of a pattern language and a description of *SBVR-Structured English* to express rules in a form that can be checked by human experts. Attempto Controlled English has been applied to legal and clinical language with limited success (Shiffman et al., 2010; Wyner et al., 2010b; Wyner and van Engers, 2010).

ACE, OPM, and SBVR try to systematize the NL to CL translation by proposing alternative formulations for unwanted constructions. However, when the source regulations get more complex, the NL to CL translation either fails or gives a logic-like result, with explicit scopes and qualifiers, which is difficult to read, and even harder to adjudicate, for experts. Moreover, these tools require that the source text is entirely and manually transformed into the CNL standard, which is time- and labor-intensive. In addition, the tools are intended for use in decision making rather than semantic web applications.

A third approach relies on the semantic annotation of legal texts, without engaging with the detailed syntactic complexity of legal sentences. Asooja et al. (2015) annotates at the paragraph level, making use both of a high-level legal ontology and a specific domain ontology. Francesconi (To appear) or Wyner and Peters (2011) annotate at the provision level, relying on a general model of relationships between normative provisions. The provision collection is encoded in RDF-OWL and can be queried using SPARQL. However, such tools have not achieved general functionality. In this vein, the LegalRuleML mark-up language is designed to represent legal rules for the semantic web (Athan et al., 2013), though it does not provide the means to analyze natural language.

In Lévy et al. (2015), a high-level controlled language (hCL) is proposed, expressing the content of semantic annotations and fixing the interpretation of underlying fragments of legal sources. This approach builds on the SemEx methodology, which was designed to annotate business regulations by business rules through an iterative rewriting process, ideally until a CL form is obtained (Guissé et al., 2012). The current position paper explicates the underlying vision of Lévy et al. (2015) and argues for an underspecified approach to controlling legal language which does not require transformation of the source language into a CNL yet is useful for semantic web applications.

3. Combining annotations and high-level controlled language

Formalization of legal documents yields representations that support content management (indexing and search, merge, comparison, update of documents) and legal reasoning (*Is it necessary to test X for Y?*). However, completely formalizing the content of legal documents is a distant goal, due to legal and domain-specific terminology, long, complex and possibly ambiguous sentences. Legal professionals often use complex sentences to express the subtleties or generalities of regulations and the range of facts and situations.

We pragmatically address the formalization dilemma through the annotation of legal texts and the standardization of the structure of the rules. In Figure 1, we have an analysis of our running example:

> *(b)* To test for evidence of infection due to communicable disease agents designated in paragraph (a) of this section, you must use screening tests that the Food and Drug Administration (FDA) has approved for such use, in accordance with the manufacturer's instructions.

The controlled representations serve the following purposes:

- Add to and enrich the source text, but do not replace it.
- Make explicit high-level constructions of rules, even if the source text is not parsed.
- Provide simplified and semantically more explicit versions of the source rule statements.
- Annotate rule statements with *form-based semantic structures*.

Interpretation is important in legal reasoning, for there always remains room for interpretation of legal texts. Since the controlled representations only make explicit high-level constructions of rules and perhaps disambiguate some aspects of a rule, alternative interpretations need not be fixed. Figure 2 shows how the annotation in hCL highlights two alternative readings of an ambiguous text fragment but note that the original terms (*e.g.* "screening tests" in Fig. 2) are preserved in the representations, so that their applicability to actual cases can be discussed in legal terms, e.g. just what legally counts as a "screen test", which is a matter for lawyers to determine.

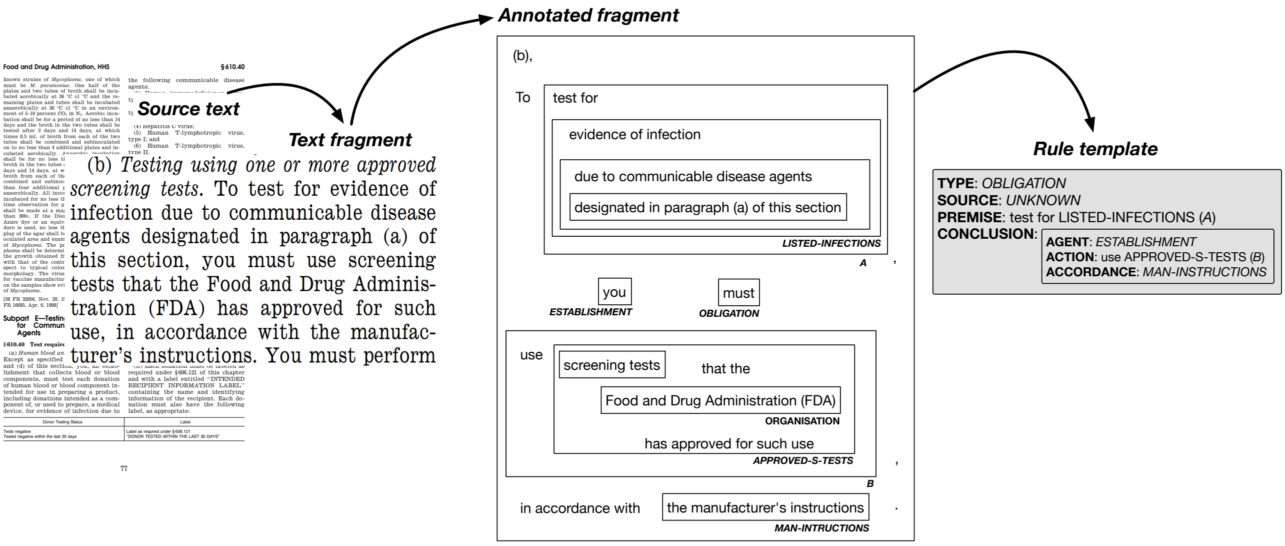


Figure 1: Example of annotation

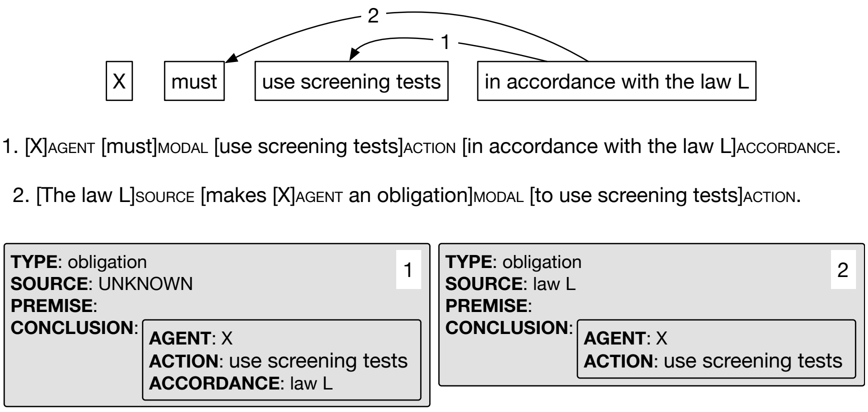


Figure 2: Alternative annotation of an ambiguous source sentence: The ambiguity concerns the attachment of the prepositional phrase "in accordance with the law L" to the modal or main verb (Readings 1 and 2, resp.).

The controlled language focuses on the high-level structure of the rule statements, which are thus associated with an explicit and unambiguous semantics (see Figure 2), leaving aside the detailed parsing of the constituents. These may remain unanalyzed (*e.g.* "use screening tests" in Figure 2). We argue that it is both possible and useful to define a controlled language specifying all the acceptable high-level rule structures even if some parts remain unspecified.

The parts correspond to annotations, which associate semantic tags to actual text fragments, and relates them to roles in the semantic form. The granularity of annotations may vary and there may be several annotations on top of each other. For instance, "LISTED-INFECTIONS" in Figure 1 stands for "infections due to communicable disease agents designated in paragraph (a) of this section". This approach hides ambiguities and complexities of the lower-level of analysis (*e.g.* the anaphoric expression "this section") to highlight the main structure of the rule statements, but it remains flexible.

The different levels of annotations can be exploited for indexing documents and mining legal content on a rule rather than a keyword basis. This allows for answering queries like "Which are the rules that have been emphasized in that document?", "Do the analysts agree on the reading of a specific rule statement or more generally on a section of a document?", and "Find all the rules that concern infections cases" (*i.e.* that contain a reference to infections in the premise part). These queries can be answered from the hCL structures associated to the text.

4. Incrementally annotating the source texts

The annotations play a key role in our approach to controlled legal language. We aim to provide a semantic annotation of the source text, which maintains the structure of the original text, while supporting the analyst's interpretive work. The analyst proceeds incrementally and interactively through a succession of annotations. The analysis can combine top-down and bottom-up approaches, whereby a legal analyst first selects a relevant rule statement in the text and directly annotates it with a hCL formula, or he/she performs a detailed analysis of the selected sentences, recursively tagging sub-components and components until the overall structure of the rule is explicit. In the course of the analysis the interpretation process is documented.

4.1. Annotation method

There are two major levels in the annotation process.

The first one identifies small fragments, keywords of phrases, such as discourse markers (*in accordance with*, modalities (*must*), named entities (*Food and Drug Administration*) or domain terminology (*manufacturer's instructions*) that play an important role in the interpretation. Fixed keywords or phrases are identifiable with some consistency and with little to no ambiguity. NLP tools can be used to locate these key elements and 'trigger' annotations, so that candidate annotations are offered to the analyst, who can accept or reject them. Lower-level annotations may be reused to create intermediate-level annotations. In this way, annotation is done interactively and incrementally over the text.

At the second level, the low- or intermediate-level annotations are used to create high-level structures, *e.g.* rule constructions, which therefore admit of a large degree of open-ended variation.

The approach depends on the corpus of text and domain having some relatively consistent patterns of expressions. While these are limitations, legislative and regulatory texts are known to be highly structured given editorial guidelines imposed on their composition as well as the often formulaic expression of law. Moreover, it is reasonable to expect that some patterns will appear in other contexts, while other patterns will be revised as the analysis expands.

The next subsection illustrates the incremental process on the example of Paragraph (b) in Section 3.

4.2. Example

In the first phase of an analysis, key elements such as discourse markers or modalities should be identified. They appear underlined in the following.

> To test for evidence of infection due to communicable disease agents designated in paragraph (a) of this section, you must use screening tests that the Food and Drug Administration (FDA) has approved for such use, in accordance with the manufacturer's instructions.

In parallel, we can apply part-of-speech and chunking analysis to the text, which are known to be highly reliable. Terms that are relevant to the text such as noun phrases are identified (represented between brackets).

> To test for [evidence of infection] due to [communicable disease agents] designated in [paragraph (a) of this section], you must use [screening tests] that the [Food and Drug Administration (FDA)] has approved for such use, in accordance with [the manufacturer's instructions].

The terms can be associated with annotations, which are provided from a preset list of annotations or provided on-the-fly by the analyst, and which may be propagated both to other analysts (online dynamic dictionary) as well as throughout the text (similar strings are similarly annotated).

The annotations are indicated with subscripts. The annotation `ACTION-ANAPHOR` indicates an anaphoric expression, that is, an expression that depends for its reference on a (generally) preceding explicit expression. We treat this further below:

> To test for [evidence of infection]$_{INFECTION}$ due to [communicable disease agents]$_{DISEASE-AGENTS}$ designated in [paragraph (a) of this section]$_{LIST-OF-\ PAR-A}$, you must use [screening tests]$_{S-TEST}$ that the [Food and Drug Administration (FDA)]$_{FDA}$ has approved for [such use]$_{ACTION-ANAPHOR}$, in accordance with [the manufacturer's instructions]$_{MAN-INSTRUCTIONS}$.

At each stage of the analysis, we can substitute the annotations in for the terms (which we suppress further below) to get a simplified and more structured view of the source fragment:

> To test for INFECTION due to DISEASE-AGENTS designated in LIST-OF-PAR-A, you must use S-TEST that the FDA has approved for ACTION-ANAPHOR, in accordance with MAN-INSTRUCTIONS.

Over and above these terms, several annotations and annotation patterns appeared to be relevant, so that annotations can be themselves further annotated and larger chunks can be identified and tagged. *e.g.* the annotation `FDA` is annotated as `NAMED-ORGANISATION` and 'designated in LIST-OF-PAR-A' is annotated `DOCUMENT-REFERENCE`.

> To test for INFECTION due to DISEASE-AGENTS designated in [LIST-OF-PAR-A]$_{DOCUMENT-REFERENCE}$, [you]$_{NAMED-INDIVIDUAL}$ [must]$_{MODAL}$ use S-TEST that the [FDA]$_{NAMED-ORGANISATION}$ has approved for ACTION-ANAPHOR, [in accordance with]$_{ACCORDANCE-INDICATOR}$ [MAN-INSTRUCTIONS]$_{DOCUMENT-REFERENCE}$.

It worth emphasizing that the annotations to this point are rather straightforward markups over the string of words as they appear in the original text, but large chunks can also be identified at once without any internal analysis.

At the higher level, the annotated fragments are interpreted as fillers of a rule template. This means that the relations between the elements are identified. The following tagging corresponds to the template presented in Figure 1.

> To [test for INFECTION [due to DISEASE-AGENTS designated in DOCUMENT-REFERENCE]$_{CAUSE}$]$_{ACTION1}$, [NAMED-INDIVIDUAL]$_{AGENT2}$ MODAL [use S-TEST that the NAMED-ORGANISATION has approved for ACTION-ANAPHOR]$_{ACTION2}$, [ACCORDANCE-INDICATOR DOCUMENT-REFERENCE]$_{ACCORDANCE}$.

With reference to our analysis in Figure 2, the ACCORDANCE phrase seems to be best interpreted as a modifier that further specifies how the screening tests are used, and PREMISE introduces the overall perspective in which they are used. The template is filled in light of such interpretations.

5. Tools

There are widely available component tools to support some of the essential tasks involved in the approach we have discussed. However, we do not envisage at this point a fully automatic end-to-end tool. Rather, we would propose a workbench that interacts with the analyst and applies incrementally over the course of the text. In this way, the analyst is supported in identifying the relevant constructions as well as having a consistent "palette" of components available (See Wyner and Peters (2011) for a tool somewhat along the lines as proposed here, but with less structure).

In the first phase of an analysis, dictionary look-up can help to identify key elements such as domain terminology, discourse markers, modal operators (e.g. obligation), or indicators of subordination. Such look up expressions can be helpful in highlighting relevant "gross" textual structure. A sample of these are underlined in the following:

> To test for evidence of infection due to communicable disease agents designated in$_{verb}$ paragraph (a) of this section, you must$_{modal}$ use screening tests that the Food and Drug Administration$_{term}$ (FDA) has approved for such use, in accordance with$_{subordinate}$ the manufacturer's instructions.

For further structure, standard highly reliable NLP tools such as part-of-speech tagging and NP or VP chunking can be applied to the text. [1] It will be important to identify recurrent textual patterns such as entities or actions that are mentioned several times in the body of the corpus. Tools such as TermRaider (Maynard et al., 2008) in GATE and Termostat (Drouin, 2003) can be used to automatically identify such frequently occurring terms, which are usually noun phrases.

Prototype tools have been developed to identify rule components such as premises, conclusions, or exceptions Wyner and Peters (2011). Further development is required to ensure they are reliable and have a high coverage.

Given such linguistically oriented information, NLP tools would identify potentially relevant passages and offer the analyst a context-sensitive menu of alternative annotations which could fill the components of the template structure as discussed above. In this way, the analyst is supported to fill the structure, but given options on alternatives.

6. Discussion

Legal language found in legislation, regulations, case law, and elsewhere poses unusual issues with respect to the standardization of controlled natural languages. It remains unclear to what extent the issues can be engineered away so as to satisfy the requirements of the legal community. Some of the challenges are sentence length, complexity, ambiguity, and domain terminology as well as unusual formatting (e.g. list structures). Some of these aspects are deliberate linguistic strategies to not only "cover all the bases" on some matter but also to leave strategic "loopholes" that allow some flexibility in the applicability of the law. It is unclear whether the legal profession wishes to increase readability and also reduce complexity and ambiguity (movements to simplified English aside, which themselves have only gained relatively minor traction). It also raises issues for legal professionals, for whom the text as it is written by legislators or judges, remains itself the gold standard - textual transformations, reductions, additions, and simplifications all may give rise to some legally liable distortion, which a legal professional wishes to avoid at all costs. Indeed, legal professionals are highly conservative concerning the quality and usability of the text, and convincing the community to adopt and adapt to a legal CNL is a socio-linguistic problem that does not appear to have a straightforward solution. Legal professionals wish the tools to adapt fully to their style, they do not wish to fulfill the conventions of some other profession. While it is possible to manually transform the source legal language to a CNL and to process it further (e.g. Oracle's Policy Automation), it does not keep the source language intact, nor make it reusable in semantic web applications. In addition, it introduces interpretations on the source, which may or may not be acceptable. Other approaches to legal CNL (vocabularies, generic CNLs such as Attempto, and others) have related problems

In contrast to an approach to a CNL, the paper has taken a position on providing a controlled language for the analysis and annotation of complex legal texts. Rather than providing a CNL that is a normalized expression of some source text, the proposal argues for leaving the source text intact, whilst providing a bridge controlled language that has a structure approximating key legal rule statements. Our approach combines the benefits of controlled languages – to give manageable although simplified descriptions of legal content – and of semantic annotation – to maintain a tight correlation with the source texts. It was pragmatically designed to help analysts publish legal sources for semantic web applications. The examples represent an initial fragment which can clearly be extended to other constructions, *e.g.* exceptions and conditionals (Wyner and Peters, 2011). To evaluate the advantages or weaknesses of the fragment language, we can qualitatively apply it to the larger regulation from which the sample is drawn, modifying it as required. Tool support, *e.g.* contextually relevant pop-up annotation alternatives along with the option to create new, would be essential to control for annotation variation and to measure inter-annotator agreement.

Acknowledgements

This work is facilitated by the French National Research Agency (ANR-10-LABX-0083) in the context of the Labex EFL (Strand 5 "Computational semantic analysis").

