ACL Lifetime Achievement Award

On Whose Shoulders?

Yorick Wilks[*]
University of Sheffield

Introduction

The title of this piece refers to Newton's only known modest remark: "If I have seen farther than other men, it was because I was standing on the shoulders of giants." Since he himself was so much greater than his predecessors, he was in fact standing on the shoulders of dwarfs, a much less attractive metaphor. I intend no comparisons with Newton in what follows: NLP/CL has no Newtons and no Nobel Prizes so far, and quite rightly. I intend only to draw attention to a tendency in our field to ignore its intellectual inheritance and debt; I intend to discharge a little of this debt in this article, partly as an encouragement to others to improve our lack of scholarship and knowledge of our own roots, often driven by the desire for novelty and to name our own systems. Roger Schank used to argue that it was crucial to name your own NLP system and then have lots of students to colonize all major CS departments, although time has not been kind to his many achievements and originalities, even though he did build just such an Empire. But to me one of the most striking losses from our corporate memory is the man who is to me the greatest of the first generation and still with us: Vic Yngve. This is the man who gave us COMIT, the first NLP programming language; the first random generation of sentences; and the first direct link from syntactic structure to parsing processes and storage (the depth hypothesis). I find students now rarely recognize his name, and find that incredible.

This phenomenon is more than corporate bad memory, or being too busy with engineering to do the scholarship. It is something endemic in the wider field of Computer Science and Artificial Intelligence, although bottom-up wiki techniques are now filling many historical gaps for those who know where to look, as the generation of pioneers has time to reminisce in retirement.[1] There are costs to us from this general lack of awareness, though: a difficulty of "standing on the shoulders" of others and acknowledging debts, let alone passing on software packages. Alan Bundy used to highlight this in the *AISB Quarterly* with a regular column where he located and pilloried reinventions in the field of AI; he also recommended giving obituaries for one's own work, and this paper could be seen in that way, too.

[*] Department of Computer Science, The University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK. E-mail: Y.Wilks@dcs.shef.ac.uk. This article is the text of the talk given on receipt of the ACL's Lifetime Achievement Award in 2008.

[1] See the video interview with Victor Yngve on my Web site at `http://www.dcs.shef.ac.uk/~yorick/YngveInterview.html`.

Early Academic Life

My overwhelming emotion on getting this honor was, after surprise, a feeling of inadequacy in measuring up to previous honorees, but nonetheless, I want to grasp at this moment of autobiography, or at what in his own acceptance paper Martin Kay called: "but one chance for such gross indulgence." I was born in 1939 in London at just about the moment the Second World War started in Europe; this was, briefly, a severe career slowdown. However, the British Government had a policy of exporting most children out of the range of bombs and I was sent to Torquay, a seaside town in southwest England that happened to have palm trees on all the main streets, a fact it is often difficult to convince outsiders of. The town had, and has, a Grammar School for Boys, which had a very good Cambridge-trained mathematician as its headmaster, and eventually I made my way back across England to Pembroke College, Cambridge, to study mathematics, a college now for ever associated with my comedian contemporaries: Peter Cook, Clive James, Eric Idle, Tim Brooke-Taylor, and similar wastrels. I began a series of changes of subject of study, downhill towards easier and easier ones: from mathematics to philosophy to (what in the end after graduation became) NLP/AI. It was not that I could not do the mathematics, but rather that I experienced the shock that many do of finding how wide the range of talent in mathematics is, and that being very good in a provincial grammar school does not make one very good at Cambridge. This is a feeling peculiar to mathematics, I think, because the talent range is so much wider than in most subjects, even at the top level.

Margaret Masterman, who was to become the main intellectual influence in my life, was the philosophy tutor for my college, although her main vocation was running the Institute she had founded, outside the University in a Cambridge suburb: CLRU, the Cambridge Language Research Unit. It was an eccentric and informal outfit, housed in what had been a museum of Buddhist art, some of whose sculptures were built into the walls. MMB (as she was known) ran the CLRU from the mid 1950s to the early 1980s on a mix of US, UK, and EU grants and did pioneering work in MT, AI, and IR. Of those honored by the ACL with this award over the last five years, three have been graduates of that little Buddhist shed, and include Martin Kay and Karen Spärck Jones, a remarkable tribute to MMB. The lives and work of we three have been quite different but all in different ways stem from MMB's interests and vision: She had been a pupil of Wittgenstein and, had she known it, would have approved of Longuet-Higgins's remark that "AI is the pursuit of metaphysics by other means." She believed that practical research into the structure of language could give insight into metaphysics, but was in no way other-worldly: She was the daughter of a Cabinet Minister and knew what it was to command.

In a final twist, I found after her death in 1986 that she had made me her literary executor: She had never written a book and wanted me to construct one from her papers posthumously. It took me twenty years to get the required permissions but the volume finally appeared in 2005 (Masterman et al. 2005).

