The Semantic Web as the apotheosis of annotation, but what are its semantics?

Yorick Wilks
Oxford Internet Institute
and
University of Sheffield
www.dcs.shef.ac.uk/~yorick

> "In the middle of a cloudy thing is another cloudy thing, and within that another cloudy thing, inside which is yet another cloudy thing……and in that is yet another cloudy thing, inside which is something **perfectly clear and definite."**
>
> ----------Ancient Sufi saying

Abstract

The paper discusses what kind of entity the proposed Semantic Web (SW) is, and does so principally by reference to the relationship of natural language structure to knowledge representation (KR). It argues that there are three distinct views on the issue: first, that the SW is basically a renaming of the traditional AI knowledge representation task, with all its problems and challenges. Secondly, there is a view that the SW will be, at a minimum, the WorldWideWeb (WWW) with its constituent documents annotated so as to yield their content, or meaning structure, more directly. This view of the SW makes natural language processing central as the procedural bridge from texts to KR, usually via some form of automated Information Extraction. This view is discussed in some detail and it is argued that this can also be seen as a way of justifying the structures used as KR for the SW. There is a third view, possibly Berners-Lee's own, that the SW is about trusted databases as the foundation of a system of web processes and services, but it is argued that this ignores the whole history of the web as a textual system, and gives no better guarantee of agreed meanings for terms than the other two approaches. There is also a fourth view, much harder to define and discuss, which is that if the SW just keeps moving as an engineering development and is lucky (as the successful scale-up of the WWW seems to have been luckier, or better designed, than many cynics expected) then real problems will not arise

Introduction

This paper is concerned with the issue of what kind of object the Semantic Web is to be and, in particular, to ask about its semantics in the context of the relationship between knowledge representations and natural language itself, a relationship concerning which this paper expresses a view which will appear below. This is a vast, and possibly ill-formed, issue but the Semantic Web is no longer simply an aspiration in a magazine article (2001) but a serious research subject world-wide, with its own conferences and journal. So, even though it may not yet exist in a demonstrable form, in the way the WWW itself plainly does, it is a topic for research about which fundamental questions can be asked, as to its representations, their meanings and their groundings, if any.

The position adopted in this paper is that the concept of the Semantic Web (SW) has two distinct origins, and this persists now in two differing lines of SW research: one, closely allied to notions of documents and natural language processing (NLP) and one not. These differences of emphasis or content carry with them quite different commitments about what it is to interpret a knowledge representation, and what the method of interpretation has to do with meaning in natural language.

We shall attempt to explore both these strands here, but our assumptions will be consistent with the first branch of the bifurcation above, the view that assumes that natural language is, in some clear sense, humans' primary method of conveying meaning and that other methods of conveying meaning (formalisms, science, mathematics, codes etc.) are parasitic upon it. This is not a novel view: it was once associated firmly with the philosophy of Wittgenstein (1953), who we shall claim is slightly more relevant to these issu than Hirst's (2000) immortal, and satirical, line that "The solution to any problem in AI may be found in the writings of Wittgenstein, though the details of the implementation are sometimes rather sketchy." Before continuing, it must be made clear, too, that the quotation at the head of the paper is intended to suggest, not a skeptical position, but one where the SW will become a reality. Many popular criticisms of the SW (e.g. http://halfanhour.blogspot.com/2007/03/why-semantic-web-will-fail.html) do not examine foundational issues with any care and, moreover, fail to see that the thrust of their criticisms---e.g. that agreed ontologies in a field are difficult to obtain (see below)---would imply that science and medicine cannot be formalized at all, quite independently of the SW's existence, a view completely at odds with current developments in e-Science (see Wilks and van Besten, in press) practice, and indeed the whole history of science itself.

1 The Semantic Web and AI

The Hirst quotation above serves to show that any relation between philosophies of meaning, such as Wittgenstein's, and classic AI (or GOFAI as it is often known: Good Old Fashioned AI) is not an easy one. GOFAI remains committed to some form of logical representation for the expression of meanings and inferences, even if not the standard forms of the predicate calculus. Most issues of the AI Journal consist of papers within this genre.

Some have taken the initial presentation (2001) of the SW by Berners-Lee, Hendler and Lassila to be a restatement of the GOFAI agenda in new and fashionable WWW terms. In that article, the three authors describe a system of services, such as fixing up a doctor's appointment for an elderly relative, which would require planning and access to the databases of both the doctor's and relative's diaries and synchronizing them. This kind of planning behaviour was at the heart of GOFAI, and there has been a direct transition (quite outside the discussion of the SW) from decades of work on formal knowledge representation in AI to the modern discussion of ontologies. This is clearest in work on formal ontologies representing the content of science (e.g. Patel-Schneider, Hayes and Horrocks, 2004; Horrocks, 2005), where many of the same individuals have transferred discussion and research from one paradigm to the other. All this has been done within what one could call the standard KR assumption within AI, and one that goes back to the earliest work on systematic KR by McCarthy and Hayes (1969), a work we could take as defining core GOFAI. A key assumption of all such work was that the predicates in such representations merely look like English words but are in fact formal objects, loosely related to the corresponding English, but without its ambiguity, vagueness and ability to acquire new senses with use. We shall return to this assumption below, one which has certainly been apparent in both the original SW paper and some of what has flowed from it.

But it must also be noted that very few of the complex theories of knowledge representation in GOFAI actually appear within SW contributions so far: from McCarthy and Hayes fluents (McCarthy and Hayes, 1969) , McCarthy's later autoepistemic logic (1990), Hayes' Naïve Physics (1979), Bobrow and Winograd's KRL (1977), to name but a few prominent examples. A continuity of goals between GOFAI and the SW has not meant continuity of research traditions and this is both a gain and a loss: the gain of simpler schemes of representation which are probably computable; a loss because of the lack of sophistication in current schemes of the DAML/OIL (http://www.w3.org/TR/daml+oil-reference) family, and the problem of whether they now have the representational power for the complexity of the world, common sense or scientific, a point we shall return to later. There have been at least two other traditions of input to what we now call the SW, and I shall discuss one in some detail: namely, the way in which the SW concept has grown from the traditions of document annotation.

