A Wittgensteinian computational linguistics?

Yorick Wilks
University of Sheffield

Abstract. The paper tries to relate Wittgenstein's later writings about language with the history and content of Artificial Intelligence (AI), and in particular, its sub-area normally called Computational Linguistics, or Natural Language Processing. It argues that the shift, since 1990, from rule-driven approaches to computational language and logic, associated with traditional AI and the linguistics of Chomsky, to more statistical models of language have made those connections more plausible, in particular because there is good reason to think the latter is a better model of use than the former. What statistical language models are not, of course, are immediately plausible models of meaning. Moreover, a statistical model seeking a model of a whole language, one can now look at the world wide web (WWW) as an encapsulation of the usage of a whole a language, open to computational exploration, and of a kind never before available. I describe a recent empirical effort to give sense to the notion of a model of a whole language derived from the web, but whose disadvantage is that that model could never be available to a language user because of the sheer size of the WWW. The problematic issue in such an analogy (Wittgenstein and NLP) is how one can go beyond the anti-rule aspect of both to some view of how concepts can even appear to exist, whatever their true status.

"….in philosophy we often compare the use of words with games and calculi which have fixed rules, but cannot say that someone who is using language must be playing such a game. But if you say that our languages only approximate to such calculi you are standing on the very brink of a misunderstanding. For then it may look as if what we are talking about is an ideal language. As if our logic were, so to speak, a logic for a vacuum. - Whereas logic does not treat of language - or of thought - in the sense in which a natural science treats a natural phenomenon, and the

most that can be said is that we construct ideal languages. But here the word "ideal" is liable to mislead, for it sounds as if these languages were better, more perfect, than our everyday language. and as if it took the logician to shew people at last what a proper sentence looked like." Wittgenstein: Philosophical Investigations $81 (my emphasis).

"A main source of our failure to understand is that we do not command a clear view of the use of our words - ---our grammar is lacking in this sort of perspicuity. A perspicuous representation produces just that understanding which consists in "seeing connexions". Hence the importance of finding and inventing intermediate cases. The concept of a perspicuous representation is of fundamental significance for us. It earmarks the form of account we give, the way we look at things." Wittgenstein: Philosophical Investigations $122.

Introduction

Seeking out its intellectual roots or scholarly ancestors is not an activity popular or respected in Natural Language Processing (NLP, alias Computational Linguistics). Many people have some vague notion that logical predicate representation, now almost a form of shorthand in NLP, owes a lot to Frege and Russell in the late 19th Century, but few know or care that, long before Chomsky (if we agree to allow him by courtesy into the history of NLP) Carnap, Chomsky's teacher, set up in the 1930s what he called The Logical Syntax of Language (1936) with formation and transformation rules whose function was to separate meaningful from meaningless expressions by means of rules. Carnap's driving role behind all that has been utterly forgotten and Chomsky's own work has now simply filled in all the intellectual space.

Another contemporary of Carnap, also now lost to view, is Wittgenstein, whom Russell took to be a greater man than himself, and whose long campaign against simple-minded notions of linguistic rules was largely

provoked by Carnap. His life preceded Chomsky and NLP, though his influence lived on as the staple of Anglo-Saxon "linguistic philosophy" for many decades, but whose practitioners had little time or patience for what they saw as Chomsky's simplicities and certainties.

Some attempt to rectify this omission thirty years ago was Brown's "Wittgensteinian Linguistics" (1974), but his main concern was to contrast Wittgenstein with Chomsky's views, which were more central to language studies then than they are now; our concern here will be to contrast and compare Wittgenstein with developments specifically in NLP and computational linguistics, which has become more central within linguistics as a whole, as Chomsky's influence has declined.

Brown noted that Wittgenstein had much in common with Chomsky's anthropological predecessors, from whom he separated himself so clearly with his rule-driven, Carnap-inspired linguistics. Malinowski's observation (1923:287ff) that language is "a mode of action, rather than a counter-sign of thought" is a sentiment that Wittgenstein could have expressed (see REF), and the latter's notion of communities of use who share assumptions and language forms, however bizarre, is not far from anthropological views (often associated with Whorf and Sapir) on the language and belief systems valid in their own terms. Quine (1960) later took up the same scenario, that of remote languages, unknown to the observer, and the non-veridical nature of any communication based on translation or supposed meaning equivalence: how could we ever know definitively, he asked, what "Gavagai" meant simply from the utterances (and pointings) we observed?

Wittgenstein seemed less sceptical about translation than Quine; perhaps living in two languages and cultures, as he did, made it seem more natural to him: classic Wittgenstein apothegms like "the limits of my language mean the limits of my world" do not imply that one cannot be in two or more such worlds. He listed (PI pp.11-12) translation as among normal human activities, and he seemed sceptical about the nature and function of none of his list. It also seems clear that Wittgenstein did

believe in some conceptual world over and above surface use, but the problem is knowing what that was, and how it was grounded within usage. In his early work, what he called forms of facts (1961) were separate from language and identified with "pictures of fact" and it is not clear that he ever rejected the explanatory power of diagrams and pictures: he continued to use them, even though he was unsure how they "worked" (cf. The problem of knowing why the arrow so obviously points the way it does; PI $129). Pictures and drawings remained important to Wittgenstein because they expressed intention in a way that objects in the world do not.

In spite of many things he says that appear to be classic behaviourism – e.g. the apparent denial of the possibility of a private language (PI $243)---Wittgenstein was not an empiricist in the sense that Chomsky intended by that word, as is someone like Sampson (2004) who insists that we have no evidence that anything more is innate in humans than a learning mechanism. Wittgenstein could never have written "It is conceivable....that all the processes of understanding, obeying, etc. should have happened without the person ever having been taught the language" (PI $12) had that been his position. Moreover, Chomsky seems to have no understanding whatever of Wittgenstein's overall position, given remarks like: (Chomsky 1984: 60) "[For Wittgenstein] meanings of words must not only be learned, but also taught (the only means being drill, explanation. Or the supplying of rules….." . Chomsky has no feeling at all for Wittgenstein's investigation of how we could know that someone was following [a linguistic] rule, and for the simple reason that Chomsky always appears to know that we are following rules, and when, and to see no problem about a statement that a rule is being followed by a speaker.