Thesis Building and CLRU

When I started work at CLRU in 1962 to do a doctorate, it had no computer in the normal sense, only a Hollerith card sorter of the sort built for the US census half a century before. Basically, you put a stack of punched cards into one of these things—which looked like a metal horse on four legs—and the cards fell into (I think) 10 slots depending on how you had plugged in a set of wires at the back to identify destination slots for sorted cards with hole patterns on the cards. With some effort, these could be turned into quite interesting Boolean machines; my first task was to take a notion of Fred Parker-Rhodes that a Hallidayan grammar could be expressed as a lattice of typed classes, and then program the card sorter so that repeated sorts of punched cards could be used to parse a sentence. It was triumph of ingenuity over practicality. Later the CLRU owned an ICL 1202 computer with 1,200 registers on a drum, but it was a so-called bini-ten machine designed for UK cash transactions when there were still 12 pennies to a shilling, and so the 1202 has print wheel characters for 10, 11, and 12 (as well as 0–9), a fact on which Parker-Rhodes built a whole world of novel print conventions for his research. This was the period at CLRU when Karen Spärck Jones was completing her highly original thesis (published twenty years later as Jones [1986]) on unsupervised clustering of thesaurus terms—whose goal was to produce primitives for MT, it is often forgotten—until she had to move her computations to a real computer at the University Computing Laboratory, where she eventually created a new career in IR, essentially using the same clump algorithms—created by Parker-Rhodes and her husband Roger Needham—to do IR.

My own interests shifted to notions in an early Masterman paper titled "Semantic message detection using an interlingua" (Masterman 1961), an area in which Martin Kay had also originally worked on an interlingua for MT. My thesis computation was done in LISP 1.6 on an IBM360 (under a one-man US Air Force contract, administered by E. Mark Gold, who later became famous as the founder of learnability theory), at SDC in Santa Monica, where I was attached loosely in 1966 to the NLP group there run by Bob Simmons. My thesis was to be entitled "Argument and proof in Metaphysics from an empirical point of view" and my advisor was MMB's husband, Richard Braithwaite, Knightbridge Professor of Moral Philosophy at the University. He was a philosopher of science and a logician, and was given the chair of moral philosophy—a subject about which he knew nothing—because it was the only one available at Cambridge at the time. This produced an extraordinary inaugural lecture in which he effectively founded a new subject: "The theory of games as a tool for the moral philosopher."

Unfortunately for me he was not interested in my thesis, and took me on only as a favor to MMB. My interest was the demarcation of metaphysical text: what it was, if anything, that distinguished it from ordinary language text. Wittgenstein had once said that words were "on holiday" in metaphysical text, but also that he wanted to "bring words back from their metaphysical to their everyday usage" (Wittgenstein 1973). This is exactly what I wanted to capture with computation, and the thesis was eventually submitted to the Cambridge Philosophy faculty in 1967—then called Moral Sciences—with a large appendix of LISP program code at the back, something they had never seen before, or since. The thesis was bound in yellow, though the regulations stipulated black or brown bindings; I must have had some extraordinary idea that someone might cruise the long corridors of Cambridge theses looking for one that stood out by color—the arrogance of youth!

The thesis's starting point was Carnap's monumental *Logische Syntax der Sprache* (1937) and his claim that meaningfulness in text could be determined by "logical syntax"—rules of formation and transformation (a notion which may well sound familiar; Chomsky was a student of Carnap). My claim was that this was a bad demarcation and a better criterion of meaningfulness would be to have one interpretation rather than many, namely, that word-sense discrimination (WSD) was possible for a given text. On that view, the "meaningless" text had too many interpretations rather than none (or one). A word in isolation is thus often meaningless. Preference Semantics was a WSD program to do just that, and to provide a new sense where WSD failed.

The other starting point of the thesis was a slim paper by Bosanquet on the nature of metaphysical discourse, entitled "Some Remarks on Spinoza's Ethics." He argued that Spinoza's logical arguments are all false, but that what Spinoza was actually doing is *rhetorical*, *not logical*: imposing a new sense on the reader. The system as implemented was, of course, a toy system, in the sense that all symbolic NLP systems were in that era. It consisted of an analysis of five metaphysical texts (by Wittgenstein, Spinoza, Descartes, Kant, and Leibniz) along with five randomly chosen passages from editorials in the London *Times*, as some sort of control texts.

The vocabulary was only about 500 words, but this was many years before Boguraev declared the average size of vocabularies in working NLP systems to be 36 words. The semantic structures derived—via what we would now call chunk parsing—consisted of tree structures of primitives (from a set of about 80), one tree for each participating word sense in the text chunk, that fitted into preformed triples called **templates**. These templates were subject–predicate–object triples that defined well-formed sequences of the triples of trees (i.e., the first tree for the sense of the subject, the second for the action and so on), whose tree-heads had to fit those of the template's three primitive items in order. The overall system selected the word senses that fitted into these structures by means of a notion of "semantic preference" (see subsequent discussion), and then declared those to be the appropriate senses for the words, thus doing a primitive kind of WSD.

There was in the thesis an additional "sense constructor" mode, called if the WSD did not work, which tried to identify some sense of a word in the text whose representation would fit in the overall structure derived, and so could be declared a suitable "new" sense for the word which had previously failed to fit in. Unsurprisingly, it identified, say, a sense of "God" in the Spinoza text with an existing sense of "Nature" so that, after this substitution, the whole thing fitted together and WSD could proceed, and thus the passage be declared meaningful, given the criterion of having a single, ambiguity-free, interpretation. This was the toy procedure that allowed me to argue that Spinoza's real aim, whether he knew it or not, was to persuade us that the word "God" could have the sense of "Nature" and that this was the real point of his philosophy—exactly in line with what Bosanquet had predicted.

The philosophy work was never really published, outside an obscure McGill University philosophy journal, although the meaningfulness criterion appeared in *Mind* in 1971 under the title "Decidability and Natural Language" (Wilks 1971). Since publishing in *Mind* was, at the time, the ambition of every young philosopher, I was now satisfied and could move to the simpler world of NLP. The thesis, shorn of the metaphysics, appeared as my first book, *Grammar, Meaning and the Machine Analysis of Language* (Wilks 1972); the title was intended as a variation on the title of some strange German play, popular at the time, and whose actual name I can no longer remember.