2.1 Natural language and the SW: Annotation and the lower end of the Semantic Web diagram

If one looks at the classic SW diagram from the original *Scientific American* paper (see Figure 1 below), the tendency is always to look at the upper levels: rules, logic framework and proof, and it is these, and their traditional interpretations, that have caused both critics and admirers of the SW to say that it is the GOFAI project by another name. But if one looks at the lower levels one finds Namespaces and XML, which are all the products of what we may broadly call NLP (natural language processing) obtained from the annotation of texts by a range of NLP technologies we may conveniently gather under the name IE (information extraction, see e.g. Cowie and Wilks, 2000).

It is useful to remember that available information for science, business and everyday life, still exists overwhelmingly as text; 85% of business data still exists as unstructured data (i.e. text). So, too, of course does the WorldWideWeb, though the proportion of it that is text is almost certainly falling. And how can the WWW be absorbed into the SW except by information being extracted from natural text and stored in some other form, such as a database of facts extracted from text or annotations on text items, stored as metadata either with or separate from the texts themselves. These forms are, of course, just those provided by large-scale Information Extraction (IE) (e.g. Cunningham et al., 1997). If, on the other hand, we were to take the view that the WWW will not become part of the SW, one is faced with an implausible evolutionary situation of a new structure starting up with no reference to its vast, functioning, but more primitive, predecessor. Things just do not happen like that.

XML, the annotation standard which has fragmented into a range of cognates for particular domains (e.g. TimeML, VoiceML etc.) is the only the latest standard in a history of annotation languages. These attach codings to individual text items so as to indicate information about them, or what should be done with them in some process, such as printing. Indeed, annotation languages grew from origins as metadata for publishing documents (the Stanford roff languages, and then Knuth's Tex, later Latex), as well as semi-independently in the humanities community as a way of formalizing the process of scholarly annotation of text. The Text Encoding Initiative (TEI) adopted SGML, a development of Goldfarb's (1997, publication date) original GML. SGML in turn became the origin of HTML (as a proper subset), which then gave rise to XML as well as being the genesis of the annotation movement in NLP that initially underpinned IE technology. There were early divisions over exactly how and where the annotation of text was to be stored for computational purposes; particularly between SGML, on the one hand, where annotations were infixed into the text with additional characters (as in Latex), and which had the effect of making the annotated text harder for humans to read. The DARPA research community, on the other hand, produced a functioning IE technology based on the storage of annotations (indexed by spans of characters in the text) separately as metadata, a tradition preserved in the GATE language processing platform from Sheffield (Cunningham et al., 1997), for example, and which now underpins many of the SW projects in Europe (e.g. Boncheva et al., 2003; Ciravegna et al., 2003.) This was one of the two origins of the metadata concept, the other being the index terms that were the basis of the standard information retrieval approach to document relevance.

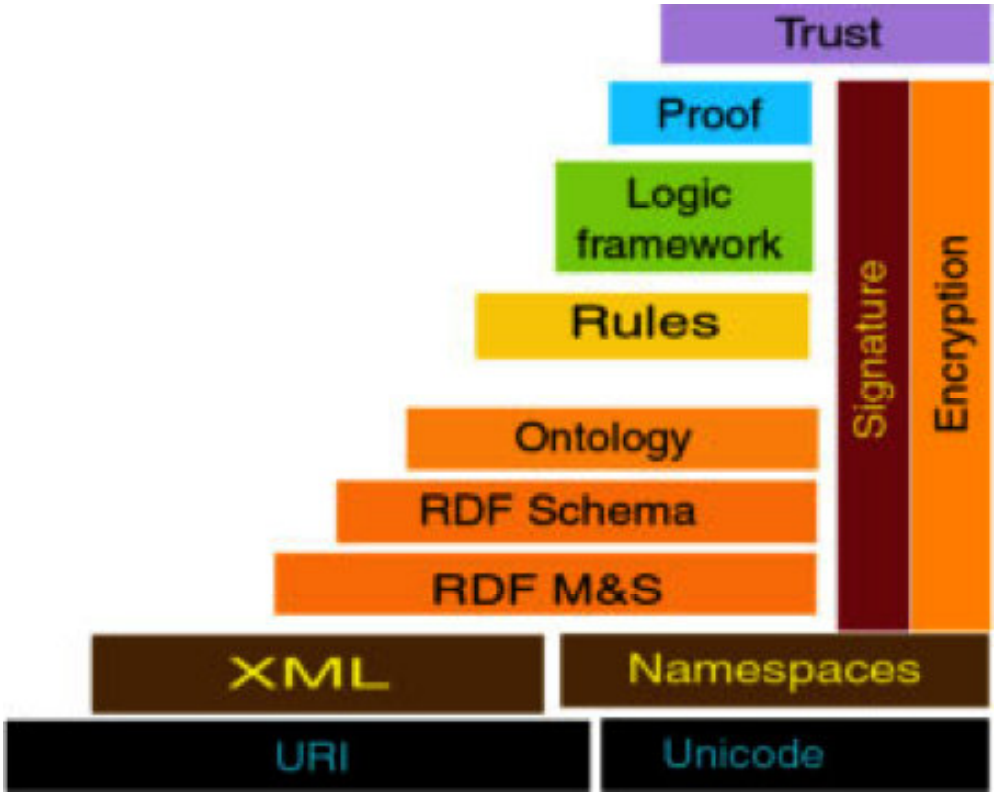


Figure 1: Levels of annotation and objects in the Semantic Web (from Berners Lee et al. 2001)

IE is now a technology with some twenty-five years of history, one which began with the hand-coded approach of Sager (1973) and DeJong (1979), but which then moved to a fully automatic system with tools like Leech's CLAWS4 program (Leech et al., 1994) for part-of-speech tagging in 1966. This was the first program systematically to add to a text "what it meant" even at the low level of interpretation that such tags represent. IE now reliably locates names in text, their semantic types, and relates them together by means of learned structures called templates into forms of fact and events, objects virtually identical to the RDF triple stores at the basis of the SW: which are not quite logic, but very like IE output. IE began by automating annotation but now has what we may call annotation engines based on machine learning (Ciravegna, 2003) which learn to annotate in any form and in any domain.