[1]NLTK http://www.nltk.org, GATE https://gate.ac.uk, or UIMA https://uima.apache.org

7. Bibliographical References

Asooja, K., Bordea, G., Vulcu, G., O'Brien, L., Espinoza, A., Abi-Lahoud, E., Buitelaar, P., and Butler, T. (2015). Semantic annotation of finance regulatory text using multilabel classification. In *Legal Domain And Semantic Web Applications (LeDA-SWAn)*. To appear.

Athan, T., Boley, H., Governatori, G., Palmirani, M., Paschke, A., and Wyner, A. (2013). OASIS LegalRuleML. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (ICAIL 2013)*, pages 3–12, Rome, Italy.

Bos, J. (2008). Wide-coverage semantic analysis with boxer. In Johan Bos et al., editors, *Proceedings of Semantics in Text Processing*, Research in Computational Semantics, pages 277–286. College Publications.

Dayal, S. and Johnson, P. (2000). A web-based revolution in Australian public administration. *Journal of Information, Law, and Technology*, 1. Online.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115, January.

Francesconi, E. (To appear). Semantic model for legal resources: Annotation and reasoning over normative provisions. *Semantic Web Journal*.

Fuchs, N. E., Kaljurand, K., and Kuhn, T. (2008). Attempto controlled english for knowledge representation. In *Reasoning Web*, pages 104–124.

Guissé, A., Lévy, F., and Nazarenko, A. (2012). From regulatory texts to BRMS: how to guide the acquisition of business rules? In *RuleML 2012*, Montpellier, France.

Hoefler, S. (2012). Legislative drafting guidelines: How different are they from controlled language rules for technical writing? In *CNL 2012*, volume 7427 of *Lecture Notes in Computer Science*, pages 138–151.

Kuhn, T. (2014). A survey and classification of controlled natural languages. *Computational Linguistics*, 40(1):121–170, March.

Lévy, F., Nazarenko, A., and Wyner, A. (2015). Towards a high-level controlled language for legal sources on the semantic web. In *Workshop on Legal Domain And Semantic Web Applications (LeDA-SWAn 2015)*. To appear.

Maynard, D., Li, Y., and Peters, W. (2008). NLP techniques for term extraction and ontology population. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*, pages 107–127, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Mariette Meunier, et al., editors. (2013). *La traduction juridique : Points de vue didactiques et linguistiques (Actes de colloque international 2010)*. Publications du Centre d'Etudes Linguistiques. 333 pages.

Marion Meunier-Crespo et al., editors. (2011). *Faut-il simplifier la langue du droit?* GREJA, Nov.

OMG. (2008). Semantics of business vocabulary and business rules (sbvr). formal specification, v1.0. Technical report, The Object Management Group.

Shiffman, R. N., Michel, G., Krauthammer, M., Fuchs, N. E., Kaljurand, K., and Kuhn, T. (2010). Writing clinical practice guidelines in controlled natural language. In *Proceedings of the 2009 conference on Controlled natural language*, CNL'09, pages 265–280, Berlin, Heidelberg. Springer-Verlag.

U.S. Government. (2015). Plain language: Improving communications from the federal government to the public, `http://www.plainlanguage.gov/index.cfm`.

Wyner, A. and Peters, W. (2011). On rule extraction from regulations. In Katie Atkinson, editor, *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference*, pages 113–122. IOS Press.

Wyner, A. and van Engers, T. (2010). A framework for enriched, controlled on-line discussion forums for e-government policy-making. In Jean-Loub Chappelet, et al., editors, *Electronic Government and Electronic Participation*, pages 357–364, Linz, Austria. Trauner Verlag.

Wyner, A., Angelov, K., Barzdins, G., Damljanovic, D., Davis, B., Fuchs, N., Hoefler, S., Jones, K., Kaljurand, K., Kuhn, T., Luts, M., Pool, J., Rosner, M., Schwitter, R., and Sowa, J. (2010a). On controlled natural languages: properties and prospects. In *Proceedings of the 2009 conference on Controlled natural language*, CNL'09, pages 281–289, Berlin, Heidelberg. Springer-Verlag.

Wyner, A., van Engers, T., and Bahreini, K. (2010b). From policy-making statements to first-order logic. In *EGOVIS*, pages 47–61.

Wyner, A., Bos, J., Basile, V., and Quaresma, P. (2012). An empirical approach to the semantic representation of law. In *Proceedings of 25th International Conference on Legal Knowledge and Information Systems (JURIX 2012)*, pages 177–180, Amsterdam, The Netherlands. IOS Press.

Rule-Based Technical Writing: A Meta-Standard on Controlled Language Extended towards Controlled Communication

Christian Galinski1, Blanca Stella Giraldo Pérez2
International Information Centre for Terminology (Infoterm)
Gumpendorfer Strasse 65/1, 1060 Vienna, Austria
[1]christian.galinski@chello.at, [2]blancaese@gmail.com

Standardization of rule-based technical writing (RBTW) emerged in English in certain industries. It started with Simplified Technical English (STE), or Simplified English which is the original name of a controlled language standard originally developed for aerospace industry maintenance manuals. Formerly called AECMA Simplified English, ASD (Aerospace and Defence Industries Association of Europe) renamed it to ASD Simplified Technical English. ASD-STE became so widely used by other industries and for a wide range of document types, that 'simplified English' is often used as a generic term for 'controlled language'. Today the controlled language approach is applied in probably about a hundred languages, particularly in user instructions of all sorts. Increasingly such user instructions have to be rendered eAccessible, as the Convention on the Rights of Persons with Disabilities (CPRD) has been adopted into national legislation by numerous countries. As the needs of persons with disabilities (PwD) should be taken into account, whether in paper form or on websites, a systematic approach is commended for the development of such content on paper and as equivalent web content. For this purpose, a meta-standard with rules for the formulation of RBTW guides or standards would be useful.

Keywords: Simplified English, rule-based technical writing (RBTW), multilingual and multimodal user instructions, controlled language, controlled communication, standardization, meta-standard

1. History of Standard 'Simplified English'

Standardization of rule-based technical writing (RBTW) started with Simplified Technical English (STE), or 'Simplified English' which is the original name of a controlled language standard developed for aerospace industry maintenance manuals. AECMA [1] originally created the standard in the 1980s, based on a standard of the aircraft producer Fokker, which in turn had borrowed from earlier controlled languages, especially Caterpillar Fundamental English. In 2005, the Aerospace and Defence Industries Association of Europe (ASD), after subsuming AECMA in 2005, renamed the standard to *ASD Simplified Technical English* or ASD-STE100. (ASD, 2013) Although it was not intended for use as a general writing standard, it has been successfully adopted by other industries and for a wide range of document types. It became so widely used, that 'simplified English' is often used as a generic term for a 'controlled language'. However, similar approaches are now used in many languages – not to mention in many more application fields as before.

Complementary to ASD-STE100, ASD developed the XML specification S1000D for preparing, managing, and using equipment maintenance and operations information for use with military aircraft. It has since been modified for use with land, sea, and commercial equipment. S1000D (2007) requires a document to be broken down into individual data items (called data modules) which can be marked with individual XML labels and metadata, and be part of a hierarchical XML structure.

RBTW influenced 'rule-based technical communication' in various written and spoken forms, such as common and necessary in aviation communication (ICAO, s.a.) on the one hand; in other application fields, such as military or emergency services, it had already existed in some form or the other sometimes for decades, on the other hand.

With respect to multilingual user instructions in printed, electronic form or in the form of structured web content, it will need a comprehensive approach including from the outset requirements of

- Multilinguality: covering also cultural diversity to be dealt with by localization (L10N) and internationalization (I18N) techniques,
- Multimodality: technically implemented through multimedia,
- eInclusion and eAccessibility: technically implemented through assistive technology (AT),
- Multi-channel presentations, in order to cope with many display formats and sizes.

These requirements were formulated concisely by the *Recommendation on software and content development principles 2010* (MoU/MG, 2012) which should be considered at the earliest stage of the software design process and data modeling (including definition of the metadata), and hereafter throughout all the iterative development cycles. It was adopted in 2012 by the Management Group (MoU/MG) of the ITU-ISO-IEC-UN/ECE Memorandum of Understanding concerning eBusiness standards.

[1] AECMA: French acronym for the Association Européenne des Constructeurs de Matériel Aérospatial, in English: European Association of Aerospace Manufacturers

2. New Developments and Requirements

The following developments certainly will have an impact on the development of ASD-STE100 and similar standards:

- The use of mobile devices for technical documentation (TD) – and especially user instructions implying the application of responsive web design (RWD) – provides for many new combinations of written information with non-written features (incl. the compression of information in order to increase readability and comprehensibility).
- The adaptation of ICT systems and content presentation to comply with standards related to eAccessibility&eInclusion, such as ISO/IEC 40500:2012 Information technology – W3C Web Content Accessibility Guidelines (WCAG) 2.0 (ISO/IEC, 2012) is becoming a legal requirement.
- The further development of the approaches of internationalization (I18N) and localization (L10N) towards more languages and language varieties, in order to reach more communities and also provide more and more diverse user communities with content in their respective language variety – as much as necessary/appropriate.

In this connection, the use of spoken language and other means of communication will increase and will have to be integrated into existing approaches and technologies.

ASD-STE100 has been adapted to other languages, such as in the German-speaking region with *Regelbasiertes Schreiben – Deutsch für die Technische Kommunikation* (Tekom 2013).

Given the fact that quite a number on specific language oriented RBTW standards or guidelines exist, there seems to be a need for an international meta-standard on RBTW.

3. Meta-standard for RBTW

An international meta-standard on RBTW should applies to more or less

- All languages (sometimes also certain language varieties), where technical documentation or technical communication is needed – e.g. in user instructions,
- All domains and subjects and their applications (particularly scientific-technical fields, but also beyond where applicable),
- Other linguistic and communicative aspects, such as language and communication proficiency of user,
- Controlled communication means, in order to take into account among others the needs of persons with disabilities (PwD), particularly those having certain kinds of 'communication disorder'.

Such a meta-standard would also take the educational level and other factors of different target groups into account, including considerations for aspects of:

- Localization (L10N) in the meaning of "the process of modifying products or services to account for differences in distinct markets" (The Globalization Industry Standards Association LISA, 2007)
- Internationalization (I18N) in the meaning of "the process of enabling a product at a technical level for localization" (The Globalization Industry Standards Association LISA, 2007)

Such a meta-standard could be for instance an extension of ISO/TS 24620-1:2015 (ISO, 2015) *Language resource management – Controlled natural language – Part 1: Basic concepts and general principles*. This new part of 24620 about RBTW focusing on user instructions, would facilitate the development of tools for:

- Measuring the degree of text readability and text comprehensibility (of written and spoken texts),
- Authoring rule-based user instructions,
- Checking individual content elements or processes.

As proven in other system developments – for instance the experience with machine translation approaches and systems – such dedicated tools would be much more effective than general purpose controlled language systems, because of the focus on RBTW.

4. Definitions

In connection with rule-based technical writing the following main concepts have to be defined:

eAccessibility: approaches to ensure that all citizens have access to Information Society services.

NOTE: eAccessibility is about removing the technical, legal and other barriers that some people encounter when using ICT-related services. It also concerns people with disabilities (PwD) and certain types of elderly people with impairments. (European Commmission EUR-Lex, 2005)

eInclusion: approaches to achieve that "no one is left behind" in enjoying the benefits of ICT

NOTE: eInclusion means both inclusive ICT and the use of ICT to achieve wider inclusion objectives. It focuses on participation of all individuals and communities in all aspects of the information society. eInclusion policy, therefore, aims at reducing gaps in ICT usage and promoting the use of ICT to overcome exclusion, and improve economic performance, employment opportunities, quality of life, social participation and cohesion. (European Commission, 2010)

As the two terms eAccessiblity and eInclusion are overlapping they are used in this contribution in the combined form of eAccessibility&eInclusion.

controlled language, CL: approaches to apply restrictions on vocabulary, grammar and/or semantics on natural language for the purpose to improve the readability and comprehension of texts duly considering the target user group

NOTE: CL includes among others approaches that have been called simplified language, plain language, formalized language, processable language, conceptual authoring, language generation, and guided natural language interfaces, etc. CL in the singular refers to the principles, rules and methods applied to languages and language variations. In the plural CL refers to the individual language or language variation subjected to CL approaches.

controlled communication: <in analogy to controlled language> approaches to apply restrictions on elements of communication for the purpose to improve the understanding taking into account the needs of the target group and making use of the whole range of multimodality

NOTE: Elements of communication can comprise those at the level of lexical semantics (e.g. hand signs, animated graphs) or at higher semantic levels equivalent to messages.
multimodality: <theory of communication and social semiotics> communication practices in terms of the textual, aural, linguistic, spatial, and visual resources – or modes – used to compose messages
NOTE: Where media are concerned, multimodality is the use of several modes (media) to create a single artifact.
technical writing, TW: any written form of writing or drafting technical communication used in a variety of technical and occupational fields, such as computer hardware and software, engineering, chemistry, aeronautics, robotics, finance, consumer electronics, and biotechnology
NOTE: TW encompasses the largest sub-field within technical communication. The Society for Technical Communication defines technical communication as any form of communication that exhibits one or more of the following characteristics: "(1) communicating about technical or specialized topics, such as computer applications, medical procedures, or environmental regulations; (2) communicating through technology, such as web pages, help files, or social media sites; or (3) providing instructions about how to do something, regardless of the task's technical nature". (STC, 2016)
rule-based technical writing, RBTW: technical writing carried out based on the principles and rules of controlled language
NOTE: The applications of principles and rules of RBTW in present standards does not compel the use of tools. However, they should be sufficiently concise and granular, in order to facilitate the development of the respective CL-based authoring tools.
rule-based technical communication, RBTC: <in analogy to RBTW> technical writing carried out based on the principles and rules of controlled communication
NOTE: RBTC exists in various written and spoken forms, such as necessary in aviation communication, military or emergency services. It needs further extension to cover the whole range of multimodality, such as necessary in augmentative and alternative communication (AAC) or with/among persons with disabilities (PwD) in general.
(text) readability: degree of easiness of reading texts which in turn indicates the degree of text comprehension (Yi, Park, & Cho, 2015)
NOTE: The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success that a group of readers have with it. (Dale & Chall, 1949)
(text) comprehensibility: degree of easiness of extracting and constructing meaning from written words, sentences, and text
NOTE: Text comprehension is the process of extracting and constructing meaning from written words, sentences, and text. (Yi, Park, & Cho, 2015)
Depending on the scientific approach, comprehensibility largely overlaps with readability (the latter including or not including legibility).
communication disorder: impairment in the ability to receive, send, process, and comprehend concepts or verbal, nonverbal and graphic symbol systems
NOTE: "A communication disorder may be evident in the processes of hearing, language, and/or speech. A communication disorder may range in severity from mild to profound. It may be developmental or acquired. Individuals may demonstrate one or any combination of communication disorders. A communication disorder may result in a primary disability or it may be secondary to other disabilities." (ASHA, 1992)

5. Content of a meta-standard on RBTW

We will now elaborate on structure and content of the proposed meta-standard for the development of RBTW guides or standards based on experience with existing guidelines in several languages. Its main purpose is to give guidance to developers of RBTW guides or standards in whatever language (sometimes also in certain language varieties) and for whatever purpose where a controlled communication approach is needed. If a technical committee would head for an ISO standard, the proposed RBTW meta-standard would comprise (in addition to the standard Foreword):

0 Introduction: states the background, objectives, related standards and approaches
1 Scope: states the extent to which the specifications of the RBTW standard apply
2 Normative references: lists references to other standards or normative documents referred to in the RBTW standard
3 Terms and definitions: explains the main concepts occurring in the RBTW standard represented by terms and definitions
4 Rules and specifications
4.1 General rules concerning RBTW:

- Aims of the rules, such as saving of space, writing with translation/localization in mind
- How to formulate the rules
- Interconnection between the rules
- How to deal with alternative rules
- Quality criteria and automatic checking
- Legal considerations
- How to introduce RBTW

4.2 Rules referring to the text (as a whole):

- Consider the target users
- Formulate principles on how to structure the text
- Rules concerning headings for parts of the text: short and clear, no repetition
- One topic in a paragraph is given one (sub)heading
- How to use cross-references
- Relation between textual information and non-verbal representations
- How to mark keywords in text and design the keyword index
- Recommendation for preferences for certain parts of speech (if applicable)
- Formulate principles for using tables
- Rules for enumerations
- Rules for highlighting: by using color, typefaces …

- Rules for the glossary (if applicable)
- Explanations and pointing to an explanation

4.3 Rules for sentences:
- Completeness of sentences
- Rules for the relation between sentences
- Preference for simple sentence structures
- Rules for the relations between sentence elements
- Rules for various parts of speech (PoS)
- Rules for the use of parentheses
- What style to use to address a target user group

4.4 Rules concerning words and terms
- General rules for using certain words or terms
- Rules concerning word or term formation
- Rules on abbreviations
- Rules for using synonyms
- Rules for (obligatory or arbitrary) special signs: diacritics etc.
- Use of syntactic signs, numbers, mathematical symbols etc. in words or terms
- Rules for loanwords or borrowed terms
- How to use collocations and phraseology

4.5 Orthography:
- Which standards or laws to follow
- How to deal with alternative rules

4.6 Punctuation:
- Which standards or laws to follow
- How to deal with alternatives

4.7 Typographical features: for example for
- Highlighting
- Other purposes

4.8 Annexes

Part 4.4 for example could be formulated in a generic way as presented in the column on the right.

Of course, it has to be stated at several places that different language – and different writing systems – are subject to specific requirements. One and the same language may be written with different scripts, or may be subject to different orthographic or other regulations.

In this way the meta-standard on RBTW would be concise and provide guidance to those who want to formulate a RBTW guide or standard in a language (or language variety) or domain or application, where such a document is needed but does not yet exits. In addition, it can be used for benchmarking of such documents.