Preference Semantics

I returned from California to CLRU but left again for the Stanford AI Lab in 1969. I had fantasized at CLRU about all the things one could do with a methodology of trying to base a fairly complex compositional semantics on a foundation of superficial pattern matching. This had earlier produced speculations like my 1964 CLRU paper "Text searching with templates," procedures that we could not possibly have carried

```
I.1 ((*ANI 1)((SELF IN)(MOVE CAUSE))(*REAL 2))→(1(*JUDG) 2)
      Or, in semi-English:
      [animate-1 cause-to-move-in-self real-object-2]→[1 *judges 2]
I.2 (1 BE (GOOD KIND))↔((*ANI 2) WANT 1)
      Or, again:
      [1 is good]↔[animate-2 wants 1]
```

Figure 1
Inference rules in Preference Semantics.

out with the machines then available, but which I now choose to see as wanting to do Information Extraction: though, of course, it was Naomi Sager who did IE first on medical texts at NYU (see Sager and Grishman 1975).

At Stanford as a post-doc, I was on the same corridor as Winograd, just arrived from MIT; Schank, then starting to build his Conceptual Dependency empire; and Colby and his large team building the PARRY dialogue system, which included Larry Tesler, later the Apple software architect. Schank and I agreed on far more than we disagreed on and saw that we would be stronger together than separately, but neither of us wanted to give up our notation: He realized, rightly, that there was more persuasive power in diagrams than in talk of processes like "preference." It was an extraordinary period, when AI and NLP were probably closer than ever before or since: Around 1972 Colmerauer passed through the Stanford AI Lab, describing Prolog for the first time but, as you may or may not remember, as a tool for machine translation! I spent my time there defining and expanding the coherence-based semantics underlying my thesis, calling it "Preference Semantics" (PS), adding larger scale structures such as inference rules (see Figure 1) and thesauri, and building it into the core of a small semantics-based English-to-French machine translation system programmed in LISP. At one point the code of this MT system ended up in the Boston Computer Museum, but I have no idea where it is now. The principles behind PS were as follows:

- an emphasis on processes, not diagrams;
- the notion of affinity and repulsion between sense representations (cf. Waltz and Pollack's WSD connectionism [1985]);
- seeking the "best fit" interpretation—the one with most satisfied preferences (normally of verbs, prepositions, and adjectives);
- yielding the least informative/effort interpretation;
- using no explicit syntax, only segmentation and order of items;
- meaningfulness as being connected to a unique interpretation/sense choice;
- meaning seen as represented in other words, since no other equivalent for the notion works (e.g., objects or concepts);
- gists or templates of utterances as core underlying entities; and
- there is no *correct* interpretation or set of primitive concepts, only the best available.

One could put some of these, admittedly programmatic and imprecise, points as follows:

- Semantics is not necessarily deep but also superficial (see more recent results on the interrelations between WSD, POS, and IE, e.g. Stevenson and Wilks [2001]).
- Quantitative phenomena are unavoidable in language: John McCarthy thought they had no place anywhere in AI, except perhaps in low-level computer vision.
- Reference structures (like lexicons) are only temporary snapshots of a language in a particular state (of expansion or contraction).
- What is important is to locate the update mechanism of language, including crucially the creation of new word senses, which is not Chomsky's sense of the creativity of language.

Constructible Belief Systems

I returned to Europe in the mid 1970s, first to the ISSCO institute in Lugano, where Charniak was and Schank had just left, and then to Edinburgh as a visitor before taking a job at Essex. I began a long period of interest in belief systems, in particular seeking some representation of the beliefs of others, down to any required degree of nesting—for example, A's belief about B's belief about C—that could be constructed recursively at need, rather than being set out in advance, as in the pioneering systems emerging from the Toronto group under Ray Perrault (Allen and Perrault 1980). I began thinking about this with Janusz Bien of the University of Warsaw, who had also published a paper arguing that CL/NLP should consider "least effort" methods: in the sense that the brain might well, due to evolution, be a lazy processor and seek methods for understanding that minimized some value that could be identified with processing effort. I had argued in PS for choosing shortest chains of inferences between templates, and that the most connected/preferred template structure for a piece of text should be the one found first. I am not sure we ever proved any of this: It was just speculation, as was the preference for the most semantically connected representation, and the representation with the least information. All this is really only elementary information theory: a random string of words contains the maximum information, but that is not very helpful. Clearly, the preferred interpretation of "He was named after his father" (i.e., named *the same* rather than *later in time*) is not the least informative, since the latter contains no information at all—being necessarily true—so one would have to adapt any such slogan to: "prefer the interpretation with the least information, unless it is zero!"

The belief work, first with Bien, later with Afzal Ballim (Wilks and Ballim 1987) and John Barnden, has not been a successful paradigm in terms of take-up, in that it has not got into the general discourse, even in the way that Fauconnier's "Mental Spaces" (Fauconnier 1985) has. That approach uses the same spatial metaphor, but for strictly linguistic rather than belief and knowledge purposes. But I think the VIEWGEN belief paradigm, as it became, had virtues, and I want to exploit this opportunity to remind people of it. It was meant to capture the intuition that if we want, for language understanding purposes, to construct X's beliefs about Y's beliefs—what I called the environment of Y-for-X—then:

1. It must be a construction that can be done in real time to any level of nesting required, because we cannot imagine it pre-stored for all future nestings, as Perrault el al. in effect assumed.
2. It must capture the intuition that much of our belief is accepted by default from others: As VIEWGEN expresses it, I will accept as a belief what you say, because I have normally no way of checking, or experimenting on, let alone refuting, the things you tell me, e.g., that you had eggs for breakfast yesterday. As someone in politics once put it, "There is no alternative." Unless, that is, what you say contradicts something I believe or can easily prove from what I believe.
3. We must be able to maintain apparently contradictory beliefs, provided they are held in separate spaces and will never meet as contradictions. I can thus maintain within my-space-for-you beliefs of yours (according to me) that I do not in fact hold.