Extensions of this technology have led to effective question-answering systems trained from text corpora in well-controlled competitions and, more recently, the use of IE patterns to build ontologies directly from texts (e.g. Brewster et al., 2005). Ontologies can be thought of as conceptual knowledge structures, which organize facts derived from IE

at a higher level. They are very close both to the traditional Knowledge Representation goal of AI, and occupy the middle level in the original SW diagram. I shall return to ontologies later, but from now I only want to draw attention to the obvious fact that the SW inevitably rests on some technology with the scope of IE to annotate raw texts simply to derive names, then semantic typings of entities, fact databases, and later ontologies. Where would lists of names, and namable objects, come from if not automatically from texts; are we to imagine such inventories as simply made up by researchers?

On this view of the SW, which is not the only one, as I emphasised at the beginning, but it is the one that underlies most work on the SW and webservices in Europe (e.g. Norton et al., 2005); on such a view, the SW can be seen at its base level as a conversion from the WWW of texts by means of an annotation process of increasing grasp and vision, one that projects notions of meaning up the classic SW diagram from the bottom. Richard Braithwaite (1956) wrote a classic book on how scientific theories get the semantic interpretation of "high level" abstract entities (like neutrinos or bosons) from low level data; he named the process one of *semantic ascent* up a hierarchically-ordered scientific theory. The view of the SW under discussion here, which sees NLP and IE as among its foundational processes, bears a striking resemblance to that view of scientific theories in general.

2.2 The SW blurs the text-program distinction

The view of the SW sketched above has been that the IE technologies at its base (i.e. on the classic 2001 diagram above), are technologies that add "the meaning of a text" to web content in varying degrees and forms. These also constitute a blurring of the distinction between language and knowledge representation, because the annotations are themselves forms of language, sometimes very close indeed to language they annotate.This process at the same time blurs the distinction between programs and language itself, a distinction that has already been blurred historically from both directions, by two contrary assertions:

1. Texts are really programs (which is one form of GOFAI)
2. Programs are really texts

As to the first, there is Hewitt's (1972) claim that "language is essentially a side effect" in AI programming and knowledge manipulation. Longuet-Higgins (1972) also devoted a paper to the claim that English was essentially a high-level programming language. Dijkstra's view of natural language (personal communication) was essentially that natural languages were really not up to the job they had to do, and would be better replaced by precise programs, which is close to being a form of the first view.

Opposing this is a smaller group, what one might term the Wittgensteinian opposition, and I will cite my own version (2005), which is the view that natural language is and always must be the primary knowledge representation device, and all other representations, no matter what their purported precision, are in fact parasitic upon language—in the sense that they could not exist if language did not. The reverse is not true, of course, and was not for most of human history. Such representations can never be

wholly divorced from language, in terms of their interpretation and use. The paper is intended as a modest contribution to that tradition: but a great deal more can be found in a dialogue with Nirenburg in (Nirenburg and Wilks, 2001).

On such a view, systematic annotations are just the most recent bridge from language to programs and logic, and it is important to remember that, not long ago, it was perfectly acceptable to assume that a knowledge representation must be derivable from an unstructured form, i.e. natural language. Thus Woods in 1975:

"A KR language must unambiguously represent any interpretation of a sentence (logical adequacy), have a method for translating from natural language to that representation, and must be usable for reasoning."

The emphasis there is on a method of going from the less to the more formal, a process which inevitably imposes a relation of dependency between the two representational forms (language and logic). This gap has opened and closed in different research periods: in the original McCarthy and Hayes (1969) writings on KR in AI, it is clear, as with Hewitt and Dijkstra's views (mentioned earlier), that language was thought vague and dispensable. The annotation movement associated with the SW can be seen as closing the gap in the way in which Woods described.

The separation of the annotations into metadata (as opposed to leaving them within a text, as in Latex or SGMLstyle annotations) has strengthened the view that the original language from which the annotation was derived is dispensible, whereas the infixing of annotations in a text suggests that the whole (original plus annotations) still forms some kind of object. Notice here that the "dispensability of the text" view is not dependent on the type of representation derived, in particular to logical or quasi-logical representations. Schank (1972) certainly considered the text dispensable after his Conceptual Dependency representations had been derived, because he believed them to contain the whole meaning of the text, implicit and explicit, even though his representations would not be considered any kind of formal KR. This is a key issue that divides opinion here: can we know that any representation whatsoever contains all and only the meaning content of a text, and what would it be like to know that?

Standard philosophical problems, like this one, may or may not vanish as we push ahead with annotations to bridge the gap from text to meaning representations, whether or not we then throw away the original text. David Lewis in his (1972) critique of Fodor and Katz, and of any similar non-formal semantics, would have castigated all such annotations as "markerese": his name for any mark up coding with objects still recognizably within natural language (NL), and thus not reaching to any meaning outside language. The Semantic Web movement, at least as described in this section of the paper , takes this criticism head on and continues onward, hoping URIs and what some call « popping out of the virtual world » (e.g. by giving a web representation your concrete phone number!) will solve semantic problems. That is to say, it accepts that a SW. even if based on language via annotations, will provide sufficient "inferential traction" with which to run web-services. But is this plausible? Can all you want to know be put in RDF triples, and can they then support the subsequent reasoning required? Even when

agents thus based seem to work in practice, nothing will satisfy a critic like Lewis except a web based on a firm (i.e. formal and extra-symbolic) semantics and effectively unrelated to language at all. But a century of experience with computational logic has by now shown us that this cannot be had outside narrow and complete domains, and so the SW may be the best way of showing that a non-formal semantics can work effectively, just as language itself does, and in some of the same ways.

2.3 An Information Retrieval (IR) critique of the semantics of the SW

Sparck Jones (2004) in a critique of the SW, characterized much as we have above, returned to a theme she had deployed before against much non-empirically based NLP, such as ontology building and used her key phrase "words stand for themselves" and not for anything else, a claim has been the basis of successful IR search in the WWW and elsewhere. Content, for her, cannot be recoded in any general way, especially if it is general content as opposed to that from some very specific domain, such as medicine, where she seemed to believe technical ontologies may be possible as representations of content. As she put it mischeviously: IR has gained from "decreasing ontological expressiveness".

Her position is a restatement of the traditional problem of "recoding content" by means of other words (or symbols closely related to words, such as thesauri, semantic categories, features, primitives etc). This task is what automated annotation attempts to do on an industrial scale. Sparck Jones' key example is (in part): "A Charles II parcel-gilt cagework cup, circa 1670". What, she asks, can be recoded there, into any other formalism, beyond the relatively trivial form: {object type: CUP}?

