**ROUND TABLE - Ontologies**

**Contribution from Belinda Maia (University of Porto)**

**Introduction**

When I was invited to take part in this round table, I made it an excuse to rush out and order Lakoff & Johnson and Lakoff & Núñez’s latest best sellers - *Philosophy in the Flesh* and *Where Mathematics Comes from*, together with other suggestive titles like – *How Brains make up their Minds*, with the idea of checking out the latest philosophical and neurological developments in the cognitive debate. However, while I waited for them to arrive, I set out to discover why this word *ontologies*, the plural of a term associated with philosophy, was of such interest to a debate in a conference devoted to corpora.

If one looks for the singular, *ontology*, in corpora, it will appear in texts that discuss the existence, or otherwise, of God, and other metaphysical debates. A quick search of the British National Corpus (BNC) will return plenty of examples that refer to the ideas of Hegel, Husserl, Heidegger, Lacan, Sartre, Levinas, Merleau-Ponty and others. The Portuguese corpus CETEMPúblico also produced several philosophy-related examples. The plural form, however, appears only once in both cases, which is itself indicative of the fact that this usage has developed over the last eight years, or since the BNC was finished and after most of the material in CETEMPúblico was published. Yet a search with the Google browser returned about 80,900 answers to the singular and 32,900 to the plural – so something must be happening.

*Ontology* leads one to [http://www.formalontology.it/index.htm](http://www.formalontology.it/index.htm) a site on which one can do a whole course on Ontology from the point of view of the philosopher. This site provides information and links on everything from a historical view of the subject, the philosophers who have discussed the existence (or otherwise) of God, to logic, artificial intelligence and…. Multi-lingual Information Management (MLIM) – a report published in April 1999 at [http://www.cs.cmu.edu/~ref/mlim/index.html](http://www.cs.cmu.edu/~ref/mlim/index.html) which provides the definition:

> **An ontology can be viewed as an inventory of concepts, organized under some internal structuring principle.**

Together with a whole report on the subject of MLIM in which ontologies play a very important part.

*Ontologies* led me quickly from the odd academic site to plenty of semi- or fully commercial sites dealing with knowledge engineering and offering software which promised help with organising any area of knowledge at all kinds of levels, from the
hierarchies of large areas like medicine or engineering - or large companies - to small and very specific sub-areas - or even parts of companies. The academic areas are more familiar to those of us taking part in the debate – the commercial side less so.

The fact is that, ontologies is a new term that has been chosen to go beyond conceptual systems or frameworks as people recognise that the data used to express concepts includes more than language – e.g. formulas, diagrams, and pictures. The recent use of the term can be seen from the fact that Wright & Budin Volume I (1997) includes a whole section on ‘Concept Systems’, and uses the term consistently throughout this volume, but their Volume II (2001) has a special information box on ‘Ontologies’ (p. 888-91). Another term, semantic networks, is used by computer scientists like Sowa (see reference) to refer to the formalisation of ontologies or concept systems for use in artificial intelligence and related projects. The Semantic Web (see reference) is developing on this latter set of principles.

Whatever one wants to call it, the systematic organisation of knowledge is not new. What is relatively new, perhaps, is the realisation that this sort of work has so many applications and is so relevant to organisational strategies today at so many levels of everyday life. And, of course, the Internet and its organisation – or lack of it – has made everyone even more anxious to find ways of systematising knowledge – and being seen to do so in an easily intelligible way.

At the more academic level, we are talking of creating methods of supplying lexical and semantic information to AI, knowledge engineering, machine translation, and various forms of language technology, in which the uses of corpora are just one aspect of the whole picture. In this context, terminology work becomes more than simply creating databases or glossaries for the use of those interested in a specialised subject.

Specialised language
Terminologists do more than just list lexical items for specific purposes. They are also expected to help provide a framework that permits the organisation and categorisation of the knowledge structures involved. Although many believe that the possibility of de-limiting terms and knowledge is wishful thinking on the part of those who endeavour to do this, one cannot ignore the fact that such systematisation provides an intellectual challenge few can resist.

