



## Contents

<b>Formal representation of concepts: The Suggested Upper Merged Ontology and its use in linguistics</b>	
<i>Adam Pease</i>	1
<b>Index of names</b>	13



## **Formal representation of concepts: The Suggested Upper Merged Ontology and its use in linguistics**

*Adam Pease*

We believe that human language can be meaningfully mapped to a formal ontology for use in computational understanding of natural language expressions. We have created a formal ontology in a first order logic language called the Suggested Upper Merged **Ontology (SUMO)** (Niles and Pease 2001) composed of roughly 1000 terms and several thousand formal statements about those terms. We have also created an index (Niles and Pease 2003) linking all 66,000 noun synsets, 12,000 verb synsets and 18,000 adjective synsets from **WordNet** (Fellbaum 1998) to terms in SUMO. The links have been made to version 1.6 and then ported to 2.0. The links were made one synset at a time, by hand, over the course of a year, rather than by an automatic process. We are working on using this index for **natural language understanding** tasks.

In this chapter we describe SUMO, its WordNet mappings and use in natural language understanding and inference. We also contrast SUMO to other ontologies.

### **1. Other ontologies**

The only other publicly available axiomatized ontology is the Descriptive Ontology for Linguistic and Cognitive Engineering (**DOLCE**) (Masolo et al. 2003).

SUMO is a **formal ontology**, in that it is not simply a collection of terms and English definitions, but rather a fully axiomatized ontology, with definitions for terms provided in first order logic. Note that although the terms in SUMO were initially created as English labels, they have no inherent linguistically dependent content. The labels are simply convenient mnemonics for the human ontologist, much like the names of variables in procedural software code. Each term name could be replaced with a meaningless unique code and still retain its meaning, since the meaning of a term is given solely by its formal **axioms**. SUMO is very different from **taxonomies**, lightweight

ontologies developed in frame systems like **Protege**, or early versions semantic web languages such as **DAML+OIL** language, all of which lack entirely or have only very restricted axioms, which limit the use of such representations for inference. The most notable proprietary formal ontology is **Cyc** (Lenat 1995), for which the top one percent of the taxonomy, but not the rules, has been released publicly, and called OpenCyc.

DOLCE has a similar purpose and business process to SUMO in that it is a free research project for use in both natural language tasks and inference. DOLCE has been carefully crafted with respect to strong principles. It is reported that DOLCE is also being mapped to a portion of WordNet, although this content has not been released at the time of this writing. DOLCE is described by its authors as an “ontology of particulars” which the authors explain as meaning an ontology of instances, rather than an ontology of universals or properties. DOLCE does in fact have universals (classes and properties), but the claim is that they are only employed in the service of describing particulars. In contrast, SUMO could be described as an ontology of both particulars and universals. It has a hierarchy of properties as well as classes. This is a very important feature for practical knowledge engineering, as it allows common features like transitivity to be applied to a set of properties, with an axiom that is written once and inherited by those properties, rather than having to be rewritten, specific to each property. Other differences include DOLCE’s use of a set of meta-properties as a guiding methodology, as opposed to SUMO’s use and formal definition of such meta-properties directly in the ontology itself. Currently, DOLCE is much smaller than SUMO, with 103 terms, and a similar number of axioms, and lacking such items as a hierarchy of process types, physical objects, organisms, units and measures, and event roles.

## **2. Mapping WordNet to SUMO**

WordNet is described in more detail in Fellbaum (this vol.), but for the purposes of this effort, we can describe WordNet as a very large electronic **dictionary**, where synonymous word senses are grouped together and called “**synsets**”. When the mapping to WordNet was performed with WordNet version 1.6, there were approximately 100,000 synsets. There are four primary data files, covering nouns, verbs, adjectives and adverbs, (called NOUN.DAT etc.) in which each set of synonymous word senses is accompanied by a brief

definition and example usage. We went through each file, looking at each synset and attempting to find the formal SUMO term that most closely captures the meaning of the synset. These mappings were completed and released in 2002.