> **EXAMPLE: 4.4 Rules concerning words and terms** (draft)
> Depending on the language (or language variety) there may be different rules for word separation, word formation, abbreviations, synonyms, non-verbal elements in words, loanwords and compounding of words. There may be special requirements concerning scientific or technical terms.
> The RBTW guide or standard should define how to identify (e.g. for mark-up purposes) words or terms and how they are to be used in running text, headings, glossaries, etc. The use of words or terms shall be consistent as much as possible.
> The following phenomena should be avoided:
> - Unnecessary synonyms
> - Abbreviations representing other concepts
> - …
>
> If there are no – or not generally agreed upon – rules concerning words, the RBTW guide or standard should formulate a set of rules or specifications – preferably in analogy of similar guides or standards in other languages.
> After each rule or specification, it should be stated, whether there are automatic means to check lexical elements or processes (such as spell check, word count, etc.).
> Depending on the language, domain or application, there may be specific technical or legal regulations which must be observed.

6. New Interoperability Requirements

For the sake of re-usability, devices, communication and human communication resources (HCR) should also comply with the requirements of multilingualism and cultural diversity. There are no formal standards concerning human communication (h-h, h-m, m-m) – especially under the requirements of eAccessibility&eInclusion. A number of requirements and specifications for RBTW also apply – possibly in adapted form – to controlled communication approaches or a combination thereof for the communication with/among PwD which may require communication modalities other than written or spoken language.

In the framework of the EU project IN LIFE (INdependent LIving support Functions for the Elderly), human communication refers to the communication with/among PwD, between PwD and their carers, between PwD&carers and devices, among the devices. IN LIFE plans a major initiative in this direction with the aim to help:

- To formulate rules for reducing the potential number of utterances (and lexical items) depending on the kind and degree of communication disorders,
- To facilitate the conversion of any utterance (and lexical items) into other modalities and vice versa,
- To facilitate the adaptation of utterances or other modes of representing meaning under a cultural diversity perspective,
- To allow the development of human communication resources (HCR) fully complying to the requirements of multilingualism and multiculturalism.

In addition, the combination of the above-mentioned approaches will make more efficient:

- the use of devices supporting human communication,
- the training of speech (and human communication at large) systems to support human communication.

If RBTW would take these approaches and requirements to comply with the needs of PwD, it could have a methodologically speaking positive effect on RBTW itself. As PwD have to use many devices, products and services, RBTW has to comply with legal requirements concerning eAccessibility&eInclusion.

7. Conclusion

Given the degree of detail of several existing guides or standards on RBTW, it is considered to be possible to formulate generic provisions in a meta-standard on RBTW without referring to specific natural languages. However, if necessary, major language types may be referred to.

Given legal requirements stemming from the Convention on the Rights of Persons with disabilities (CRPD), aspects of eAccessibility&eInclusion have to be taken into account also in technical documentation/communication. Therefore, provisions to this extent have to be introduced in RBTW guides and standards – at least in countries that have signed the Convention.

8. Acknowledgements

Thanks are due to the EU Commission co-financing the IN LIFE project.

9. Bibliographical References

Ad Hoc Committee on Service Delivery in the Schools (1992). *Definitions of Communication Disorders and Variations.* ASHA, American Speech-Language-Hearing Association Guidelines.

ASD. (2013). Simplified Technical English - Specification ASD-STE100:2013. Issue 6. Retrieved from http://guiseppegetto.com/pwr393/wp-content/uploads/2013/02/ASD-STE100-ISSUE-6.pdf

Basic S1000D Comparison with Traditional Documentation Methods (2007). *Technical Publications Specification Management Group (TPSMG)*. Retrieved from http://www.s1000d.net/s1000d-comparison.pdf

Dale, E., & Chall, J. (1949). *The concept of readability*. The concept of readability - Elementary English.

European Commission (2010) *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - A Digital Agenda for Europe*. Brussels.

European Commmission EUR-Lex. (2005). EUROPA: EU law and Publications, EUR-Lex 52005DC0425: *Communication from the Commission to the Council, the European Parliament and the European Economic and Social Committee and the Committee of the Regions - eAccessibility*. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52005DC0425

ISO. (2015). ISO/TS 24620-1:2015 *Language resource management -- Controlled natural language (CNL) -- Part 1: Basic concepts and principles*. Retrieved from http://www.iso.org/iso/catalogue_detail.htm?csnumber=37334

ISO/IEC. (2012). ISO/IEC 40500:2012 *Information technology – W3C Web Content Accessibility Guidelines (WCAG) 2.0*

The Localization Industry Standards Association. (LISA). (2007). *The Globalization Industry Primer: An Introduction to preparing your business and products for success in international markets*.

MoU/MG. (2012). MoU/MG/12 N 476 Rev.1 *Recommendation on software and content development principles* 2010. Retrieved from http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/MoU-MG/Moumg476Rev.1.pdf

STC Society for technical communication. (2016). *Defining technical communication*. Retrieved from http://www.stc.org/about-stc/the-profession-all-about-technical-communication/defining-tc

Tekom. 2013. *Regelbasiertes Schreiben – Deutsch für die Technische Kommunikation*. 2nd enlarged edition.

United Nations (UN). (2006/2008). Convention on the Rights of Persons with Disabilities (CRPD). Retrieved from https://en.wikipedia.org/wiki/Convention_on_the_Rights_of_Persons_with_Disabilities#Committee_on_the_Rights_of_Persons_with_Disabilities

Yi, W., Park, E., & Cho, K. (2015). *E-Book Readability, Comprehensibility and Satisfaction.* Retrieved from https://www.researchgate.net/publication/221089846_E-book_readability_comprehensibility_and_satisfaction

10. Language Resource References

International Civil Aviation Organization (ICAO). (s.a.). *ICAO Standard Phraseology. A Quick Reference Guide for Commercial Air Transport Pilots*. Retrieved from http://www.skybrary.aero/bookshelf/books/115.pdf

A Controlled Language for Sense Mining and Machine Translation for Applications in Mission-Critical Domains

Sylviane Cardey
Centre Tesnière
UFR SLHS, 30 rue Mégevand, F-25030 Besançon Cedex, France
E-mail: sylviane.cardey@univ-fcomte.fr

Abstract

In this paper we present methodologies as well as the theoretical contributions involving the analysis and generation of texts for the application of controlled languages in multilingual mission-critical domains particularly safety-critical such as aeronautics, medicine and civil protection, where reliable results are obligatory. We show that the analysis involves the extraction of sense, that is sense-mining, and the generation that of controlled texts and their machine translation. This work has involved language modelling based on micro-systemic linguistic analysis, this itself being underpinned by a formal mathematical model, and which also inherently provides traceability, mandatory in safety-critical applications. A norms based approach is described involving extraction and application of norms in order to use them in the methodologies, both for analysis and generation. Applications, application domains and applicability are discussed.

Keywords: controlled language, machine translation, mission-critical application, safety-critical application, sense mining

1. Introduction

In this paper we present research results involving methodologies which have been developed enabling three applications: sense mining, controlled languages and machine translation and where all three are for use in mission critical domains particularly safety-critical such as aeronautics, medicine and civil protection which means that the work undertaken must lead to reliable results. Our work has involved language modelling (Cardey, 2013) based on micro-systemic linguistic analysis developed in Centre Tesnière (Cardey & Greenfield, 2006), this itself being underpinned by a formal mathematical model (Cardey & Greenfield, 2005).

Analysis and generation of texts involving a controlled language are first presented. Theoretical considerations both from the linguistics and computational points of view are addressed. Representations for sense mining and machine translation are then presented. Traceability, inherent in micro-systemic linguistic analysis, is discussed, this being mandatory in safety-critical applications. Our norms based approach for a controlled language is then presented. The applications and domains of application are then addressed commencing with a theoretical and architectural schema encompassing sense mining, controlled language and machine translation, and this is followed by a conclusion.

2. Analysis and Generation

2.1 Analysis

Analysis concerns the extraction of sense, be this for sense mining or for machine translation. The methodology which is applied is the same for both. The descriptive model uses the same rule format. The methodology is indeed generally applicable to diverse applications.

Figure 1 shows 2 rules, one for sense mining and the other for machine translation. The formal representation is the same for both.

Sense mining:

```
l(d) + '['(_) + 从 + l(chiffres)
+ 到 / 至 + l(chiffres) + l(temps)
+ ']'(_) + l(f)
```

Machine translation:

```
opt(neg1) + lexis('يجب') +
opt(neg2) + nver + arg1(acc) +
opt(opt(prep_comp1),comp1(n)) +
opt(opt(prep_comp2),comp2(n)) +
pt
```

Figure 1: Example of rules for sense mining and machine translation

2.2 Generation

Generation here concerns utterances produced by the controlled language and output by the machine translation system, these utterances being intimately linked because the machine translation results depend not only on the translation model, but also on the control effected on both the source and also the target language, this point being discussed later in the paper. Concerning controlling per se, the goal is not only to improve the quality of the utterances, but also their comprehensibility in suppressing, amongst others, any ambiguities.

2.2.1. Lack of Precision and Controlled Languages

The following example has been extracted from the French Red Cross (Croix-Rouge) first aid guide for home use (Guide des premiers secours à la maison).

Initial text:

Disposez le lien en double sous le membre blessé alors que vous maintenez le point de compression.
(*Place the link doubled under the injured limb while you maintain the pressure point.*)

Because of the use of certain terms, the lack of precision and the non-compliance concerning the chronology, this extract seems difficult to understand, even for a learner, but particularly so when it has to be used in a real situation. We have conducted tests during the ‘LiSe’ project (Linguistique et Sécurité ANR-06-SECU-007) (Cardey, Anantalapochai et al., 2010) involving this text which have enabled verifying the improvement due to the LiSe controlled language. The participants reported having had to reread the original text several times before understanding it and being able to establish the precise chronology to respect, problems which did not occur with the text reformulated in the LiSe controlled language.

Reformulated text:

> Pendant la pose du garrot :
>
> > Maintenir le point de compression.
>
> Pour poser le garrot :
>
> > Plier le lien en 2.
> >
> > > *But : obtenir une boucle*.
> >
> > Passer le lien sous le membre inférieur de la victime.
>
> (While placing the tourniquet:
>
> > Maintain the pressure point.
>
> To apply the tourniquet:
>
> > Fold the link in 2.
> >
> > > *Purpose: to obtain a loop.*
> >
> > Place the link under the lower limb of the patient.)

2.2.2. Interferences

One of the LiSe controlled language rules is the attribution of a unique sense to each lexical entry. However certain polysemic terms can occur in the everyday lexicon as well as in one or even several specialty domains; we call such a phenomenon ‘domain interference’. A strict application of our aforementioned rule results in reducing the scope of the LiSe controlled language to a particular domain. So as to ensure our controlled language’s application to diverse domains, specific senses can be attributed to the same lexical entry according to the application domain or the intended audience (general public, professional etc.). Thus the term plateau will be generally used with its most common meaning support plat (serving to put or transport objects) which is translated in English by tray, in Arabic by طبق and in Chinese by 盘子. In a medical protocol addressed at specialists, this same term plateau could designate the support upon which one places the instruments required for carrying out an operation, tray in English, صينية in Arabic and 托盘 in Chinese. However, when plateau is qualified as technique (e.g. plateau technique de radiologie), the set of equipment needed to perform an examination is translated by technical wherewithal in English, معدات تقنية in Arabic and 器械盘 in Chinese. If we pass to the aeronautical domain, here the term plateau will designate a relatively flat area of country which dominates its surroundings, plateau in English, هضبة in Arabic and 高原 in Chinese.

In respect of interferences, there is also the problem of phoneme confusion; we simply cite the French dessus and dessous (on and underneath) which will both be pronounced the same by an Anglophone with the phoneme [u] (ou) for both ou and u, the phoneme [y] (u) not existing in English.

3. Our Methodology/Other Methodologies

3.1 Our Methodology for Machine Translation

Our methodology requires no pre-edition in the conventional sense of the term.

As we use a controlled language which firstly serves to provide a good interpretation of the information to be transmitted by suppressing any ambiguities and all which is detrimental to the intelligibility of the information, the writing guide that we have developed and which is incorporated into the user interface of the controlled language machine translation system, is an aid to the user when entering the sentence to be translated (see Figure 6). However, we quickly discovered that even in controlling the source language, the results were still far from what we had hoped. We therefore decided to control also the target languages; this signifies that a very fine comparative analysis of the target languages and French, the source language, was undertaken. We were thus able to extract mega and micro structures which were similar not only for French and the target languages, but also between the target languages. We have constructed our translation system from these resemblances, each divergence being subsequently treated at the specific transfer level for each language.

The following example shows how and why this control is necessary. We take the non-controlled sentence:

> *Refroidir immédiatement la brûlure, en l'arrosant avec de l'eau froide durant 5 minutes*.
>
> (*Immediately cool the burn, watering it with cold water for 5 minutes.*)

(example taken from P. Cassan, C. Cross, (2005), « Guide des premiers secours à la maison », Editions d'Organisation, Eyrolles pratique, ISBN-13: 978-2708135789, 182 p.).

After controlling we obtain:

> *Verser de l'eau froide sur la brûlure immédiatement durant 5 minutes.*
>
> (*Pour cold water on the burn immediately for 5 minutes.*)

The reasons for the control are as follows:

- The sentence contains two distinct pieces of information: 1) the injunction *arroser* and 2) the explanation *refroidir*. It seems more logical to state first the action to be done and only afterwards the motives for this action. Furthermore, in order to ensure understandability as well as obtaining a good translation, the controlled language imposes a unique verb for each sentence, and also forbids the use of the gerundive, here *arrosant*.

- The control is required also as a result of the three target languages, Arabic, Chinese and English, because of the verb *arroser*. These three target languages use this verb only when it is followed by an argument which is of vegetable type.
- So as to avoid any error in identifying a pronoun's antecedent, pronouns are forbidden in the controlled language.

Controlling resolves numerous linguistic problems; nevertheless some remaining problems are observed when controlled sentences are translated by machine translation systems which are available on the market. In the examples that follow, elements exhibiting problems are underligned.

Chinese Reverso:

在 5 分钟期间在烧伤上立刻倒冷水 (preposition 在 5min pendant brûlure maintenant verser froide eau) (preposition在 5min during burn now pour cold water)

The problem with the Chinese is at the structural level, and there is also a lexical inexactitude concerning *pendant*.

'在 …期间' can only be used for a duration much longer than 'minute', for 'année' (year) for example. For the error concerning *brûlure* (burn), in this context one would use rather *blessure* (wound) in Chinese.

Arabic Reverso:

الدفع (صب) بعض الماء البارد على أن تحترق بعد5 دقائق

(poussée (verser) quelque eau froide pourvu que tu brûles après 5 minutes) (pushed (pour) some water cold provided you burn after 5 minutes)

English Reverso:

Cool at once the burn, by spraying him(it) with some cold water for 5 minutes.

There are 2 problems in the English – the location of the complement *at once* which is understandable but non standard, and the problem of the pronoun.

We give below the results produced by our machine translation system:

Chinese:

立刻在伤口上浇冷水 5 分钟，

Arabic:

يجب صبّ الماء البارد على الحرق فورا لمدّة 5 دقائق .

English:

Pour cold water on the burn immediately for 5 minutes.

3.2 Our Methodology for Sense Mining

We work at the level of sense in general, that is to say with all the various elements (morphemes (syntactic or derivational flexions (lexical), simple or compound lexes, etc.) and with their organisation and distribution in the sentence, or in the lexes for the morphemes. Current methodologies based on the use of keywords therefore operate on part of the lexis, and not the lexis in its totality. Thus lists of 'non important words' are created (also called 'empty words') so as to enable recognising only the words (terms) called 'important' words or 'keywords'. The problem is what is an 'important word'? The principal question is what is a word? Take the example: *Ce produit aurait dû parfait* (This product should have been perfect), where there is the understatement *il ne l'est plus* (it is no longer perfect). So, if we keep only *parfait* and *produit*, we obtain a bad interpretation. The same bad interpretation occurs in Arabic if we only consider the two words '**المنتوج**/produit' and '**ممتازا**/parfait'. It has to be added that current methods require training and/or pre-edition, and this is not possible in crisis situations due to the lack of time.

Our methodology, sense mining (Cardey et al., 2006), which was developed in the context of a project involving classifying and interpreting a food industry enterprise's customer verbatims uses not only the lexicon in its totality, principally its morphology, but also and most importantly syntax and of course semantics together with their morpho-syntactic, lexico-syntactico-semantic etc. intersections represented by rules and sets structured in systems functioning in interrelation. Sense mining interprets a text even if it contains no word said to be a keyword, and it analyses all the text.

4. The Theoretical Point of View

4.1 From the Linguistic Point of View

Micro-systemic linguistic analysis (Cardey, 2013), developed in Centre Tesnière, does not have as goal describing the whole of a language by means of some global representation of the different 'layers': lexis, syntax, morphology, semantics separately. Rather, this analysis method advocates firstly delimiting the problem or the analyses' needs concerning some specific application. According to the needs, a specific system is constructed that represents and resolves the problem. This system can be manipulated and represented, which cannot be done for a language in its totality which latter can neither be delimited nor manipulated. Thus only the necessary elements, be these lexical, morphological, syntactic, are represented in a single system, this latter being able to be related to other such systems. We do not even mention semantics because, finally, this is the only 'layer' that interests us; whatever the operations one does, in reality these are to be able to access the sense. Thus we do not need a complete description of the language, or languages, concerning their lexis or their morphology as habitually one tries to do. This allows us to resolve problems thanks to analyses which are much lighter in size and in time spent.

