Unlocking Medical Ontologies for Non-Ontology Experts

Shao Fen Liang
School of Computer Science,
The University of Manchester, Oxford Road,
Manchester, M13 9PL, UK
Fennie.Liang@cs.man.ac.uk

Donia Scott
School of Informatics,
The University of Sussex, Falmer,
Brighton, BN1 9QH, UK
D.R.Scott@sussex.ac.uk

Robert Stevens
School of Computer Science,
The University of Manchester, Oxford Road,
Manchester, M13 9PL, UK
Robert.Stevens@cs.man.ac.uk

Alan Rector
School of Computer Science,
The University of Manchester, Oxford Road,
Manchester, M13 9PL, UK
Rector@cs.man.ac.uk

Abstract

Ontology authoring is a specialised task requiring amongst other things a deep knowledge of the ontology language being used. Understanding and reusing ontologies can thus be difficult for domain experts, who tend not to be ontology experts. To address this problem, we have developed a Natural Language Generation system for transforming the axioms that form the definitions of ontology classes into Natural Language paragraphs. Our method relies on deploying ontology axioms into a top-level Rhetorical Structure Theory schema. Axioms are ordered and structured with specific rhetorical relations under rhetorical structure trees. We describe here an implementation that focuses on a sub-module of SNOMED CT. With some refinements on articles and layout, the resulting paragraphs are fluent and coherent, offering a way for subject specialists to understand an ontology's content without need to understand its logical representation.

1 Introduction

SNOMED CT (Spackman and Campbell, 1998) is widely mandated and promoted as a controlled vocabulary for electronic health records in several countries including the USA, UK, Canada and Australia. It is managed by the International Health Terminology Standards Development Organisation (IHTSDO) [1] . SNOMED describes diagnoses, procedures, and the necessary anatomy, biological process (morphology[2]) and the relevant organisms that cause disease for over 400,000 distinct concepts. It is formulated using a Description Logic (DL) (Baader et al., 2005). Description logics, usually in the form of the Web Ontology Language (OWL)[3] have become a common means of representing ontologies. Description logics in general and SNOMED in particular have been recognised as difficult to understand and reuse (Namgoong and Kim, 2007;Power et al., 2009). Even with the more or less human readable, Manchester OWL Syntax (Horridge et al., 2006) and using tools such as Protégé (Knublauch et al., 2004) the task of understanding ontologies remains non-trivial for most domain experts.

Consider, for example, a clinician seeking information about the concept of *thoracic cavity structure* [4] (i.e., anything in the chest cavity). SNOMED provides the following six axioms:

1. <Structure of thoracic viscus>
   SubClassOf <Thoracic cavity structure>
2. <Intrathoracic cardiovascular structure>
   SubClassOf <Thoracic cavity structure>
3. <Mediastinal structure>
   SubClassOf <Thoracic cavity structure>
4. <Thoracic cavity structure>
   SubClassOf <Structure of respiratory system and/or intrathoracic structure>
5. <Thoracic cavity structure>
   SubClassOf <Thoracic structure>
6. <Thoracic cavity structure>
   SubClassOf <Body cavity structure>

[1] http://www.w3.org/TR/owl-features/
[2] Literally, the altered structure as seen by the pathologist, but usually the evidence for the process that gave rise to it.
[3] http://www.w3.org/TR/owl-features/
[4] The SNOMED identifier for this class is ID: SCT_43799004

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 174–181,
Portland, Oregon, USA, June 23-24, 2011.

Although these axioms are shown with the more readable Manchester OWL syntax, the represented meaning of *Thoracic cavity structure* will not be easy for the typical clinician to decode.