What, she asks, of the rest of that (perfectly real and useful) description of an artifact in an auction catalogue, can be rendered other than in the exact words of the catalogue (and of course their associated positional information in the phrase)? This is a powerful argument, even though the persuasiveness of this example may rest more than she would admit on it being one of a special class of cases. But the fact remains that content can in general be expressed in other words: it is what dictionaries, translations and summaries routinely do. Where she is right is that GOFAI researchers are wrong to ignore the continuity of their predicates and classifiers with the language words they clearly resemble, and often differ from only by being written in upper case (an issue discussed at length in Nirenburg and Wilks, 2001). What can be done to ameliorate this impasse?

One method is that of empirical ontology construction from corpora (Brewster et al., 2001, 2005), now a well-established technology, even if not yet capable of creating complete ontologies. This is a version of the Woods quote above, according to which a knowledge representation (an ontological one in this case) must be linked to some natural language text to be justifiably derived. The derivation process itself can then be considered to give meaning to the conceptual classifier terms in the ontology, in a way that just writing them down a priori does not. An analogy here would be with grammars: when linguists wrote these down "out of their heads" they were never much use as input

to programs to parse language into structures. Now that grammar rules can be effectively derived from corpora, parsers can, in their turn, produce better structures from sentences by making use of such rules in parsers.

A second method for dealing with the impasse is to return to the observation that we must take "words as they stand" (Sparck Jones). But perhaps, to adapt Orwell, not all words are equal; perhaps some are aristocrats, not democrats. On that view, what were traditionally called "semantic primitives" remain just words but are also special words: a set that form a special language for translation or coding, albeit one whose members remain ambiguous, like all language words. If there are such "privileged" words, perhaps we can have explanations, innateness (even definitions) alongside an empiricism of use. It has been known since (Olney et al., 1968) that counts over the words used in definitions in actual dictionaries (Webster's Third, in his case) reveal a very clear set of primitives on which all the dictionary's definitions rest.

By the term "empiricism of use", I mean the approach that has been standard in NLP since the work of Jelinek (Jelinek and Lafferty, 1991) and which has effectively driven GOFAI-style approaches based on logic to the periphery of NLP. It will be remembered that Jelinek attempted to build a machine translation system at IBM based entirely on machine learning from bilingual corpora. He was not ultimately successful—in the sense the his results never beat those from the leading hand-crafted system, SYSTRAN--- but he changed the direction of the field of NLP as researchers tried to reconstruct, by empirical methods, the linguistic objects on which NLP had traditionally rested: lexicons, grammars etc. The barrier to further advances in NLP by these methods seems to be the "data sparsity" problem to which Jelinek originally drew attention, namely that language is "a system of rare events" and a complete model, at say the trigram level, for a language seems impossibly difficult to derive, and so much of any new, unseen, text corpus will always remain uncovered by such a model.

2.4 The whole Web as a corpus and a move to much larger language models

It may now be possible, using the whole web----and thus reducing data sparsity-----to produce much larger models of a language and to come far closer to the full language model that will be needed for tasks like complete annotation and automatically generated ontologies. The Wittgensteinian will always want to look for the use rather the meaning, and nowhere has more use available than the whole web itself, even if it could not possibly be the usage of a single individual. Work will be briefly described here that seeks to make data for a language much less sparse, and without loss, by means of *skip-grams*. These results are as yet only suggestive and not complete, but they do seem to offer a way forward.

Kilgarriff and Grefenstette (2001) were among the first to point out that the web itself can now become a language corpus in principle, even though that corpus is far larger than any human could read in a lifetime as a basis for language learning. A rough computation

shows that it would take about 60,000 years of constant reading for a person to read all the English documents on the WWW at the time of writing. But the issue here is not building a psychological model of an individual and so this fact about size need not deter us: Moore (2004) has noted that current speech learning methods would entail that a baby could only learn to speak after a hundred years of exposure to data. But this fact has been no drawback to the development of effective speech technology ----in the absence of anything better. A simple and striking demonstration of the value of treating the whole web as a corpus has been shown in experiments by e.g. Grefenstette (2004) who demonstrated that the most web-frequent translation of a word pair-------- from among all possible translation equivalent word pairs in combination------ is invariably also the correct translation.

What follows is a very brief description of the kind of results coming from the REVEAL project (Guthrie et al. 2006), which takes large corpora, such as a 1.5 billion word corpus from the web, and asks how much of a test corpus is covered by the trigrams present in that large training corpus. The project considers both regular trigrams and *skipgrams*: which are trigrams consisting of any discontinuity of items with a maximum window of four skips between any of the members of a trigram. So, if we take the sentence:

Chelsea celebrate Premiership success.

Then the two standard tri-grams in that sequence will be:

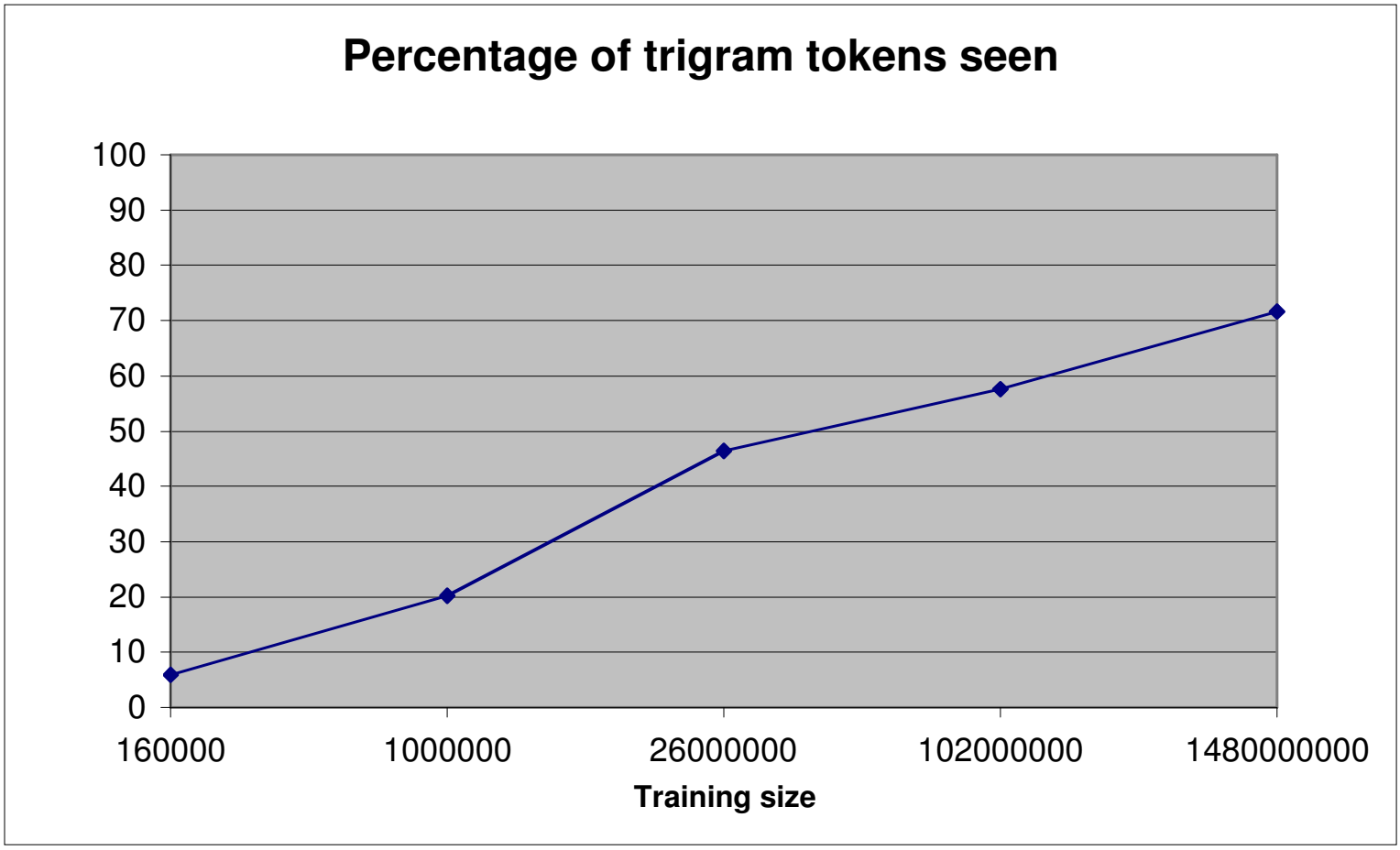
Chelsea celebrate Premiership
celebrate Premiership success

But the one-skip tri-grams will be:

Chelsea celebrate success
Chelsea Premiership success

Which seem at least as informative, intuitively, as the original trigrams and our experiments suggest that, surprisingly, skipgrams do not buy additional coverage at the expense of producing nonsense. Recent work shows the use of skip-grams can be more effective than increasing the corpus size. In the case of a 50 million word corpus, similar results (in terms of coverage of test texts) are achieved using skip-grams as by quadrupling corpus size. This illustrates a possible use of skip-grams to expand contextual information to get something closer to 100% coverage with a (skip) trigram model, combining greater coverage with little degradation, and thus achieving something much closer to Jelinek's original goal for an empirical corpus linguistics.

The 1.5 billion word training corpus gives a 67%+ coverage by such trigrams of randomly chosen 1000 word test texts in English, which is to say 67% of the trigrams found in any random 1000 passage of English were already found in the gigaword corpus. But we obtained 74% coverage with 4skiptrigrams, which suggests, by extrapolation, that it would need 75x10*10 words to give 100% trigram coverage (including skipgrams up to 4grams). Our corpus giving 74% coverage was 15x10*8 words, and Greffenstette (2003) calculated there were over 10*11 words of English on the web in 2003 (I.e. about 12 times what Google indexed at that time), so the corpus needed for complete coverage of training texts by trigrams would be about seven times the full English web in 2003, which is somewhat closer to the size of today's (2007) English web.



All this is, again, preliminary and tentative, but it suggests that an empiricism of usage may now be more accessible (with corpora closer to the whole web) than Jelinek thought at the time (1990) of his major MT work at IBM.

Figure 2: percentage of trigrams seen with training corpus size

Such modern web corpora are so vast they cannot conceivably offer a model of how humans process semantics, so a cognitive semantics based on such usage remains an open question. However, one possible way forward would be to adapt skipgrams so as to make them able (perhaps with the aid of a large-scale fast surface parser of the kind already applied to large chunks of the WWW) to pick up Agent-Action-Object triples capturing proto-facts in very large numbers. This is a old dream going back at least to (Wilks, 1968) where they were seen as trivial Wittgensteinian "forms of fact", later revived by Greffenstette (Kilgarriff and Grefenstette, 2001) as a "massive lexicon" and now available as inventories of surface facts at ISI (Hovy, 2005). These objects will not be very different from standard RDF triples, and might offer a way to deriving massive SW content on the cheap, even simpler than that now offered by machine learning-based IE.

If anything were possible along these lines, then NLP would be able to provide the base semantics of the SW more effectively than it does now, by making using of some very large portion of the WWW as its corpus. If one finds this notion unattractive, one should demonstrate in its place some other plausible technique for deriving the massive RDF content the SW will require. Can anyone seriously believe that can be done other than by NLP techniques of some type like the one described above?

3. A third view of the what the SW is: trusted data-bases

There is a third view of the SW, different from both the GOFAI and NLP views that I have contrasted so far in this paper. That is, in my view, one close to Berners-Lee's own vision of the SW, as expressed in (Berners-Lee et al., 2001), one that emphasizes databases as the core of the SW: databases, the meanings of whose features are kept constant and trustworthy by a cadre of guardians of their integrity, a matter quite separate from both logical representations (dear to GOFAI) and to any language-based methodology of the kind described in this paper. Berners-Lee's view deserves extended discussion and consideration that cannot be given here, but it will inevitably suffer from the difficulty of any view (like GOFAI) that seeks to preserve predicates, features, facets or whatever from the NLP vagaries of sense change and drift with time. We still "dial numbers" when we phone even though that no longer means the action it did a few decades ago; hence not even number-associated concepts are safe from time. The long-running CyC project (Lenat, 1996), one of the predecessors of the SW as a universal repository of formalized knowledge, suffered from precisely this difficulty of "predicate drift": that predicates did not mean this year what coders meant by them 20 years earlier. The SW has at present no solution to offer to this problem.