Several Arabic utterances from the simplest to the most complex, and providing important semantic elements, can be for example simply represented by the macro-structure shown in Figure 2 (Mikati, 2009). We do not need a transformational analysis to enables us, amongst other things, to pass from an affirmation to a negation. One can observe that syntax, morphology and certain categories that have been defined according to the needs are presented all together in this macro-structure. This structure is linked to micro-systems that have been defined for the type of problem to be processed.

```
opt(particule(s))+ (…) + verbe +
opt(particule(s) + (…) + sujet +
opt(particule(s)) + (…) + cod +
opt(particule(s)) + (…)
```

Figure 2: Macro-structure covering several Arabic utterances from the simplest to the most complex.

4.2 From the Calculability and Computational Points of view

Micro-systemic linguistic analysis is based on discrete mathematics and in particular on constructive logic and model theory, set theory, relations and partitions. This basis also serves as the language of communication between the linguist and the software engineer as, independent of applications, it is understood by both. In respect of the calculations to be performed for some application, a representation such as those presented in Figures 1 and. 2 can be interpreted by a computer program that has been specified and implemented according to the linguistic model. Thus the linguist does not have to be concerned with the computer programming; instead it is the software engineer who, according to the linguistic model, constructs the computational model and the subsequent program. The advantage here is that instead of having some predefined computational model, which unfortunately in practice is in reality nearly always the case, and with which the linguist(s) must either 'bend' linguistics or 'bend' the language(s), each linguist performs *linguistic programming* by entering his or her data in, for example, a spread-sheet table (see Figure 3), this latter being then interpreted by the computer program produce by the software engineer, the only constraint being that of consistency. Another advantage is that one can add as many languages as one wishes, and also as many linguists as is necessary.

4.2.1. Representation for Machine Translation

A spread-sheet file serves as the link between the linguist who is a specialist in the source and target languages, and the software engineer. This file is designed so as to enable simple and rapid formalisations; additions and modifications are carried out on the different tables making up the file.

Take for example the French (source language) sentence from the aeronautical domain *maintenir le train d'atterrissage en position DOWN* (keep the landing gear on DOWN position) so as to show the data input process for its translation to English (target language). The spread-sheet file includes several tables which correspond to the formalisation model established during the LiSe project. The linguist enters the source language *verb* together with its target language translation in the verbal group table *anC_groupesVerbaux_frC* (*an*glais – English). Identifiers with a numerical component are associated with each verb according to its macro-structure. The French verb *maintenir* will be ascribed by the linguist *fr*C_7 which refers to a particular structure in French. For this identifier, the linguist will make correspond another identifier which represents the target structure, which for this example is *anC*_1.7. A second table *anC_groupes* retakes the two macro-structures (source and target) with the already attributed identifiers. Once this second table has been completed, the other tables which depend on it will be filled by the linguist according to the content of the two macro-structures. The structure of our verb *maintenir* imposes going to the table *anC_args_frC* which lists all the possible micro-structures with all the transfer rules which are associated with them. In our example, the two micro-structures which represent respectively *le train d'atterrissage* as arg1 and *position DOWN* as arg2 will be entered in this table. It is important to underline that another table *anC_comps_frC* could intervene if our structure includes complements. Two other tables exist which act as an inventory of all the parts of speech that have been determined according to the source language and the target language, and which are necessary respectively for segmentation and generation. The final table is *anC_dictionnaireLexical_frC* which contains the dictionary; this table is invoked during the construction of the target sentence, the final phase of the translation process.

4.2.2. Representation for Sense Mining

In Figure 3 we show as an example an extract from a spread-sheet table which has been programmed by the linguist; this table is part of a system for the automatic recognition of acronyms which has been implemented during a project with Airbus France in the context of safety-critical technical documentation (Cardey et al., 2009).

To avoid confusing acronyms and conventional lexis, it seemed to us judicious that the controlled acronyms do not contain certain sequences of graphemes present in the host language, namely American English. Our technique, as mentioned in the Introduction of this article, is based on micro-systemic linguistic analysis, itself underpinned by a formal mathematical model. We observe here that in terms of legal sequences of graphemes in English, we have partitioned the English lexis over sequences of 2 and of 3 letters contained in each lexical unit; thus each non-empty cell (not containing "Ø") which contains a couple (sequence of letters, attestation) gives rise to a distinct equivalence class of English lexical units which share this same sequence, the couple being in effect the name of the equivalence class. In particular, the hapax attestations correspond to equivalence classes that are singletons. As traceability is mandatory due to the safety-critical nature of the domain, the attestations act not only as a static trace of justification, but also as determining factors during the algorithmic interpretation of the table by the automatic recognition program in producing a dynamic trace.

ACRONYMS_Legal_Sequences_Graphemes					
Start_References					
Ø	COMPETENCE	Nov - Dec 2007			
M	www.merriam-webster.com	Nov - Dec 2007			
…	…	…			
X	HAPAX	Jan - Feb 2008			
End_References					
Maximum_Length	6				
Start_Attestations					
2 Letters	plus Ø	plus A	plus B	…	plus Z
AA	Ø	Ø	Ø	…	Ø
AB	STAB	ABATE-D	ABBEY	…	Ø
…	…	…	…	…	…
AI	CHAIN	NAIAD-M-X	Ø	…	BAIZE-M-X
…	…	…	…	…	…
ZZ	NOZZLE	PIZZA	Ø	…	Ø
End_Attestations					

Legend

Attestation cell content	Meaning
Ø	Indicates no attestation
WORD	Indicates by its presence an attestation. Must be capital letters. Must be < or = Legal_Sequences_Graphemes_Maximum_Length
WORD with no suffix	Attested by competence
WORD with a suffix '–' followed by a letter other than X	The letter is a reference. Must appear in the Reference cells
WORD with a suffix '–X'	The word is a hapax attested by competence
WORD with a suffix '–' followed by a letter other than X followed by '–X'	The letter is a reference and the word is a hapax

Figure 3: Example of our linguistic programming using a spread-sheet table.

What is important is that the operations performed are traceable (see for example Figure 4, here a dynamic trace for machine translation). This advantage enables us during testing to correct errors that have been found very rapidly. Traceability is in any case mandatory in safety-critical domains and this is certainly the case in the aeronautical domain.

5. Global Results

5.1 LiSe Project Schema

The theoretical and architectural schema encompassing the 'LiSe' project sense mining, controlled language and machine translation is shown in Figure 5.

```
RegroupageEnSyntagms_LS =
[['',neg1],['',neg2],[maintenir,vinf],['le
train
d'atterrissage',arg1],[en,prep_v],['positi
on
DOWN',arg2],[comp('',''),comp1],[comp('','
'),comp2],['.',pt]]
LS = frC, LC = anC
LC_GroupeVerbal_LS =
[maintenir,frC_7,'',arg,'',en,pos,'','',''
,'','','',keep,'anC_1.7','',on,prep_v,'','
','','','','']
Regroupage_En_Arguments_LS =
[neg1    -    [],neg2    -    [],vinf    -
[[maintenir,frC_7,keep,'anC_1.7']],arg1  -
([[le,adms,the,ad,'','',''],['train
d'atterrissage',nms,'landing
gear',ns,'','','']]                        -
[[art,n],[art,n],'','','']),prep_v         -
[[en,prep_v,on,prep_v]],arg2               -
([[position,nfs,position,ns,'','',''],['DO
WN',adjs,'DOWN',adj,'','','']]             -
[[n,adj],[adj,n],'','','']),comp1  - ([]  -
[]),comp2    -    ([]    -    []),pt       -
[['.',pt,'.',pt,'','','']]]
Unites_Source = frC
[maintenir  -  frC_7,le  -  adms,'train
d'atterrissage' - nms,en - prep_v,position -
nfs,'DOWN' - adjs,'.' - pt]
Regroupage_En_Arguments_LC =
[neg1    -    [],neg2    -    [],vinf    -
[[maintenir,frC_7,keep,'anC_1.7']],arg1  -
([[le,adms,the,art],['train
d'atterrissage',nms,'landing  gear',n]]  -
[[art,n],[art,n],'','','']),prep_v         -
[[en,prep_v,on,prep_v]],arg2               -
([['DOWN',adjs,'DOWN',adj],[position,nfs,p
osition,n]]                                -
[[n,adj],[adj,n],'','','']),comp1  - ([]  -
_9884308),comp2  -  ([]  -  _9926514),pt  -
[['.',pt,'.',pt,'','','']]]
Unites_Cible = anC
[keep - 'anC_1.7',the - art,'landing gear' -
n,on - prep_v,'DOWN' - adj,position - n,'.'
- pt]
Traduction = 'keep the landing gear on DOWN
position.' ;
```

Figure 4: Dynamic trace for the machine translation of *maintenir le train d'atterrissage en position DOWN* (keep the landing gear on DOWN position) to English.

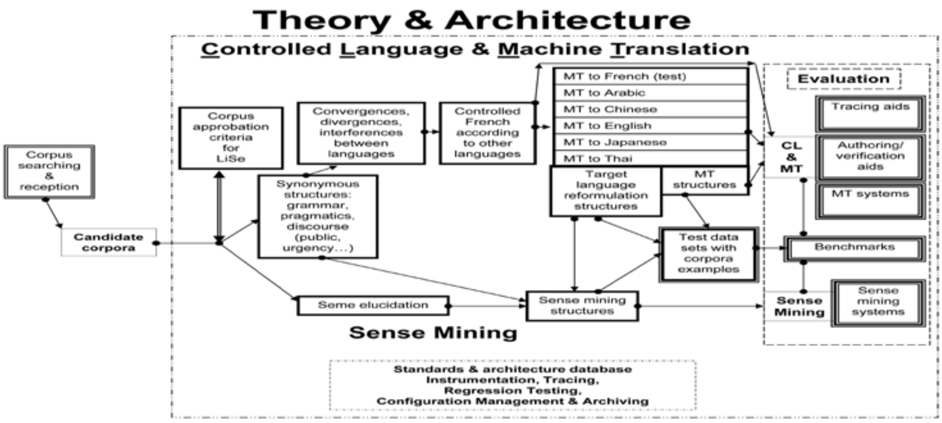


Figure 5: Schema encompassing sense mining, controlled language and machine translation

5.2 Norms

The idea at the outset was to extract and apply norms in order to use them in our methodologies, both for analysis and generation, in the applications of machine translation and sense mining. A sense can effectively be expressed by means of various written or spoken sequences (synonymous structures and lexica).

A controlled language writing guide has been created together with a user interface which latter facilitates entering normalised text. For example there are different ways of expressing an injunction, both in the same language and in different languages (Dziadkiewicz, 2007). In French, if the imperative and the infinitive are frequently used as written injunctive moods, our study of the corpus has enabled us firstly to note that, paradoxically, the passive is often used when indicating some action to be executed, and secondly we found numerous other injunctive constructions, of which we give some examples: *il convient de, il est recommandé, il n'est pas nécessaire de,* etc. However, an injunction in the passive voice often induces confusion with a purely informative type of content, and the other injunctive constructions for which we have given examples above cast doubt on the real need to execute an action, and very often perturb the reading of the text. It is for these reasons that we authorise only the infinitive mood for expressing injunctions.

We then tried to put the norms of each of the languages that we treat in relation with each other and we kept only those which enabled us to obtain the best translations (Cardey et al., 2008).

For example, for the French sentence:

Réduire la vitesse en dessous de 205/.55.

we have the structure:

frC_7

`opt(neg1) + opt(neg2) + vinf + arg1 + prep_v + arg2 + opt(opt(prep_comp), comp1(n)) + opt(opt(prep_comp), comp2(n)) + pt`

and the corresponding sentence in Arabic:

يجب تقليص السرعة تحت 205/.55.

with its structure:

arC_7a

`opt(neg1) + lexis('يجب') +opt(neg2) + nver + arg1(acc) + prep_v + arg2(acc) + opt(opt(prep_comp1), comp1(n)) + opt(opt(prep_comp2),comp2(n)) + pt`

For the French sentence:

Signaler le cas à l'Institut de Veille Sanitaire immédiatement.

we have the same structures as before:

frC_7

`opt(neg1) + opt(neg2) + vinf + arg1 + prep_v + arg2 + opt(opt(prep_comp), comp1(n)) + opt(opt(prep_comp), comp2(n)) + pt`

However for the translation into Arabic:

يجب إعلام مركز المراقبة الصحية بالحالة فورا.

we have a different structure:

*ar*C_7*b*

`opt(neg1) + lexis('يجب') +opt(neg2) + nver + arg2(acc) + prep_v + arg1(acc) + opt(opt(prep_comp1), comp1(n)) + opt(opt(prep_comp2),comp2(n)) + pt`

By means of these two examples, we observe that the same French structure *frC_7* which refers to the two French verbs *réduire* and *signaler* gives two different Arabic structures *arC_7a* and *arC_7b*. If the structure of the first verb *réduire* is nearly identical to that of the Arabic, the same structure *frC_7* of the verb *signaler* requires a permutation of the two arguments arg1 and arg2 in Arabic which gives the Arabic structure *arC_7b* which is totally different to that of the French.

Concerning the application of sense mining, here the norms and the divergences are retained in the same structure because the goal is to find all the different ways to say the same thing.

5.3 Applications and Application Domains

We now present extracts of results from various applications in the domains of aeronautics, medicine, and civil security (Cardey, 2009).

The user interface of the controlled language and machine translation system has 4 parts, which are indicated in Figure 6:

(1) enables making various choices which influence the form of the output
(2) guides the user in entering his or her text
(3) gives explanations as to the choice of rules provided by the user guide
(4) presents the output, and thus the drafted alert or protocol.

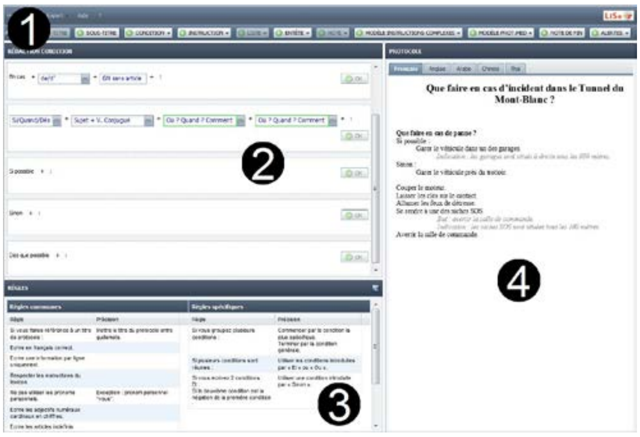


Figure 6: User interface.

In part (4), if one clicks on one of the language buttons, one obtains the translation in the chosen language, as shown in Figure 7.

Figure 7: Translations output.

5.4 Applicability

In order to show the flexibility and scalability of our controlled language and machine translation system, we have tested it adding target languages. Take as example Thai, which is a language which is distant from French, the source language, and which has as a characteristic the presence of classifiers (the function of a classifier is the determination of the type of the noun it qualifies (human, pointed object, a fruit etc.). Normally these classifiers are obligatory but they can be avoided in certain contexts. Thus the Thai structures corresponding to the sentence *Ne pas brancher plusieurs prises* (Do not connect many plugs) are for example:

a) อย่า เสียบ ปลั๊ก หลาย อัน
Neg V N Adj cl. du N
(Ne pas brancher prises plusieurs + classifier for objects in general)

b) อย่า เสียบ ปลั๊ก หลาย ปลั๊ก
Neg V N Adj cl. du N
(Ne pas brancher prises plusieurs + 'prises' as classifier)

c) อย่า เสียบ ปลั๊ก จำนวนมาก
Neg V N Adj
(Ne pas brancher prises plusieurs)

For this case, we control the Thai by choosing c) as the canonical transfer structure because it resembles the French sentence structure the most. The other paraphrases are prohibited. In this manner we have eliminated the classifier variants which can eventually provoke ambiguities, whilst having an exact translation which is not ambiguous, either lexically or syntactically. This methodology has also been applied to Japanese.

In respect of our controlled language machine translation methodology, this has also been applied to Japanese, Russian, Spanish and Turkish. Russian (Jin & Khatseyeva, 2012), Spanish and Turkish (both in 2015) have been added to our controlled language machine translation system as target languages.

In respect of extensions of our methodology to other mission-critical domains, we cite systems for controlled language for business rule specifications (Feuto Njonko et al., 2014) and software requirements specifications (Thongglin et al., 2012).

Within the MESSAGE project (Alert Messages and Protocols) JLS/2007/CIPS/022, our methodology for controlled languages has been the object of standards for its transfer to English, Polish and Spanish, these for a wide diversity of mission-critical domains where the specific target groups concerned were aeronautics, chemistry, civil protection, emergency medical personnel, fire fighting, law enforcement, local government, meteorology and transport (Cardey, Bogacki et al., 2010), MESSAGE project consortium (2010).

6. Conclusion

To conclude, we can say that without the theoretical support provided by micro-systemic linguistic analysis, and the diverse methodologies which all respect the same formal model, it would have been impossible to have obtained such reliable results and, furthermore, to have been able to extend the methodologies and the application domains.

7. Acknowledgements

In respect of funding we wish to acknowledge the Agence Nationale de la Recherche (French National Research Agency): (Projet LiSe Linguistique et Sécurité ANR-06-SECU-007), and the European Commission: (MESSAGE Alert Messages and Protocols project JLS/2007/CIPS/022).

8. Bibliographical References

Cardey S., Anantalapochai R., Beddar M., Cornally T., Devitre D., Greenfield P., Jin G., Mikati Z., Renahy J., Kampeera W., Melian C., Spaggiari L., Vuitton D. (2010) Le projet LiSe, "Linguistique, normes, traitement automatique des langues et sécurité" : du data et sense mining aux langues contrôlées, In *Actes du WISG 2010, Workshop Interdisciplinaire sur la*

Sécurité Globale, Université de Technologie de Troyes, 26 & 27 Janvier 2010, 10 pages.

Cardey S., Bogacki K., Blanco X., Mitkov R. (2010) Resources for Controlled Languages for Alert Messages and Protocols In the European Perspective, In *Proceedings of LREC 2010*, 17-23 May 2010, Valetta, Malta, ISBN 2-9517408-6-7.