In VIEWGEN, belief construction is done in terms of a "push down" metaphor: A permeable "container" of your beliefs is pushed into a "container" of my beliefs and what percolates through the membrane, from me to you, will be believed and ascribed to you, unless it is explicitly contradicted, namely, by some contrary belief I already ascribe to you, and which, as it were, keeps mine from percolating through. The idea is to construct the appropriate "inner belief space" at the relevant level of nesting, so that inference can be done, and to derive consequences (within that constrained content space) that also serve to model, in this case, you the belief holder in terms of goals and desires, in addition to beliefs. This approach is quite different not only from the Perrault/Toronto system of belief-relevant plans but also to AI theories that make use of sets-of-support premises, since this is about belief-inheritance-by-default. It is also quite distinct from linguistic theories like Wilson and Sperber's Relevance Theory, which take no account at all of belief as relative to individuals, but perform all operations in some space that is the same for everyone, which is an essentially Chomskyan ideal competence-style notion of belief that is not relative to individuals—which is of course absurd.

Mark Lee and a number of my students have created implementations of this approach and linked it to dialogue and other applications, but there has been no major application showing its essential role in a functioning conversational theory where complex belief states are created in real time. However, the field is, I believe, now moving in that direction (e.g., with POMDP theories [Williams and Young 2007]) since the possibility of populating belief theories with a realistic base from text by means of Information Extraction or Semantic Web parsing to RDF format is now real (a matter we shall return to subsequently).

There were, for me at least, two connections between the VIEWGEN belief work and Preference Semantics, in terms of meaning and its relation to processes. First, there was the role of choice and alternatives, crucial to PS, in that an assigned meaning interpretation for a text was no more than a choice of the best available among alternatives, because preference implies choice, in a way that generative linguistics—though not of course traditions like Halliday's—always displayed alternatives but considered choice between them a matter for mere performance. What was dispensable to generative linguistics was the heart of the matter, I argued, to NLP/CL. Secondly, VIEWGEN suggested a view of meaning, consistent locally with PS, dependent on which individuals or classes one chose to see in terms of each other—the key notion here was seeing one thing as another and its consequences for meaning. So, if one chose to identify (as being the same person under two names) Joe (and what one believed about him) with Fred's father (and what one knew about him), the hypothesis was that a belief environment should be constructed for Joe-as-Fred's-father by percolating one set of beliefs into the other, just as was done by the basic algorithm for creating A's-beliefs-about-B's-beliefs from the component beliefs of A and B. This process created a hybrid entity, with intensional meaning captured by the set of propositions in that inner environment of belief space, but which was now neither Joe nor Fred's father but rather the system's point of view of their directional amalgamation: Joe-as-Fred's-father (which might contain different propositions from the result of Fred's-father-as-Joe).

More natural, and fundable, scenarios were constructed for this technique in those days, such as knowledge representations for Navy ships' captains genuinely uncertain as to whether ship-in-my-viewfinder-now was or was not to be identified with the stored representation for enemy-ship-number-X. The important underlying notion was one going back to Frege, and which first had an outing in Winograd's thesis (Winograd 1972), where he showed you could have representations for blocks that did not in fact exist on the Blocks World table. A semantics must be able to represent things without knowing whether they exist or not; that is a basic requirement.

Later, and working with John Barnden and Afzal Ballim, this same underlying process of conflating two belief objects was extended to the representation of "metaphorical objects," which could be described, quite traditionally in the literature, as A-viewed-as-B (e.g., an atom viewed as a billiard ball). The metaphorical object atom-as-billiard-ball was again created by the same push-down or fusion of belief sets as in the basic belief point-of-view procedure. All this may well have been fanciful, and was never fully exploited in published work with programs, but it did have a certain intellectual appeal in wanting to treat belief, points of view, metaphor and identification of intensional individuals—normally quite separate issues in semantics—as being modellable by the same simple underlying process (see Ballim, Wilks, and Barnden 1991). One novel element that did emerge from this analysis was that, in the construction of these complex intensional identifications, such as between "today's Wimbledon winner" and "the top male tennis seed," one could choose directions of "viewing as" with the belief sets that led to objects which were neither the classic *de re* nor *de dicto* outcomes: Those became just two among a range of choices, and the others of course had no handy Latin names.

Adapting to the "Empirical Wave" in NLP

For me, as with many others, especially in Europe, the beginning of the empirical wave in NLP was the work of Leech and his colleagues at Lancaster: CLAWS4 (a name which hides a UK political joke), their part-of-speech tagger based on large-scale annotation of corpora. Such tagging is now the standard first stage of almost every NLP process and it may be hard for some to realize the skepticsm its arrival provoked: "What could anyone want that for?" was a common reaction from those still preoccupied by computational syntax or semantics. That system was sold to IBM, whose speech group, under Jelinek, Mercer, and Brown, subsequently astonished the CL/NLP world with their statistical machine translation system CANDIDE. I wrote critical papers about it at the time, not totally unconnected to the fact that I was funded by DARPA on the PANGLOSS project at NMSU (along with CMU and ISI/USC) to do MT by competing, but non-statistical, methods.