Systematisation of knowledge has been going on for centuries. Nowadays we tend to think immediately of the Universal Decimal Classification, first created in 1895, and which was developed to provide ways of organising the documents librarians and others have to deal with. The UDC was obviously built on the accumulated theories of centuries, but it has become an international standard, and it is hardly surprising that many of the browsers on the Internet are built on similar lines.
Domain specialists are often very interested in providing more detailed classifications for their knowledge and some have developed quite complicated thesauri for this effect. Subject specialists in chemistry or medicine, for example, often rely on the categories of their thesauri, and then on the recognisable specificity of its technical terminology to do the rest. A chemical compound, for example, will generally be expressed by an internationally recognised term that restricts its usage, but this is not as simple as it may seem. For example, few people will object to something like carbon dioxide being considered a term in the area of chemistry, although we may also consider its special connotations in relation to the greenhouse effect, and possibly consider it a candidate for inclusion in a glossary on the environment.

The advantage of obviously specialised language is that one can make filters that allow one to make limited ontologies and to only search specific areas in any particular database. When this is feasible, the possibilities offered by hypertext and relational databases are indeed exciting. The role of ontologies here is to supply a well-constructed ‘architecture’ behind them, even if it is not immediately visible to the uninitiated. Making ontologies ‘comprehensible’ to artificial intelligence (AI) and machine translation (MT) is one of the big intellectual challenges today, yet not everyone involved in the endeavour is aware of the difficulties involved.

Standardisation institutions like ISO and DIN spend their time categorising and standardising all manner of things – and the terms used to describe them. Terminology itself is subject to classifications as can be seen from ISO/DIS standard 1087-1.2 (2000), where indications are given on how to account for semantic relationships between terms/words. The ISO/DIS standard 1087-1.2 on Terminology work -- Vocabulary -- Part 1: Theory and application (2000) gives various types of relationships between concepts and the terms that represent them. According to this document, a concept is seen as ‘a unit of knowledge created by a unique combination of characteristics’ that can be explained either by its ‘designation’ or representation of a concept by the sign which denotes it’ or a ‘definition’ or ‘representation of a concept by a statement which describes it’. Concepts are seen as individual and general, superordinate - sub-divided into generic and comprehensive concepts, and subordinate – sub-divided into specific, partitive and co-ordinate concepts. Apart from the usual considerations of the intensional and extensional nature of concepts and their characteristics, the standard also draws attention to other considerations like the relations between concepts – hierarchical (generic and partitive) and associative (sequential, temporal and causal) – as well as general aspects of terminology work.

This classical type of categorisation may help, but it does not resolve every problem in the organisation of knowledge. The section on concept systems in Wright & Budin (1997) gives plenty of examples of how one can progress from quite simple to multi-
dimensional structures of the type designed by Bowker (1995). These latter structures are difficult to represent in an easily understandable form but they attempt to show how we can possibly ‘structure’ knowledge. Traditional lexicographers and terminologists have always had to publish in the linear form of printed matter, but they were still aware of the difficulties behind constructing thesauri and special dictionaries in this way, as instructions like ‘see also’ demonstrate. Today this ‘see also’ is substituted by a link in hypertext or a database which takes us to the related area of knowledge without our necessarily being aware of how we got there.

The Systemization of General Language
The classification of what we like to call ‘general’ language has provided scholars with plenty of work over the centuries and anyone who has attempted any form of lexical analysis will be well aware of the problems of categorisation. Roget, the author of the famous Thesaurus, was a product of the 19th century when people could still dream on such ambitious lines, but people have been building on his categories ever since. At least, lexicographers and others interested in the lexicon have been doing their best, even if mainstream linguists have tended to ignore the problems raised during most of the 20th century.

Most people at the PALC 2001 conference do not need to be reminded of the problems involved in studying the lexicon. Let us only remember here the attempts that have been made with projects like WordNet, which provides a network of relationships between words, Framenet, which is an example of studying words within semantic frames, and other attempts at systematisation, such as Wordtree and Wordsmyth (for references see below). The results of all these projects has been a lot of valuable reference material and some useful theoretical constructs, but no one would claim that they represent a definitive classification of the entire English language, nor can they be considered the ultimate blueprint for the analysis of other languages, which may not allow for the same type of lexical organisation.

The point is that we all realise that general language, for all the dictionaries, thesauri and other reference material we draw up, is always changing and developing new meanings. We accept this fact as part of life, even when we are dedicated lexicographers. One of the driving forces behind corpora research has always been to find the many possibilities for word usage that had escaped previous searches. The subject of much current research is the way in which we can move forward from finding this knowledge to usefully systematising it.