It is this kind of claim, above all, that has made it hard for philosophers of language to take Chomsky seriously; he just does not see the problem that one cannot know, from linguistics or psychology, or even from physiology what rule a speaker is using, if any. And, contrary to Chomsky's whole edifice built on a speaker's intuition, a speaker is in no

special position to know what rule he is following, a point Wittgenstein demonstrated time after time (see below).

These arguments, that effectively separate Wittgenstein in every way from the Chomskyan enterprise, can be found in Brown's work, but I wish to add here that Chomsky and classic Artificial Intelligence (AI, e.g. McCarthy and Hayes, 1969)—with its emphasis on the role of logic as a "mental representation" are not in different positions here in contrast to Wittgenstein.

Since Brown, one can hear new hints of Wittgenstein's influence, as when Veronis called recently for looking "not for the meaning but the use" (1993), thus reviving one of the best known Wittgensteinian slogans.One could hear it, too, in Sinclair's call to let a corpus "speak to one" (Quoted in Moon, 2007), without the use of analytical devices and in Hanks' claim (1989) that a dictionary could be written consisting only of use citations. We suspect that this last is false, but it does have the authentic Wittgensteinian demand to look at language data, though not at all in the way a linguist would mean who gave the same exhortation (i.e. to form a generalization from it, in the linguist's case).

Wittgenstein, of course, knew nothing of computers in the modern sense, although he trained as an engineer. All I can do in this paper is to set out (1) the core of his doctrine on the nature of language was (2) what movements in modern NLP that doctrine is closer to and farther from, and (3) why his arguments and insights should still be taken account of by those concerned to process language by machine. This paper will not be about scholarly claims of direct influence, for there are probably few to be found. Margaret Masterman (see below), and perhaps the present author, are two of the very few NLP researchers who acknowledged his influence and referred to him often. One thinks here, too, of Graeme Hirst's immortal and not wholly serious (2000) "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.", which only serves to show how much remedial work there is to do.

Berlitz are currently running an advertisement in the US that explicitly uses another well known quotation from Wittgenstein: "The limits of my language are the limits of my world" and then goes on to offer courses in French that will help you order in a restaurant. To quote him, more or less correctly, is clearly not to understand him, given that those new food ordering powers, even in France, would not normally be said to move the limits of one's world. Yet perhaps that is an unfair and snobbish reaction: had the issue not been new restaurant behaviour but the acquisition of a new, and more exotic, language by an Anglophone, say Japanese, then the shift of world limits might have been more plausible and the force of the quotation would partly return, as when the Romanian poet Manea said he could leave Romania but not Romanian.

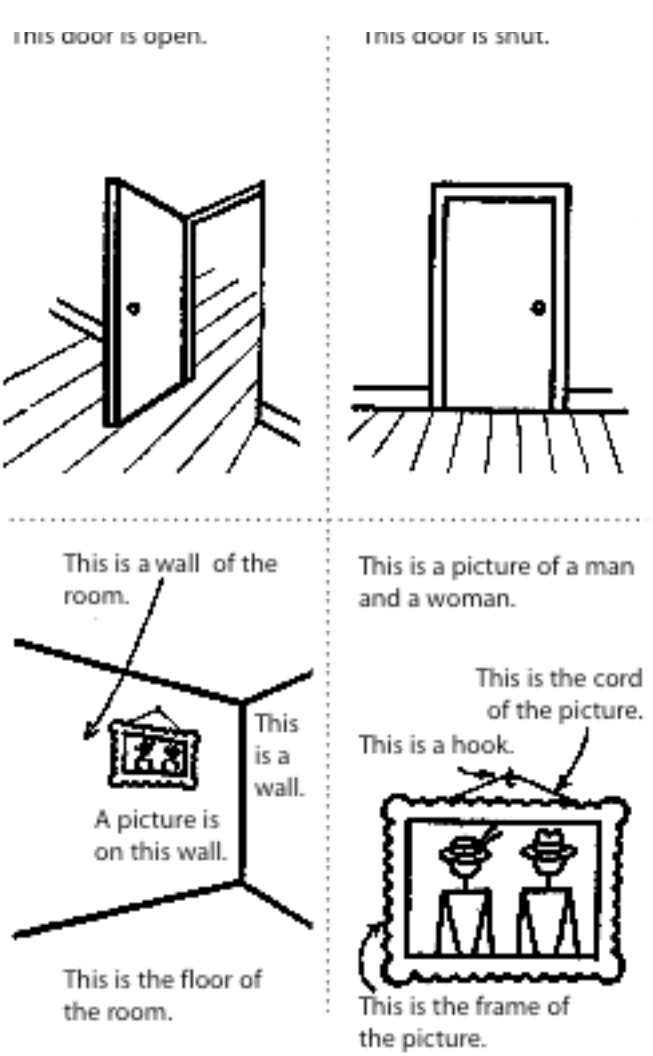
Which Wittgenstein?

It used to be conventional to distinguish an earlier from a later Wittgenstein: the earlier man wrote the Tractatus (reprinted: 1961) and believed in formalisms and a world of discrete facts; the later, and wiser, man wrote the Investigations (1958) to question all his earlier beliefs. Without necessarily endorsing this over-simple view of two-philosophers-in-one, we shall restrict ourselves in this paper largely to connections between NLP and the later Wittgenstein, if only because the connections between the logic-orientated earlier phase and the growth of the formalisms that led to much of modern AI and linguistics are all too clear, through Carnap in both cases.