In one paper, I used the metaphor of "stone soup" (Wilks 1996): A reference to the old peasant folk-tale of the traveler who arrives at a house seeking food and claiming to have a stone that makes soup from water. He begs a ham bone to stir the water and stone and eventually cons out of his hosts all the ingredients for real soup. The aspect of the story I was focusing on was that, in the CANDIDE system, I was not sure that the "stone," namely IBM's "fundamental equation of MT," was in fact producing the results, and suggested that something else they were doing was giving them their remarkable success rate of about 50% of sentences correctly translated. As their general methodology has penetrated the whole of NLP/CL, I no longer stand by my early criticisms; IBM was of course right, and had everything to teach the rest of us.

Early critics of data-driven, alias empirical, CL found it hard to accept, whatever its successes in, say, POS tagging, that its methods could extend to the heartland of semantics and pragmatics. Like others, I came to see this assumption was quite untrue, and myself moved towards Machine Learning (ML) approaches to word-sense disambiguation (e.g., Stevenson and Wilks 2001) and I now work in ML methods applied to dialogue corpora (as I shall mention subsequently). But the overall shift in approaches to semantics since 1990 has not only been in the introduction of statistical methods, and ML in particular, but also in the unexpected advantages that have been gained from what one might call non-statistical empirical linguistics, and in particular Information Extraction (IE; see Wilks 1997).

I referred earlier to the fact that my early work could be called, in a general sense, semantic parsing, and that it was in fact some form of superficial pattern matching onto language chunks that was then transformed to different layers of compositional semantic representation. There were obvious relations between that general approach and what emerged from the DARPA competitions in the early 1990s as IE, a technology that, when honed by many teams, and especially when ML techniques were added to it later, had remarkable success and a range of applications; it also expanded out into other, traditionally separate, NLP areas such as question answering and summarization. This approach is not in essence statistical at all, however, although it is in a clear sense "superficial," with the assumption that semantics is not necessarily a "deep" phenomenon but present on the language surface. I believe the IE movement is also one of the drivers behind the Semantic Web movement, to which I now turn, and which I think has brought NLP back to a position nearer the core of AI, from which it drifted away in the 1980s.

Meaning and the Semantic Web

The Semantic Web (SW; Berners-Lee, Hendler, and Lassila 2001) is what one could call Berners-Lee's second big idea, after the World Wide Web; it can be described briefly as turning the Web into something that can also be understood by computers in the way that it is understood by people now, as a web of texts and pictures. Depending on one's attitude to this enterprise, already well-funded by the European Commission at least, it can be described as any of the following:

1. As a revival of the traditional AI goal (at least since McCarthy and Hayes [1969]) of replacing language, with all its vagueness, by some form of logical representation upon which inference can be done.

2. As a hierarchy of forms of annotation—or what I shall call augmentation of content—reaching up from simple POS tagging to semantic class annotation (e.g. CITY, PERSON-NAME) to ontology membership and logical forms. DARPA/MUC/NIST competitions have worked their way up precisely this hierarchy over the years and many now consider that content can be "annotated onto language" reliably up to any required level. This can be thought of as extending IE techniques to any linguistic level by varieties of ML and annotation.

3. As a system of access to trusted databases that ground the meanings of terms in language; your telephone or social security number might ground you uniquely (in what is called a URI), or better still—and this is now the standard view—a unique identifying object number for you over and above phones and social systems. This is very much Tim Berners-Lee's own view of the SW.

There is also a fourth view, much harder to express, that says roughly that, if we keep our heads, the SW can come into being with any system of coding that will tolerate the expansion of scale of the system, in the way that, miraculously, the hardware underpinnings of the World Wide Web have tolerated its extraordinary expansion without major breakdown. This is an engineering view that believes there are no fundamental problems about the meanings and reference of SW terms in, for example, the ontologies within the SW, and everything will be all right if we just hold tight.

This view may turn out to be true but it is impossible to discuss it. Similarly, view (3) has no special privilege because it is the World Wide Web founder's own view: Marx was notoriously not a very consistent Marxist, and one can find multiple examples of this phenomenon. View (3) is highly interesting and close to philosophical views of meaning expressed over many years by Putnam, which can be summarized as the idea that scientists (and Berners-Lee was by origin a database expert and physicist) are "guardians of meaning" in some sense because they know what terms really mean, in a way that ordinary speakers do not. Putnam's standard example is that of metals like molybdenum and aluminum, which look alike and, to the man in the street, have the same conceptual, intensional meaning, namely light, white, shiny metal. But only the scientist (says Putnam) knows the real meanings of those words because he knows the atomic weights of the two metals and methods for distinguishing them.

No one who takes Wittgenstein—and his view that we, the users of the language, are in charge of what terms mean, and not any expert—at all seriously can even consider such a view. On the view we are attributing to Wittgenstein, the terms are synonymous in a public language, just as *water* and *heavy water* are, and any evidence to the contrary is a private matter for science, not for meaning.

View (1) of the Semantic Web is a well-supported one, particularly by recycled AI researchers: They have, of course, changed tack considerably and produced formalisms for the SW, some of which are far closer to the surface of language than logic (what is known as RDF triples), as well as inference mechanisms like DAML-OIL that gain advantages over traditional AI methods on the large and practical scale the SW is intended to work over. On the other hand there are those in AI who say they have ignored much of the last 40 years of AI research that would have helped them. This dispute has a conventional flavor and it must be admitted that, in more than 40 years, AI itself did not come up with such formalisms that stood any chance at all of working on a large scale on unstructured material (i.e., text).