Cardey S., Greenfield P., Bioud M., Dziadkiewicz H., Kuroda K., Marcelino I., Melian C., Morgadinho H., Robardet G., Vienney S. (2006) The Classificatim Sense-Mining System. In *Advances in Natural Language Processing*, Springer-Verlag – LNAI 4139, ISBN 3-540-37334-9, 674-684, pp. 674--684.

Cardey, S. (2009) *Proceedings of ISMTCL*, Ed. S. Cardey, Presses universitaires de Franche-Comté, ISSN 0758 6787, ISBN 978-2-84867-261-8.

Cardey, S. (2013) *Modelling Language*, John Benjamins, Amsterdam/Philadelphia, ISBN 9789027249968.

Cardey, S., Devitre, D., Greenfield, P. and. Spaggiari, L. (2009) Recognising Acronyms in the Context of Safety Critical Technical Documentation. In *Proceedings of ISMTCL*, Presses universitaires de Franche-Comté, ISSN 0758 6787, ISBN 978-2-84867-261-8, pp. 56--61.

Cardey, S., Greenfield, P. (2005) A Core Model of Systemic Linguistic Analysis. In *Proceedings of the International Conference RANLP-2005 Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 21-23 September 2005, pp. 134--138.

Cardey, S., Greenfield, P. (2006). Systemic Linguistics with Applications. In Eloína Miyares Bermúdez and Leonel Ruiz Miyares (Eds.), *Linguistics in the Twenty First Century*. Cambridge Scholars Press, United Kingdom, ISBN 1904303862, 2006, pp. 261--271.

Cardey, S., Greenfield, P., Anantalapochai, R., Beddar, M., DeVitre, D., Jin, G. (2008) Modelling of Multiple Target Machine Translation of Controlled Languages Based on Language Norms and Divergences. In *Proceedings of ISUC2008 (Second International Symposium on Universal Communication)*, Osaka, Japan, December 15-16, 2008. Proceedings published by the IEEE Computer Society, ISBN 978-0-7695-3433-6, pp. 322--329.

Dziadkiewicz, A. (2007) *Vers une reconnaissance et une traduction automatique de phraséologismes pragmatiques (application du français vers le polonais)*, thèse de doctorat, Centre Tesnière, Besançon.

Feuto Njonko, P. B., Cardey S., Greenfield P. and El Abed W. (2014) RuleCNL: A Controlled Natural Language for Business Rule Specifications. In: *Proceedings of the 4th International Workshop on Controlled Natural Language (CNL 2014)*, LNCS, vol 8625, Galway, Ireland, August 20-22, ISBN: 978-3-319-10222-1, pp. 66--77.

Jin, G.., Khatseyeva, N. (2012) A Reliable Communication System to Maximize the Communication Quality, In *Proceedings of the 8th International Conference on Natural Language Processing, JapTAL 2012*, Kanazawa, Japan, October 22 - 24, Springer-Verlag Berlin Heidelberg, LNCS/LNAI 7614, Vol 7614, ISBN: 978-3-642-33982-0, pp. 52-–63.

Mikati,I Z. (2009) Data and Sense Mining and their Application to Emergencies and to Safety Critical Domains, In *ISMTCL Proceedings*, International Review BULAG, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8, pp. 179--184.

Thongglin K., Cardey S., Greenfield P. (2012), *Controlled syntax for Thai software requirements specification*, In *Proceedings of the 24th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2012)*, Athens, Greece, November 7-9, pp. 964--969.

9. Language Resource References

MESSAGE project consortium (2010) Resources EU, ELRA-ELDA, LREC2010 Map of Language Resources, Technologies and Evaluation, http://www.resourcebook.eu: Resource Names: “Standards for Controlled Languages”, “Courses on the writing of safe and safely translatable alert messages and protocols”.

A Corpus of German Clinical Reports for ICD and OPS-based Language Modeling

Christina Lohr, Robert Herms
Chair Media Informatics, Technische Universität Chemnitz, Germany
{christina.lohr,robert.herms}@cs.tu-chemnitz.de

Abstract

In the field of health care and corresponding clinical institutions all occurring treatments need to be registered and documented in a comprehensive manner. For clinical documentation, complex reports (e.g., surgical interventions) are dictated by doctors and subsequently typed by secretaries. These reports are annotated with standardized codes for diagnosed diseases (ICD) and executed procedures (OPS). In this paper, we present a corpus of 450 German written clinical reports constructed for evaluation purposes, in particular for language modeling. We investigated the potential of the hierarchical structures of ICD and OPS codes in order to construct content-based language models for the clinical context. Experimental results show that OPS-based language modeling performed best using the highest level of the corresponding standard.

Keywords: medical language processing, medical corpus collection

1. Introduction

In the field of health care and corresponding clinical institutions all occurring treatments need to be registered and documented in a comprehensive manner. The process of clinical documentation includes complex reports which are dictated by doctors and subsequently typed by secretaries, e.g., see (Suominen et al., 2015) and (Herms et al., 2015). These reports have to be annotated with adapted standardized codes concerning diagnosed deseases (ICD: International Statistical Classification of Diseases and Related Health Problems, see (Graubner, 2015a)) and executed procedures (OPS - for Germany: Operationen- und Prozedurenschlüssel, see (Graubner, 2015b)). These standards are administrated by the WHO (World Health Organization) and additionally as part of the Federal Ministry of Health in Germany by the DIMDI (Deutsches Institut für Medizinische Dokumentation und Information).

In this connection, some previous works have been done for the optimization of the documentation process using diverse techniques. (Botsis et al., 2011) describes how English clinical texts are classified by text mining algorithms. The work demonstrates the automatic recognition of vaccine using the MedDRA (The Medical Dictionary for Regulatory Activities). Clinical text contains many synonym words and acronyms. (Zhou et al., 2006) describes how terms of clinical text are tagged by ontologies and the standardized Unified Medial Language System (UMLS). (De Vine et al., 2014) constructed language models based on the OHSUMED corpus (Hersh et al., 1994) which includes English journal abstracts of medical publications. Most publications respectively research concerning clinical language processing is dealing with English language. In (Schulz and López-García, 2015) is shown potential for language processing in the clinical context, especially for German institutions. However, the OPS and ICD standards are still under investigation for automatic clinical language processing.

In this paper, we present a corpus of 450 German written clinical reports constructed for evaluation purposes, in particular for language modeling. These reports are already annotated regarding the codes ICD and OPS. The corpus can be used by the scientific community for natural language processing, e.g., content based language modeling and document clustering. Additionally, we built a generic corpus comprising data of German newspaper articles with clinical background.

We investigated the potential of the hierarchical structures of ICD and OPS codes in order to construct content-based language models for the clinical context. For the experiments, the generic corpus was used in order counteract the generalization problem of the standard specifications.

This paper is organized as follows: In section 2. we report the construction of our corpus as well as the methodology of collecting and processing the data. In section 3. we verify the corpus concerning language modeling and describe the experimental setup and results. Finally, we conclude this paper in section 4. and give some future directions.

2. Corpus Construction

2.1. Reports of Surgical Interventions

In cooperation with the clinical center Klinikum Chemnitz (department Klinik für Allgmein- und Viszeralchirurgie) we collected 450 different German written clinical reports of surgical interventions originated from the years 2008 to 2014.

As a general rule in practice, the reports of surgical interventions contain the following different information: name and date of birth of the patient, the room and the day with time of the surgical intervention, names of doctors and nurses, diagnoses with ICD-codes, procedures with OPS-codes and a procedure description, which is dictated by the leading doctor. We obtained reports as anonymous data without name and date of birth of the patient and without date, time and the room of the intervention.

The clinical center provided the following types of surgical interventions as a structured documentation: "thyroid", "rectectomy", "rectum amputation", "right hemicolectomy", "sigmoid", "stomach", "cholecystectomy" and "pancreas". We partitioned the reports using the OPS-schema into the following topics ("5-42…5-54" with

400 reports are associated to "operations on digestive tract"):

- 50 reports "thyroid" (OPS-code starts with "5-06")
- 60 reports "rectectomy" ("5-484")
- 25 reports "rectum amputation" ("5-485")
- 6 repors with starting OPS-code "5-48", without "5-484" and without "5-485"
- 42 reports "right hemicolectomy" ("5-455.4")
- 44 reports "sigmoid" ("5-455.7")
- 25 reports starting OPS-code "5-45", without "5-455.4" or "5-455.7"
- 50 reports "stomach" ("5-43")
- 70 reports "cholecystectomy" ("5-511")
- 51 reports "pancreas" ("5-52")
- 27 reports cannot be classify in the mentioned OPS-code but ordered into the OPS-chapter "5-42. . . 5-54".

For further investigation, we processed these reports as follows: First, we segmented the textual data into separate sentences using the corresponding punctuation. Next, acronyms were detected using a collected list derived from (Dräger, 2006). As there are different notations of terms we resolved acronyms, e.g., "V." to "Vena". Moreover, there are different notations with the same content, e.g., we resolved "Colon" to "Kolon". There were many typographical errors that have been fixed by hand, e.g "Blutnugen" instead of "Blutungen", which means "bleeding" in English. Medical terms that contain letters as well as numbers (e.g., "R1") were splitted and all numbers and dates were transformed into words. Since punctuations are typically dictated in the medical domain, all punctuation marks were transformed into words, e.g., "." into "punkt".
The corpus contains 22,427 documents, 266,390 tokens and 11,008 types, see Table 1 and Table 2. For evaluation purposes, we partitioned the 450 reports into the following datasets with a balanced distribution of the OPS-codes for each set:

- training: 225 reports
- developing: 113 reports
- testing: 112 reports

Table 1: Corpora of medical reports of 450 surgical interventions – storage, tokens, documents and types.

corpus	storage	tokens	documents	types
training	1.0 MB	131.8 K	11.1 K	7.9 K
development	0.5 MB	69.8 K	5.9 K	5.9 K
evaluation	0.5 MB	64.8 K	5.4 K	5.9 K
full corpus	2.1 MB	266.4 K	22. 4 K	11.0 K

Table 2: Corpora of medical reports of 450 surgical interventions – n-grams (n=2,3,4,5).

corpus	n=2	n=3	n=4	n=5
training	37.2 K	62.0 K	72.3 K	73.3 K
development	24.7 K	38.5 K	43.0 K	42.6 K
evaluation	24.3 K	37.6 K	41.8 K	41.3 K
full corpus	59.6 K	107.1 K	129.8 K	134.7K

2.2. A Generic Corpus for Clinical Purposes

The DWDS (Digitales Wörterbuch der Deutschen Sprache) provides an interface for the retrieval of articles (Didakowski and Geyken, 2014). The corpus contains a collection of German newspapers ("Berliner Zeitung", "Der Tagesspiegel', "Potsdamer Neueste Nachrichten" (PNN) and "DIE ZEIT") and books ("Kernkorpus 20" from 20th century (KK 20) and "Kernkorpus 21" from 21th century (KK 21)). We composed as set of 400 medical terms, for example "Ambulanz", "Operation" and "Patient", and downloaded text with three sentences around these terms. (The 400 used terms in this work are available on request.) We processed the retrieved textual data in the same way as the reports: we segmented the textual data into separate sentences. Acronyms and different notations with the same content were resolved. Moreover, typographical errors have been fixed and sentences were deleted. The corpus has a size of 809 MB, 125,913,596 tokens, 1,697,868 types and 5,756,010 documents, see Table 3 and Table 4.

Table 3: A generic corpus for clinical purposes from DWDS – storage, tokens, documents and types.

corpus	storage	tokens	doc.	types
Berl. Zeitung	202 MB	1.6 M	31.3 M	0.7 M
DIE ZEIT	248 MB	38.2 M	1.9 M	0.8 M
PNN	33 MB	5.2 M	0.2 M	0.2 M
Tagesspiegel	175 MB	27.2 M	1.3 M	0.6 M
KK 20	145 MB	24.4 M	0.7 M	0.7 M
KK 21	2 MB	0.4 M	0.02 M	0.03 M
full corpus	809 MB	125.9 M	5.8 M	1.7 M

Table 4: A generic corpus for clinical purposes from DWDS– n-grams (n=2,3,4,5).

corpus	n=2	n=3	n=4	n=5
Berl. Zeitung	7.1 M	17.7 M	24.6 M	26.4 M
DIE ZEIT	8.7 M	21.8 M	30.4 M	32.5 M
PNN	1.7 M	3.5 M	4.3 M	4.4 M
Tagesspiegel	0.6 M	6.4 M	15.7 M	21.6 M
KK 20	6.2 M	14.8 M	20.0 M	21.1 M
KK 21	0.2 M	0.3 M	0.3 M	0.3 M
full corpus	21.8 M	61.7 M	93.2 M	104.5 M

3. Language Model Experiments

The goal of language model experiments is to find the optimal annotation code for building content based language models. There is searched a configuration for language models which perplexity values has the smallest average and standard derivation. Another goal is to find interpolation weights for the models of DWDS and the clinical reports.

3.1. Experimental Setup

Our approach for language modeling is based on the assumption that each type of code has its own hierarchical level. Some codes of OPS and ICD can be summarized: “5-455.4” and “5-455.7” into “5-455”; “5-484” and “5-485” into “5-48”. The structure of both codes is like a tree, see Figures 1 and 2.

Figure 1: Hierarchical levels of ICD-codes for diseases

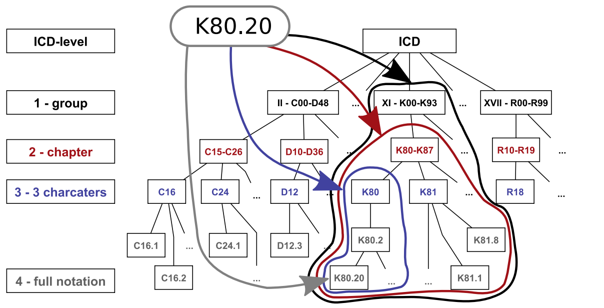
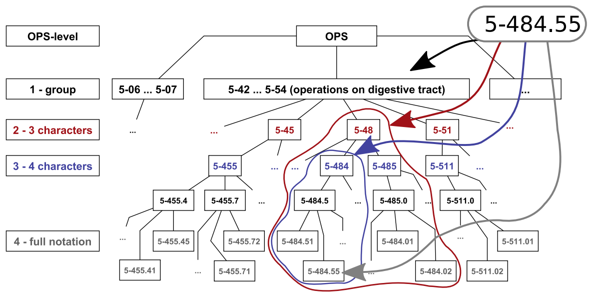


Figure 2: Hierarchical levels of OPS-codes for procedures and operations



The ICD-code “K80.20” is classified into “Chapter XI Diseases of the digestive system” (level 1), “Disorders of gallbladder, biliary tract and pancreas (K80-K87)” (level 2), “K80.2” “Calculus of gallbladder without cholecystitis” (level 3) and “0” defines “unspecified or without cholecystitis” (level 4). The OPS-code for all operations is “5”, “5-42…5-54” defines “operations on digestive tract” (level 1), “5-51” “operations on gallbladder und biliary tract” (level 2), “5-511” “cholecystectomy’ (level 3) and “.11” “without laparoscopic revision of the bile ducts” (level 4).
There are four ways creating text for language models from annotated reports from one of the coding systems. At each level of these codes there is a different count of text. There are eight ways for building language models. For every hierarchical level we took the code definitions and built text from the training dataset. We used the toolkit SRILM (Stolcke and others, 2002) for estimating and evaluating language models with 3-grams.

3.2. Results

The perplexity of training data was for the development set 26.2 and for the test set 29.4. The perplexity of the DWDS-corpus arose for the development set 1426.9 and for test set 1333.0.
Different content based language models were built only by the corpus of clinical reports using the annotation of OPS and ICD. The lowest average perplexity μ=26.4 with standard deviation σ=7.0 for the development set and perplexity μ=28.0 with σ=6.9 for the test set evolved from the value of the first OPS-level, see Table 5. This level is the definition of the OPS-chapter where the most of adapted text is collected.

Table 5: Perplexities of language models built by anotations of codes OPS and ICD without a background model. The best results are highlighted in bold.

	development data		test data	
	perplexity		perplexity	
level of anotation	μ	σ	μ	σ
OPS level 1	**26.4**	7.0	**28.0**	6.9
OPS level 2	29.4	6.2	39.9	11.6
OPS level 3	29.2	11.4	38.5	12.9
OPS level 4	32.7	15.0	44.4	16.1
ICD level 1	31.6	13.1	38.4	14.1
ICD level 2	35.2	15.5	38.6	14.0
ICD level 3	35.7	16.8	41.6	15.5
ICD level 4	40.9	16.1	41.8	15.9

We built content based language models by OPS and ICD with the DWDS-model. The goal was to find an optimal interpolation weight for the language models. We used the interface `compute-best-mix` of the toolkit SRILM for estimating optimal weights for language models of the different levels of OPS and ICD. We took the average weight λ_{dwds} of every hierarchy of the codes and estimated the perplexity from development set.

Table 6: Perplexities of language models built by anotations of codes OPS and ICD with a background model by DWDS data. The best results are highlighted in bold.

		development data		test data	
		perplexity		perplexity	
level	λ_{dwds}	μ	σ	μ	σ
OPS 1	0.07	**31.9**	7.8	**34.7**	6.9
OPS 2	0.10	40.6	9.4	61.9	40.0
OPS 3	0.16	135.8	494.6	106.0	148.8
OPS 4	0.22	138.7	222.9	158.7	180.1
ICD 1	0.11	73.4	122.1	124.2	181.9
ICD 2	0.14	71.9	87.5	112.9	134.8
ICD 3	0.22	124.6	190.4	121.3	122.8
ICD 4	0.30	171.9	168.3	150.9	119.8

The best average perplexity arose μ=31.9 with standard deviation σ=7.8 for development set with the interpolation weight $\lambda_{dwds} = 0.07$ for the background model. The values arose on the first level of OPS-code, which is the definition for the OPS-chapter and the most text is collected.