One little noticed connection to the earlier Tractatus phase of Wittgenstein's work is the clear link between his metaphor of pictures as facts

(the so-called Picture Theory of Truth) and the stick-picture situations in Richards and Gibson's (1958) language teaching books, which were used for decades to teach languages without meta-explanations, using stick pictures as unambiguous situations, each expressing a simple proposition. Since Richards was close to Wittgenstein at Cambridge, the relationship is an obvious one. Masterman made much of the link, and used stickpictures from Richards as a grounding for a notion of sameness of meaning---a relationship usually discussed by Quine and others (1960and below) in terms of substitutions of words in sentences. Masterman grounded her overall sameness of meaning, resulting from the substitution of semi-synonyms, not in any philosophical or linguistic notion but in the notion of the "same situation" which she identified with individual stickpictures.

"....not very clever differing-language speakers with minimal sign-apparatus can understand one another-----that is, translate to one another-----if and only if they can both recognise and react to situations common to both of them in real life. What I want to say is that, even when we know one another's languages, we still do the same thing. It is important to side with the language-teachers, and not the psychologists or linguists on this; for either of these last two groups....can talk one into thinking that translation, in the ordinary sense, is impossible. But language-teachers who teach translation know how it is that it can occur; the right hotel room is engaged, the puncture in the left back tyre is mended, the telegram is sent, the friend's (unknown) friend is safely met at the station, all because...they know a very great deal about the relevant situation. And in so far as this knowledge of a common stock of situations breaks down, as between us and the termites, or between us and sulphur-breathing beings from another planet, then it becomes evident that, whatever the languages involved, translation becomes impossible." (Masterman 1961)



The stick-picture technique is very similar to the illustrations used by Barwise and Perry to explain their influential Situation Semantics (1983, a book in which, curiously, no reference to Wittgenstein appears). The importance of pictures, again, is that unlike objects and even photographs, they can be said to show intention directly.

How one characterises the essentially critical attitudes of the Wittgenstein of the Philosophical Investigations, critical (as we shall see below) of rules, of definitions, of logic and of limited, primitive, sub-languages, determines to some extent what NLP developments seem compatible with, if not influenced by, his thinking. One very broad characterisation would be a rejection of all metadescription of words, that is to say, by entities that are not themselves words. Understanding this, in so far as I did, led me once to formulate the (Wilks, 1970)

proposition “Meaning is other words” which was intended as explicitly Wittgensteinian in motivation, in the spirit of his giving “explanations by other words” (Blue Book, 1958 p.27), a notion which is crucially not the same as giving definitions and can be taken to mean words are not explained except by words, and is consistent with denying the veridical role of ostensive acts, though it can lead to a view that the world is a closed one of texts and possibly dictionaries.

Later in the paper, we shall turn to the notion of words as referring directly to things: Wittgenstein did not deny that we could sometimes do that to explain the meaning of words, but he emphasised that it was to do something *special*, and to understand it, the hearer has to know that is what is going on, since it is not a normal, everyday, concomitant of talking to people. We shall also focus on attempts to add language-like items to texts (or annotations, as they would now be called, an idea that actually goes back to medieval scholarship) to explain their meaning, while claiming that the items added are not simply more explanatory words, but rather semantic features, types or markers, or logical descriptions. A classic example would be Katz and Fodor’s (1963) semantic marker HUMAN to attach to words like “Bertrand Russell” to show they referred to a human being. Wittgenstein came to distrust such metadescriptions if it was suggested they were in some special space, or logic or semantics, rather than being in the space of words, which is what they certainly appeared to be in most cases.

What the rejection of metadescription is also close to is Sparck Jones’ description (2003) of the basis of statistical information retrieval (IR): as she puts it, in IR the key idea is that of “taking words as they stand”and not decorated with primitives from another realm of predicates as in logical formulae, or formal semantics or even linguistic “features”. One could put this, with an almost theological flavour, that nothing stands between a word (or wordstring) and the mind in the act of understanding. It is a notion not far from the, once popular now neglected, “linguistic field” theory, a view in which words take on meaning from their neighbours (or co-field members) by oppositions and contrasts, and not

by any kind of CODING of the sort that became standard in the linguistics and AI literature.

In that sense, then, the statistical, or "bag of words" approach, as non-linguistic approaches are sometimes described, satisfies Wittgenstein's distaste for logic-style explanations unless (and we shall pursue this possibility later) one can make a case that, in spite of appearances, formal and semantic/linguistic languages are all miniature, functional, natural languages, whether or not their designers intended them to be, and in that case such meta-descriptions are no more than the translation of one language into another. As we noted above, Wittgenstein, unlike Quine, never really questioned the notion of translation as a meaning-preserving device---his notion of explanation by other words does not require that they be in the same language, after all.

The Web as a corpus of use

Wittgenstein's appeal to look for the use rather than the meaning is not, on its face, a clear injunction: as we noted, he writes of giving meanings by means of explanations and one may reasonably infer that the meanings NOT to look for are pointings at objects, and that when meanings are to be given they are in terms of more words, paraphrases (and not, he makes clear elsewhere, definitions) rather than an artificial coded language for meaning expression, such as that traditionally offered by logic, and later by linguistics and AI.

All this suggests an approach to actual language use more sympathetic than that usually associated with philosophers, and that was indeed the movement he created. Later, Quine, who made many of the same assumptions as Wittgenstein, explicitly linked looking at language use with the methods of structural linguistics, seeking data in languages not understood by the researcher, and drew a range of conclusions (1960)

very close to those of Wittgenstein, in particular that it was not mere language data that would do the trick but data in a language that was understood, by whatever process.

This also shows how wary one must be of trying, as Brown did, as somehow closer to the anthropological-empirical tradition than to Chomsky. It is true that Wittgenstein had something in common with the earlier writers, as Brown noted, but his emphasis on seeing language "from the inside', as something already understood and distinctively human, rather than as an object for scientific observation, brings him closer to Chomsky's emphasis on the native speaker and intuition. The truth is that, while Chomsky was a committed anti-behaviourist, Wittgenstein maintained an ambiguous position, one which declined to give the speaker veridicality on what he meant, so that he could not be wrong, a certainty Wittgenstein considered vacuous (REF).