Ontology concepts can be much more complex than those shown above. Not only can there be more axioms, but there can be nested axioms to an arbitrary depth. So the comprehension problem facing the typical clinician is even greater than that just described. It should be reduced, however, if the ontological content were presented in a more coherent, fluent and natural way – for example as:

> A *thoracic cavity structure* is a kind of *structure of* the *respiratory system and/or intrathoracic structure*, *thoracic structure* and *body cavity structure*. It includes a *structure of* the *thoracic viscus*, an *intrathoracic cardiovascular structure* and a *mediastinal structure*.

or, with added layout, as:

> A *thoracic cavity structure* is a kind of
> - *structure of* the *respiratory system and/or intrathoracic structure,*
> - *thoracic structure,*

> and
> - *body cavity structure.*

> It includes
> - a *structure of* the *thoracic viscus,*
> - an *intrathoracic cardiovascular structure*

> and
> - a *mediastinal structure.*

In these (human-generated) texts, the author has chosen to retain the general form of the anatomical terms as they appear in SNOMED, signalling them through the use of italics and introducing in places a definite article (e.g., "*structure of* the *thoracic viscus*". While these terms (particularly in the peculiar form they take in SNOMED names[5]) still present a barrier to non-subject-specialists, nevertheless the ontological content rendered as natural language is now much more accessible to non-ontology specialists.

Using natural language descriptions is obviously one way of improving the transparency of ontologies. However, authoring such descriptions is tedious and time-consuming to achieve by hand. This is clearly an area where automatic generation could be beneficial. With this in mind, we have built a verbaliser that renders SNOMED concepts as fluent natural language paragraphs.

2 Mapping SNOMED to a Representation of Coherent Discourse

Our goal is to use standard techniques for natural language generation (NLG) to generate fluent paragraph-sized texts for SNOMED concepts automatically.

Verbalisation is a two-staged process of deciding *what to say* and then *how to say it*. In our work the first of these is a non-issue: the content of our verbalisation will be SNOMED concepts. Our focus is therefore on deciding how to express the content.

As with any NLG system, our task begins by organising the input content in such a way as to provide a structure that will lead to coherent text, as opposed to a string of apparently disconnected sentences. Given the nature of our problem, we need to focus on the semantics of the discourse that can accommodate the nature of ontology axioms. For this purpose, we have chosen to use Rhetorical Structure Theory (RST) (Mann and Thompson, 1987;Mann and Thompson, 1988), as a mechanism for organising the ontological content of the SNOMED input.

RST is a theory of discourse that addresses issues of semantics, communication and the nature of the coherence of texts, and plays an important role in computational methods for generating natural language texts (Hovy, 1990;Scott and Souza, 1990;Mellish et al., 1998;Power et al., 2003). According to the theory, a text is coherent when it can be described as a hierarchical structure composed of text spans linked by rhetorical relations that represent the relevance relation that holds between them (among the set of 23 relations are EVIDENCE, MOTIVATION, CONTRAST, ELABORATION, RESULT, CAUSE, CONDITION, ANTITHESIS, ALTERNATIVE, LIST, CONCESSION and JUSTIFICATION). Relations can be left implicit in the text, but are more often signalled through *discourse markers* – words or phrases such as "because" for EVIDENCE, "in order to" for ENABLEMENT, "although" for ANTITHESIS,

[5] To reduce this problem somewhat, we use here the 'preferred term' for given SNOMED names, but even these can be quite peculiar, e.g., "*renal hypertension complicating pregnancy, childbirth and the puerperium - delivered with postnatal complication*".

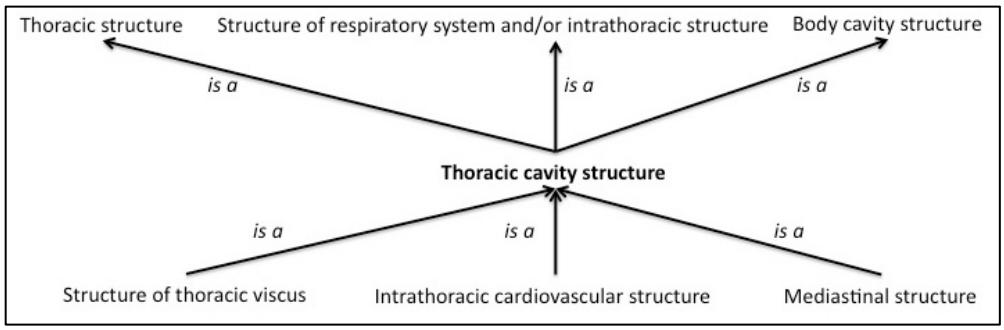

Figure 1: axioms and their relations to the class *Thoracic cavity structure*

"but" for CONCESSION, "and" for LIST, "or" for ALTERNATIVE, etc. (Sporleder and Lascarides, 2008;Callaway, 2003). They can also be signalled by punctuation (e.g., a colon for ELABORATION, comma between the elements of LIST, etc.).