This leaves us with View (2), which is my own: namely, that we should see the SW partially in NLP terms, however much Berners-Lee rejects such a view and says NLP is irrelevant to the SW. The whole trend of SW research, in Europe at least, has been to build up to higher and higher levels of semantic annotation—a technology that has grown directly out of IE's success in NLP—as a way of adding content to surface text. It seems to me obvious that any new SW will evolve from the existing WWW of text by some such method, and that method is basically a form of large-scale NLP, which now takes the form of transducers from text to RDF (such as the recently advertised Reuters API). The idea that the SW can start from scratch in some other place, ignoring the existing World Wide Web, seems to me unthinkable; successful natural evolution always adapts the function of what is available and almost never starts again afresh.

I have set out my views on this recently in more detail (Wilks 2008), but it is important to see that the SW movement—at least as I interpret it herein, and that does seem pretty close to the way research in it is currently being funded, under calls and titles like "semantic content"—is one that links to the themes already developed in this paper in several ways, and which correspond closely to issues in my own early work, but which have not gone away:

1. The SW takes semantic annotation of content as being a method—whether done by humans or after machine learning—of recoding content with special terms, terms close to what have traditionally been called semantic primitives. It is exactly this that was denied by the early forms of, say, statistical MT, where there was nothing available to the mechanism except the words themselves. This is also quite explicit in traditional IR, where, for example, Karen Spärck Jones consistently argued against any form of content recoding, including the SW. As she put it: "One of these [simple, revolutionary IR] ideas is taking words as they stand" (Spärck Jones 2003).
2. The SW accords a key role to ontologies as knowledge structures: partially hierarchical structures containing key terms—primitives again under another guise—whose meanings must be made clear, particularly at the more abstract levels. The old AI tradition in logic-based knowledge structuring—descending from McCarthy and Hayes (1969)—was simply to declare what these primitive predicates meant. The problem was that predicates, normally English words written in capital letters (as all linguistic primitives in the end seem to be), became affected by their inferential roles over time and the process of coding itself. This became very clear in the long-term Cyc project (Lenat 1995) where the key predicates changed their meanings over 30 years of coding, but there was no way of describing that fact within the system, so as to guarantee consistency. In Nirenburg and Wilks (2000), Nirenburg and I debate this issue in depth, and I defend the position that one cannot simply maintain the meanings of such terms by fiat and independent of their usage—they look like words and they function like words because, in the end, they are words. The SW offers a way out of this classic AI dilemma by building up the hierarchy of annotations with empirical processes like ontology induction from corpora (e.g., ABRAXAS; see Iria et al. 2006); in this way the meanings of higher level terms are connected back directly to text usage. Braithwaite, my thesis advisor, described in his classic "Scientific explanation" (Braithwaite 1953) a process in the philosophy of science he

called “semantic ascent” by which the abstract high-level terms in a scientific theory, seen as a logical hierarchy of deductive processes—terms such as “neutron,“ possibly corresponding to unobservables—acquired meaning by an ascent of semantic interpretation up the theory hierarchy from meanings grounded in experimental terms at the bottom. It is some such grounding process I envisage the SW as providing for the meanings of primitive ontological terms in a knowledge structure.

3. The RDF forms, based on triples of surface items, as a knowledge base—usually with subject–action–object as basic form—can provide a less formal but more tractable base for knowledge than traditional First Order Predicate Logic (FOPL). They have a clear relationship back to the crude templates of my early work and the later templates of IE. I claim no precedence here, but only note the return of a functioning but plausible notion of “superficial semantics.” It seems to me not untrue historically to claim that RDF, the representational base of the SW, is a return of the level of representation that Schank (under the name Conceptual Dependency, in Schank [1975]) and I (under the name Preference Semantics) developed in the late 1960s and early 1970s (Wilks 1975). I remember that at the Stanford AI Lab at that time, John McCarthy, a strong advocate of FOPL as the right level of representation of language content, would comment that formalisms like these two might have a role as a halfway house on a route from language to a full logic representation. On one view of the SW that intermediate stage may prove to be the right stage, because full AI representations have never been able to deliver in terms of scale and tractability. Time will tell, and fairly soon.

The most important interest of the SW, from the point of view of this paper, is that it provides at last a real possibility of a large-scale test of semantic and knowledge coding: One thing the empirical movement has taught us is the vital importance of scale and the need to move away from toy systems and illustrative examples. I mentioned earlier the freely available Reuters API for RDF translation which Slashdot advertised under the title “Is the Semantic Web a Reality at Last?” This is exactly the kind of move to the large scale that we can hope will settle definitively some of these ancient issues about meaning and knowledge.

A Late Interest in Dialogue: The Companions Project

My only early exposure to dialogue systems was Colby’s PARRY: As I noted earlier, his team was on the same corridor as me at Stanford AI Lab in the early 1970s. I was a great admirer of the PARRY system: It seemed to me then, and still does, probably the most robust dialogue system ever written. It was available over the early ARPANET and tried out by thousands, usually at night: It was written in LISP and never broke down; making allowances for the fact it was supposed to be paranoid, it was plausible and sometimes almost intelligent. In any case it was infinitely more interesting than ELIZA, and it is one of the great ironies of our subject that ELIZA is so much better known. PARRY remembered what you had said, had elementary emotion parameters and, above all, had something to say, which chatbots never do. John McCarthy, who ran the AI Lab, would never admit that PARRY was AI, even though he tolerated it under his roof, as it were, for many years; he would say “It doesn’t even know who the President is," as if most of the world's population did! PARRY was in fact a semi-refutation of the claim that you need knowledge to understand and converse, because it plainly knew nothing; what it had was primitive "intentionality," in the sense that it had things "it wanted to say."