For test data the best average perplexity μ=34.7 with standard deviation σ=6.9 arose on the first level of OPS-code. It may be that the background model and the content based models from the both codes are not good enough and the corpus of clinical reports is too small, because the simplest experiment with the full training data and the development set has shown the smallest perplexity (development data: μ=26.22, test data: μ=29.4) and the experiments with the DWDS-corpus were not better than those without.

4. Summary and Outlook

In this paper, we presented a corpus of 450 German written clinical reports. These reports were already annotated regarding the standards ICD and OPS. The motivation for the corpus development were evaluation purposes, in particular for language modeling as well as document clustering in the medical domain. Additionally, we built a generic corpus comprising data of German newspaper articles with clinical background. We described the procedure of collecting and developing both corpora. Furthermore, we investigated the potential of the hierarchical structures of ICD and OPS codes in order to construct content-based language models for the clinical context. Experimental results show that OPS-based language modeling performed best using the highest level of the corresponding standard.

For the future, we try to extend the corpus of clinical reports by a higher range of ICD and OPS codes to better reflect the real world scenario. Our goal is to apply the most appropriate language models for automatic speech recognition to further boost established systems, such as (Herms et al., 2015). Moreover, the automatic assignment of ICD and OPS standards to reports using Text Mining and classification algorithms could support the comprehensive clinical workflow.

5. Bibliographical References

Botsis, T., Nguyen, M. D., Woo, E. J., Markatou, M., and Ball, R. (2011). Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *Journal of the American Medical Informatics Association*, 18(5):631–638.

De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., and Bruza, P. (2014). Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1819–1822. ACM.

Didakowski, J. and Geyken, A. (2014). From DWDS corpora to a German Word Profile – methodological problems and solutions. *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, 417:39–42.

Dräger, H. (2006). *Medizinische Abkürzungen*. Thieme.

Graubner, B. (2015a). *ICD-10-GM 2015 Alphabetisches Verzeichnis: Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme*. Deutscher Ärzte-Verlag, Köln, 1. edition.

Graubner, B. (2015b). *OPS 2015 Alphabetisches Verzeichnis: Operationen- und Prozedurenschlüssel – Internationale Klassifikation der Prozeduren in der Medizin*. Deutscher Ärzte-Verlag, Köln, 1. edition.

Herms, R., Richter, D., Eibl, M., and Ritter, M. (2015). Unsupervised Language Model Adaptation using Utterance-based Web Search for Clinical Speech Recognition. In *Conference and Labs of the Evaluation Forum (CLEF)*.

Hersh, W., Buckley, C., Leone, T., and Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*, pages 192–201. Springer.

Schulz, S. and López-García, P. (2015). Big Data, medizinische Sprache und biomedizinische Ordnungssysteme. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz*, pages 1–9.

Stolcke, A. et al. (2002). SRILM – an extensible language modeling toolkit. In *INTERSPEECH*, volume 2002, page 2002.

Suominen, H., Johnson, M., Zhou, L., Sanchez, P., Sirel, R., Basilakis, J., Hanlen, L., Estival, D., Dawson, L., and Kelly, B. (2015). Capturing patient information at nursing shift changes: methodological evaluation of speech recognition and information extraction. *Journal of the American Medical Informatics Association*, 22(e1):e48–e66.

Zhou, X., Han, H., Chankai, I., Prestrud, A., and Brooks, A. (2006). Approaches to text mining for clinical medical records. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 235–239. ACM.

ProphetMT: Controlled Language Authoring Aid System Description

Xiaofeng Wu, Liangyou Li, Jinhua Du, Andy Way
ADAPT Centre, School of Computing,
Dublin City University, Ireland
xiaofengwu,liangyouli,jdu,away@computing.dcu.ie

Abstract

This paper presents ProphetMT, a monolingual Controlled Language (CL) authoring tool which allows users to easily compose an in-domain sentence with the help of tree-based SMT-driven auto-suggestions. The interface also visualizes target-language sentences as they are built by the SMT system. When the user is finished composing, the final translation(s) are generated by a tree-based SMT system using the text and structural information provided by the user. With this domain-specific controlled language, ProphetMT will produce highly reliable translations. The contributions of this work are: 1) we develop a user-friendly auto-completion-based editor which guarantees that the vocabulary and grammar chosen by a user are compatible with a tree-based SMT model; 2) by applying a shift-reduce-like parsing feature, this editor allows users to write from left-to-right and generates the parsing results on the fly. Accordingly, with this in-domain composing restriction as well as the gold-standard parsing result, a highly reliable translation can be generated.

Keywords: Controlled Language, Authoring Tool, Statistical Machine Translation

1. Introduction

Although current machine translation (MT) methods have improved rapidly in the past decade, SMT is still not reliable enough to be considered human-quality without significant post-editing (O'Brien, 2005). The primary reason is that natural languages are full of ambiguities.

A Controlled Language (CL) is widely used in professional authoring where the aim is to write for a certain standard and style demanded by a particular profession, such as law, medicine, patent, technique etc (Gough and Way, 2004; Gough and Way, 2003). For multilingual documents, CL has been shown to improve the quality of the translation output, whether the translation is done by humans or machines (Nyberg et al., 2003).

The advantages of applying CL are self-evident: clear and consistent composition guidelines as well as less ambiguity in translation. However, the problems are also obvious: design of the rules usually requires human linguists, and rules may be difficult for end-users to grasp. In addition, the sentences that can be generated are often limited in length and complexity (O'Brien, 2003).

This paper presents ProphetMT,[1] a tree-based SMT-driven CL authoring tool. ProphetMT employs the source-side rules in a translation model and provides them as auto-suggestions to users. Accordingly, one might say that users are writing in a 'Controlled Language' that is 'understood' by the computer.

2. Related Work

All existing computer-aided authoring tools within a translation context employ a kind of interactive paradigm with a CL. Mitamura (1999) allow users to compose from scratch, and discuss the issues in designing a CL for rule based machine translation. Power et al. (2003) describes a CL authoring tool for multilingual generation. Marti et al. (2010) present a rule-based rewriting tool which performs syntactic analysis. Mirkin et al. (2013) introduce a confidence-driven rewriting tool which is inspired by Callison-Burch et al. (2006) and Du et al. (2010) that paraphrases the out-of-vocabulary words (OOV) or the "hard-to-translate-part" of the source side in order to improve SMT performance.

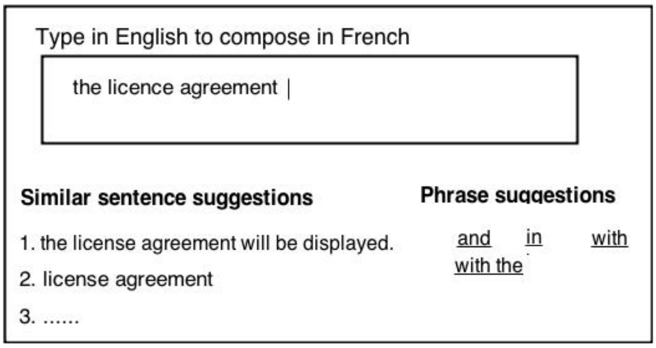


Figure 1: SMT-driven Authoring Tool by (Venkatapathy and Mirkin, 2012)

To our knowledge, Venkatapathy and Mirkin (2012) is the first interface that could be called an SMT-driven CL authoring tool: the main interface screen shot is shown in Figure 1. Their tool provides users with the word, phrase, even sentence level auto-suggestions which are obtained from an existing translation model. Nevertheless, it lacks syntactically-informed suggestion and constraints.

Sentences in all languages contain recursive structure. Synchronous context-free grammars (SCFG) (Chiang, 2005) and stochastic inversion transduction grammars (ITG) (Wu, 1997) have been widely used in SMT and achieve impressive performance. However, MT systems which make use of SCFG tend to generate an enormous phrase table containing many erroneous rules. This huge search space not only leads to an unreliable output, but also restricts the input sentence length that the system can handle. Other tree-based SMT models like Liu et al. (2006) and shen2008new depend heavily on the accuracy of the parsing algorithm

[1]ProphetMT has been granted Enterprise Ireland Feasibility Study Funding 2016.

which introduces noise upstream to the MT system.
Our method, ProphetMT, allows monolingual users to easily and naturally write correct in-domain sentences while also providing the structural metadata needed to make the parsing of the sentence unambiguous. The set of structural templates is provided by the tree-based MT system itself, meaning that highly reliable MT results can be generated directly from the user's composition.
Syntactic annotation is a tedious task which has traditionally required specialised training. In order to maintain a natural and easy writing style, ProphetMT makes use of auto-suggestion both for syntactic templates and for terms. A shift-reduce like (Aho, 2003) authoring interface, which allows users to easily parse the "already composed part" of the sentence, is also applied to maintain the structural correctness and unambiguous parsing while the source sentence is being composed.

3. ProphetMT: Syntactical SMT-Driven Authoring

3.1. An Overview of ProphetMT

ProphetMT is a client-server application. There are three main components involved:

1. A website client provides a structural writing user interface.
2. A web-service provides source-language rule/term auto-completion.
3. A web-service provides hierarchical phrase-based machine translation.

The main interface is shown in Figure 2. The 4 areas are:

1. the input area (upper)
2. the source tree structural area (middle left)
3. the target tree structural area (middle right)
4. the composed sentence and the translation (bottom)

The behavior of the ProphetMT can be defined by Algorithm 1, explained below are some terminology:

- NodeBox: the recursive (nestable) editing unit
- Non-Terminal Rule (NTR): rules like "*X is X*", "*one of X*", "*has X with X*" which have variables
- Non-Terminal(NT): the "*X*" in the NTR
- Terminal Rule(TR): rules without NT

While the user is inputting text, both TR auto completion and NTR auto completion are provided. Autocompletion candidates are automatically selected from the normal tree-based MT model according to the guidance introduced in Section 4.. When the user finishes composing the sentence, the result is sent to a tree-based MT engine, and the target translation(s) are generated according to the source- side rules decided by the user.

With the following example, we further explain Algorithm 1 in detail.

3.2. An Example

Suppose the user wants to input the sentence "*Australia is one of the few countries that have diplomatic relationships with North Korea*", which is shown in Figure 3 together with the NodeBox numbers.

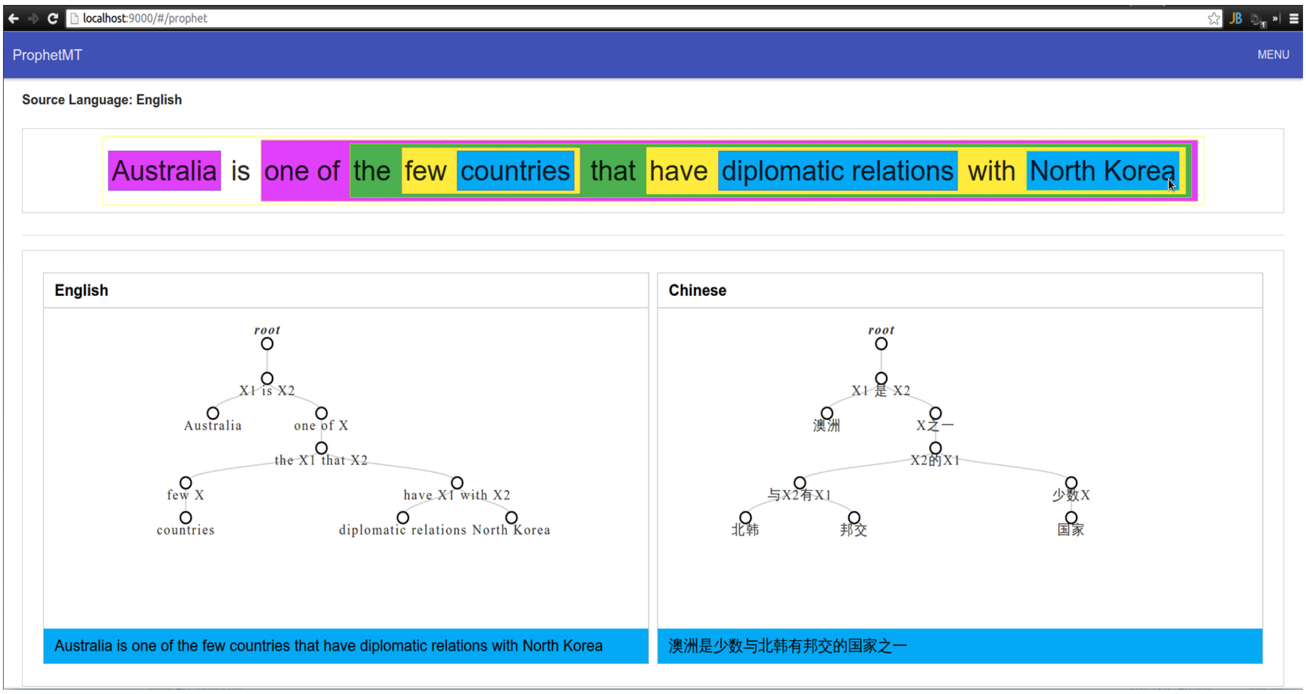


Figure 2: ProphetMT Main Interface Screen Shot

```
Initialize: ProphetMT opens an empty NodeBox;
while User is typing in an NodeBox do
    Provide TR auto suggestions;
    if There is a left adjacent NodeBox then
        Provide all NTR suggestions;
    else
        Provide the NTRs which DO NOT have NT at the
          beginning position;
    end
    if User selects "translate" then
        Finish the source and target parse trees;
        Translate and output the results;
        Go to stop;
    end
    if User Chooses an NTR then
        Generate the according NodeBoxs;
        if The current selected NTR has an NT at the
          beginning position then
            The corresponding NodeBox will
              automatically merge with the left adjacent
              NodeBox ;
        end
        Focus goes to the first NodeBox which is empty;
        Continue;
    end
    if User starts a new NodeBox then
        Stop the current NodeBox editing;
        Focus goes to the new NodeBox;
        Continue;
    end
end
Stop:
```

Algorithm 1: ProphetMT Main Workflow

The following steps will be performed:

1. ProphetMT starts a new NodeBox1 andthe user types in "*Australia*"; NodeBox1 is finished.
2. User starts a new NodeBox0 to the right of the NodeBox1 and types in "*is*"; The user chooses an NTR "*X is X*" , then two new NodeBoxes will be generated within NodeBox0 by "*NodeBox is NodeBox*";The left NodeBox within NodeBox0 will automatically **merge** with the NodeBox1 which is left adjacent to NodeBox0; NodeBox0 is finished.
3. The user selects the second NodeBox within NodeBox0, which is NodeBox2, and types in "*one of*" and selects the NTR "*one of X*". NodeBox2 is finished.
4. In the generated NodeBox3, th user types in "*the*" and chooses "*the X that X*". The NodeBox3 is finished and two new NodeBoxes are generated as NodeBox4 and NodeBox6
5. In NodeBox4, the user types in "*few*" and chooses "*few X*". NodeBox4 is finished, NodeBox5 is generated.
6. In NodeBox5 the user types in "*countries*". NodeBox5 is finished
7. In NodeBox6, the user types in "*have*" and chooses NTR "*have X with X*". NodeBox7 is finished, two new NodeBoxes, NodeBox7 and NodeBox8 are generated.
8. In NodeBox7, the user types in "*diplomatic relations*". NodeBox7 is finished.
9. In NodeBox8, the user types in "*North Korea*". NodeBox8 is finished.
10. The user finishes editing the sentence and then the translation for the specific languages as well as the parsing trees are generated.

3.3. Merging

The merging process which happens in step 2 is shown in Figure 4. This merging process allows the user to compose the sentence from left-to-right while keeping the partially parsed structure intact.

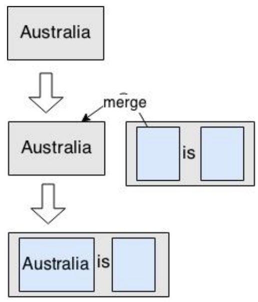


Figure 4: The Merging Process

3.4. NodeBox Starting Points Selection

Figure 5 further illustrates how the user starts a new NodeBox and how ProphetMT maintains the syntactic structure by adopting a shift-reduce-like strategy. Suppose the user has written "Australia ... and China ...". Figure 5a shows the current state in the input area and the arrows "A" and "B" are the possible inserting point options. Figure 5b shows the corresponding partially parsed tree shown in the source parsing area which also indicates the two insertion positions. Figure 5c shows the parsing area when the user wants to further describe China and chooses the rule "*X which is X*" in position "A". Because there is a left-adjacent NodeBox and there is a NodeBox at the start position of the selected rule, so a merging process takes place. Figure 5d

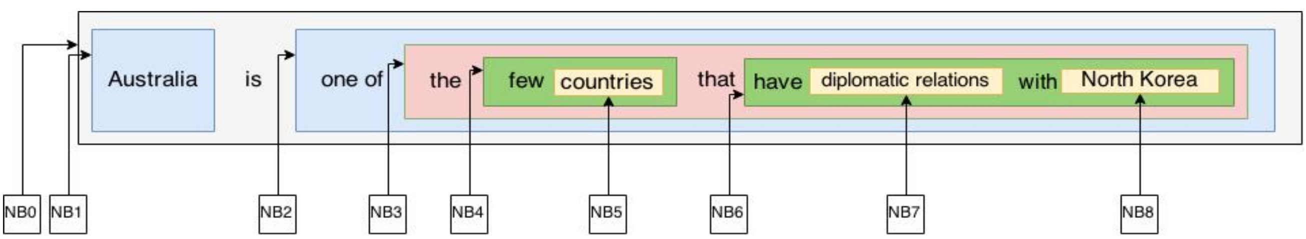


Figure 3: Example of Using ProphetMT Together With The NodeBox Numbers

shows the parsing area when the user wants to keep "Australia .. and China .." as a unit and chooses the rule "*X are X*" at position "B". As shown, a similar merging process happens.
We can see that this shift-reduce strategy allows composition to proceed from left-to-right while at the same time maintaining a correct parse of the existing text.