In RST, text spans are divided into a schema, containing either a nucleus (N) and satellite (S), or two or more nuclei. Nuclei contain the information that is critical to the communicative message; satellites contain less critical information, which support the statements of their nuclei. The relations among nuclei and satellites are often expressed as:

RELATION(N,N)

RELATION(N,S)

These expressions conveniently take the same form as those expressing the types of ontology axiom, e.g.:

SubClassOf(A, B)
EquivalentClasses(C, D)

where, SubClassOf and EquivalentClasses express relations between A and B, and C and D. This suggests that with careful selection of RST relations, and applying appropriate discourse markers, ontologies can be represented as RST structures, and generated as natural language paragraphs that are not far from human written text.

To investigate the feasibility of this proposal, we have experimented with feeding axioms into RST trees, and have achieved a positive outcome. For example, the six axioms of the *thoracic cavity structure* concept that we have seen earlier can be organised into two groups of relations as shown in Figure 1. In the upper group are the super-classes of the *thoracic cavity structure* class, and in the lower are the sub-classes. This way of grouping the axioms can better present their relations to the class.

This structure can now be transformed into the RST tree shown in Figure 2, where the most important element of the message is the class *Thoracic cavity structure*, and this forms the main nucleus of the RST tree. The remaining content is related to this through an ELABORATION relation, the satellite of which is composed of two items of a multinucleus LIST, each of which is itself a LIST. This structure can be expressed textually as (among others) the two natural language descriptions we have shown earlier. These texts satisfy the requirement of coherence (as defined by RST), since each part bears a rhetorical relation to the other, and the entire text is itself spanned by a single rhetorical relation.

Our exploration of RST has shown that some relations map well to the characteristic features of ontology axioms. For example:

- the LIST relation captures well those cases where a group of axioms in the ontology bear the same level of relation to a given class;
- the ELABORATION relation applies generally to connect different notions of axioms to a class (i.e., super-, sub- and defining- classes), in order to provide additional descriptive information to the class;
- the CONDITION relation generally applies in cases where an axiom has property restrictions.

We also found that some rhetorical relations appear to bear a one-to-one mapping with logical forms of axioms, such as ALTERNATIVE to the logical `or`, and LIST to the logical `and`.

Our experience and the evidence over many practical cases have indicated that the full set of rhetorical relations is unlikely to be applied for ontology verbalisation. In particular, the set of so-called *presentational* relations are unlikely to apply, as ontology authors do not normally

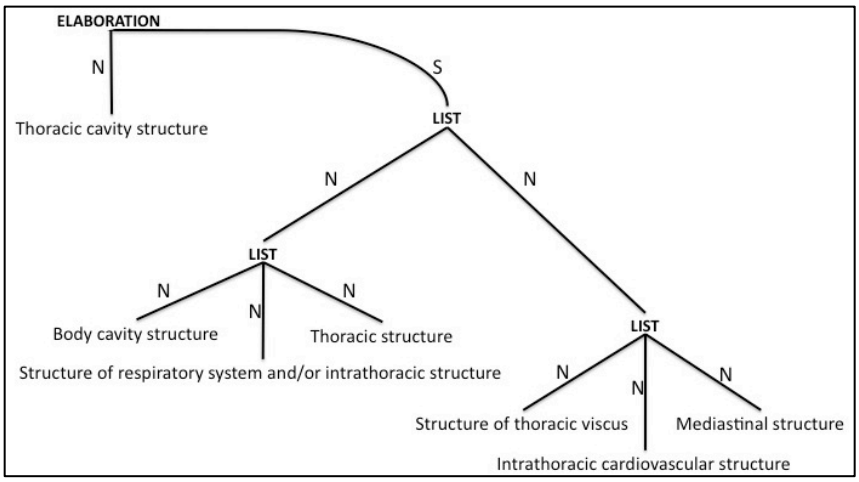

Figure 2: RST tree of the class *Thoracic cavity structure* with six axioms

create comparisons or attempt to state preferences amongst classes. (For example, SNOMED has no comparison operator between different treatments of diseases).