My own introduction to practical work on dialogue was when I was contacted in the late 1990s by David Levy, who had written 40 books on chess and ran a company that made chess machines. He already had a footnote in AI as the man who had bet McCarthy, Michie, and other AI leaders that a chess machine would not beat him within ten years, and he won the bet more than once. In the 1990s he conceived a desire to win the Loebner Prize[2] for the best dialogue program of the year, and came to us at Sheffield to fund a team to win it for him, which we did in 1997. I designed the system and drew upon my memories of PARRY, along with obvious advances in the role of knowledge bases and inference, and the importance of corpora and machine learning. For example, we took the whole set of winning Loebner dialogues off the Web so as to learn the kinds of things that the journalist-testers actually said to the trial systems to see if they were really humans or machines.

Our system, called CONVERSE (see Levy et al. 1997), claimed to be Catherine, a 34-year old female British journalist living in New York, and it owed something to PARRY, certainly in Catherine's desire to tell people things. It was driven by frames corresponding to each of about 80 topics that such a person might want to discuss; death, God, clothes, make-up, sex, abortion, and so on. It was far too top-down and unwilling to shift from topic to topic but it could seem quite smart on a good day, and probably won because we had built in news from the night before the competition of a meeting Bill Clinton had had that day at the White House with Ellen de Generes, a lesbian actress. This gave a certain immediacy to the responses intended to sway the judges, as in "Did you see that meeting Ellen had with Clinton last night?"

This was all great fun and gave me an interest in modeling dialogue that has persisted for a decade and is now exercised through COMPANIONS (Wilks 2004), a large EU 15-site four-year project that I run. COMPANIONS aims to change the way we think about the relationships of people to computers and the Internet by developing a virtual conversational "Companion." This will be an agent or "presence" that stays with the user for long periods of time, developing a relationship and "knowing" its owner's preferences and wishes. It will communicate with the user primarily by using and understanding speech, but also using other technologies such as touch screens and sensors.

Another general motivation for the project is the belief that the current Internet cannot serve all social groups well, and it is one of our objectives to empower citizens (including the non-technical, the disabled, and the elderly) with a new kind of interface based on language technologies. The vision of the Senior Companion—currently our main prototype—is that of an artificial agent that communicates with its user on a long-term basis, adapting to their voice, needs, and interests: A companion that would entertain, inform, and react to emergencies. It aims to provide access to information and services as well as company for the elderly by chatting, remembering past conversations, and organizing (and making sense of) the owner's photographic and image memories. This Companion would assume a user with a low level of technical knowledge, and who might have lost the ability to read or produce documents themselves unaided, but who might need help dealing with letters, messages, bills, and getting information from the Internet. During its conversations with its user or owner, the system

[2] See http://www.loebner.net/Prizef/loebner-prize.html.

builds up a knowledge inventory of family relations, family events in photos, places visited, and so on. This knowledge base is currently stored in RDF, the Semantic Web format, which has two advantages: first, a very simple inference scheme with which to drive further conversational inferences, and second, the possibility, not yet fulfilled, of accessing arbitrary amounts of world information from Wikipedia, already available in RDF, which could not possibly have been pre-coded in the dialogue manager, nor elicited in a conversation of reasonable length. So, if the user says a photo was taken in Paris, the Companion should be able to ask a question about Paris without needing that knowledge pre-coded, but only using rapidly accessed Wikipedia RDFs about Paris. An ultimate aim of this aspect of the Senior Companion is the provision of a life narrative, an assisted autobiography for everyone, one that could be given to relatives later if the owner chose to leave it to them. There is a lot of technical stuff in the Senior Companion: script-like structures—called DAFs or Dialogue Action Forms—designed to capture the course of dialogues on specific topics or individuals or images, and these DAFs we are trying to learn from tiled corpora. The DAFs are pushed and popped on a single stack, and that simple virtual machine is the Dialogue Manager, where DAFs being pushed, popped, or reentered at a lower stack point are intended to capture the exits from, and returns to, abandoned topics and the movement of conversational initiative between the system and the user. We are halfway through the project and currently have two prototype Companions: The other, based not at Sheffield but at Tampere, is a Health and Fitness Companion (HFC).[3] It is more task-oriented than the Senior Companion and aims to advise on exercise and diet. The HFC is on a mobile phone architecture as well as a PC, and we may seek to combine the two prototypes later. The central notion of a Companion is that of the same "personality," with its memory and voice being present no matter what the platform. It is not a robot, and could be embodied later in something like a chatty furry handbag, being held on a sofa and perhaps reminding you about the previous episodes of your favorite TV program.

Finale

This article has had something of the form of a life story, and everyone wants to believe their life is some kind of narrative rather than a random chase from funding agency to funding agency, with occasional pauses to carry out a successful proposal. But let us return to Newton for a moment in closing; for us in CL he is the great counter-example, of why we do not do science or engineering in that classic solitary manner:

> . . . where the statue stood
> Of Newton, with his prism and silent face,
> The marble index of a mind for ever
> Voyaging through strange seas of Thought, alone.
>
> —William Wordsworth (1770–1850)
> *The Prelude*, book iii, line 61

The emphasis there for me is on *alone*, which is pretty much unthinkable in our research world of teams and research groups. Our form of research is essentially corporate and cooperative; we may not be sure whose shoulders we are standing on, but we know whose hands we are holding. I have worked in such a way since my thirties and, at

[3] An early demo of a Companion can be seen on YouTube at `http://www.youtube.com/watch?v=SqIP6sTt1Dw`.

