

**Workshop ‘Toward the Interoperability of Language Resources’
held at Stanford University, July 13-15**

Report of Working Group 3 ‘Ontologies’

Co-chairs: Dunstan Brown and Andreas Witt.

Members: Michael Appleby, Östen Dahl, Helen Aristar-Dry, Arendse Bernth, Alexis Dimitriadis, Elizabeth Lowe, Michael McCord.

1. Introduction

The Working Group was tasked with answering a number of questions:

- Identify lacunae and difficulties with the different approaches to creating ontologies.
- Consider the complexity of format
- Change process
- GOLD (Farrar & Langendoen 2003) enable applications
- Integration of annotation and corpora
- What is required to standardize annotation?

2. Reality check

In the guidance notes for WG3 at least three strategies for establishment of a common set of concepts and associated terminology had been identified: unifying termsets, the construction of an ontology in a top-down manner (as in the case of GOLD), or the attempt to use existing data to construct ontologies bottom-up. Approaches to find the best annotation scheme were proposed in the 90’s, but it is not clear how successful they were.

There was some initial discussion of how strictly the term ‘ontology’ should be applied. For computer scientists the term would have to involve some notion of a domain of concepts and the relationships which hold between them so that it is possible to reason about these. At least the following distinctions can be made:

- term set
- super term set¹
- taxonomy
- ontology

At least the use of term-sets is necessary. Within the looser ontological grouping (term set ... ontology) the Leipzig Glossing Rules (LGR), for example, fit as a term set. There has to be some degree of inclusiveness when discussing ontologies, as elements which do not fit within the definition may be useful for helping to create them. There was some level of ontology-scepticism in that it was felt that the creation of a generally applicable, and accepted, linguistic ontology in its strict sense may currently be too ambitious an

¹ This distinction was suggested by Will Lewis for a customised term set within a COPE.

undertaking. However, it was generally felt to be a good thing to try and develop ontologies. The WG noted a variety of uses: machine translation, terminology, parsing, and at the other end, typology and linguistic theory. This spectrum of uses brought with it a fundamental question. We must not confuse ontologies with the truth, but for more typologically and theoretically oriented uses, there is a stronger requirement for approximation to the truth. What is more, it is not clear that an ontology could be created which would apply readily across the different uses.

From the foregoing the question naturally arose whether we need the advanced properties of ontologies. One answer might be that they would be required for reasoning about the properties of languages, and also about checking inconsistencies in the conceptual apparatus of linguistics. The advanced properties of ontologies could also be used for determining the substantive similarities and dissimilarities of different theoretical approaches. Of course, if this were ever to be possible, there would also need to be a change process in place to deal with the inconsistencies found etc.

3. *Change process*

One possible element of a change process is to have a registration authority (RA) and maintenance authority (MA) as in the framework described by Sue Ellen Wright (2004) for the Global Data Category Registry. For GOLD this type of approach could be too bureaucratic and there should perhaps be a more flexible system, whereby GOLD is sensitive to innovations which occur within particular COPES. As part of the change process there needs to be a clearer understanding of what the specific tasks of the ontology are: search or theorising.

Another aspect of the change process involves a division between the annotations and ontology. The mapping of the concepts and the termset is a narrowing. A refinement of the system could also involve confidence indexes for the mapping so that a user could have some idea of the degree of truthfulness that is being claimed.

It was also felt that the ontology works best where the typological work has already been done. This speaks in favour of some top-down element.

4. *Ontologies and why we need them*

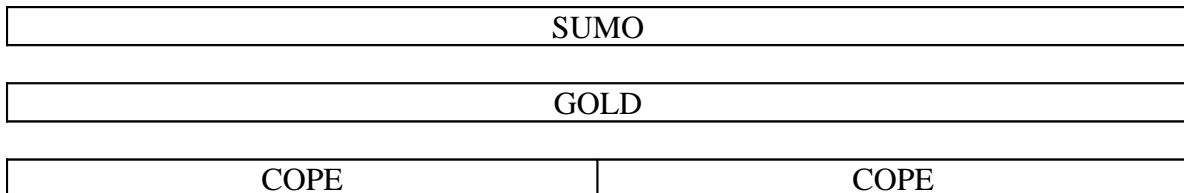
There was discussion of why ontologies are needed. A number of suggestions were made:

- Standardising markup across corpora.
- Discourse structure – anaphoric relations – knowledge about interaction/gestures
- Annotation layers could be mapped to multiple ontology layers.
- The tension between two totally different approaches to the ontology.
 - ◆ One linked in to the typologist side. Building toward something which is almost a language description.
 - ◆ People who have been using ontologies successfully have been adapting them to the task.

Seems like what ontologies are needed right now: information on selectional restrictions (the main use would be for a parser) and it would also help with anaphoric reference.

5. *Different concepts of how the ontology should look*

In figure 1 we see that GOLD is located between the general SUMO ontology and the more specific COPEs.



L1... L2... L3

Figure 1: Where GOLD is located

Specific descriptions of a languages then inherit information from GOLD via a particular COPE. It was noted in the discussion that the way SUMO defines certain linguistic concepts, such as word, or language, is wrong. So potentially making GOLD part of SUMO would lead to the discovery of some contradictions.

An alternative view of the ontology is one in which there are parallel structures.

References

Farrar, Scott & D. Terence Langendoen 2003. A linguistic ontology for the Semantic Web. *GLOT International*. 7 (3). 2003. pp. 97--100.

Wright, Sue Ellen. 2004. A Global Data Category Registry for Interoperable Language Resources.