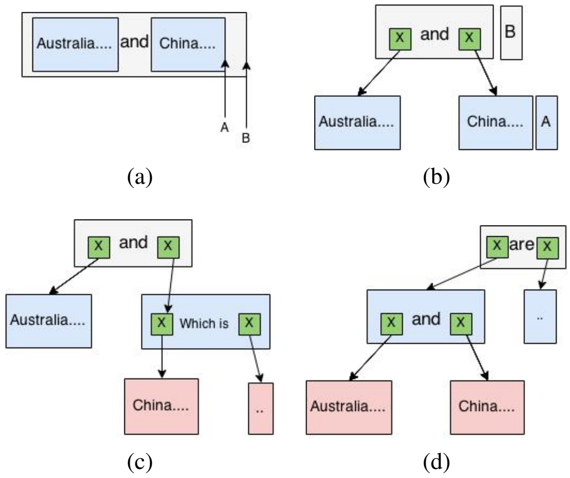


Figure 5: NodeBox Starting Points Selection

4. Auto-suggestions

In this section we introduce the auto-suggestion web-service employed in ProphetMT, including the *phrase level*, *rule level*, and *paraphrase* suggestion engines.

4.1. Terminal Rule (Phrase) Auto-suggestions

Following the work of Venkatapathy and Mirkin (2012), the phrase level auto-suggestions are also guided by three factors:

- **Fluency**: What the user has input will influence the incoming auto-suggestion. We use the SRILM (Stolcke and others, 2002) toolkit to rescore the phrase auto-completion.
- **Translatability**: The phrase pairs in the phrase table (i.e. the SMT model) are be sorted according to the four translation possibility features.
- **Semantic Distance**: The semantic distance of the suggested phrases must be close to the already composed part.

The final rank of the proposed phrases is based on the minimization of the Semantic Distance and maximization of the Fluency and Translatability.

4.2. Non-Terminal Rule (NTR) Auto-suggestions

In order to extract NTRs which are meaningful to humans, we parse the source side of the training corpus with the Berkeley Parser[2]. Then we wrap the parsed result with xml for Moses[3] (Koehn et al., 2007) hierarchical phrase-based model to extract rules. The ranking of NTR suggestions follows the same methodology that was employed for phrase suggestions.

4.3. Filtering

To further reduce the size, the NTRs containing content words like nouns, pronouns and numbers can be removed. This filtering is based on the observation that the structure of a sentence is primarily dictated by function words as well as verbs. The phrase-level auto-suggestions are responsible for providing the content words that fill the leaf nodes in the hierarchical templates. Because most NTRs will be discarded, and because the source side is already parsed when fed to the decoder, the normal restrictions of tree-based models, such as the maximum span (which is $<= 20$) and the NT numbers (which is usually $<= 2$), can be removed.

4.4. Paraphrase Auto-suggestions

If the user inputs an OOV, a paraphrase engine will be queried to try to suggest terms within the current SMT model. Paraphrases are obtained from PPDB.[4] If the OOV is not found in PPDB, than the user will be forced to choose another word.

5. Tree-based SMT

When the user finishes composing the sentence and commands ProphetMT to translate, a tree-based SMT engine will take the sentence and the user parsing results as input to generate the final target language selected by the user.

6. Preliminary Evaluation

Currently we have the UI implementation and the backend server. In this section we provide a preliminary evaluation. The following example was downloaded from Moses.[5] We use the model in the tree-to-tree folder and replace the non-terminal tag with "X" in order to be compatible with Moses hierarchical phrase-based (HPB) model. The whole model is shown below.

```
overhead [X] ||| toudingshangde [X]
overhead [X][X] [X] ||| toudingshangde [X][X] [X]
masks [X] ||| mianzhao [X]
oxygen [X] ||| yangqi [X]
[X][X]1 in the [X][X]2 [X] ||| [X][X]2 de [X][X]1 [X]
cabin section [X] ||| chuancang qu [X]
section [X] ||| qu [X]
[X][X] had [X][X] [X] ||| [X][X] yi [X][X] [X]
had [X] ||| yi [X]
dropped into place [X] ||| hualuo [X]
. [X] |||  . [X]
```

Given the following input sentence

```
overhead oxygen masks in the cabin section
had dropped into place .
```

the Chinese translation by HPB is

```
toudingshangde yangqi chuancang qu de
mianzhao yi hualuo .
```

[2]https://code.google.com/p/berkeleyparser/
[3]http://www.statmt.org/moses/
[4]http://www.cis.upenn.edu/ ccb/ppdb/
[5]http://www.statmt.org/moses/download/sample-models.tgz

It is wrong ordered which leads to a totally wrong meaning. However, were a native Chinese speaker to read the translation, it is very hard to see if the sentence is correct or not. This is due to the fact that the left NT of the rule "X in the X" wrongly covers only "masks", not the "over head oxygen masks". This is a very common mistake in the HPB model.

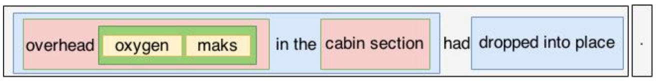


Figure 6: Authored With ProphetMT

However, with the following input generated by using ProphetMT, shown in Figure 6 by an native English speaker without any training, the correct translation can be generated using the same HPB model as:

```
chuancangqu de toudingshangde yangqi
mianzhao yi hualuo .
```

7. Discussion and Future Work

In this paper we describe ProphetMT, which is, to the best of our knowledge, the first syntactic SMT-driven CL authoring tool. ProphetMT provides both in-domain phrase-based and syntactic suggestions, which are compatible with an existing tree-based SMT system, to the user via auto-completion. By employing a novel shift-reduce-like scheme, users can naturally write from left-to-right while also parsing the composed parts and visualising the parsing results. With the help of the NodeBox component, the syntactic structure is rendered to the user in a simple format which does not require linguistic expertise to understand. Using the gold standard parsing results from the interface, a highly reliable SMT output can be generated which will reduce the post-editing efforts required later in the translation pipeline.

We hypothesise that ProphetMT will be suitable for monolingual CL writers of technical reports and manuals, patents, as well as contracts which are domain-specific and intented to be translated downstream. Furthermore, the tool is useful for more than just SMT — ProphetMT can also be helpful in maintaining consistency in writing styles across organizations, and in training beginners.

Future Work: In order to provide efficient auto-suggestions, we want NTRs which are relatively short, and "meaningful" to humans. The following improvements to the existing system can be implemented:

- Filtering the rules with syntactic parsing results as shown in Li et al. (2012) would be another way to prune the hierarchical phrase table.
- Dependency tree-to-string SMT model (Xie et al., 2011) as well as constituency tree-based model (Liu et al., 2006; Zhang et al., 2007) could also be interesting approaches for filtering out ungrammatical rules.
- Manual filtering may need to be performed for some domains, especially when ambiguity is costly.

The final evaluation of ProphetMT involves two different criteria:

- Measuring the composition time or the average edits as reported in the interactive machine translation tools (Foster et al., 2002; Alabau et al., 2014).
- Measuring the final translation quality relative to unaided composition.

The evaluation can be conducted by allowing human native speakers using ProphetMT to paraphrase test data.

8. References

Aho, A. V. (2003). *Compilers: Principles, Techniques and Tools (for Anna University), 2/e*. Pearson Education India.

Alabau, V., Buck, C., Carl, M., Casacuberta, F., Garcıa-Martınez, M., Germann, U., González-Rubio, J., Hill, R., Koehn, P., Leiva, L., et al. (2014). Casmacat: A computer-assisted translation workbench. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 25–28.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24. Association for Computational Linguistics.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

Du, J., Jiang, J., and Way, A. (2010). Facilitating translation using source language paraphrase lattices. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 420–429. Association for Computational Linguistics.

Foster, G., Langlais, P., and Lapalme, G. (2002). User-friendly text prediction for translators. In *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, jul. TT2 TransType2.

Gough, N. and Way, A. (2003). Controlled generation in example-based machine translation. In *MT Summit IX*. New Orleans, LO.

Gough, N. and Way, A. (2004). Example-based controlled translation. In *In Proceedings of the Ninth Workshop of the European Association for Machine Translation*. EAMT, Valetta, Malta.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Li, J., Tu, Z., Zhou, G., and van Genabith, J. (2012). Using syntactic head information in hierarchical phrase-based translation. In *Proceedings of the Seventh Workshop on*

Statistical Machine Translation, pages 232–242. Association for Computational Linguistics.

Liu, Y., Liu, Q., and Lin, S. (2006). Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 609–616. Association for Computational Linguistics.

Marti, J., Ahs, D., Lee, B., Falkena, J., Nelson, J., Kohlmeier, B., Liger, F., Pamarthi, R., Lerum, C., Mody, V., et al. (2010). User interface for machine aided authoring and translation, May 4. US Patent 7,711,546.

Mirkin, S., Venkatapathy, S., Dymetman, M., and Calapodescu, I. (2013). Sort: An interactive source-rewriting tool for improved translation. In *ACL (Conference System Demonstrations)*, pages 85–90. Citeseer.

Mitamura, T. (1999). Controlled language for multilingual machine translation. In *Proceedings of Machine Translation Summit VII, Singapore*, pages 46–52.

Nyberg, E., Mitamura, T., and Huijsen, W.-O. (2003). for authoring and translation. *Computers and Translation: A Translator's Guide*, 35:245.

O'Brien, S. (2003). Controlling controlled english. an analysis of several controlled language rule sets. *Proceedings of EAMT-CLAW*, 3:105–114.

O'Brien, S. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation*, 19(1):37–58.

Power, R., Scott, D., and Hartley, A. (2003). Multilingual generation of controlled languages.

Stolcke, A. et al. (2002). Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.

Venkatapathy, S. and Mirkin, S. (2012). An smt-driven authoring tool. In *COLING (Demos)*, pages 459–466. Citeseer.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, 23(3):377–403.

Xie, J., Mi, H., and Liu, Q. (2011). A novel dependency-to-string model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 216–226. Association for Computational Linguistics.

Zhang, M., Jiang, H., Aw, A., Sun, J., Li, S., and Tan, C. L. (2007). A tree-to-tree alignment-based model for statistical machine translation. *MT-Summit-07*, pages 535–542.

Evaluating and Implementing a Controlled Language Checker

Rei Miyata[1], **Anthony Hartley**[2], **Cécile Paris**[3], **Kyo Kageura**[1]
[1]The University of Tokyo, [2]Rikkyo University, [3]CSIRO
{rei, kyo}@p.u-tokyo.ac.jp, a.hartley@rikkyo.ac.jp, Cecile.Paris@csiro.au

Abstract

This paper describes the evaluation of a detection component of a controlled language (CL) checker designed to assist non-professional writers in creating Japanese source texts that conform to a set of writing rules. We selected 23 Japanese CL rules shown to be effective for two Japanese to English machine translation systems as well as human readability, and implemented them with simple pattern matching using an existing morphological analyser. To benchmark the performance of the component, we created a comprehensive test set of non-compliant and compliant sentences from Japanese municipal websites, with all rule violations manually annotated. The results showed that 15 rules achieved high F-measure scores (more than 0.8) with nine obtaining the score of 1.0, while the precision scores of eight rules are low (less than 0.7). A detailed analysis of the results indicated ways to improve performance. Finally, based on the evaluation, we created an interface designed to alleviate the low-precision issue and implemented a prototype CL checker that is operational online.

Keywords: controlled language checker, machine translation, evaluation

1. Introduction

Recent years have witnessed the increased public use of machine translation (MT) in Japan. In particular, Japanese municipalities typically rely solely on MT to disseminate their website content in multiple languages, given their limited budgets for hiring translators or post-editors. However, MT between languages with as greatly differing structures as Japanese and English is still of generally poor quality.

To help with this situation, we are developing a Web-based controlled authoring system, called *MuTUAL*, designed to support non-professional writers in Japanese municipal departments in creating multilingual municipal documents using MT systems.[1] MuTUAL consists of a suite of modules for document structuring, controlled writing and multilingualisation.

The core module for controlled writing is the controlled language (CL) checker. Although our experimental results from human evaluation showed that Japanese CL rules tuned to particular MT systems can achieve more than 15% increase in cases where an MT output is judged as both understandable and accurate (Miyata et al., 2015), the feasibility of application of the rules by writers, especially non-professional writers, still needs to be examined. Thus, providing writing support tools and assessing their effectiveness are crucially important from the point of view of their deployment.

In this study, as a starting point, we implemented a function to detect violations of Japanese CL rules which had previously been shown to be effective for source readability and for two Japanese-English MT systems. We also gauged performance using test data extracted from Japanese municipal website texts. Based on these results, we designed a Japanese CL checker and implemented a prototype system.

The remainder of this paper is structured as follows. In Section 2, we give an overview of CLs and writing support methods, referring to related studies and existing tools. Section 3 explains CL rules which we included in our CL violation detector, our first step towards a comprehensive CL checking environment. Section 4 presents the results of a pilot evaluation of the CL violation detection function and discusses ways of improving its performance. Section 5 describes our Japanese CL checker, showing the interface of the prototype system, while Section 6 sketches out our future plans towards a practical implementation of the tool.

2. Controlled language and writing support

A controlled language (CL) is a constructed language with restrictions on lexicon, grammar, and style of a natural language for the purposes of improving machine tractability as well as facilitating human communication (Kuhn, 2014). A number of English-based CLs have been proposed and actually implemented in technical documentation. Evidence of improved machine translatability and post-editing productivity has also been provided (Pym, 1990; Bernth and Gdaniec, 2001; Aikawa et al., 2007).

In the case of Japanese-based CL, several recent research projects have attempted to design a CL focusing on documents for business (Hartley et al., 2012; Matsuda, 2014) or for government services (Miyata et al., 2015). Despite the proven effectiveness of Japanese CLs for both machine translatability and human readability, and a growing need for CL in technical communication, practical deployment of CL is not yet advanced.

Since writing source texts in accordance with a CL is not an easy task, particularly for non-professional writers, offering software support for applying a CL is essential for its practical use in the workplace. An extreme solution is the fully-automatic rewriting of source texts, which was explored not only for English (Mitamura and Nyberg, 2001) but also for Japanese (Shirai et al., 1998). However, rewriting without human intervention is a difficult task[2] and could induce other errors in grammar and style.

A more moderate solution is human-machine interactive writing. When the number of rules to be consulted is large and the application of rules stated verbally is difficult, sup-

[1]MuTUAL Project, http://mutual-project.com

[2]Just consider, for example, disambiguating sentence structure or word sense, and inferring omitted elements.

porting human decision making in authoring becomes essential. CL writing is divided into four processes and support mechanisms can be defined as in Table 1.

Writing process	Support mechanism
Notice violations of rules	Detect violations
Find alternatives	Suggest alternatives / Provide examples
Decide the best one	Rank alternatives / Provide information for dicision making
Rewrite text	Correct text

Table 1: CL writing processes and support mechanisms

A leading example of CL writing support linked to MT is the KANTOO Controlled Language Checker (Mitamura et al., 2003; Nyberg et al., 2003), which detects problems in input source texts and provides diagnostic information for authors. Several others have also been proposed for languages other than English, such as Greek (Karkaletsis et al., 2001) and German (Rascu, 2006). Although the pioneering work by Nagao et al. (1984) proposes a tool to disambiguate the construction of Japanese sentences and some commercial text checking tools for Japanese have recently become available,[3] to date few practical implementation and evaluation results of Japanese CL tools have been provided.

3. Controlled language rules

We defined two requirements of our CL rules: (1) to raise the quality of the MT outputs, and (2) to improve, or at least maintain, the human readability of the source texts (ST).
Table 2 shows the 23 rules we chose to implement in our CL violation detector and whether they meet the above requirements. A '–' means there is no significant effect.
With a practical implementation of the system in mind, we focused at this stage on two MT systems: TransGateway,[4] a commercial rule-based MT (RBMT) system widely used in Japanese municipalities (hereafter, MT1), and TexTra,[5] a freely available state-of-the-art statistical MT (SMT) system (hereafter, MT2).
In previous research, we performed a human evaluation experiment to assess the MT quality and ST readability of 38 CL rules for municipal documents (Miyata et al., 2015). The results revealed some rules have positive effects on MT quality but degrade the readability of the ST. We decided to postpone this problem of occasional incompatibility between MT and ST quality to the later stage of system implementation, and, based on the evaluation results, selected rules which met at least one of the requirements above. This is why we retained rules 4, 14 and 18, which are shown to be effective only for ST readability, and rules 3 and 16, which improve MT quality but degrade ST readability. We further discuss means to solve the problem of this incompatibility in Section 6.

[3]For example, Acrolinx caters for several languages including Japanese. http://www.acrolinx.com/

[4]Kodensha CO., http://www.kodensha.jp

[5]NICT, https://mt-auto-minhon-mlt.ucri.jgn-x.jp

No	Rule	MT1	MT2	ST
1	Do not omit subject.	✓	✓	✓
2	Do not omit object.	✓	✓	✓
3	Do not use comma for connecting noun phrase enumeration.	✓	✓	×
4	Avoid using particle Ga (が) for object.	–	–	✓
5	Avoid using Te-kuru (てくる) / Te-iku (ていく).	–	✓	–
6	Avoid inserted adverbial clause.	–	✓	–
7	Do not end clause with noun.	✓	✓	–
8	Avoid using Sahen-noun + auxiliary verb Desu (です).	✓	✓	✓
9	Avoid using attributive use of Shika-Nai (しか-ない).	✓	✓	–
10	Avoid using verb + You-ni (ように).	–	✓	–
11	Avoid using Sahen-noun + honorific Sare-ru (される).	✓	✓	–
12	Avoid using particle Nado (など/等).	–	✓	–
13	Avoid using giving and receiving verb.	✓	–	✓
14	Avoid using verbose word.	–	–	✓
15	Avoid using compound word.	✓	✓	–
16	Do not omit parts of words in enumeration.	✓	✓	×
17	Do not omit expression to mean 'per A'.	✓	✓	✓
18	Avoid using conjunctive particle Te (て).	–	–	✓
19	Avoid using if particle To (と).	✓	×	–
20	Use Chinese Kanji characters for verb as much as possible instead of Japanese Kana characters.	–	✓	✓
21	Avoid leaving bullet mark in texts.	✓	✓	–
22	Avoid using machine dependent characters.	✓	✓	–
23	Avoid using square bracket for emphasis.	✓	✓	–

Table 2: The list of CL rules implemented in our CL violation detector

4. Pilot evaluation

Given our aim of developing a CL tool which helps writers create CL-compliant municipal texts, we initially implemented a sub-component to detect violations for the 23 selected CL rules. We then benchmarked its performance and analysed the results.