Our initial concentration has been on simple mappings, where there is a direct correspondence between an English noun and a class membership statement in **logic**, or where a verb can be mapped directly to a SUMO event type. Note that in the examples below, words that are the object of discussion in this paper appear in an italic font. Words that are reified terms in the SUMO and logical **formulas** are given in monospaced bold font. For example, “The dog bites the man” shows both simple noun mappings and a simple verb mapping:

```
(exists (?D ?M ?E)
  (and
    (instance ?E Biting)
    (instance ?D Canine)
    (instance ?M MalePerson)
    (agent ?E ?D)
    (patient ?E ?M)))
```

The verb *bite* maps directly to the statement that there is an event of type *Biting*. The noun *Dog* maps directly to the statement that there is an instance of the class *Canine* that participates in the event etc. Note that we have adopted a basically Davidsonian approach to action semantics (Davidson 1967). One of the challenges of that approach is in an over-generalization to all verbs. Many verbs refer to states rather than **actions**, and only actions are appropriately modeled with the Davidsonian approach. **Copula** sentences are treated differently for example.

The above example shows a simple mapping from a word to a logical concept. It is the case however that many possible sentences do not have this sort of simple mapping to concepts in an ontology. The next simplest case is where a word does not have a reified equivalent in the ontology, primarily due to the need for clarity and simplicity in the ontology, independent of the concerns of natural language understanding issues. Note that this is an important factor, since we intend that the ontology be appropriate for theorem proving tasks, and each of the terms in the ontology that are used to state the formal equivalent of English sentences must have an associated logical definition. Those terms and definitions must be stated in a manner that makes logical inference possible, and efficient.

There are several other considerations that may result in an ontology not having a term that directly corresponds to a word in a **lexicon**. **Logic** has a great degree of flexibility for expressing the same information, much in the same way that human natural languages do. However, while people are adept at understanding the similarity be-

tween semantically equivalent but syntactically different statements, machines are not. Proving the equivalence of logical formulas is a process that is not guaranteed to terminate in the general case. For that reason, it is highly beneficial for a **knowledge engineer** to state semantically equivalent statements in syntactically identical ways.

Another consideration is the reusability of knowledge-based content. In (Pease et al. 2000), we discussed the need for having *compositional* expressions in order to facilitate the reuse of knowledge, as well as for efficiency concerns. Compositional expressions are those that state a concept by employing a set of more basic expressions, rather than encapsulating a single notion in a reified term. This factor may also encourage the use of expressions rather than single terms to express a particular concept. In “The man began walking to the market” there is no need for a *BeginWalking* concept. The information contained by that phrases is only the narrative **temporal** information that the following text is likely to refer to events after the beginning and before the end of the walk.

Finally, many expressions in natural languages may have very minimal semantic content. In many contexts, the notion of starting or continuing to do something is equivalent to simply performing the **action**. There is no need to reify the concept of a starting action, although there is the need to express parts of **processes** or relate a process to its starting time or the time of another process.

A typical case of the lack of direct correspondence between English words and terms in the ontology is where the ontology models roles that agents play as relationships between the agent and a type of role, rather than reifying a subclass of agent filling the role. For example, “The pilot lands the plane” results in

```
(exists (?P ?PL ?E)
  (and
    (instance ?E Landing)
    (instance ?P Person)
    (attribute ?P Pilot)
    (instance ?PL Airplane)
    (agent ?E ?P)
    (patient ?E ?PL)))
```

We are also dealing with more complex relationships where an entire phrase forms a pattern that has an equivalent template structure for a logical expression.

While many nouns map to instances of reified classes, and many verbs map to instances of *Process* types, the situation for mapping adjectives is more often problematic. Some adjectives can map to instances of reified classes, for example *dying* → *Death*, and *potential* → *Possibility*. Many adjectives map to a particular class of subjective assessments in SUMO, which can be related to the noun being modified by the relation *attribute*. Some examples are *absolute* and *pure*. A more detailed and specific ontology could profitably extend SUMO’s existing set of **Attributes**. Some adjectives map to implied values in relations with *spongy*, *thirsty* and *repellent* all expressing a certain value for the second argument of the SUMO capability

relation. A spongy object is something that is relatively easy to compress. That is, it has the capability of participating as a **patient** in a Compression process. Some **adjectives** map to logical values for a sentence that may be implicit in the common sense information related to the sentence being interpreted. *Faulty* and *unfaithful* both relate to the falsehood of a particular proposition. Some adjectives indicate that a process has occurred at some point in the past to the noun being modified. *Mounted*, *paneled* and *studded* all imply the occurrence of a **Putting event**.