In addition, even within the set of *informational* relations (Moser and Moore, 1996), there are several that will not be found in ontologies. For example, since each axiom in an ontology is assumed to be true, using one axiom as an EVIDENCE of another axiom would be redundant. Similarly, using one axiom to JUSTIFY another axiom is not a conventional way of building ontologies.

3 Applying RST

Our investigations have shown that it is possible to build a top-level RST schema to cover all axioms with different meanings related to a class (see Figure 3). In SNOMED, axioms relating to a concept (i.e., class) can be either *direct* or *indirect*. Direct axioms describe the topic class directly, in which the topic class is the first class appearing in those axioms. Indirect axioms provide extra information, typically about how a class is used with other classes. For example, the axiom

<Structure of thoracic viscus>
SubClassOf(<Structure of viscus> and <Thoracic cavity structure>)

can be placed as direct information about *structure of thoracic viscus*; it can also be placed as indirect information about *Structure of viscus* or *Thoracic cavity structure*.

Within the categories of direct and indirect information, axioms are also classified as either *simple* or *complex*. This distinction allows us to control the length of the verbalisation, since most complex axioms tend to be translated into longer sentences, involving as they do more properties and value restrictions. Simple axioms, on the other hand, describe only class relations, the length of which can be better controlled.

For a given SNOMED class, our verbalisation process starts with its super-, sub- and equivalent-classes, within an ELABORATION relation. The use of the ELABORATION relation allows the first part of the text to connect all classes relating to the topic class; the second part then starts to introduce more complex information directly related to the topic class. The ELABORATION relation is used until all the direct information has been included. Next the CONCESSION relation is applied to connect direct and indirect information.

Additionally, each indirect axiom should have its own subject, and therefore, they cannot be combined smoothly into a single sentence. We therefore use LIST as the relation for these axioms, since they are equally weighted, and changing the order among them does not affect the meaning of the whole paragraph.

Every complex axiom is translated using a CONDITION relation. This is because complex axioms contain conditional information to their subject class. For example:

<Disorder of soft tissue of thoracic cavity>
EquivalentTo(<Disorder of soft tissue of body cavity>
and (<RoleGroup> some

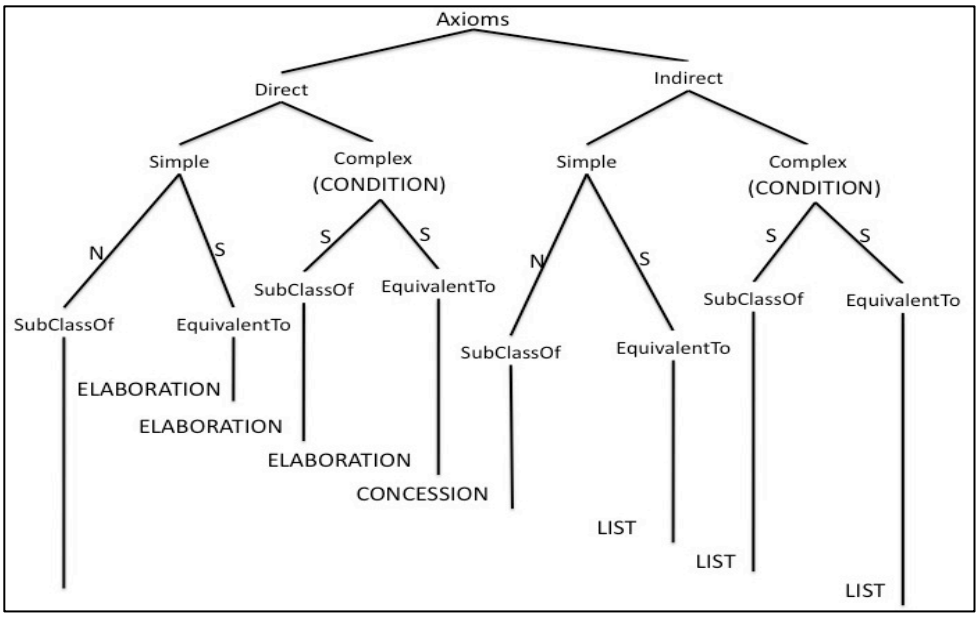

Figure 3: Top-level RST schema for SNOMED

(<Finding site> some
<Thoracic cavity structure>))
and (<RoleGroup> some
(<Finding site> some
<Soft tissues>)))

