# Ontological Help for a Lexical Semantics of Basic English 

Jim Cunningham ${ }^{1}$


#### Abstract

This paper began as an exercise to see how lexicon and ontology might combine to help underpin human-agent communication, using Ogdens Basic English as a case study [11]. It became an attempt to clarify, through exemplification in Basic English, the relationships between ontology, lexical semantics, and common sense reasoning, reporting some tentative steps towards an executable semantics.


## 1 INTRODUCTION

It has long been recognized that commonsense categorial distinctions can be be used to help disambiguate word senses, as well as case and referential ambiguities. Manifestations of the idea occur in Katz \& Fodor 1964 [6], Wilks 1975 [21], and Dalgren \& McDowell 1986 [3], re-appearing in recent attempts to exploit WordNet for ontological annotation in the semantic web (Sanfilippo et al. 2005). But there is no standard for ontological categories, and even if one were to emerge, an ontological category alone does not give meaning to a word; it can help with so-called semantic parsing, but it is less clear how such information is to used for common sense reasoning.

There are many potential sources of world knowledge that might help a software agent comprehend natural language; these include formal logical systems, dictionaries, encyclopaediae, and less structured corpora, linguistic resources such as WORDNET [9, 10] and Framenet [1], "Ontologies" of commonsense knowledge such as CYC [7], SUMO [13], and WonDERWEB [8], and many other attempts model commonsense reasoning through naive physics, commonsense psychology, and domain-specific, concrete "ontologies" for web services.

But the sheer scale of attempts to codify commonsense knowledge make its integration with natural language a massive project. Stephen Pinker's recent book [15] emphasizes the degree to which language seems to encode its own model of the world. AI researchers and Computer Scientists, like engineers through history, have had to use simple but imperfect models for thinking and invention. An investigation for the concise 850 word BASIC lexicon is a more modest task than addressing the full language. One retrospective view of this paper is that it is part of an endeavour to decode the world model of BASIC.

## 2 BASIC ENGLISH

As a controlled, or restricted language, Ogdens Basic English (BASIC) is a historical curio, proposed as an alternative to Esperanto, but still supported by a small active interest community. BASIC was designed to be adequate for everyday language, drawing on experience of teaching English to students from the Far East. From the early promotional translations from ordinary English into BASIC, it

[^0]is evident that its semantic scope for everyday use is fairly complete, even if the original lexicon can be considered in need of revision for the computer age. (There are 50-word lexical extensions for special domains).

The BASIC lexicon is unusually small because all but 18 syntactic verbs have been eliminated. The language emphasises the use of particles to supplement these verbs in providing expressiveness (as in: come in, go out, keep off etc.). A significant proportion of the words in the lexicon could be used as a verb in normal English, but the verbal interpretation in BASIC needs to be extracted from nominal or adjectival forms (as in: its sudden growth, the painted wall, etc.). The language is less concise than normal English, but it is not unnatural. It is obviously harder to find synonyms (so a BASIC word like thing is much over-used), and a specific sense may need to be defined implicitly with adjectival phrases, or by explicit extension. But authoring in a syntax-controlled language is arguably easier than in a sense-controlled language.

For a modern linguist the informal grammar and absence of overt attention to word sense may make BASIC seem naive. But this itself, together with the modest lexicon, make it an interesting model for experimentation. Its relatively smaller scale facilitates practical consideration of the lexical semantics, where not only category, but both function and lexical meaning can be addressed.

## 3 SEMANTIC TYPES

One starting point for this work comes from Schneider (2001), [18, 19] where commonsense reasoning about sentences over a small but artificial lexicon was facilitated by incorporating thematic roles in a semantic ontology for natural language and distinguishing these roles from the types of individuals, taking into account the recommendations of Guarino and Welty (1998, 2002)[4, 5]. The top level categories were then as illustrated in figure 1.