4.1. Implementation

To implement the CL violation detection component, we created simple matching rules based on Part-of-Speech information, using the Japanese morphological analyser MeCab.[6] As Table 2 shows, some rules are defined only

[6]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

broadly (e.g., rule 14, 15 and 17). We thus first provided more detailed specifications for each rule. For instance, for rule 15 (Avoid using compound words) we decided to define 'compound word' as 'a sequence of more than two nouns (or noun-equivalents)', given that a compound word minimally consists of a sequence of two nouns. Two-noun compounds are as frequent in technical Japanese as they are in English and proscribing them would severely degrade both expressivity and naturalness. Moreover, they are easily represented in MT systems that accept lexicons and are generally well captured in the n-gram models of SMT systems.

4.2. Set up

For our evaluation, we collected test data from the website of Toyohashi City,[7] a Japanese local government. We first selected four Japanese sentences violating each of our 23 CL rules, i.e., a total of 92 sentences. We then checked whether each sentence violated more than one rule. A total of 223 CL rule violations were manually detected, and 71 of 92 sentences exhibited multiple violations. We also added eight sentences without any violations. Our dataset thus consists of 100 sentences.

For each rule, we counted the number of cases where: violations were correctly detected (True Positives: TP); non-violations were mistakenly detected as violations (False Positives: FP); and violations were not detected (False Negatives: FN). This enabled us to compute the metrics of precision, recall and F-measure[8] for each rule and overall.

4.3. Results

Table 3 shows the results of the evaluation of the detection function for each rule. We can see that 15 rules — rules 3, 4, 5, 8, 9, 10, 12, 13, 14, 15, 17, 18, 19, 21 and 22 — achieve more than 0.8 F-measure, with nine rules — rules 5, 8, 9, 10, 12, 13, 14, 17 and 22 — obtaining the perfect score of 1.0. This is an encouraging result for further implementation, although a larger-scale evaluation will be needed to confirm the results.

Overall, the total recall attains a high score of 0.870. In contrast, the overall precision is 0.676, meaning that about one third of the detections are false. We further observe that, focusing on the rules for which the F-measure is less than 0.5, precision is much lower than recall.

To find ways of improving the performance of the system (in preparation for a practical implementation of the rules), we further analysed the results of individual rules. This is presented next.

4.4. Discussion

We look at the rules with good performances first (both precision and recall higher than 0.7), followed by the rules with poor performance (both precision and recall lower than 0.7) and, finally, the rules with mixed performance (precision above 0.7, recall below 0.7).

We note that for the rules with both high precision and high recall (more than 0.7) — rules 3, 4, 5, 8, 9, 10, 12, 13, 14, 15, 17, 18, 19, 21 and 22 — deviations from rules can be captured by using Part-of-Speech information alone and the range of deviations is limited. Therefore, it is rather easy to formulate corresponding matching rules that are comprehensive. The following is an example of a sentence violating rule 9, for which the violation was correctly identified (thus a true positive (TP) example); we also provide the human reference translation (RT).

Rule No	Violation (cnt)	Precision	Recall	F-measure
1	26	0.630	0.654	0.642
2	15	0.333	0.667	0.444
3	20	**0.740**	**1.000**	**0.851**
4	5	**1.000**	**0.800**	**0.889**
5	6	**1.000**	**1.000**	**1.000**
6	7	0.286	0.571	0.381
7	4	0.111	**0.750**	0.194
8	6	**1.000**	**1.000**	**1.000**
9	4	**1.000**	**1.000**	**1.000**
10	5	**1.000**	**1.000**	**1.000**
11	4	0.500	**1.000**	0.667
12	22	**1.000**	**1.000**	**1.000**
13	4	**1.000**	**1.000**	**1.000**
14	5	**1.000**	**1.000**	**1.000**
15	35	**0.897**	**1.000**	**0.946**
16	5	0.429	0.600	0.500
17	5	**1.000**	**1.000**	**1.000**
18	14	**1.000**	**0.857**	**0.923**
19	5	**1.000**	**0.800**	**0.889**
20	4	0.364	**1.000**	0.533
21	9	**1.000**	**0.889**	**0.941**
22	7	**1.000**	**1.000**	**1.000**
23	6	0.500	0.500	0.500
Total	223	0.676	0.870	0.761

Table 3: Results of the evaluation

Rule 9: Avoid using attributive use of Shika-Nai (しか-ない)

TP 自生地には観察会の 2 日間しか入れません
(*Jisei-chi ni wa kansatsu-kai no futsuka-kan* ***shika*** *haire* ***masen***)

RT One **can only** enter the wildlife area during the two days of the observation event.

This sentence consists of a variant form of 'Shika-Nai' construction; 'masen' (ません) is a honorific mode of an auxiliary verb 'Nai' (ない). These kinds of variants are easily and reliably covered by a small number of matching rules using the morphological analyser.

In contrast, rules for which both precision and recall are below 0.7 — rules 1, 2, 6, 16 and 23 — are not easily handled with morphological information alone. We illustrate this below with rule 1.

Rule 1: Do not omit subject

In Japanese sentences, subjects are apt to be followed by particle 'Ga' (が) or 'Wa' (は).

TP 今後、広報等による啓発活動などで認定事業を応援していきます

[7] http://www.city.toyohashi.lg.jp

[8] $F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall}$

(*Kongo, kouhou-tou ni yoru keihatsu katsudou nado de nintei jugyo o ouen shite iki-masu*)

RT In the future, **we** will support certified business by educational activities through advertisements, etc.

In this true positive case, the lack of 'Ga' or 'Wa' correctly corresponds to an omission of the subject.[9] However this is not always the case, as the false positive (FP) example below shows:

FP 実行委員会一同努力しています
(***Jikkou-iinkai*** *ichidou doryoku shite imasu*)

RT **The executive committee** is working hard as one

Though the subject 'the executive committee' (実行委員会) is present in the sentence, the system mistakenly detected an omission of subject because this sentence lacks the particle.
In the false negative (FN) case below, the sentence lacks a subject in the latter clause. The human translator inferred 'they' as a subject, but the system failed to detect this subject omission, because the sentence includes both clue particles 'Wa' and 'Ga'.

FN 家庭や地域は、子どもが多くの時間を日常的に過ごす場所であり、生活の中で様々なことを学んでいきます。
(*Katei ya chiiki* ***wa****, kodomo* ***ga*** *ooku no jikan o nichijo-teki ni sugosu basho de ari, seikatsu no naka de samazama-na koto o manan-de iki-masu*)

RT Homes and communities are places where children spend a lot of time every day, and where **they** learn many things about life.

Generally speaking, the detection of missing elements — subject (rule 1), object (rule 2) and parts of words in enumeration (rule 16) — requires a deep analysis of sentence structure in addition to surface morphological information. To do so, we need to incorporate other tools, such as a parser and chunker, and techniques such as machine learning.
Finally, for the rest of the rules — rules 7, 11 and 20 — precision is high but recall is low. The surface Part-of-Speech constructions are simple, but correct identification requires fine-grained distinctions, as illustrated below.

Rule 11: Avoid using Sahen-noun + honorific Sare-ru (される)
'Sare-ru' has two grammatical usages: honorific or passive. Since distinguishing the two usages is quite difficult, we decided to simply detect Sahen-noun + 'Sare-ru' constructions without considering its usage.

TP すでに請求された方は対象になりません
(*Sude-ni* ***seikyu sare-ta*** *kata wa taisho ni nari-masen*)

RT This does not apply to those who have already **claimed**

In this case, 'sare-ta' happens to be honorific, and it is correctly detected. There are, however, many cases where 'Sare-ru' is used as passive:

FP 在留期間が3か月を超えて適法に在留する外国人の方も、住民票に**記載される**ようになります
(*Zairyu kikan ga san-kagetsu o koete tekihou ni zairyu suru gaikoku-jin no kata mo, jyumin-hyo ni* ***kisai sareru*** *you-ni nari-masu*)

RT Foreigners who remain in the country legally for a residential period of more than three months will also begin to **be recorded** in the resident register

Considering the low precision (0.500) of this rule, further improvement is necessary. We plan to apply machine learning methods as they have been shown to be effective in this kind of disambiguation task.
Similarly, rule 20 (Use Chinese Kanji characters for verbs as much as possible instead of Japanese Kana characters) seems easy to implement at first sight. However, we found that there are (auxiliary) verbs which tend to be, or should be, written in Japanese Kana characters. In this experiment, we saw seven false positive detections from 100 sentences and observed that some of the verbs are common ones, such as 'Miru' (see), 'Kakeru' (put) and 'Iu' (say). This implies that false positive detections could occur frequently unless we rule out these verbs when implementing the rule. Therefore, it is necessary to supply a list of verbs commonly written in Kana.
In summary, while we are able to achieve a perfect detection score for some rules, there is still much room for improvement by making use of (1) language tools such as parser and chunker, (2) machine learning techniques, and (3) lexical resources, as described above. We should, however, remain aware of the difficulty of achieving the perfect benchmark scores for all rules. We aim to alleviate some of the problems through our interface design (Section 5.2).

5. Controlled language checker

5.1. Concept

Our CL checker is intended for writing source texts from scratch, which led us to design a real-time interactive system to continuously check the conformity to CL. Whenever writers enter input violating any of the working CL rules, the system detects it and helps them to make corrections.
The target users of our system are non-professional writers who are, in many cases, not accustomed to the principle of controlled writing and writing support tools. Unlike common spell and grammar checkers, what our system detects is grammatically correct but, according to the pre-defined writing rules, should be avoided. Writers need to change their usual writing styles that are allowed in non-technical writing. Moreover, some CL rules require linguistic knowledge even native speakers may be unfamiliar with, such as 'Sahen-noun' and 'giving and receiving verbs'. Thus it is of particular importance to provide adequate descriptions of the rules and editing instructions.

5.2. System components and prototype implementation

Our CL checker consists of four components for *detection*, *suggestion*, *ranking* and *correction* based on the Table 1 in Section 2.

[9]A human translator inferred 'we' as a subject in the RT.

Figure 1: CL checker

Figure 1 shows a prototype interface of the CL checker.[10] The use scenario for this checker is as follows:

1. Authors enter Japanese text in the text box, guided by the instructions about implemented CL rules.
2. The system automatically analyses each sentence and displays in red highlighting any segment that violates the CL (*detection*), together with diagnostic comments and advice for rewriting.
3. For particular highlighted segments, the function offers alternative expressions displayed on mouse-over (*suggestion*). If there are more than one, suggestions are presented in the order of priority (*ranking*).
4. If the author clicks one of the suggestions, the segment in the text box above is automatically replaced (*correction*).

In step 2, false detection alerts generated by the system could annoy and even misguide writers. From the viewpoint of usability, we take two measures to address this issue: (1) allow users to select which rules to run; (2) display a 'confidence score', which tells writers how accurate the detection might be.
Here, our experimental results in Section 4 give us indicators for the performance of the CL detection component. For example, a confidence score for rule 11 (Avoid using Sahen-noun + honorific Sare-ru) can be defined as, say, '50%' based on the precision score in Table 3. Users can simply switch off the particular rule if they are sure of checking the rule by themselves, or keep it active while being fully aware of possible mis-detections.

6. Future plans

6.1. Full implementation of CL checker

In this study, we evaluated the performance of the detection sub-component of our CL checker. The results showed that 15 rules with high benchmark scores are ready to be implemented in the system, and gave us insights into how we can improve the performance of the remaining rules. The results also enabled us to design an interface to mitigate the false positive detections.

The next step is to develop the other components and deliver a fully functional system. Offering suggestions is a critical part of supporting writers, as they may not always think of proper alternatives even if they notice the violations. We assume that for some rules candidate suggestions can be clearly defined, such as 'Use expressions Dake (だけ) or Nomi (のみ)' for rule 9, and 'Delete Nado (など/等)' for rule 12. On the other hand, others — such as predicting omitted subject (rule 1) and object (rule 2) — are challenging. In these cases, it is effective to present several model examples of rewriting that authors can generalise to the particular cases they are dealing with.
Furthermore, we need to address the incompatibility issue raised by those rules which, while enhancing translation quality, actually degrade the readability of the source text (rules 3 and 16 in Section 3). Understandably, human writers may be reluctant to produce 'less readable' texts.
Accordingly, we envisage two approaches for handling such rules: (1) automatically pre-edit as a background process prior to MT if it is easily handled by machine; (2) devise an additional interface for asking authors to further rewrite the texts for MT use only, which revisions will not be provided to ST readers.

6.2. Terminology checker

Controlling terminology, more broadly vocabulary, is also an important issue to be tackled alongside the formal, grammatical and stylistic control offered by our CL checker. For example, there are several competing expressions for 'health check-up', such as 健康診査 (Kenkou shinsa) and 健康診断 (Kenkou shindan). Consistent use of terminology improves not only source readability but also — in the form of MT dictionaries — machine translatability.
Importantly, a terminology checking component can be implemented by simple string matching rules and seamlessly integrated into the CL checker. What is first needed is to formulate synsets of preferred and prohibited terms (Warburton, 2014). We are now collecting Japanese and English terms from parallel texts of the municipal domain, and specifying variant forms for the same referents.

6.3. Usability evaluation

Writing CL-compliant texts would be an arduous task for non-professional municipal writers. Our system supports and facilitates writers in each process of writing, incorporating mechanisms to reduce the adverse effects of false alerts of the system and to resolve, where they occur, incompatibilities in ST phrasing between human readability and machine translatability.
The question then is whether the checker is actually helpful in the workplace. We plan to conduct a usability evaluation to assess to what extent the system enhances the writing process, by comparing the time and effort taken in writing texts with and without the help of the system, and by asking users for feedback. This will eventually help us to refine the functions and interface towards a workable system.

[10] We are developing it as a Web-based application. The prototype system is functional online, implementing several rules.

7. Acknowledgements

This work was supported by the Research Grant Program of KDDI Foundation, Japan, and the Research Grant of Tokyo Institute of Technology, Japan. The MT system J-SERVER Professional TransGateway V3 was offered by Kodensha Co. Paris's stay in Japan to work with Miyata, Kageura and Hartley was funded by the Japanese Society for the Promotion of Science and CSIRO.

Aikawa, T., Schwartz, L., King, R., Corston-Oliver, M., and Lozano, C. (2007). Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In *Proceedings of the Machine Translation Summit XI*, pages 1–7, Copenhagen, Denmark.

Bernth, A. and Gdaniec, C. (2001). MTranslatability. *Machine Translation*, 16(3):175–218.

Hartley, A., Tatsumi, M., Isahara, H., Kageura, K., and Miyata, R. (2012). Readability and translatability judgments for 'Controlled Japanese'. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 237–244, Trento, Italy.

Karkaletsis, V., Samaritakis, G., Petasis, G., Farmakiotou, D., Androutsopoulos, I., Markantonatou, S., and Spyropoulos, C. D. (2001). A controlled language checker based on the ellogon text engineering platform. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 90–103, Pennsylvania, USA.

Kuhn, T. (2014). A survey and classification of controlled natural languages. *Computational Linguistics*, 40(1):121–170.

Matsuda, S. (2014). Efforts for Technical Japanese: Focusing mainly on the 'Patent Documents Writing Manual'. *Journal of Information Processing and Management*, 57(6):387–394. (in Japanese).

Mitamura, T. and Nyberg, E. (2001). Automatic rewriting for controlled language translation. In *Proceedings of the NLPRS2001 Workshop on Automatic Paraphrasing: Theory and Application*, pages 1–12, Tokyo, Japan.

Mitamura, T., Baker, K., Nyberg, E., and Svoboda, D. (2003). Diagnostics for interactive controlled language checking. In *Proceedings of the EAMT2003 Workshop on Controlled Language Applications*, pages 237–244, Dublin, Ireland.

Miyata, R., Hartley, A., Paris, C., Tatsumi, M., and Kageura, K. (2015). Japanese controlled language rules to improve machine translatability of municipal documents. In *Proceedings of Machine Translation Summit XV*, pages 90–103, Miami, USA.

Nagao, M., Tanaka, N., and Tsujii, J. (1984). Support system for writing texts based on controlled grammar. *IPSJ SIG Technical Reports*, NL(44):33–40. (in Japanese).

Nyberg, E., Mitamura, T., and Huijsen, W.-O. (2003). Controlled language for authoring and translation. In Somers, H., editor, *Computers and the Translator*, pages 245–281. Benjamins, Amsterdam.

Pym, P. (1990). Pre-editing and the use of simplified writing for MT. In Mayorcas, P., editor, *Translating and the Computer 10*, pages 80–95. Aslib, London.

Rascu, E. (2006). A controlled language aproach to text optimization in technical documentation. In *Proceedings of KONVENS 2006*, pages 107–114, Konstanz, Germany.

Shirai, S., Ikehara, S., Yokoo, A., and Ooyama, Y. (1998). Automatic rewriting method for internal expressions in Japanese to English MT and its effects. In *Proceedings of the 2nd International Workshop on Controlled Language Applications*, pages 62–75, Pennsylvania, USA.

Warburton, K. (2014). Developing lexical resources for controlled authoring purposes. In *Proceedings of LREC 2014 Workshop: Controlled Natural Language Simplifying Language Use*, pages 90–103, Reykjavik, Iceland.