Among those who traditionally drew the attention of NLP researchers to data in large quantities were lexicographers, of linguistic or computational bent, as the remarks of Sinclair and Hanks above show. Since the return of machine learning and statistical methods to NLP, applied to large corpus data bases since the early 1990s, and following their proven success in speech recognition, NLP has taken large collections of text seriously as its databases; recently Kilgarriff and Grefenstette have based a journal issue on the notion of "web as corpus": the use of the whole web in a given language as a corpus for NLP and, given Grefenstette's estimates (2004), it is now clear the total of pages in English is up to forty times the number indexed by Google (currently in excess of 12 billion).

A corpus of that size is of course a data base of use/usage, one far greater than any human could encounter in a lifetime, even if it is not structured in the way any human would encounter language, e.g. as dialogue, rather than prose, and graded appropriately for age on encountering it. But of course that is just a search problem, too, for there must be, in those 300 billion pages of English, a great deal of dialogue and child language at

all levels. We must give up any idea that such a vast corpus could be a cognitive model of any kind: it would take a reader, reading constantly, at least 60,000 years to train on the current English web corpus, if we make plausible assumptions. One can compare this with Roger Moore's observation that (2007) if a baby learned to speak using the best models of speech acquisition currently available, it would take 100 years to learn to talk.

The question we can now ask is, does that access to the whole web as a corpus by NLP research bring us closer to an ability to compute over uses, to language surveyed in its full variety, rather than the examples an individual might think up, or generate from rules or whatever? The odd answer seems to be that, although a web corpus, even now, only ten years after its inception, is so vast in human-life (of reading) terms, it is still no kind of full survey of language possibilities and never can be, and the reason for that is not any kind of Chomskyan notion of novelty to do with the infinite number of sentences that can be generated from a finite base of rules.

But there is no finite base in any straightforward sense: as far as words (unigrams) are concerned, it is clear they will continue to occur at a steady rate no matter how large the corpus (cf. Dunning 1993). This fact also hold for all forms of combinations of words. These are only examples of what is known as "data sparseness", and maybe no more than a statistical/combinatoric updating of Chomsky's point: as Jelinek has put it "language is as system of rare events". But is it vital to emphasise (since this whole discussion will have to be brought back to the notion of rules in due course) how wrong that finite base assumption was. In work at Sheffield Krotov induced all possible phrase-structure rules explicitly for the Penn Tree Bank (PTB) and graphed them against the corpus length. What was clear and astonishing was that at the end of the process –i.e. training on the whole of the PTB ----the number of rules found (over 18K) was still rising linearly with the length of the corpus! It is quite unclear that there is any empirical justification for the idea of a finite syntactic base, at least for English.

There is no reason to think this tendency will change with much longer corpora; given that fact, assuming it is one, it is one hard to grasp within the history of modern formal linguistics. Chomsky took it simply as an article of faith that there was a finite set of rules underlying a language, if only they could be written or found. (Krotov et al. 2001) suggests this is simply not so.

We are approaching a paradox here: we began with the opposition, clear in Wittgenstein, to the notion of boundedness in language implied by the rule-driven approach to a natural language he found in Carnap, and which continued in Chomsky's work. Wittgenstein wanted to question both that we could be said to be using any such rules and that any set of them could bound the language and determine well-formedness. Goedel's results on undecidability in mathematics (1986) must have seemed to him analogues from that world, and this is explicit in the Remarks on the Foundations of Mathematics (1978).

However, just as it may be the case that the rule set for a language , like its sentence set, is not finite at all, so it may be the case that the corpus itself cannot be bounded, no matter how large it grows; or, rather, there is no corpus that captures the whole language, and so usage/use itself it not something finite that can be appealed to. One could, presumably, restrict oneself to all the sentences of English up to, say, 15 words long and bound that by permutations, but the problem remains that the word set itself is shifting all the time: e.g. more than 900 words a year are being added to non-scientific English (The Times, 9/10/03).

Can Wittgenstein's appeal to use be related to the fact that NLP over the web now surveys enormously more use than it did? It is clear that there can now be real experiments that appeal to use in a very satisfying way: Grefenstette, for example, (2004) has described a novel algorithm for machine translation---following an earlier example due to Dagan (1994)----in which a Spanish bigram XY is translated into, say, English by taking the n senses of a Spanish word X in a Spanish-English bilingual

dictionary, and making a Cartesian product with the m senses of Spanish word Y, and then seeking the n x m resulting English bigrams in an English corpus and ranking them by frequency of occurrence. One may be confident that the most frequent one is always the correct translation.

This algorithm is in fact quite hard to explain and justify a priori: it feels exactly like "Asking the audience" in the popular quiz show "Who wants to be a Millionaire?" where, again, the most frequent answer from the audience is usually, but not always, correct, a phenomenon very close to what some would call the Google-view-of-truth, or what is now referred to as the "Wisdom of Crowds" (Surowiecki, 2004). But whatever is the case about that, there is no doubt this algorithm is precisely an appeal to use rather than meaning and a model for the future deployment of the web-as-corpus to solve linguistic problems.

A constant theme in Wittgenstein is that language is not actually the way it is conventionally considered to be by researchers: we touched above on its not being bound by reference to physical objects by acts of ostension, but we should also consider, in the light of NLP developments, issues of the relationship of language to:

1) rules
2) definitions and essential properties
3) formal and primitive languages
4) reference
5) mini languages
6) language and the world as a whole
7) understanding as a feeling

and we now shall look briefly at each of these in turn.

1) Rules

We argued above that rule sets for languages are probably not finite with expanding corpora, as the Chomsky approach had always implicitly

assumed. But there is another, more fundamental, issue concerning rules determining language forms, one which can be put in terms of decidability. Since the earliest days of transformational grammar (TG) (Chomsky, 1957), the question had arisen as to whether a system of such rules was decidable, in the sense that, for any sentence, a TG could decide whether or not it was well-formed, and there was a consensus that Greibach (1966). (and later Peters, REF) had shown it was not.

