# *Sovereign* - Natural Language Interfaces for authoring, annotating and accessing Knowledge on the Semantic Desktop

Brian Davis

Digital Enterprise Research Institute, National University of Ireland, Galway
brian.davis@deri.org

## 1 Research Problem

Semantic Technologies are currently inaccessible to non-expert users wishing to author(1), annotate(2) and or access(3) knowledge. Our research investigates how Human Language Technology (HLT) Interfaces; specifically Controlled Natural Languages (CNL) and applied Natural Language Generation(NLG) can provide a user friendly means for the non-expert users or small organisations to exploit Semantic Web technologies specifically the Social Semantic Desktop.

## 2 Proposed Approaches and State of the Art

### 2.1 Ontology Authoring

With respect to **authoring(1)** - specifically ontology authoring, while there are many ontology editing tools aimed at expert users, there are very few which are accessible to users wishing to create simple structures without delving into the intricacies of knowledge representation languages. The CLIE (Controlled Language for Information Extraction ) approach however allows users to create and edit ontologies quite simply by using a restricted version of the English language. This 'controlled natural language' is based on an open vocabulary and a restricted set of grammatical constructs. Sentences written in this language unambiguously map into a number of knowledge representation formats including OWL and RDF-S). Previous work [2] has used CLIE to generate ontologies from the input Controlled Natural Language (CNL) called CLOnE (Controlled Language for Ontology Engineering). The reverse of the process involves the generation of CLOnE from an existing ontology using Natural Language Generation (NLG), specifically shallow NLG. The NL generator and the authoring process both combine to form a RoundTrip Ontology Authoring (ROA) environment: one can start with an existing or empty ontology, create an Ontology using CLIE, reproduce CLOnE from the newly using the NL generator, modify or edit the text as required and subsequently parse the text back into the ontology using the CLIE environment. The process can be repeated as necessary until the required result is obtained. ROA raises the following research questions:

– Can NLG effectively substitute for CNL reference guides?
– Can NLG help ease the learning curve associated with CNLs?
– Can NLG improve on previous evaluation results for ontology authoring?

CNLs are "subsets of natural language whose grammars and dictionaries have been restricted in order to reduce or eliminate both ambiguity and complexity.[1]"The use of CNLs for ontology authoring and population is by no means a new concept and it has already evolved into quite an active research area[7]. The majority of existing tools do not employ NLG at all or at least not in the same manner as ROA. Furthermore, where this is the case, no empirical evaluation is provided[2].

### 2.2 Semantic Annotation

Concerning **annotation(2)**, richly interlinked, machine-understandable data constitute the basis for the Semantic Web, and by extension the Social Semantic Desktop[3]. Manual semantic annotation is a complex and arduous task both time-consuming and costly often requiring specialist annotators. (Semi)-automatic annotation tools attempt to ease this process by detecting instances of classes within text and relationships between classes, however their usage often requires knowledge of Natural Language Processing(NLP) and/or formal ontological descriptions. This challenges researchers to develop user-friendly annotation environments within the knowledge acquisition process. CNLs offer an incentive to the novice user to annotate, while simultaneously authoring, his/her respective documents in a user-friendly manner, but simultaneously shielding him/her from the underlying complex knowledge representation formalisms. A natural overlap exists between tools, used for both ontology creation and semantic annotation. However, there is a subtle difference between both processes. Semantic annotation is described as "a process, as well as the outcome of the process. Hence it describes i) the process of addition of semantic data or metadata to the content given an agreed ontology and ii) it describes the semantic data or metadata itself as a result of this process"[4]. Of particular importance here is the notion of the addition or association of semantic data or metadata to *content*. This raises the following research questions:

– Can a CNL make an effective semantic annotation tool?
– Can authoring and annotation be successfully merged using CNL?
– How can we effectively evaluate a CNL for semantic annotation?

CNLs have already been successfully applied within the context of ontology authoring, yet very little research has focused on CNLs for semantic annotation. For instance, Project HALO[2] was a research venture sponsored by Vulcan Inc[3]. It aimed to develop, a "Digital Aristotle" - a comprehensive, automated tutor

[1] http://www.ics.mq.edu.au/~rolfs/controlled-natural-languages/
[2] http://www.projecthalo.com/
[3] http://www.vulcan.com

and research assistant. A CNL for semantic annotation was implemented as part of the project, yet no public material describing the CNL is available for scientific scrutiny.

### 2.3 Applied NLG for Knowledge Access

With regard to **knowledge access(3)**, NLG can act as a human-readable window into the otherwise formally structured database. Rather than summarising the contents of an Ontology for the purposes of modification or quality assessment as is specified above in the context of ROA(1), the use case of NLG in our context will be to provide user friendly means of presenting semantically annotated knowledge captured within the Social Semantic Desktop in a human readable form and furthermore at the appropriate level of detail to the non-expert user, based on his/her specific queries. Consequently, knowledge will be outputted to the user in the form of natural language text. NLG allows for productions in NL text to be tailored to the presentational context and needs of the target reader. Based on user profiling, appropriate presentation strategies and appropriate levels of detail can be selected for presenting the same data to different users. This is an important problem because textual documentation is more readable than the corresponding formal notations and thus helps users who are not knowledge engineers to understand and use ontologies. In other words, NLG can be used to present structured information in a user-friendly way. NLG raises the following research questions:

- What balance between shallow and deep NLG techniques to choose?
- Or should we attempt to implement a Hybrid system?
- How do we effectively evaluate our NLG system?