A topic suggested for the workshop that this volume is partially a result of (Schalley and Zaefferer 2003) is another case of this issue reflecting literal word correspondence problems instead of problems with the correspondence between a language and an ontology, or one language to another language. The stated problem is that *stellen*, *setzen*, *legen*, and *stecken* do not each simply map to a single English word, much as the single English word ‘pilot’ does not map to a single term in the ontology above. English, however, is rich enough to communicate the notion of putting an object in an upright position without having to have a single word for it. In the same way, an ontology can be rich enough to express a concept like ‘pilot’ without having to reify that notion as a named term in the ontology.

Natural languages and human constructed ‘languages’ like ontologies both exhibit a certain kind of organic growth in response to the pressures of use, although on a different scale. One should neither expect nor require them to parallel each other, nor should the degree of coincidence be a mark of quality, much in the same way that one would not assess the value of a human language on the basis of its similarity to a romance language for example.

### 3. SUMO and **domain ontologies**

SUMO consists of eleven modules with the dependency structure given in Figure 1.

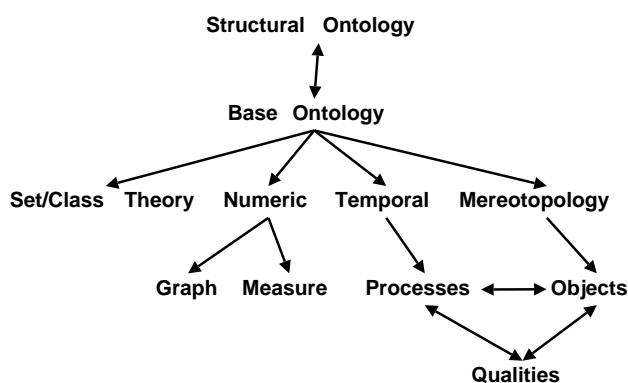


Figure 1. Hierarchy of SUMO theories

The Structural Ontology consists of fundamental relations, with associated formal definitions, such as `instance`, `subclass` and `disjoint` that are used to define most other concepts in SUMO. The Base Ontology consists of fundamental classes such as `Abstract` and `Object` from which all other classes `inherit`, as well as classes of relations, such as `TransitiveRelation`, and very frequently used relations which are instances of `CaseRole(s)`. The Numeric ontology contains common numeric operators such as `AdditionFn`. The Set/Class Theory contains set operators such as `UnionFn` and `IntersectionFn`. The Graph ontology has basic graph theoretic notions such as `DirectedGraph`. The Measure ontology consists of classes of measures such as `LengthMeasure` and `TemperatureMeasure` as well as specific units such as `Meter`. The Temporal ontology contains notions such as duration, as well as the standard 13 Allen relations (Allen 1984) (which are: `before`, `meets`, `overlaps`, `during`, `starts`, `finishes` and their inverses plus `equals`). The `Mereotopology`, or theory of parts and places contains relations such as `partiallyOverlaps` and `height`. The ontology of Processes is quite extensive, consisting of roughly 175 process types, some of which are shown in Figure 2.



Figure 2. Top two levels of the Process hierarchy

The ontology of Objects contains a biological taxonomy with associated formal definitions, along with concepts such as `Nation`. The Qualities ontology has concepts such as compass directions (e.g. `North`), `SocialRole` and `NormativeAttribute(s)`.

There are a number of ontologies that extend SUMO. Together, they number some 20,000 terms and 60,000 axioms, making the SUMO family the largest publicly available formal ontology. These additional ontologies include a mid-level ontology comprising some 5000 terms with associated formal definitions, for concepts that are still quite general, but are arguably too low-level for SUMO itself. Domain ontologies include theories of geography, terrorism, political systems, economic systems, as well as the GOLD linguistic ontology described in Farrar (this vol.).

#### **4. Phrases**

We have recently proposed (Pease and Fellbaum 2004) creating a new corpus of phrase elements with mappings to the template logical expressions that are entailed by each phrase. A good example of what is required in this area is the semantics of the English **possessive**. “John’s car” refers to a car that is owned by John and should be represented with the SUMO relation *possesseses*. “John’s nose” however refers to a physical part of John and is best represented by the SUMO relation *part*. “John’s company” is ambiguous, and while it can mean a company that is owned by John, it more often represents a company of which John is an employee, which would use the term *member*. In each case, the relation used is dependent not only on the grammatical construction, but on the class membership of the grammatical elements. A **possession** relation between an Agent and a BodyPart entails *part*. A relation between an Agent and an Object (other than an Organization) entails *possesseses*.