The condition in this axiom starts from the first "and" in the fourth line and extends to the end of the axiom. This condition needs to be attached to the class *Disorder of soft tissue of body cavity* to be equivalent to the *Disorder of soft tissue of thoracic cavity* class. We apply this rule to all complex axioms in an ontology.

4 Verbalising Individual Axioms

We use a template-based technique for verbalising the axioms as sentences. We have carefully selected translations of the SNOMED expressions. Our choice has been driven by an attempt to translate each axiom so as to preserve the meaning in the ontology and to avoid introducing misleading information. For example, the convention within ontologies is to conceptualise super-classes as an "is a" relation. However, translating this term as the English string "is a" can lead to misunderstanding, since the English expression can also be used to mean "equal to". Clearly, though, a class is not equal to its super-class. In this context, a more accurate translation is "is a kind of". We show some of these translations in Table 1.

Relation to the topic class X	**Translation wording**
With its simple super-class	X is a kind of …
With its complex super-class	X is a kind of … that …
With its simple sub-class	X includes …
With its simple equivalent class	X is defined as …
With its complex equivalent class	X is defined as … that

Table 1: Translations for axiom types

Consider for example, the SNOMED content:

<Benign hypertensive renal disease>
SubClassOf <Hypertensive renal disease>

<Benign arteriolar nephrosclerosis>
SubClassOf <Benign hypertensive renal disease>

<Benign hypertensive heart AND renal disease>
SubClassOf <Benign hypertensive renal disease>

<Benign hypertensive renal disease>
SubClassOf <Hypertensive renal disease>

and (<Finding site> some
<Kidney structure>)

<Benign arteriolar nephrosclerosis>
SubClassOf <Benign hypertensive renal disease>
and <Arteriolar nephrosclerosis>

<Benign hypertensive heart AND renal disease>
EquivalentTo<Benign hypertensive renal disease>
and <Benign hypertensive heart disease>
and <Hypertensive heart AND renal disease>

Our generator describes *Benign hypertensive renal disease* with its super-class as

> "*Benign hypertensive renal disease* is a kind of *hypertensive renal disease*."

and with its sub-classes as

> "*Benign hypertensive renal disease* includes *benign arteriolar nephrosclerosis* and *benign hypertensive heart and renal disease*."

There are two sub-classes in the above sentence, and we have signalled their connection (in a LIST relation) with "and" as the discourse marker. In those cases where there are more than two sub-classes, we use instead a comma "," except for the last mentioned, where we introduce "and". The same approach is applied to super-classes.

In those cases where a class has both super- and sub-classes to describe, we introduce the second sentence with "It" thus achieving better linguistic cohesion by avoiding having to repeat the same subject from the first sentence.

To bridge simple-direct and complex-direct axioms, we use "Additionally" to signal the introduction of more information relevant to the topic. For example to continue from the above two sentences, we have

> "Additionally, *benign hypertensive renal disease* is a kind of *hypertensive renal disease* that has a *finding site* in a *kidney structure*."

All direct information should have been consumed at this point, and we now need some bridging expression to signal the introduction of the indirect axioms. For this we use "Another relevant aspect of" or "Other relevant aspects of", depending on the number of axioms in the set. Continuing with our example, we now have

> "Other relevant aspects of *benign hypertensive renal disease* include the following: *benign arteriolar nephrosclerosis* is a kind of *benign hypertensive renal disease* and *arteriolar nephrosclerosis*; *benign hypertensive heart and renal disease* is defined as *benign hypertensive renal disease*, *benign hypertensive heart disease* and *hypertensive heart and renal disease*."