Thematic roles [12] are only a partial categorisation of the semantic ingredients of a sentence. However the subclassification of substances in figure 1 also allows for the distinction between agentive (person) and objective (body), between stative (state) and dynamic (event) properties, and between an agent's internal (private) knowledge and external (public) knowledge. Thus a typed composition of logical properties could support a small experimental demonstration of semantic parsing and query answering by an electronic agent. Integration with more general commonsense knowledge or extension to a full lexicon needs much more insight. While extensions of our experiment to incorporate further types and roles for legal reasoning did suggest that the approach could be helpful, it also introduced the problem of managing the multitude of morphological derivatives (own, owner, ownership etc), something which one would like to be automatic in use of a lexical ontology ${ }^{2}$.

[^1]Using quite different approaches to gathering lexical semantic information, both FrAMENET and WORDNET have also built ontological categories from lexical data, but neither result appears to fit well with independent categorizations of commonsense knowledge. Scheffczyk, Pease and Ellsworth [17] report intricate problems in seeking to facilitate semantic parsing and common sense reasoning by supplementing the shallow semantic types of FRAMENET by the ontological domains of SUMO. One might perhaps expect the use of ontological categories to distinguish word sense to be more successful exercise with an ontology like DOLCE, that part of WONDERWEB supposedly intended for Natural Language, but Prévot, Borgo, and Oltramari [16] report inconclusive results, with the need to re-work lexical data which WordNet has provided by peer resourcing.


Figure 1. Top-level divisions in Schneider 2001

### 3.1 Word sense disambiguation

Amongst the many unresolved issues in establishing a set of semantic types for language, one problem stands out: for an ordinary user of language many of the different word senses of a conventional dictionary go unnoticed, let alone the fine-grained sense distinctions of WordNet. This suggests that the need for an activity such as that by Peters et al. [14] on automatic sense-clustering is an artifact of the technology being pursued for classifying word senses rather than intrinsic in the problem of capturing a word's lexical semantics. But what positive evidence is there for such a claim? In part it is inconclusive experimental evidence of utility from including semantic types like events and states, and linguistic classifications such as thematic roles to provide a more typed semantic composition. In part it
is specific BASIC examples which have arison, such as the many dictionary senses of the verb make(construct, create), as in make a wall, or derived from this sense. Allowing for exceptional hard-to-deduce deduceable coloquialisms such as on the make, made of money, make sail, the majority of listed distinctions appear to be a commonsense, or sometimes default consequent of object type and context. Simple examples are make a story (a process), make coffee (product of a process), and make a day (impossible, so schedule one). This basic semantic type analysis then leaves the common but less obviously derived sense of force, as in Jill made Jack (do it). But here the object is an event to be composed, and the agentive Jill has imposed her will on Jack by making the event happen.

While verbal sense ambiguity may need many cases with a BASIC verb like make, there are well known cases of regular noun sense ambiguity that can be determined by grammatical context. Example in BASIC include:

- the distinction between a substance and a quantity of it (made of chalk versus take some chalk).
- the distinction between a substance and its qualities (made of chalk versus the look of chalk).
- the distinction between a substance and a man-made thing which consists of the substance (made of chalk versus bit of chalk).
- the distinction between a substance and a thing which is identified by that name (made of iron versus make flat with the iron).
- the distinction between a thing and its place for some event (pick up the chalk versus come on the chalk).
- the distinction between an event and a time it determines (sound of the bell versus come at the bell)

It would clutter a classification of a lexicon by type to insist that a word have a separate entry for each such ambiguity, even if duplicate entries might help in some cases. Substances are known because they exist in our environment and have uses which depend on their characteristics; their qualities distingush them. In contrast, countable entities like buildings and meals are typically composite, and in the case of a material thing like a building it has parts which are made of substances. One can assume and require that the rules for composition of meaning will disambiguate word use by context, possibly with the aid of supplementary ontological relations such as use, part or product to assist in reasoning about the meaning, But if all else fails, a agent may enter into a dialogue to clarify meaning.