Sheffield, my work would not have been possible without a wide range of colleagues and former students in the NLP group there over many years and including Louise Guthrie, Rob Gaizauskas, Hamish Cunningham, Fabio Ciravegna, Mark Stevenson, Mark Hepple, Kalina Bontcheva, Roberta Catizone, Nick Webb, and many others. In recent years, what one could call "DARPA culture"—of competitions and cooperation subtly mixed—as well as the great repositories of software and data like LDC and ELRA, have gone a long way to mitigate the personal and group isolation in the field.

But we do have to face the fact that, in many ways, we do not do classic science: We have no Newtons and will never have any. That is not to deny that we need real ideas and innovations, and now may be a time for fresh ones. We have stood on the shoulders of Fred Jelinek, Ken Church, and others for nearly two decades now, and the strain is beginning to tell as papers still strive to gain that extra 1% in their scores on some small task. We know that some change is in the air and I have tried to hint in this article as to some of the places where that might be, even if that will mean a partial return to older, unfashionable ideas; for there is nothing new under the sun. But locating them and exploiting them will not be in my hands but in yours, readers of *Computational Linguistics*!

Acknowledgments

First of course to all those who have worked with me over many years and to whom I owe so much, particularly in connection with this award. Then to my current sponsor: This work was funded 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

Allen, James F. and C. Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence*, 15:143–178.

Ballim, Afzal, Yorick Wilks, and John A. Barnden. 1991. Belief ascription, metaphor, and intensional identification. *Cognitive Science*, 15(1):133–171.

Berners-Lee, T., J. Hendler, and O. Lassila. 2001, September. The semantic web. *Scientific American*, 28–37.

Braithwaite, Richard Bevan. 1953. *Scientific Explanation. A Study of the Function of Theory, Probability and Law in Science*. Cambridge University Press, Cambridge, UK.

Carnap, Rudolf. 1937. *The Logical Syntax of Language*. Kegan Paul, London.

Fauconnier, Gilles. 1985. *Mental Spaces*. Cambridge University Press, Cambridge, UK.

Iria, José, Christopher Brewster, Fabio Ciravegna, and Yorick Wilks. 2006. An incremental tri-partite approach to ontology learning. In *Proceedings of the Language Resources and Evaluation Conference (LREC-06)*, 22–28 May.

Lenat, Douglas B. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Levy, D., R. Catizone, B. Battacharia, A. Krotov, and Y. Wilks. 1997. Converse: A conversational companion. In *Proceedings of the First International Workshop of Human-Computer Conversation*. Bellagio, Italy.

Masterman, Margaret. 1961. Semantic message detection for machine translation, using an interlingua. In *Proceedings of the First International Conference on Machine Translation of Languages and Applied Language Analysis*, pages 438–475. HMSO, Teddington, Middlesex, UK.

Masterman, Margaret. 2005. In Yorick Wilks, editor, *Language, Cohesion and Form (Studies in Natural Language Processing)*. Cambridge University Press, New York.

McCarthy, J. and P. J. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence*, volume 4. Edinburgh University Press, Edinburgh, pages 463–502.

Nirenburg, Sergei and Yorick Wilks. 2000. Machine translation. *Advances in Computers*, 52:160–189.

Sager, Naomi and Ralph Grishman. 1975. The restriction language for computer grammars of natural language. *Communications of the ACM*, 18(7):390–400.

Schank, Roger C. 1975. *Conceptual Information Processing*. Elsevier Science Inc., New York.

Spärck Jones, Karen. 1986. *Synonymy and semantic classification*. Edinburgh University Press, Edinburgh, Scotland.

Spärck Jones, Karen. 2003. Document retrieval: Shallow data, deep theories; historical reflections, potential directions. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, 1–11. Springer, Berlin/Heidelberg.

Stevenson, Mark and Yorick Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349.

Waltz, David L. and Jordan B. Pollack. 1985. Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*, 9(1):51–74.

Wilks, Y. 1975. Preference semantics. In E. L. Keenan, editor, *Formal Semantics of Natural Language*. Cambridge University Press, Cambridge, pages 329–348.

Wilks, Yorick. 1971. Decidability and natural language. *Mind*, 80:497–520.

Wilks, Yorick. 1972. *Grammar, Meaning and Machine Analysis of Language*. Routledge and Kegan Paul, London.

Wilks, Yorick. 1996. Statistical versus knowledge-based machine translation. *IEEE Expert: Intelligent Systems and Their Applications*, 11(2):12–18.

Wilks, Yorick. 1997. Information extraction as a core language technology. In *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, volume 1299 of *Lecture Notes In Computer Science*, pages 1–9, Springer, Berlin.

Wilks, Yorick. 2004. Artificial companions. In *Machine Learning for Multimodal Interaction: First International Workshop*, pages 36–45.

Wilks, Yorick. 2008. The semantic web: Apotheosis of annotation, but what are its semantics? *IEEE Intelligent Systems*, 23(3):41–49.

Wilks, Yorick and Afzal Ballim. 1987. Multiple agents and the heuristic ascription of belief. In *Proceedings of the International Joint Conference Artificial Intelligence (IJCAI-87)*, pages 118–124, Milan, Italy.

Williams, Jason D. and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422.

Winograd, Terry. 1972. *Understanding Natural Language*. Academic Press, Orlando, FL.

Wittgenstein, Ludwig. 1973. *Philosophical Investigations*. Blackwell Publishers, Oxford, UK.