Berners-Lee's view has the virtues and defects of Putnam's later theory of meaning (Putnam, 1975/1985), mentioned earlier, one where scientists become the guardians of meaning, since only they know the true chemical nature of, say, molybdenum and how it differs from the phenomenally similar aluminium. Hence only these guardians know the *meaning* of molybdenum, independently of how it appears (which is just like aliminum!). It was essential to his theory that the scientists did not allow the criteria of meaning to leak out to the general public, lest they became subject to change. For Putnam, only scientists know the distinguishing criteria for water and deuterium dioxide (heavy water) which seem the same to most of the population but are not. Many observers, including this author (1975, and see Mellor 1977), have argued this separation cannot be made, in principle or in practice, since scientists are only language users in lab coats.

4. The SW and the representation of tractable scientific knowledge

The issues concerned with Berners-Lee's "scientific data-base" view of the SW can be illustrated concretely by turning some to questions of meaning and interpretation of

formal knowledge raised first by Kazic in connection with biological data-bases, which could be expected to form part of any SW wide enough to cover scientific and technical knowledge. Kazic (2006) has posed a number of issues close in spirit to those of this paper, but against a background of expert knowledge of biology that is hard to capture here without more exposition than was needed in a Biocomputing proceedings, where she published. Broadly, and using arbitrary names for terms like "thymidine phosphorylase", she draws attention to two symmetric chemical reactions of "cleavage" we may write as:

A <-> B
and
C <-> D

An enzyme Z (actually EC 2.4.2.4) catalyzes both reactions above according to the standard knowledge structures in the field (KEGGs maps: http:// www.genome.-ad.jp/kegg/kegg1.html). But Z is not in the class Y (a purine nucleoside) and so should not, in standard theory, be able to catalyze the two reactions above, ormally the province of Y compounds, *yet it does*. There is a comment in the KEGG maps saying that Z can catalyze reactions like those of another enzyme Z' (EC 2.4.2.6) under some circumstances, where Z' actually is a Y, although its reactions are quite different from Z, and they cannot be substituted for each other, and neither can be rewritten as the other. Moreover, Z has apparently contradictory properties, being both a statin (which stops growth) and a growth factor. Kazic asks "so how can the same enzyme stimulate the growth of one cell and inhibit the growth of another?" (p.2)

This is an inadequate attempt to state the biological facts in this non-specialist form, but it is clear that something very odd is going on here, something that Marxists might once have hailed as a dialectical or contradictory relationship. It is certainly an abstract structure that challenges conventional knowledge representations, and is far more complex than the standard form of default reasoning in AI, on which view, if anything is an elephant it has four legs even though Clyde, undoubtedly an elephant, has only three.

The flavour of the phenomena here is that of extreme context dependence, that is to say, that an entity behaves quite differently--------- indeed in opposite fashions------ in the presence of certain other entities. Languages are, of course, full of such phenomena, such as when "cleave to the Lord" and "cleave a log" mean exactly opposite things, and we have structures in language representation for describing and representing such phenomena, though there is no reason at the moment to believe they are of any assistance here.

The point Kazic is making is that it will be a requirement on any SW that represents biological information (and licences correct inferences) that it can deal with phenomena as complex as this. At first sight such phenomena seem beyond those within a standard ontology dependent on context-free relations of inclusion and the other standard relations:

"To ensure the scientific validity of the Semantic Web's computations, it must

sufficiently capture and use the semantics of the domain's data and computations ."(p.2)

In connection with the initial translation into RDF for, she continues:

"Building a tree of phrases to emulate binding….forces one to say explicitly something one may not know (e.g. whether the binding is random or sequential, what the order of any sequential binding is….). By expanding the detail to accommodate the phrasal structure, essential and useful ambiguities have been lost." (ibid.)

The last quotation is revealing about the structure of science, and the degree to which it remains in parts a craft skill, even in the most technical modern areas. Even if that were not the case, being forced to be more explicit and to remove ambiguities could only be a positive influence. The quotation brings out the dilemma in some parts of advanced science that intend to make use of the SW: that of whether the science is yet explicit enough and well understood enough to be formally coded, a question quite separate from issues of whether the proposed codings (from RDF to DAML/OIL) have the representational power to express what is to be made explicit. If it is not, then Biology may not be so different from ordinary life as we may have thought, certainly not so different from the language of auction house catalogues, in Sparck Jones' example, where the semantics remains implicit, in the sense of resting on our human interpretation of the words of annotations or comments (in this case in the margins of KEGG maps).

The analogy here is not precise, of course: the representational styles in the current SW effort have, to some degree, sacrificed representational sophistication to computational tractability (as, in a different way, the WWW itself did in the early 90s). It may be that, when some of the greater representational powers in traditional GOFAI work are brought to bear, the KEGG-style comments may be translated from English phrases, with an implicit semantics, to the explict semantics of ontologies and rules. It is what we must all hope for. But in the case of Sparck Jones' C16 cup description, the problem does not lie in any knowledge representation, but only in the fact that the terms involved are all so precise and specific that no generalizations ---no imaginable "auction ontology"---- would provide a coding that would enable the original English to be thrown away. The possibility always remains of translation into another language, or an explicit numbering of all the concepts in the passage, but there is no representational saving to be made there in either case (see here, as always, McDermott 1981 for a classic demolition of that very possibility).

Kazic goes on to argue that one effect of these difficulties about explicitness is that "most of the semantics are pushed onto the applications" (p. 7), where the web agents may work or not, but there is insufficient explicitness to know why in either case. This is a traditional situation: as when a major AI objection to the connectionist/neural net movement was that, whether it worked or not, nothing was served scientifically if what it did was not understood, that is to say, transparent and explicit. There is not yet enough SW data yet to be sure, but it is completely against the spirit of the SW that its operations should be unnecessarily opaque or covert. That becomes even clearer if one sees the SW as the WWW " plus the meanings" where only additional, rather than less explicit,

information would be expected.

Discussions in this area normally resile themselves from more traditional ontological enquiry, namely what things there are in the world. Ancient questions have a habit of returning to bite one at the end though, in this paper we have taken a robust position, in the spirit of Quine (1953) that whatever we put into our representations---concepts, sets, etc.—has existence, at least as a polite convention. But it may be that a fully explicit SW has to make ontological commitments of a more traditional sort, at least as regards the URIs: the points where the SW meets the world of unique descriptions of real things. But scientific examples of this interface in the world of genes are by no means straightforward.