### 3.2 Semantic types for BASIC

Given the discussion above the present work has reverted first to a more elementary challenge: discover the semantic types that are distinguished in Basic by building a (loose) is-a taxonomy of subsorts - something is an $A$ if it is a B or a C or a $\ldots$. The taxonomy is loose, in that the "or $a$ " is not exclusive unless indicated (through negation). The (unfinished) experiment is to test the effectiveness of the semantic sorts as a basis for a lexical semantics with commonsense distinctions. (Here the word sort is used because it appears in Basic, whereas type does not, indeed the meta-language is intentionally from Basic). The top-level divisions of the Basic sort taxonomy are at present as indicated in figure 2, where a substance is a naive classification of the constituents of material things, and numerables, becomes a classification of both countable material things - living, man-made, and more abstract (but still countable) notions. So already
there is a computational notion which reflects the semantic idea underlying the grammatical distinction between mass and count nouns.


Figure 2. Top-level divisions of the loose ontology for sorts of BASIC "Thing"

The tables appended to the text of this paper are fragments of the initial extensions of figure 2, giving some idea of the language and interim classification categories to be validated through experiment. There is no claim of faithfulness to any particular perception of Basic; its service here as an experimental word base will include inaccuracies as well as deliberate simplifications. Not only are inflections ignored in the sort taxonomy, but lexical derivatives are too. Both may be considered to have ontological significance: for example, a plural is also a sort of aggregation, while Ogden's use of nominalized verbs (like addition) and adjectival participles (painted) may be considered, respectively, to objectify a process, and to extract a state as the product of a process. One can even presume that Ogden considered that such sleights of meaning were less of a problem for human learners than more formal syntax. It is assumed here that a reasoning engine which can deal with the language should also deal with them. But then any sort ontology for a language should also deal with relations between underlying sorts, The distinction between adjectival and pronominal uses of the word own in standard English are a further similar reminder that the grammatical language is semantically determined.

The use of BASIC words and phrases for the names of divisions, the meta language of ontological categories and relations is both an oddity and an illustration of the expressiveness of the language. So the word qualities is used rather than attributes, and thing rather than object or entity. If BASIC vocabulary is minimal, but broad in application, metalinguistic uses should part of language, and any new ambiguity introduced by doing so needs to be dealt with. Thus we have a num(b)erable thing rather than an object or entity, a madething rather than an artifact, a condition rather and than a state and a use rather than a role. The selection of BASIC words for such meta level purpose is not without frustration. (One instance is the absence
of a better word than example to convey the sense of set member, element or item). However, as with other technical language, it is assumed that definitions can be made and axioms postulated. So a "math" thing can be a subdivision defined by its instances.

More seriously, the conceptual limitations of the sort taxonomy reflect the inadequateness of expression in such a bounded language. It becomes impractical to succinctly represent the evolution of scientific knowledge. The word meat appears in table 1.1 as substance which is solid, soft food, but not fruit. This may be intuitive, but it is arbitary. It is explicated for practical purposes by associating further qualities with the substance, most essentially that it is animal muscle. (The word flesh is not available). So for consistency, table 1.2 includes muscle and fat as parts or detail of an animal, along with words like body, throat, neck and nerve. But these do not fit comfortably together, they are different perspectives, and this is represented by the different lists. Such detail is clearly knowledge rather than a linguistic taxonomy, but on the grounds that a part of a material thing is also a material thing, one can introduce a further subdivision of animal, so that the parts of a head may be distinguished. It is not clear that there is similar detailed structure for parts of a building and little effort has been made in analysis of the subdivisions of man-made artififacts (house-thing, etc) because the vocabulary is too out of date for either a faithful model or a relevant one for today's English. A workable classification of less material numerables like the conditions and events of tables 1.2.3 and 1.2.4 is more pressing for the design of a linguistic computing agent. (Table 1.2.2 is omitted completely, it concerns groups: family, nation or organization).