The improved transparency of the underlying ontological content can be clearly demonstrated by comparison with the SNOMED input from which it is derived.

The output that we have shown so far has all been generated as running text with minimal formatting except for the use of italic face for SNOMED labels. This works well for simple examples, but as can be seen from the previous example, readability becomes increasingly challenged as the expressions become longer. For this reason, we have also included in our system the facility to use layout to convey the logical structure of the ontological content. For example, the content shown above can also be generated as

> "*Benign hypertensive renal disease* is a kind of *hypertensive renal disease*. It includes
> - *benign arteriolar nephrosclerosis*

> and
> - *benign hypertensive heart and renal disease*.

> Additionally, *benign hypertensive renal disease* is a kind of *hypertensive renal disease* that has a *finding site* in a *kidney structure*. Other relevant aspects of *benign hypertensive renal disease* include the following:
> - *benign arteriolar nephrosclerosis* is defined as *benign hypertensive renal disease and arteriolar nephrosclerosis*;
> - *benign hypertensive heart and renal disease* is defined as *benign hypertensive renal disease*, *benign hypertensive heart disease* and *hypertensive heart and renal disease*."

5 Issues Related to Fluency

The quality of a text, whether human- or machine-generated, is to a large extent determined by its fitness for purpose. For example, the characteristics of a scientific article, a newspaper article or a twitter will be rather different, even though they may convey the same "message". The same is true for natural language descriptions of

ontological content, which can range from the fully-fluent to the closely literal (e.g., something likely to be thought of as a kind of "SNOMED-ese"), depending on whether it is intended, say, for inclusion in a narrative summary of an electronic patient record (Hallett et al., 2006) or for ontology developers or users who want to know the precise ontological representation of some part of the ontology. So far, our aim has been to generate descriptions that fall into the latter category. For this purpose we retain the full expressions of the pseudo-English labels found in the official SNOMED Descriptions document[6], representing them within our generation process as "quotes" (Mellish et al., 2006) and signalling them through the use of italics. The texts still need to be grammatical, however, and achieving this can be challenging. In what follows we give a few examples of why this is so.

It is a convention of ontology design to treat each class as singular; we follow this convention, introducing each class with the indefinite article. So, for example, the SNOMED labels

<Intrathoracic cardiovascular structure>

and

< Structure of thoracic viscus>

can be expressed straightforwardly as "a *structure of thoracic viscus*" and "an *intratrathoracic cardiovascular structure*". However, matters are not so simple. For example,

<Heart structure>

will require the definite article ("the *heart structure*") and while

<Structure of thoracic viscus>

will attract an indefinite article at its front, it would read much better if it also had a definite article within it, giving "a *structure of* the *thoracic viscus*". A similar story holds for

<Abdomen and pelvis>

which properly should be "the *abdomen and pelvis*" or "the *abdomen and* the *pelvis*". Achieving this level of grammaticality will rely on knowledge that, for example, the human body contains only one heart and abdomen. Interestingly, this information is not captured within the SNOMED ontology, and so external resources will be required. Additionally, introducing articles within the labels (as in "*abdomen and* the *pelvis*", above) will require some level of natural language interpretation of the labels themselves.

The same applies to number. While we currently follow the SNOMED convention of describing entities in the singular, there are occasions where the plural is called for. For example:

<Abdominal vascular structure>
SubClassOf <Abdomial structure>
SubClassOf <Lower body part structure>

<Abdominal cavity structure>
SubClassOf <Abdominal structure>
SubClassOf <Lower body part structure>

would be better expressed as "Lower body part structures include all abdominal structures", instead of as currently "A lower body part structure includes an abdominal structure".

Another issue to consider is the roles of properties in SNOMED. This problem can be characterised by the following example:

<Hypertension secondary to kidney transplant>
EquivalentTo (<Hypertension associated with transplantation>
and (<After> some <Transplant of kidney>))>

which is currently verbalised as

> *Hypertension secondary to kidney transplant* is defined as *hypertension associated with transplantation* that has an *after* in a *transplant of kidney*.