At about the same time, I argued (Wilks 1971) that this result could be understood not only in formal, syntactic terms, as an aspect of the TG representation and inferential power, but also semantically in that, it followed from Goedel's approach that, if you showed that a formal system could not assign a property decidably to a sentence set (truth in Goedel's case, well-formedness in Chomsky's), then merely saying that assumed the property in question had some intuitive assignability *in advance of all formalization*. You cannot make sense of Goedel's result unless you already know that the sentences a logic cannot decide are TRUE, and that truth must be known in another way than proof—it must be obvious by inspection (as in Goedel's case) , or something like that. The paper argued, by analogy, that syntactic well-formedness was simply not a property about which people had reliable intuitions, and therefore in was no surprise it could not be decided by a TG. Half a century of psychology and linguistic practice tends to confirm that there are no reliable, general, intuitions at the edge as to what syntactic well-formedness is, contrary to everything Chomsky had based his system on.

The paper went further, and closer to our Wittgensteinian quarry: it argued that meaningfulness was, by contrast, just such a property, as a basis for decidability, in that people could, in general, say whether something was meaningful or not, even if that required making it meaningful. As Wittgenstein put it "The meaning of a word is what is explained by the explanation of the meaning." i.e., if you want to understand the use of the word "meaning", look for what are called "explanations of meaning" (PI $560): and we can see that process at

work in the way writers have given meaning/interpretation many times to Chomsky's famous sentence "Colorless green ideas sleep furiously", which he had deemed well-formed but meaningless. We give meaning by expanding the language, which may or may not mean expanding some underlying rule base. One can think of this as analogous to the way a truth can always be decided in a Goedelian system by adding it to the axioms, but this is pointless in the proof case because there will then be some new undecidable sentence as a result of doing that. This process may be useless in logic but sensible in NLP, in that the whole system is constantly expanding in this way, thus enlarging the space of what is meaningful.

The paper argued, perhaps not completely coherently or convincingly, that one could imagine how, by some such method, a set of meaning-determining rules could be decidable but without the boundary this implied being fixed, since it could be constantly augmented to form a new, larger, system by every act of meaning augmentation. These considerations in 1971 made explicit reference to Wittgenstein, and (Wilks and Catizone 2002) later described detailed methods for sense augmentation in exactly the same spirit, and compared it with sense augmentation proposals by Nirenburg and Raskin (1996), Pustejovsky (1995), Buitelaar (1997) and Briscoe et al. (1991).

The paradigm underlying the 1971 paper was called Preference Semantics (1968), though its philosophical underpinnings and link to Wittgenstein were never restated. It was a notion compatible with, though different from, Minsky's original notion of a default filler for a semantically typed slot (REF): the proposal was that active linguistic entities (usually expressed in English as verbs, but also many prepositions and adjectives) prefer to dominate or associate with certain individual objects -----or, more usually, semantic types of object----- and verbs traditionally dominate more than one such slot. At its simplest, drinking actions prefer human drinkers, but will accept non-human drinkers, organizations or even machines if no human is available to fill the slot, and they similarly prefer liquid objects. The key notion is the

word “prefer”, preferring one type but accepting others if necessary or anything at all if those remain unavailable. The preferred is like a default, but the essence of understanding that this implies is how to link the distribution of available objects to slots to satisfy the maximum number of slots overall, a quantitative notion quite different from default, as it is from the Fodor and Katz (ibid.) notion of a decidable meaningfulness given semantic conditions that are necessary and sufficient for filling slots in linguistic structures.

For the original preferences, expressed in ”semantic formulae”, the preferences were drawn, like everything else at the time, from intuition. Since then, however, a number of researchers have computed the preferences of e.g. verbs from corpora including Lehnert (REF), Grishman and Sterling (REF), and Resnik (REF) among others.

In an age of empirical, data-driven, linguistics it is tempting to believe that that movement expresses something close to Wittgenstein’s injunction to look for the use rather than the meaning. But to do no more than that ignores his belief that there were also “deep grammatical forms”, and it was from there that the metaphor of “depth” in modern linguistics took off:
“ In the use of a word one can distinguish a ‘superficial grammar’ ("Oberflächengrammatik") from a ‘deep grammar’ ("Tiefengrammatik")”. (PI $664).

Wittgenstein always resisted any attempt to formalize a theory of these deep forms, and there is no doubt he saw them as concerned with meaning, and not only what we would now call syntax. Nor should we imagine he would have been happy had he lived to see modern linguistics and AI as alternatives to the logical paradigm. But many of the concerns of modern NLP and AI are already there in his work, and his line of thinking is a powerful antidote to the naive errors with which those subjects are still riddled.

2. Definitions

The closely related notions of preference and default have been illustrated with respect to the preferences of linguistic items, such as verbs, and contrasted with the semantic features of Fodor and Katz (ibid.) which were obligatory, not preferred, and what failed their constraints was deemed ill-formed, in clear violation of all intuition about metaphors, and so on. One can look at exactly the same point in terms of definition, which traditionally assigns an essential property to a class of objects, without which the class would cease to be under the genus it is. If lemons are essentially yellow, then a green lemon is not a lemon at all.

Wittengenstein questioned this whole way of looking at the world and language and his best known example was the notion of "game", where he argued that games shared no single property that made them a game (ibid. $66). Putnam (1970) discussed the same point on the assumption that cats might be found that lacked the "essential" property of animacy, if, say, they were all found to be robots controlled from Mars. Putnam argued, plausibly, that we would not then say there were no cats, only that we had found out a new fact about cats, which he took to imply that one could not take definitions seriously, as one would be doing if one then rejected cats as cats after the discovery. However, it became clear that Putnam did take scientific structures as definitional, in some sense, since their genetic code defined lemons and cats, so one could argue he had just moved the definition to somewhere else where less people knew it, which appeared to satisfy him but not his critics (e.g. Mellor, 1977).

Wittgenstein's constructive suggestion on this issue was one which seems to have strong resonances in statistical methods in linguistics and information retrieval (IR): he produced his famous "family resemblance" metaphor, and argued that one could observe a family's pictures and spot a special kind of family face, but without implying there was any set of features shared by all those with the "family face".