## 4 LEXICAL SEMANTICS

Although progress so far has mostly addressed the the semantic types of the Basic lexicon. Further progress is dependent on experimental evidence that an executable form of lexical meaning can be specified and composed. Compositional meaning with generalised quantifiers is a standard exercise for natural language processing by logic programmers [2], but a true lexical semantics is not normally part of the exercise. This is where the ingredients of the Basic world model need to be identified, represented in the model of a processing agent, and be subject to update through communication. A suitable world model is a topic for another paper, but some essential ingredients are the agents (persons) and objects (numerable things) which exist in the world, the locations (place), and associated with each, and the knowledge, perception and other attitudes (attention, desires, feeling etc.) that each agent has regarding the other things in the world.

There are many ways to represent and run such a model, including the technologies of frames, executable logics, directly coded processes and data. The task is one where both artificial intelligence and the processing of dialogue in natural language are being engineered. But although such a task can be disparaged as "AI-hard", less ambitious models are possible. For the composition of meaning in Basic sentences, the restriction to 18 verbs is a huge simplification, for this is the core of sentential meaning, although still complicated by the need for an adequate compositional semantics for prepositions and other "operation" words of the language, and for the remaining lexicon, still hundreds of words, although not thousands, and mostly seemingly less complex "things".

A start has been made on representing the core lexical semantics of the 18 verbs and prepositions, using an interval model of time to support tense and aspect as well as event nominals, with terms to capture the sorting implied by the lexical ontology. The underlying meaning of each verb is intuitive, and together their scope seems embracing:
come and go change the place of the subject, normally another agent with respect to the utterer, put, get and send change the place of an object with respect to the agentive subject, give and take change the possessive attribute of an object. The verbs be and have form stative conditions, see and seem perceptions, might and will modal auxiliaries, while let and keep affect the dynamic frame. Finally, make, do and say force existential change, chiefly by exploiting the meaning of the residual lexicon of things. Complication comes chiefly from the interactions with particles, and the case analyses.

## 5 CONCLUSION

The sort analysis of Basic that is described here is an investigation of its semantic types with a view to sense clarification for a compositional lexical semantics. It has provided a draft taxonomy for the semantic sorts of its lexicon, altough the work is far from complete and indeed it may be seen as yet another potentially immense Natural Language Project. However there is scope to contain each stage as an exercise and publish interim results. Further evolution will be dependent on experimental effort in providing an executable semantics using (multi-threaded) logic programs.

The feasibility of applications is speculative, but some seemingly interesting features for further investigation include the (unspecified) relations between the sorts, the role of supplementary relations (including for mereology), and the relatively small but crucial lexicon of mind, including the almost formal lexicon of knowledge, and that of feeling, through which the condition of the world is expressed, some function words acquire meaning, and the "presence" of an agent might need to captured.

## ACKNOWLEDGEMENT

The author wishes to thanks the anonymous referees for indicating their concerns and interest.