Another case of semantic **ambiguity** involves **prepositions**. One gets “in a car” but “on a bus”. However, in both cases, the subject is inside the transportation device. The determining factors for the use of “on” is whether the vehicle is either not enclosed, or is a group transportation vehicle. Only enclosed, personal transportation vehicles, such as a car, require the use of “in”. Any system that attempts a deep semantic translation of such phrases must recognize the class of vehicle in order to generate the correct semantics of the spatial relation between subject and object. A somewhat more straightforward issue with “on” is its metaphorical employment with regards to times. “John arrived on the patio” meaning that his destination was the patio. “John arrived on Monday” means that the time the action was completed was Monday. The first sentence relates the event to the location with the SUMO *destination* relation. The second sentence relates the event to its temporal overlap with Monday using the function *EndFn*.

Human language is not regular with regards to its employment of surface linguistic features for significantly different semantic intent. Any software system that attempts to generate a deep semantic equivalent for natural language will need a large corpus of language-specific rules that map surface features to deep semantics. Additionally, such a system will need a large ontology which specifies the semantics of the terms used in the target representation.

## 5. Translation

We are also doing some early and arguably simplistic exploration in using SUMO for language translation. We can translate from many simple kinds of English sentences into logic. We are also able, using simple templates, to translate logic into somewhat awkward natural language sentences in English, Chinese, Czech, German, Hindi and Italian. Because the ontology serves as a “hub” for language translation we have a situation that is simpler than if we had to create specific structures for translations between any particular pair of human languages. “Understanding” language is much harder than the problem of generating merely grammatical language (albeit that problem of generating natural and idiomatic language may be equal in difficulty). A colleague has implemented a browser (Sevcenko 2003) that employs the templates to express statements from SUMO in the languages for which we have translation templates.

## 6. Language understanding

We are currently creating applications that use a restricted English language grammar as the input form for generating knowledge expressed in logic (Pease and Murray 2003). Our expectation is that these applications will test our hypothesis that we can develop a grammar and translation rules that are powerful enough to allow humans to express most thoughts, while not encountering the overwhelming problems found in machine understanding of unrestricted human language.

Our approach is similar in spirit to the tradeoff made by the Palm Pilot as opposed to the Apple Newton. The Apple Newton attempted to recognize unrestricted cursive English handwriting. Even ten years later, this problem is still too hard for a handheld computer. The Newton simply did not work enough of the time to be useful. The Palm Pilot shifted the hardest problems for the machine on the human user. In exchange for a relatively small and easily learned change in behavior a hard problem became easy, and the Palm is able to function at a high enough rate of handwriting recognition to be useful and commercially successful.

As specific examples of how we approach this tradeoff, we require user assistance for word sense disambiguation in many cases. Metaphor is not allowed. Anaphoric reference is decided by a simple algorithm. Verbs must be present tense and nouns must be singular. The use of modifiers and dependent clauses is limited to simple cases. We have started with an extremely simply constructed grammar, which we handle with deep understanding, and are building up the degree of sophistication little by little. This is in contrast to a typical approach to understanding which attempts to cover full natural language at a shallow degree of understanding. While our approach is not suitable for existing documents, it is perhaps more suitable than

other approaches for understanding commands or assertions made by a user through a software interface.

## 7. Inference

We are currently conducting experiments with formal, logical inference. While the primary goal of the effort is to develop a powerful framework for formal reasoning, a second goal is to support information retrieval and question answering with a natural language interface. This framework allows us to test our hypotheses about translation of language to logic and more specifically about efficient inference on automatically generated logical content.

First order inference presents many computational challenges. A common and general approach to addressing performance concerns is to trade space for time. That is to say, by using greater amounts of memory, it is often possible to improve the time it takes for a computational process to return results. Specifically, we are pre-computing various inferences and storing them in our knowledge base. By storing certain kinds of results, most often those having to do with reasoning about class membership, we have seen speedups of several orders of magnitude. The degree of speedup can mean the difference between tractable inference times, and no results at all.


By performing formal reasoning on logic expressions generated automatically from natural language, we will be able to test our hypotheses about language semantics in a way that is objective and verifiable.

## 8. Conclusion

We intend that the products mentioned in this paper be used as widely as possible. Most have already been released under the GNU open source license. All products will be released to the research community as they mature. We encourage other researchers either to verify our results, or develop innovative approaches to applying them differently. We equally welcome experiments that refute our hypotheses and help point us in the right direction.