Natural Language Generation (NLG) takes structured data in a knowledge base as input and produces natural language text, tailored to the presentational context and the target reader[6]. NLG systems that are specifically targeted towards Semantic Web ontologies have started to emerge only recently. Initial ones were based on templates (**Shallow NLG)**, verbalizing closely the ontology structure as is the case in ROA. More recent ones generate more fluent reports, oriented towards end-users, not ontology builders. In contrast to these applied NLG approaches, at the other end of the spectrum are sophisticated ones based on Computational Linguistic(CL) theories (**Deep NLG**), which offer tailored output based on user models. The trade-off is between applied approaches, exploring generalities in the domain ontology and with lower customization overheads, and, on the other hand, sophisticated, more flexible and expressive systems, which, however tend to be difficult to adapt by non-NLG experts[1].

## 3 Use Case, Methodology and Results

With respect to **annotation(2)**, CNLs cannot offer a panacea for semi-automatic annotation since it is unrealistic to expect users to annotate every textual resources using CNL, however there are certain use-cases where CNLs can offer

an attractive alternative as a means for semi-automatic semantic annotation, particular in contexts, where controlled vocabulary or terminology is implicit such as health care patient records or business vocabulary. Our use case focuses on administrative tasks such taking minutes during a project team meeting and weekly status reports. Very often such note taking tasks can be repetitive and boring. In our scenario the user is a member of a research group which in turn is part of an integrated EU Research project. Based on pre-defined templates, the user *simultaneously authors and annotates* his/her meeting minutes or status reports in CNL. The metadata stored on the Social Semantic Deskop is available for immediate use after creation for querying and aggregation, whereby the retrieved RDF triples can be passed to a **Natural Language Generator(3)** to produce tailored textual reports and summaries. Finally wrt **authoring(1)**, ROA supports the ontology authoring process to create a common vocabulary on which to base annotation and knowledge capture and subsequent text generation. The authoring, annotation and NLG resources are collectively called **Sovereign** - *Semantic annOtation VErbal REsources for ExtractIon and Generation of kNowledge.*

The Sovereign **authoring** resource is based on the ROA architecture, which contains a standard GATE [4]pipeline consisting of the default language processing resources in addition to both customatised finite state gazetteers and cascaded finite state transducer grammars. The text generator component is an XML based shallow NLG system. Ontologies are created, verabalised and edited using the GATE Ontology API. We refer the reader to [2] for specifics regarding the technical implementation of ROA. Building on previous methodology, we undertook a **repeated-measures, task-based** evaluation, comparing the RoundTrip Ontology Authoring process with Protégé. Where previous work required a reference guide in order to use the controlled language, the substitution of NLG can reduce the learning curve for users, while simultaneously improving upon existing results for basic ontology editing tasks[2]. See Section 2.1 for associated research questions.

With respect to annotation the Sovereign CNL Annotator is based on the ROA technology with substantial modifications to the transducer grammars and gazetteer lists and is bootstrapped via the Nepomuk Core Ontologies[5]. Currently the application populates a meeting minutes/status report ontology which references the users Personal Information Model Ontology(PIMO) [6], again using the GATE Ontology API. The CNL itself is very similar to the CLOnE language, with some modifications. (A paper describing the initial prototype was accepted to CNL09). Our evaluation of the CNL annotator will be based on the repeated-measures, task-based methodology, which was successfully applied to ROA. In addition we will add in a de facto standard Semantic Wiki and semi-automatic annotation tool into the evaluation.

[4] General Architecture for Text Engineering, See http://gate.ac.uk/
[5] http://www.semanticdesktop.org/ontologies/
[6] http://www.semanticdesktop.org/ontologies/2007/11/01/pimo/

Finally the Sovereign NLG resource is still in the early conceptualisation phase. We have however contributed to an exhaustive review of the literature[1]. Worth noting is that modern template/shallow systems are based on XML but hybrid systems are becoming more common[5]. Experience shows that knowledge management and Semantic Web ontologies tend to evolve over time, so it is essential to have an easy-to-maintain NLG approach.

## 4 Conclusion and Future Plans

Semantic Technologies are currently inaccessible to non-expert users wishing to author(1), annotate(2) and or access(3) knowledge. This paper outlines how HLT Interfaces; specifically Controlled Natural Languages (CNL) and applied Natural Language Generation(NLG) can provide a user friendly means for the non-expert users or small organisations to exploit Semantic Web technologies - in our case, the Social Semantic Desktop. Part of our work wrt to (1) has already been successfully completed and evaluated, while the CNL Annotator plugin is past the prototype stage. In addition we have successfully reviewed the literature concerning NLG for the Semantic Web. Future work will involve evaluating the CNL annotator as well commencing with development of our NLG system.

## References


1. Kalina Bontcheva and Brian Davis. Natural Language Generation from Ontologies. In John Davies, Marko Grobelnick, and Dunja Mladenić, editors, *Semantic Web Technologies: Trends and Research in Ontology-based Systems*, pages 113–127. John Wiley & Sons, July 2006.
2. Brian Davis, Ahmad Ali Iqbal, Adam Funk, Valentin Tablan, Kalina Bontcheva, Hamish Cunningham, and Siegfried Handschuh. Roundtrip ontology authoring. In *International Semantic Web Conference*, volume 5318 of *Lecture Notes in Computer Science*, pages 50–65. Springer, 2008.
3. S. Decker. The Social Semantic Desktop: Next generation collaboration infrastructure. *Information Services and Use*, 26(2), 2006.
4. Siegfried Handschuh. *Creating Ontology-based Metadata by Annotation for the Semantic Web*. PhD thesis, 2005.
5. Martin Klarner. Hybrid NLG in a Generic Dialog System. In *INLG*, pages 205–211, 2004.
6. Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, New York, 2006.
7. P. R. Smart. Controlled Natural Languages and the Semantic Web. Technical report, School of Electronics and Computer Science, University of Southampton, 2008,(Unpublished).