In the 1960s Karen Sparck Jones did her pioneering work on “Synonymy and Semantic Classification” (REF) using a statistical unsupervised clumping theory to recluster word rows from a thesaurus. It was the first statistical work on language with any constructive (as opposed to mere counting, like lexical statistics) basis, and she and her husband Roger Needham, who used the same statistical classification model in his own thesis, were quite aware of the possible connection between the classification model they were using and the “family resemblance” notion: that is, that their methods could certainly produce classifications/clumps in the data whose members did not all share any single feature used in the classification.

As Roger Needham put it at the time (1961): “The problem may be generally stated thus: Given a set of objects, and a list of properties of each, to find procedures for grouping the objects into subsets the members of which have in a defined sense a mutual `family resemblance'. We are thus concerned here with the stage before the usual classification procedures, which take a collection of objects with their properties and place them in various previously defined classes on the basis of a comparison of the objects with the defining properties of the classes. This is a sorting problem and so presents no more than technical difficulties. The problem is that of discovering, given a collection of objects, what would be a worthwhile classification for them and for similar collections that is, the problem of defining classes, not of using them once defined.

3. Formal and primitive languages

Wittgenstein created a primitive builders’ language based on terms like “block” and “slab”, which would call to mind for NLP researchers of a certain age Winograd’s box and tabletop world (1971) embodied in a simple dialogue program at MIT. Wittgenstein’s motivation seems to have been to argue that such languages are not a part of a larger language but smaller different ones, a point consistent with the “linguistic field”

theory mentioned earlier; if meaning comes from opposition and the company words keep (as in Firth's much quoted phrase) then it changes inevitably as the surrounding vocabulary changes. Or as we might now put it in NLP, toy language experiments do not scale up!

The same point seems to have been behind his critique of logical formalisms such as predicate logic: even in the early Tractatus (ibid.) he speculates as to whether one could deal with the ambiguity of words by subscripting them Word1 Word2 etc., so as to turn all word forms in logical formulas monosemic, or single sense. It should be clear by now that this cannot be done, as a way of avoiding ambiguity in formal languages or anywhere else, because there is no agreed set of senses for words, and the set is not stable over time.

It is sometimes argued, that the word forms in formulas and formal languages could all be replaced, one for one, by nonsense words (or computer-generated gensyms like G110004467) and the formal language expressions could still be interpreted by a human or a machine. This position was discussed most famously by McDermott (REF) in a much reprinted paper and he seems to have considered it seriously, at least for a while. I replied (REF) calling it the ´Gensym fallacy" and argued that humans could not in fact manipulate such substituted forms unless they learned that language fully (so as to be "inside" it, as it were), in which case, the risk of the symbols regaining ambiguity would return as in any language. The only real example of large scale logical coding like this is the twenty year Cyc project in Texas (see below), and there it is undoubtedly the case that formal predicates used up to 20 years ago have been used since in quite different ways so that inferences linking the same predicate over time are highly unreliable.The CyC project (Lenat and Guha, 1990) suggests that predicate symbols in formulas do become ambiguous with time and, further, that Wittgenstein's method for dealing with that (in the Tractatus) will not suffice to "cure" the natural-language-likens of the predicate symbols either (see below on symbols in the language LISP).

4. Reference

Wittgenstein argues that pointing or referring is in principle a vague activity. It can only be made clear by explaining from within the language what we are pointing at - i.e. useful pointing already assumes the whole language. Hence it is not that pointing explains how we mean, as the formalists thought when they defined the denotations of their symbols as objects, or sets of objects, because, argues Wittgenstein, the pointing presumes upon the language rather than explains it. Wittgenstein says we could have a language based on the referential notion (PI $2), but it would be a language more primitive than what we call natural language. The relation of this point to the referential assumption made by both formalists like Montague and many AI workers should be obvious. The general point here is very close to the one made later by Quine (ibid.) when arguing for the essential ambiguity of terms like "Gavgai" if used ostensively by people with whom we could not communicate because we shared no language, what we have called the "anthropologist scenario". The problem with this version of Wittgenstein's point is that human experience shows that total strangers can learn languages from such unpromising starts, so his point must apply only to individual acts in such situations.

$ 30. So one might say: the ostensive [i.e. pointing to] definition explains the use - the meaning - of the word when the overall role of the word in language is clear. Thus if I know that someone means to explain a color-word to me the ostensive definition "That is called 'sepia'" will help me to understand the word. (PI $30)

5. Mini languages and language games

Wittgenstein argued that we can construct mini-languages obeying any rules we like, and we can think of them as games. The important question is whether these games are sufficiently like the "whole game"

of natural language. This question does not have a definite answer any more than this question: “Can one play chess without the Queen?” Wittgenstein attributes the ostensive or “pointing” view of meaning (in (i) above) to St. Augustine and then, as we noted, proceeds to construct a mini-language of commands and objects like “block”, “slab”, and colours like “red”, etc.

$ 3. Augustine, we might say, does describe a system of communication; only not everything that we call language is this system. And one has to say this in many cases where the question arises “Is this an appropriate description or not?” The answer is: “Yes, it is appropriate, but only for this narrowly circumscribed region, not for the whole of what you were claiming to describe”. It is as if someone were to say: “A game consists in moving objects about on a surface according to certain rules ...” - and we replied: You seem to be thinking of board games, but there are others. You can make your definition correct by expressly restricting it to those games.

This mini-language that Wittgenstein constructs with “block”, “slab” and commands may remind readers strongly of Winograd's mini-language (ibid.) inherited from MIT table top robotics: it had a box, a number of blocks, a sphere and so on. The parallel is a fair one in many ways, and Wittgenstein can be seen as presenting the dangers of taking a mini-language with certain properties (definite reference to objects, for example) and assuming that they are properties of the whole natural language. One could say that it is not clear how, or whether, a Winogradian system could function in a world without definite locatable and numbered objects, such as the world of newspaper articles or of this paper.