In SNOMED, the property *after* is used to give an after-effect (i.e., "*Hypertension associated with transplantation*" is an after-effect of a kidney transplant), and for a non-SNOMED expert, this meaning is not at all clear in the generated text. This applies to many properties in SNOMED. Consider for example, the properties "*finding site*" and "*clinical course*" as in:

> "*Chronic heart disease* is defined as a *chronic disease of cardiovascular system* that is a *heart disease*, and has a *clinical course* in a *chronic*."

and

> "*Abdominal organ finding* is a *general finding of abdomen* that has a *finding site* in a *structure of abdominal viscus*."

[6] http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctarchive.html

The extent to which issues such as these are treated within the generation process will, as we mentioned before, be a matter of how fluent the text needs to be for a given purpose.

6 Conclusion

We have described a method for generating coherent and fairly fluent natural language descriptions of ontologies, and have shown how the method can be applied successfully to SNOMED CT, a medical terminology whose use is widely mandated. Through the application of Rhetorical Structure Theory, the ontological content is organised into a discourse schema that allows us to generate appropriate discourse markers, pronouns, punctuation and layout, thereby making it more easily accessible to those who are not fully familiar with the ontology language in use. In its current form, the system is aimed at readers who care how the SNOMED is constructed – for example, those wishing to know the precise meaning of a given class. We believe there is no single solution to satisfying a wider range of user interests, and thus of text types. While we continue to work towards improving the output of our system, evaluating the output with non-ontology specialists, and testing our method with other ontologies and ontology languages, achieving fully fluent natural language is beyond the scope of our system. We are not at this point overly concerned by this limitation, as the need for clarity and transparency of ontologies is, we believe, more pressing than the need for fully fluent natural language descriptions.

Acknowledgments

This work has been undertaken as part of the Semantic Web Authoring Tool (SWAT) project (see www.swatproject.org), supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/G032459/1 to the University of Manchester, the University of Sussex, and the Open University.

References

Franz Baader, Ian Horrocks and Ulrike Sattler. 2005. Description logics as ontology languages for the semantic web, Lecture Notes in Artificial Intelligence, 2605: 228-248.

Charles B. Callaway. 2003. Integrating discourse markers into a pipelined natural language generation architecture, 41st Annual Meeting on Association for Computational Linguistics, 1: 264-271.

Catalina Hallett, Richard Power and Donia Scott. 2006. Summarisation and Visualisation of e-Health Data Repositories. UK E-Science All-Hands Meeting, pages 18-21.

Matthew Horridge, Nicholas Drummond, John Goodwin, et al. 2006. The Manchester OWL syntax. 2006 OWL: Experiences and Directions (OWLED'06).

Holger Knublauch, Ray W. Fergerson, Natalya Fridman Noy, et al. 2004. The Protégé OWL plugin: an open development environment for Semantic Web applications. International Semantic Web Conference, pages 229-243.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical Structure Theory: a theory of text organization, USC/Information Sciences Institute Technical Report Number RS-87-190 Marina del Rey, CA.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: toward a functional theory of text organisation, Text, 8(3): 243-281.

Chris Mellish, Donia Scott, Lynne Cahill Daniel Paiva, et al. 2006. A reference architecture for natural language generation systems, Natural Language Engineering, 12(1): 1-34.

Megan Moser and Johanna D. Moore. 1996. Toward a synthesis of two accounts of discourse structure, Computational Linguistics, 22(3): 409-420.

Hyun Namgoong and Hong-Gee Kim. 2007. Ontology-based controlled natural language editor using CFG with lexical dependency. 6th international, the Semantic Web, and 2nd Asian Conference on Asian Semantic Web Conference, pages 353-366. Springer Verlag Berlin, Heidelberg.

Richard Power, Robert Stevens, Donia Scott, et al. 2009. Editing OWL through generated CNL. 2009 Workshop on Controlled Natural Language (CNL'09), Marettimo, Italy.

Kent A. Spackman and Keith E. Campbell. 1998. Compositional concept representation using SNOMED: Towards further convergence of clinical terminologies, Journal of the American Medical Informatics Association: 740-744.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: an assessment, Natural Language Engineering, 14(3): 369-416.