## REFERENCES

[1] Collin F. Baker, Charles J. Fillmore, and John B. Lowe, ‘The berkeley framenet project', in In Proceedings of the COLING-ACL-98, Montreal, Canada, (1998).
[2] Michael A. Covington., Natural language processing for Prolog programmers, Prentice Hall, 1994.
[3] K. Dahlgren and J. McDowell, 'Using commonsense knowledge to disambiguate prepositional phrase modifiers', in Proceedings of AAAI-86, Philadelphia, (June 1986).
[4] N. Guarino, 'Formal ontology and information systems', pp. 3-15. IOS Press, (1998).
[5] N. Guarino and C. Welty, 'Evaluating ontological decisions with ontoclean', Communications of the ACM, 45(2), 61-65, (2002).
[6] J.J. Katz and J.A. Fodor, 'The structure of a semantic theory', in in J.J. Katz and J.A. Fodor (eds) The structure of language. Readings in the philosophy of language, p. 479518. Prentice Hall, (1964).
[7] D.B. Lenat, 'Cyc: A large-scale investment in knowledge infrastructure', Communications of the ACM, 38(11), (1995).
[8] Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. http://wonderweb.semanticweb.org/ deliverables/documents/d18.pdf, 2003.
[9] George A. Miller, 'Wordnet: An on-line lexical database', International Journal of Lexicography, 3(4), (1990).
[10] George A. Miller, 'Wordnet: A lexical database for english', Communications of the ACM, 38(11), 39-41, (1995).
[11] Charles K. Ogden, Basic English: General Introduction with Rules and Grammar, Paul Treber \& Co., Ltd. London, 1930, 1940.
[12] T. Parsons, Events in the Semantics of English: A Study in Sub-atomic Semantics, MIT press, 1990.
[13] Adam Pease and Ian Niles, 'Towards a standard upper ontology', in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), (2001).
[14] W. Peters, I. Peters, and P. Vossen, 'Automatic sense clustering in eurowordnet', in in: Proc of the 1st international conference on Language Resources, (1998).
[15] S. Pinker, The stuff of Thought, Allen Lane, 2007.
[16] Laurent Prévot, Stefano Borgo, and Oltramari, 'Interfacing ontologies and lexical resources', in Proceedings of OntoLex2005, Jeju Island, South Korea, (2005).
[17] Jan Scheffczyk, Adam Pease, and Michael Ellsworth, 'Linking framenet to the suggested upper merged ontology', in "Proceedings of FOIS2006, International Conference on Formal Ontology in Information Systems, Baltimore, Maryland, USA, November 2006", (2006).
[18] Luc Schneider, Niave MetaPhysics, Master's thesis, Imperial College London, Department of Computing, 2001.
[19] Luc Schneider and Jim Cunningham, 'Ontological foundations of natural language communication in multi-agent systems', in Knowledge Based Intelligent Information and Engineering Systems, eds., Vasile Palade, Robert J.Howlett, and Lakhmi Jain, volume 2772 of LNAI, pp. 1403-1410. Springer, (2003).
[20] Jaspreet Shaheed, Alexander Yip, and Jim Cunningham, 'A top-level language-biased legal ontology', in "LOAIT Legal Ontologies and Artificial Intelligence Techniques", ed., et al. ""Jos Lehmann, "IAAIL Bologna Workshop", pp. 13-24. "Wolf Legal Publishers", (2005).
[21] Yorick Wilks, 'Preference semantics', in "in E.L. Kaenan (ed,), "Formal semantics of Natural Language", pp. 329-348, (1975).
[22] A. Yip and J. Cunningham, 'legal event reasoning for software agents', in "The Law of Electronic Agents", ed., Claudia Cevenini. "CIRSFID, Universita degli Studi di Bologna", (2003).

| Table 1 Loose ontolo | of OBE THINGS (common nouns) - Division names are | m OBE. |  |
| :---: | :---: | :---: | :---: |
| sort | new divisions and/or list of things | see below | see below |
| thing | substance, numerable |  |  |
| Table 1.1 Substances |  |  |  |
| sort - with divisions | new divisions or list of things | (some) qualities | (some) uses |
| substance | solid, threadlike, skinlike, granular, liquid, airlike, other hard, soft <br> metal, not metal <br> [brass, copper, gold, iron, lead,steel, tin] <br> [bone, coal, chalk, glass, stone, wood] | (hard touch) [yellow, brown, etc.] [white, black, etc.] |  |
| -solid |  |  |  |
| - -hard |  |  | (for man-made things) |
| ---metal |  |  |  |
| ---notmetal |  |  | [as animal-structure, for burning, as writing instrument, as window, as building face, as building structure] |
| -soft | food, other | (soft touch) | (digestion) |
| ---food | fruit, not fruit |  |  |
| ----fruit | [apple, orange] | (comes from a tree) |  |
| ----not fruit | [butter, cake, cheese, egg, fat, jelly, meat, potato] | [. . ., animal muscle, etc.] | [for bottlestop, for washing, etc] |
| ---other | [cork, soap, sponge, wax] |  |  |
| -threadlike | [cord, thread, wire] | (look of thread) | [cord for parcels, thread for cloth,etc] |
| -skinlike | [canvas, cloth, leather, paper] | (look of skin) | [(canvas) as cover, for \{curtains, clothes \}, for seat] |
| --cloth | [linen, silk, wool] | [...] |  |
| -granular | food, other |  |  |
| --food | [nut, rice, salt, sugar] | [...] | (digestion), [for food, for salt taste, for sweet taste] |
| - -other | [dust, earth, powder, sand,] | [...] |  |
| -liquid | drink, other | wet feeling |  |
| --drink | [soup, water] | [...] | [as food, as a drink] |
| - -other | [blood, ink, oil, paint, paste, polish] | [red colour, dark colour] | [part of body, for writing with, ...] |
| -air-like | [air, smoke, steam] | (feeling of air) | [for breathing, ...] |
| -other substance | [fire, flame, poison, waste] |  | [heat, light, making death,-] |
| Continuation is Table 1.2. |  |  |  |