## References

- Allen, James  
1984           Towards a general theory of action and time. *Artificial Intelligence* 23: 123–154.

- Davidson, Donald  
 1967 The logical form of action sentences. In *The Logic of Decision and Action*, Nicholas Rescher (ed.), 81–95. Pittsburgh: University of Pittsburgh Press.
- Farrar, Scott  
 this vol. Using ‘Ontolinguistics’ for language description.
- Fellbaum, Christiane  
 this vol. The ontological loneliness of verb phrase idioms.
- Fellbaum, Christiane (ed.)  
 1998 *WordNet: An Electronic Lexical Database*. (Language, Speech, and Communication.) Cambridge, MA/London: MIT Press.
- Lenat, Douglas B.  
 1995 CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38 (11): 33–38.
- Masolo, Claudio, Stefano Borgo, Aldo Gangemi, Nicola Guarino, Alessandro Oltramari, and Luc Schneider  
 2003 *The WonderWeb Library of Foundational Ontologies Preliminary Report*, WonderWeb Deliverable D17. Version 2.1 dated 29–05–2003. ISTC-CNR1 Technical Report. **TODO:PUBLICATION\_LOCATION&PUBLISHER.**
- Niles, Ian, and Adam Pease  
 2001 Toward a Standard Upper Ontology. In *Formal Ontology in Information Systems. Proceedings of the 2nd International Conference (FOIS-2001)*, Christopher Welty and Barry Smith (eds.), 2–9. New York: ACM Press.   
 2003 Linking lexicons and ontologies: Mapping WordNet to the Standard Upper Merged Ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, **TODO:EDITORS**, 412–416. **TODO:PUBLICATION\_LOCATION&PUBLISHER.**
- Pease, Adam  
 2000 *Standard Upper Ontology Knowledge Interchange Format*. Web document <http://suo.ieee.org/suo-kif.html>. This is a condensed version of the language described in Genesereth Michael R. (1991). Knowledge interchange format. In *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning (KR-91)*, James Allen, Richard Fikes, and Erik Sandewall (eds.), 238–249. San Mateo: Morgan Kaufman. See also <http://logic.stanford.edu/kif/kif.html>.
- Pease, Adam, Vinay Chaudhri, Fritz Lehmann, and Adam Farquhar  
 2000 Practical knowledge representation and the DARPA High Performance Knowledge Bases project. In *KR-2000: Proceedings of the Conference on Knowledge Representation and Reasoning, Breckenridge, CO, USA, 12–15 April 2000*, Anthony G. Cohn, Fausto Giunchiglia, and Bart Selman (eds.), 717–724. San Mateo, CA: Morgan Kaufmann.
- Pease, Adam, and Christiane Fellbaum  
 2004 Language to logic translation with PhraseBank. In *Proceedings of the Second International WordNet Conference (GWC 2004)*, Petr Sojka, Karel Pala, Pavel Smrz, Christiane Fellbaum, and Piek Vossen (eds.), 187–192. Brno: Masaryk University.

Pease, Adam, and William Murray

- 2003 An English to logic translator for ontology-based knowledge representation languages. In *Proceedings of the 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China*, TODO:EDITORS, 777–783. TODO:PUBLICATION\_LOCATION&PUBLISHER.

Schalley, Andrea C., and Dietmar Zaefferer

- 2003 Workshop on *Ontological Knowledge and Linguistic Coding* – Workshop at the 25th Annual Meeting of the German Linguistics Society (Deutsche Gesellschaft für Sprachwissenschaft), Munich, February 26–28, 2003.

Sevcenko, Michal

- 2003 Online Presentation of an upper ontology. In *Proceedings of Znalosti 2003, Ostrava, Czech Republic, February 19–21, 2003*. TODO: ANY\_EDITORS,PAGE-RANGE,PUBLICATION\_LOCATION&PUBLISHER?.  
See also <http://virtual.cvut.cz/kifb/en/>.



## **Index of names**

Allen, James, 6

Davidson, Donald, 3

Farrar, Scott, 7

Fellbaum, Christiane, 1, 2, 7

Lenat, Douglas B., 2

Masolo, Claudio, 1

Murray, William, 8

Niles, Ian, 1

Pease, Adam, 1, 4, 7, 8

Schalley, Andrea C., 5

Sevcenko, Michal, 8

Zaefferer, Dietmar, 5

