A Controlled Language Checker Based on the Ellogon Text Engineering Platform

Vangelis Karkaletsis, Georgios Samaritakis, Georgios Petasis, Dimitra Farmakiotou,
Ion Androutsopoulos and Constantine D. Spyropoulos

Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research (N.C.S.R.) "Demokritos"
P.O BOX 60228, Aghia Paraskevi, 153 10 Athens, Greece.
{vangelis, samarita, petasis, dfarmak, ionandr, costass}@iit.demokritos.gr

Abstract

We present a controlled language checker for Greek that has been implemented using *Ellogon*, a generic text engineering platform. A controlled language is a language with a restricted syntax, vocabulary and terminology that is typically applied to technical documents. The aim of using controlled languages in technical documentation is the production of texts with simple structure and restricted vocabulary that can be read and translated easier (Eijk, 1998; Vouros et al. 1997). Several software companies (e.g. Bull, IBM) as well as other companies (e.g. Caterpillar, General Motors, Boeing) are already using controlled languages during technical writing of their products. The restrictions imposed by the use of a controlled language, help to preserve uniformity in the writing style, especially in cases where authors tend to follow diverse writing approaches, and to reduce ambiguities in the resulting text. Using a controlled language makes translation more efficient and improves the quality of the translated version. Controlled language can also facilitate machine translation systems since the resources provided for the controlled language (vocabulary, terminology and syntax rules) can be embedded into the machine translation system, improving its performance.

The research prototype presented here is the first controlled language checker developed for the Greek language. It has been developed to assist Greek technical writers as well as to facilitate translation from Greek to other languages. Its lexical and grammatical resources cover technical documents from the domain of computational equipment.

Technical writers are able to call the checker through their word processor (Microsoft Word is used in the current implementation). This allows the user to check the structure and language of his/her documents in a similar way as a spelling/syntax checker. The technical document is first converted into an XML format in order to be processed by the checker. The style checker outputs the error tags in a format "understandable" by the word-processor in order to let the user see his/her errors (in the current implementations as Microsoft Word comments). The checker checks both text structure (e.g. line spacing, fonts style and size) and language (correct application of controlled language grammar and vocabulary) (see Fig. 1). The XML text is processed using linguistic resources (restricted terminology, vocabulary, grammar) and tools (tokeniser, sentence splitter, part of speech tagger, morphological analyzer, chunker) in order to apply the language checker. The text is also checked using a structure DTD (Document Type Definition) in order to locate possible errors in structure.

The current implementation of the language checker involves checks at 4 levels: paragraph, sentence, phrase, word. More specifically, it checks for the correct use of:

- paragraph size (number of sentences), sentence size (number of clauses in the sentence)
- terms (correct terms are followed by their translation in their first occurrence),
- acronyms and abbreviations (correct ones are followed by the full expression in their first occurrence)
- vocabulary (there is a list of words/phrases that should not be used)
- adjectives (upper limit in the number of adjectives used consecutively)
- verbs (restrictions in person and tense)

- pronouns, participles, conjunctions, prepositions (there is a subset of those that should not be used)
- punctuation marks

New linguistic checks will be added in the coming versions of the checker.

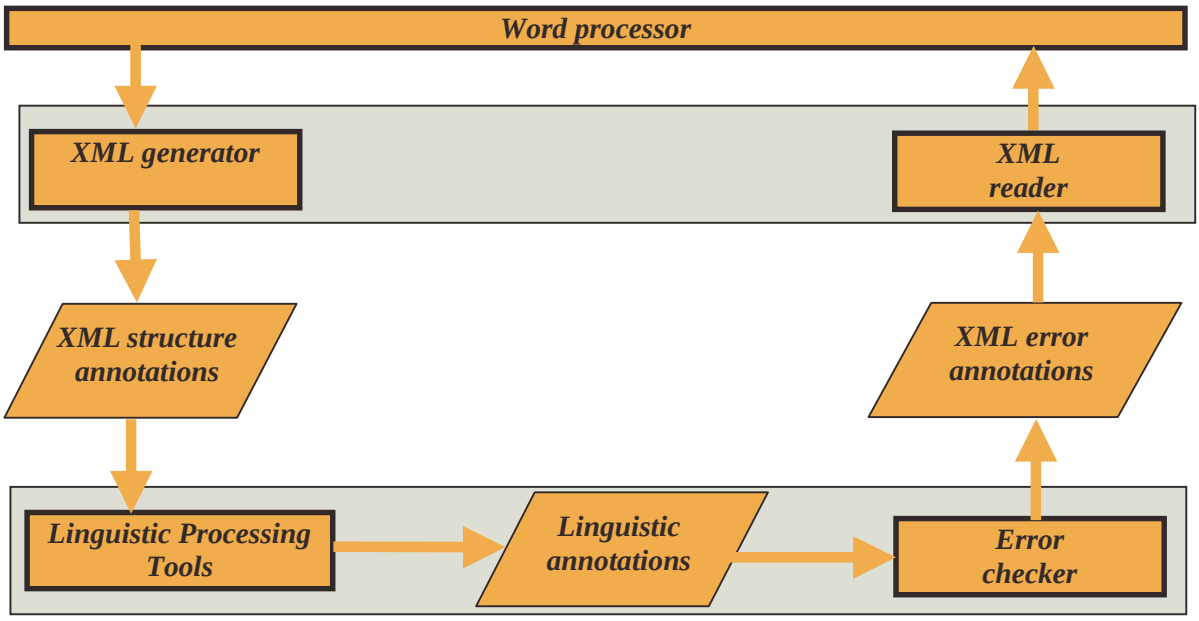

Fig.1 Processing stages of the controlled language checker

The linguistic resources and tools used have been developed using *Ellogon*, a new text engineering platform developed by our laboratory (Petasis et al. 2001). *Ellogon* provides a powerful TIPSTER-based infrastructure for managing, storing and exchanging textual data, embedding and managing text processing components as well as visualising textual data and their associated linguistic information. Among its key features are full Unicode support, an extensive multi-lingual graphical user interface and a modular architecture.

Ellogon has been used in this research work, not only for the development of the linguistic resources and tools but also for their embedding in the final application. This is achieved due to its modular architecture that allows the reuse of *Ellogon* sub-systems in order to ease the creation of applications targeting specific linguistic needs. More specifically, *Ellogon's* core component (CDM) is implemented as a separate library that can be dynamically loaded if the underlying operating system offers this ability. This library can be independently embedded inside any application that can call functions from libraries, following the C++ naming conventions. Examples of embedding CDM under various applications, apart from Microsoft Word, include Tcl and Java.

Acknowledgments

This research work has been partially supported by the R&D project SCHEMATOPOIESIS "Integrated environment for the development and exploitation of Greek controlled languages" which is funded by the Greek General Secretariat of Research & Technology.

References

(Petasis et al. 2001) Petasis G., Karkaletsis V., Paliouras G., and Spyropoulos C.D., 2001. Ellogon: A Text Engineering Platform. *NCSR "Demokritos" Technical report*.

(Eijk, 1998) Eijk P., 1998. Controlled Languages in Technical Documentation. *Elsnews, the Newsletter of the European Network in Language and Speech*, Feb. 1998, pp. 4-5.

(Vouros et al., 1997) Vouros G., Karkaletsis V., and Spyropoulos C.D., 1997. Documentation and Translation. *Software without frontiers*, P.A.Hall and R.Hudson (eds), J.Wileys & Sons, 1997, pp. 167-202.