Suppose we ask: what are the ontological "objects" in genetics, say in the classic Drosophila data base FlyBase (Morgan et al., 2003)? FlyBase ultimately grounds its gene identifiers ---the formal gene names---in the sequenced Drosophila genome and associates nucleotide sequences parsed into introns, exons, regulatory regions and so on with gene ids. However, these sequences often need modifying on the basis of new discoveries in the literature: e.g.new regulatory regions “upstream” from the gene sequence are quite frequently identified, as understanding of how genes get expressed in various biological processes increases. Thus the "referent" of the gene id. changes and with it information about the role of the `gene'. However, for most biologists the gene is still the organising concept around which knowledge is clustered, so they will continue to say quite happily that the gene `rutabaga' does so-and-so, even if they are aware that the referent of rutabaga has changed several times, and in significant ways, over the last decade. The curators and biologists are, for the most part, content with this, though the argument that the Drosophila community has been cavalier with gene naming has been made from within it.

This situation, assuming the non-expert description above is broadly correct, is of interest here because it shows there are still ontological issues in the original sense of that word: i.e. as to what there actually IS in the world. More precisely, it directly refutes Putnam's optimistic theory (1975, cited elsewhere in this paper) that meaning can ultimately be grounded in science, because, according to him, only scientists know the true criteria for selecting the referents of terms. The Drosophila case shows this is not so, and in some cases the geneticists have no more than a hunch, sometimes proved false in practice, that there are lower level objects unambiguously corresponding to a gene id.(and in the way SW URIs are intended to do), in the way that an elementary molecular structure, say, certainly does correspond to an element’s name in Mendeleev's table.

5. Conclusion

We have in this paper touched on three views of what the SW is. There is also a fourth view, much harder to define and discuss, which is that if the SW just keeps moving as an engineering development and is lucky (as the successful scale-up of the WWW seems to have been luckier, or better designed, than many cynics expected) then real problems will

not arise. This view is a hunch and not open to close analysis but one can only wish it well, as it were, without being able to discuss it in detail further at this stage. It remains the case that the SW has not yet taken off, as the WWW did, and Google-IR and iPods did; it may be that something about its semantics is holding it back, and that maybe connected, as we have argued, to its ability to generate semi-formalised material on a great scale from existing WWW material.

The main argument of the paper has been that NLP will continue to underlie the SW, including its initial construction from unstructured sources like the WWW, in several different ways, and whether it advocates realize this or not: chiefly, I argued, such NLP activity is the only way up to a defensible notion of meaning at conceptual levels (in the original SW diagram) based on lower-level empirical computations over usage. The paper's aim is definitely not to claim logic-bad, NLP-good in any simple-minded way, but to argue that the SW will be a fascinating interaction of these two methodologies, again like the WWW (which has been basically a field for statistical NLP research) but with deeper content. The paper goes on to argue that only NLP technologies (and chiefly IE) will be able to provide the requisite RDF knowledge stores for the SW from existing WWW (unstructured) text data bases, and in the vast quantities needed. There is no alternative at this point, since a wholly or mostly hand-crafted SW is also unthinkable, as is a SW built from scratch and without reference to the WWW. It also assumes that, whatever the limitations on current SW representational power we have drawn attention to here, the SW will continue to grow in a distributed manner so as to serve the needs of scientists, even if it is not perfect. The WWW has already shown how an imperfect artifact can become indispensable.

The paper also argues that contemporary statistical large-scale NLP offers new ways of looking at usage in detail and in quantity--even if the huge quantities required now show we cannot easily relate them to an underlying theory of human learning and understanding. We can see glimmerings, in machine learning studies, of something like Wittgenstein's 'language games" (1953) in action, and of the role of key concepts in the representation of a whole language. Part of this can only be done, we argued, by some automated recapitulation of the role primitive concepts play in the organization of (human-built) ontologies, thesauri, and wordnets. The heart of the issue is the creation of meaning by some interaction of (unstructured language) usage and the interpretations to be given to higher level concepts---this is a general issue, but the construction of the SW faces it crucially and it could be the critical arena for progress on a problem that goes back at least to Kant's classic formulation in terms of "concepts without percepts are empty, percepts without concepts are blind". If we see that opposition as one of language data (like percepts) to concepts, the risk is of formally defined concepts always remaining empty (see discussions of SW meaning in (Horrocks and Patel-Schneider, 2003). The answer is, of course, to find a way, upwards, from one to the other.

Acknowledgements: This paper is indebted to many discussions with colleagues within the AKT project at Sheffield and elsewhere (Aktive Knowledge Technologies: EPSRC Interdisciplinary Research Centre, 2001-

6), as well as with Arthur Thomas, Christopher Brewster, Ted Nelson and other colleagues at the Oxford Internet Institute.. The passage on Drosophila owes a great deal to conversations with Ted Briscoe, but, as always, the errors are my own. This work was funded in part by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

References

Berners-Lee, T., Hendler, J., and Lassila, O. 2001. Scientific American.

Berners-Lee, T. 2005. Keynote paper in BCS Workshop on the Science of the Web, London.

Bobrow, D., and Winograd, T., 1977. An overview of KRL, a knowledge representation language. Cognitive Science 1: 3--46.

Bontcheva, K., and Cunningham, H. 2003. Information Extraction as a Semantic Web Technology: Requirements and Promises. Adaptive Text Extraction and Mining workshop, 2003.

Braithwaite, R. 1956. Scientific Explanation. Cambridge: Cambridge Univerity Press.

Brewster, C., Ciravegna, F., Wilks, Y., 2001 Knowledge Acquisition for Knowledge Management: Position Paper in Proceedings of the IJCAI-2001 Workshop on Ontology Learning held in conjuction with the 17th International Conference on Artificial Intelligence (IJCAI-01), Seattle, August, 2001

Brewster, C., Iria, J., Ciravegna, F., and Wilks, Y. 2005. The Ontology: Chimaera or Pegasus, In Proc. Dagstuhl Seminar on Machine Learning for the Semantic Web, 13-18 February 2005

Ciravegna, F. 2003. Designing adaptive information extraction for the Semantic Web in Amilcare. In S. Handschuh and S. Staab, editors, Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications. IOS Press, 2003.

Companions, 2007, www.companions-project.org

Cowie, J. and Wilks, Y. 2000 Information Extraction. In Dale, Moisl and Somers (eds.) Handbook of Natural Language Processing. New York: Marcel Dekker.