\begin{tabular}{|c|c|c|c|}
\hline sort - with divisions \& new division \& part, detail \& process \\
\hline -event \& \multirow[t]{23}{*}{\begin{tabular}{l}
act, change, language, thought, process \\
producing, effecting, personal, simple, motion \\
\{get,put,give,take,do,make,send\}: \\
[account, addition, distribution, division, exchange, measure, observation, payment, record, reward, selection, tax] \{get,give,take,do,make\}: \\
[bite, blow, burst, burn, cover, crush, force, grip, kick, kiss, mark, pull, push, rub, shake, shock, smash, surprise, trick, use, wound] \{get,give,take,do,make\}: \\
[breath, cough, cry, drink, laugh, smile, sneeze, wave] \\
\{get,give,take,do,make\}: \\
[help, print, stitch, test, wash] \\
\{get,give,take,do,make\}: \\
[fall, fold, flight, jump, lift, move, point, slip, step, stretch, turn, roll, twist] \\
\{be, say \(\}\) \\
[answer, offer, order, question, protest, request, statement, suggestion] \\
[decision, idea, impulse] \\
\{place, condition, process\} \\
\{come, go \} to/from thing-numerable-place \\
\{take,do,make\}:[swim, walk, journey] \\
\{get,put, give,take,do,make,send\}: \\
[adjustment, attempt, effect, discovery, increase, loss, start, stop]
\end{tabular}} \& \multirow[t]{23}{*}{(} \& \multirow[t]{23}{*}{[attack, fight,]

\{get,give,take,do,make $\}:[$ transport]} <br>
\hline --act \& \& \& <br>
\hline \multirow[t]{2}{*}{---producing} \& \& \& <br>
\hline \& \& \& <br>
\hline \multirow[t]{4}{*}{---effecting} \& \& \& <br>
\hline \& \& \& <br>
\hline \& \& \& <br>
\hline \& \& \& <br>
\hline ---personal \& \& \& <br>
\hline \& \& \& <br>
\hline ---simple \& \& \& <br>
\hline \& \& \& <br>
\hline ---motion \& \& \& <br>
\hline \& \& \& <br>
\hline \& \& \& <br>
\hline - -language \& \& \& <br>
\hline \& \& \& <br>
\hline - -thought \& \& \& <br>
\hline - -change \& \& \& <br>
\hline ---place \& \& \& <br>
\hline \& \& \& <br>
\hline ---condition \& \& \& <br>
\hline trunctated Table 1.2 \& \& \& <br>
\hline
\end{tabular}


[^0]:    ${ }^{1}$ Imperial College, London SW7 2BZ, UK email: rjc @ doc.ic.ac.uk

[^1]:    ${ }^{2}$ See also Yip and Cunningham (2003), Shah et al (2005).[22, 20]