One could argue similarly that the languages of semantic primitives in Schank's -----or my own ----early systems (REFS) are also mini-languages, or language games, in a broad sense, and that there would be similar problems if they started to postulate “conceptual objects” to which the primitives refer, and Schank did in fact sometimes suggest that

his primitives referred to entities in the mind or brain. If a language of primitives is given that property then it loses one essential feature of a full natural language, and begins to look more like a “blocks world” mini-language.

Of course, it should not be thought that Wittgenstein is a defender of linguistic primitives, or primitives of any sort. Indeed, one of the attractions to him of his “truth-table” method of presenting the Propositional Calculus was that it avoided the more conventional form in terms of primitive formulas, like P IMPLIES (Q OR NOT-Q) from which all other true formulas could be derived. Yet, one could argue that a large part of what Wittgenstein found objectionable about the notion of “primitive” in logic was the idea that there is a right set of them, if only we could discover it, and which then provides an infallible starting point. But in the case of semantic primitives, it is still possible to use them without claiming that there is a single right set of them, as Schank did. The 2000 words which Longmans used as the defining vocabulary of their LDOCE (REF) dictionary can perfectly well be seen as a rival, larger, set of defining conceptual primitives.

5. The linguistic whole and confronting the world

For Wittgenstein a language is a whole and does not confront the world sentence by sentence for the testing of the truth or falsity of each individual part.

“To understand a sentence means to understand a language.” (PI $199)

This thesis is clearly incompatible both with Wittgenstein's own early “picture theory of truth”----which associated individual sentences with facts in the world in a picturing relation---- and with any theory like Montague's. where the assumption is precisely that each sentence of a language can have its truth (and its meaning, expressed as truth conditions) tested individually and in isolation. One could argue that

Wittgenstein's view is not at all inconsistent with a standard view of scientific truth. where sentences such as “This particle has spin 1/2” or “Rats are carriers of plague” cannot be tested directly, but belong only within large systems of inference that must be tested indirectly if at all. That is to say that sentences like those two can only be understood within a wider theory which, in its turn, explains complex notions like “spin”, “particle” and “carried by”. This point is also close to the heart of Quine’s philosophy, and for him followed from his analysis of the analytic-synthetic distinction (ibid.), such that if sentences could not simply be assigned to one of these classes then how much a sentence had of one property or the other would depend on its position and role in a scientific theory.

The idea that we can only understand on the basis of a whole language is clearly more attractive than the alternative of understanding within “block” and “slab” micro-worlds. But it, too. has its dangers: if taken far enough. it can lead to the view that there can be no significant generalizations about language at all, because each use of each sentence has a special relation to the language as a whole. And Wittgenstein has sometimes been accused of holding this view.

What seems more likely to have been his position was an intermediate one; namely, that there are islands of discourse, each with its own criteria of inference, truth and so on, and it is with respect to these (rather than to a micro-world or to the whole language) that utterances are to be understood. The notion of an “island of discourse” is not a self-explanatory one but, roughly speaking, it means an area defined by subject matter (say. history, or quantum physics). but still wide enough to have all the features of a full natural language , in a way that a “slab” and “block” micro-language does not, nor does a narrow domain application of the kind that constitutes much of applied NLP e.g. airline reservation worlds.

6. Understanding is not a feeling

We have the idea that “understanding” something involves, or is associated with, a special feeling of being right. But the tests of our being right are quite different from the feeling, says Wittgenstein.

$139. When someone says the word “cube” to me, for example, I know what it means. But can the whole use of the word come before my mind, when I understand it in this way?

Well, but on the other hand isn't the meaning of the word also determined by this use? And can these ways of determining meaning conflict? Can what we grasp in a flash accord with a use, fit or fail to fit it? And how can what is present to us in an instant, what comes before our mind in an instant, fit a use? What really comes before our mind when we understand a word? - Isn't it something like a picture? Can't it be a picture? Well, suppose that a picture does come before your mind when you hear the word “cube”, say the drawing of a cube. In what sense can this picture fit or fail to fit a use of the word “cube”? - Perhaps you say: “It's quite simple; if that picture occurs to me and I point to a triangular prism for instance, and say it is a cube, then this use of the word doesn't fit the picture.” -But doesn't it fit? I have purposely so chosen the example that it is quite easy to imagine a method of projection according to which the picture does fit after all.

Wittgenstein is making the point that it is dangerous to assess understanding other than in terms of actual and possible performances, and, if we take that to mean “performances with language” we will see that it argues against one sort of criticism that AI researchers have sometimes made of each other's systems: that they only “appeared to understand” but “didn't really do so”. This is often said of PARRY (REF) and Loebner type (REF) systems. Those who employ that sort of criticism are, in Wittgenstein's terms, acting as if 'understanding is a special feeling’.

There is also a very general theme running through Wittgenstein's work about the relation of knowledge and understanding to performance and to

what he calls the ability to “go on”, “to continue” ($151). This theme could be held to support those NLP and AI researchers who go beyond the assertion that our understanding of words depends on our ability to use them and to perform with them, to the much stronger and less plausible claim that our understanding of language about physical processes (say, tying our shoelaces, or stacking blocks) is closely connected with (and may even require) our ability to carry out the corresponding task. That would mean that a computer could not understand language about, say, dining in a restaurant, unless it could itself dine in a restaurant. This is a complex issue, usually discussed under terms like ´ grounding [a] or ´situatedness[a] and one where Wittgenstein's explorations are essential background.

There is often confusion between (1) what the processes actually are in our heads in carrying out a task, (2) how we feel about, or what we believe about, what the processes are, and (3) what a computer should do in program terms to carry out the same task. These are three quite separate things and arguments connecting them can be dangerous as Wittgenstein warns us in $139. Consider the following argument of Dreyfus (REF), and its similarity to Schank's position that a proper analysis system never follows a wrong path because we ourselves unhesitatingly go for the correct interpretation of an utterance:

“Of course [this human process] only looks like “narrowing down” or “disambiguation” to someone who approaches the problem from the computer's point of view. We shall see later that for a human being the situation is structured in terms of interrelated meanings so that the other possible meanings of a word or utterance never even have to be eliminated. They simply do not arise.”