Cunningham, H., Humphreys, K., Gaizauskas, R. and Y. Wilks, 1997, GATE -- a TIPSTER-based General Architecture for Text Engineering In Proc. of the TIPSTER Text Program Phase III,
Morgan Kaufmann, CA.

DeJong, G. 1979. Skimming Stories in Real Time: An Experiment in Integrated Understanding, PhD thesis, Yale University.

Goldfarb, C. F. 1997, SGML: The Reason Why and the First Published Hint, In Journal of the American Society for Information Science. (48).

Grefenstette, G. 2004. The scale of the multi-lingual Web, talk delivered at Search Engine Meeting 2004, The Hague, The Netherlands, 19-20 April 2004

Guthrie, D., Allison, B., Liu, W., Guthrie, and Wilks. Y. 2006 A Closer Look at Skip-gram Modelling.In Proceedings of the Fifth international Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy

Hayes, P, J. 1979. The Naive Physics Manifesto., in D. Michie, (ed.), Expert Systems in the Micro-Electronic Age, Edinburgh: Edinburgh University Press, 242-70

Hewitt, C. 1972. Procedural Semantics. In R. Rustin, (ed.) , Natural Language Processing. New York: Algorithmics Press

Hirst, G.. 2000. Context as a spurious concept. In Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February 2000, 273--287.

Horrocks, I. 2005. Description Logics in Ontology Applications. In KI/Tableaux 2005, Koblenz, Germany, September 2005.

Horrocks, I. and Patel-Schneider, P. 2003. Three Theses of Representation in the Semantic Web, In Proc. The Twelfth International World Wide Web Conference, Budapest, Hungary, May 2003, pages 39–47.

Hovy, E. 2005. Key Toward large-scale shallow semantics for higher-quality NLP. In Proc. 12th PACLING Conference, Tokyo, Japan.

Jelinek, F. and Lafferty, J. 1991. Computation of the Probability of Initial Substring Generation by Stochastic Context Free Grammars. Computational Linguistics 17:3: 315-323.

Kazic, T. 2006, Putting the semantics into the semantic web: how well can it capture biology? Proc. Pacific Symposium in Biocomputing, (11).

Kilgarriff, A. and Greffenstete, (eds.). 2001. Special issue of Computational Linguistics on: the Web as Corpus.

Leech, G., Garside, R., and Bryant, M. 1994. CLAWS4: The tagging of the British National Corpus. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 94) Kyoto, Japan, 622-628.

Lenat, D. 1996. CyC: A Large-Scale Investment in Knowledge Infrastructure (1995) Communications of the ACM

Lewis, D. 1972. General Semantics In D. Davidson and G. Harman, (eds.) The Semantics of Natural Language, Amsterdam: Kluwer.

Longuet-Higgins, H. 1972. The Algorithmic Description of Natural Language, In Proc. Roy. Soc. Lond. B 182: 255-276.

McCarthy, J. 1990. Formalizing common sense: papers by John McCarthy. Ablex, Norwood, N J.

McCarthy, J. and Hayes, P. 1969. Some Philosophical Problems from the Point of View of Artificial intelligence, In D. Michie (ed.) Machine Intelligence 4. Edinburgh: Edinburgh Univ. Press.

McDermott, D. 1981. Artificial Intelligence meets Natural Stupidity. In J. Haugeland (ed.). Mind Design, pp. 143 -- 160. Montgomery, VT.: Bradford.

Mellor, D. H. 1977. Natural Kinds,
in British Journal for the Philosophy of Science 28,

Moore, R. K., 2003. A comparison of data requirements for ASR systems and human listeners, In Proc. EUROSPEECH 2003.

Morgan, A., Hirschmann, L., Yeh, A., and Colosimo, M. 2003.
Gene Name Extraction Using FlyBase Resources, In ACL Workshop on Language Processing in Biomedicine, Sapporo, Japan.

Nirenburg, S., and Wilks, Y. 2001. What's in a symbol, In JETAI.(Journal of Theoretical and Empirical AI)

Norton, B., Chapman, S and Ciravegna, F. 2005. Orchestration of semantic web services for large-scale document annotation. Springer: Berlin.

Olney, J., Revard, C.., Ziff, P. 1968. Some monsters in Noah's Ark. Research memorandum, Systems Development Corp., Santa Monica, CA.

Page, R., Brin, S., Motwain, R., and Winograd, T. 1998. The pagerank citation algorithm: bringing order to the web. In 7th WWW Conference.

Patel-Schneider, P., Hayes, P.J. and Horrocks, I. .2004. OWL Web Ontology: Language Semantics and Abstract Syntax, W3C Recommendation [http://www.w3.org/TR/owl-semantics/].

Putnam, H. 1975/1985 The meaning of 'meaning'. In Philosophical Papers, Vol. 2: Mind, Language and Reality. Cambridge: Cambridge University Press.

Quine, W. V. O. 1953 From a Logical Point of View. Cambridge, MA: Harvard Univ. Press.

Sager, N. 1973. The String Parser for Scientific Literature. In Natural Language Processing (R. Rustin, ed.), pp. 61-87. Algorithmics Press, NY.
Cambridge University Press.

Schank, R. 1972. Conceptual Dependency: A Theory of Natural Language Understanding, Cognitive Psychology, (3)4.

Sparck Jones, K. 2004, What's new about the Semantic Web? ACM SIGIR Forum.

Surowiecki, J., 2004. The Wisdom of Crowds. Random House New York.

Wilks, Y. 1968. Computable Semantic Derivations. Systems Development Corporation, SP-3017.

Wilks, Y. 1975. Putnam and Clarke and Mind and Body. In The British Journal for the Philosophy of Science (26).

Wilks, Y. 2004. Companions: a new paradigm for agents. In Proc. International AMI Workshop, IDIAP, Martigny, CH.

Wilks, Y. 2005. What would a Wittgensteinian Computational Linguistics be like? In Proc. 10th International Congress on Pragmatics, Garda, Italy.

Wittgenstein, L. 1953. Philosophical Investigations, Oxford: Oxford University Press.

Woods, W. 1975. What's in a Link: Foundations for Semantic Networks. In Representation and Understanding: Studies in Cognitive Science. New York: Academic Press. 35-82.