Organising knowledge through Specialised Language
Although we accept the fluidity of general language, there is a certain collective delusion that sees specialised languages as sets of fixed terms that can be catalogued and used
without difficulty for various types of language situations, both in writing originals and
doing translations, and that these sets can be harnessed to the needs of machine (assisted)translation quite easily. Many people, including linguists, computational and otherwise,
believe one can draw up lists of specialised words in one language with links to their
equivalents in others. However, this is only fully possible with terms/words that refer
unambiguously to a specific reality, such as in the case with chemical compounds or
botanical species – and, even then, the situation is dynamic, and science discovers and
adds to the possibilities. After that, the use of words as terms to describe certain concepts
is largely a question of agreement, which may be officially standardised, but which is
more likely to be a question of social conventions that, in their turn, may vary according
to local factors that range from a geographical variety of the language (e.g. American or
British English) to a question of disagreement between rival academics or commercial
companies.

One fact that is very important, but which tends to get ignored in the middle of the
various arguments, is that much of ‘specialised’ language is made up out of general
language words that have been pressed into use to mean ‘x’ in context ‘y’. Yet anyone
who uses the Internet as a source of information realises too well how simple word
searches so often lead one in directions one had no intention of following, or even knew
existed. As one gets used to the browser, one learns to add other, more specific, words to
our requests that will help reduce the many thousands of possibilities to – at least a few
hundred!

Since most of the people discussing ‘ontologies’ nowadays are essentially interested in
facilitating information retrieval from the Internet and other digital resources, the
problem of general versus specialised language is a key factor. The researchers cannot
help but come up against the perennial problems of semantics and the way automated
systems of AI, MT and browsers cope with world knowledge. If normal human beings
have to be educated to sort out the relevant from the irrelevant in these situations, how
much more complicated it must be for the machine. The semantic networks of the kind
proposed by ONTOLINGUA and Semantic Web (references below), will no doubt go a
long way to providing help, but the computer scientists need to learn a few lessons from
the frustrations of linguists and philosophers over making language ‘stand still’ if they
are to be even partly successful.

Concluding remarks
Everyone involved in the construction of ontologies will have to become more aware of
the problems raised not just by the lexicon, but also by the several levels of analysis
required of a text as whole. They will have to devise ways of coping with polysemy,
connotation, and ambiguity – not to mention the metaphors that Lakoff et al argue show
how language reflects our physical and social connection to the world, and how we
construct our language out of our experience. As Temmermann (2000) points out, these metaphors also exist in the terms we use to describe the concepts of special subjects.

Plenty of useful work has already been done, but much of what is good is still very general in nature. Certain areas of specialisation are well categorised and documented, but tend to exist in isolation from any more general framework and in formats, and according to theoretical premises, that do not allow compatibility to be achieved with other areas. The Big Dream of a coherent systematisation of knowledge is still a dream, and it will need considerable collaborative effort on the part of computer scientists, domain specialists, linguists and others if progress is to be made beyond the experimental stage that is the present state-of-the-art.

**General Bibliography**


Electronic information
Please note: All sites last accessed 26/02/02
British National Corpus at: http://sara.natcorp.ox.ac.uk/
CETEMPúblico at: http://www.portugues.mct.pt/
DAML ontology library at: http://www.daml.org/ontologies/
DARPA Agent Markup Language (DAML) Program at: http://www.daml.org/ontologies/
Eurowordnet: http://www.hum.uva.nl/~ewn/.
Formalontology edited by Raul Corazzon at http://www.formalontology.it/.
Framenet at: http://www.icsi.berkeley.edu/~framenet/
Multi-lingual Information Management (MLIM) at:
http://www.cs.cmu.edu/~ref/mlim/index.html
ONTOLINGUA at: http://www.ksl.stanford.edu/software/ontolingua/
Semantic Web at: http://www.semanticweb.org/
Universal Decimal Classification (UDC) at: http://www.udcc.org/
Wordnet at: http://www.cogsci.princeton.edu/~wn/
Wordtree: http://www.wordtree.com/
Wordsmyth: http://www.wordsmyth.net/home.html
World Wide Web Consortium (W3C) at: http://www.w3c.org/