Are these positions not very similar to the one Wittgenstein describes and implicitly criticises in terms of “the whole use of a word coming before the mind”? ($139) Wittgenstein is suggesting there that what does, or does not, “come before the mind” is not essentially connected with our abilities to perform with, that is to use, a word correctly. One

might argue that, similarly, how we think we function (what, that is, comes before our minds about ourselves) is no sure guide to how we do function, or to how a computer program simulating us should function.

Dreyfus argued that AI is impossible because, to be intelligent, a computer would have to be exactly like us: bodies, feelings, and growing up and all that. He has often quoted Wittgenstein in support of his own position, but it can equally well be argued that Wittgenstein's clear distinction between understanding-as-performance (which AI workers believe a machine can have) and understanding-as-fee1ing (which no doubt only we have) supports exactly the opposite position.

Moving to another contemporary area, the contrast between Wittgenstein and influential formal theorists of language, such as Montague (REF), and their importance for present day discussion of AI, NLP and natural language, comes down to two issues: is there a hidden structure to natural language and is natural language itself, and its use, to be the final court of appeal. Montague gives a yes to the first and proffers a simple logic as the hidden structure; as to the second, his choice of examples, far from the concerns of ordinary speakers, such as the two interpretations of "Every man loves some woman" suggests that his answer is no. Wittgenstein's answer to the first is complex: his use of "deep grammatical structures" suggests that he thought there was structure but not one to be revealed by simple techniques, like logic, whose interests were really always somewhere else. It is, after all, the different structures we could find for "Every man loves some woman" as they can participate in proofs in the Predicate Calculus that interest the logician. The ordinary speaker rarely, if ever, sees that there is a "second interpretation" of the sentence, though such issues can distract children. On hearing Woods describe the natural interpretation of ´ There is someone in every phone box in the US ª, the child of a logician friend leaned over to me and whispered ´ and he's moving really, really, fast between them ª . Considerations like this last suggest that Wittgenstein's answer to the second question would have been a firm yes.

Back to the state of CL/NLP

Let us turn back now from exegesis of Wittgenstein to the state of computational linguistics and language processing by computer. One could generalize very rapidly as follows: in the 1970's, there arose movements such as what was then called conceptual dependency (Schank, ibid.) or preference semantics (Wilks, ibid.) which could be described as attempting to map a "deep grammar' of concepts and what I would call the preferential relations between concepts. This theme was closely allied with various forms of Fillmore's (ibid.) case grammar in linguistics, and his later work (1976) could certainly be described as a continuing search for local, but deep, grammatical relations----based on systematic substitution relations in semi-fixed phrases in English---- outside the concerns of the main thrust of work in computational syntax, which is little concerned with words themselves or local effects in language.

Fillmore's hand-coded lexicography just mentioned has been a survivor, but virtually all other attempts at conceptual mapping have been overtaken by one of the two separate movements to introduce empiricism into CL and NLP: the connectionist movement of the early 1980s, and the statistical corpus movement, driven by Jelinek's successes in speech and translation in the late 1980s. The first was not a success but the second is still continuing: a classic of the first movement would be Waltz and Pollack's (1985) neural networks showing how concepts attracted and repelled each other in terms of contexts supplied to the network, from corpora or from dialogue. The work was exciting but such networks were never able to process more than tiny fragments of language. There were more radical (or "localist") connectionists. such as (Sejnowski and Rosenberg, 1986) who went further and declined to start from explicit language symbols at all, in an attempt to show how symbols could have been reached from simpler associationist algorithms that built, rather than assumed, the symbols we use. If this had been done it might have broken through the impasse that the title of this paper suggests, namely

how can one have a theory of language which does not build in from the very start all that one seeks to explain, as intuition-based theories in linguistics, logic and AI always seem to. Connectionist theories could never give a clear account of the theory-free "simples" from which to begin, and any case they also failed to "scale up" to any reasonable sample of language use, or to confirm any strong claims about human cognition of language.

The second movement, that followed connectionism, the one we are still within, at the time of writing, was statistical associationism, driven by Jelinek with his translation work derived from trigram models of speech, and which had some success and undoubtedly used language on a very large scale indeed, too large as we noted earlier, to be cognitively plausible for human beings. This movement has been committed to an "empiricism of use" but can such approaches ever build back to reconstruct concepts empirically? This movement, as we noted earlier, shares many assumptions with the Information Retrieval (IR)(see e.g. Sparck Jones ibid.) view that language consists only of words without meta-codings, and all decorations and annotations that intuitive theories add are unexplained and unacceptable as explanatory theory. IR, it must always be remembered, underlies the successful search theories that have given us the world wide web.

We noted earlier that Jelinek became disillusioned with his first set of statistical functions and came to the view that language data is too sparse to allow the derivation of full trigram models of language, which is to say, derived from corpora so large that one could expect to have seen when training every trigram one could find in any text being tested subsequently.

In a moment I will describe some recent experimental work that suggests that Jelinek may have been too pessimistic, and that a full trigram model might now be within reach, using a device called a "skipgram". But first, what would be the point in a fuller associationist model, one that covered a language, English, say: how could that get us closer to rebuilding

concepts from all this data?

Let me give two simple examples of this, one from Jelinek's own laboratory (Brown et al., 1990) where they showed that simple association criteria could determine semantically coherent classes of objects far more easily than had been thought, provided one had enough data. One can see this most easily now on Google, where what was a research discovery ten years ago is now a toy. On labs.google.com/sets one can input any small set of objects one likes and ask Google to find more, in response to this request, from the 8 billion pages it indexes. So, if one types in Scots, Bavarian, American, German, Google replies with something like French, Chinese, Japanese etc. In other words it has "grasped the concept" of nationality from context and is, as Wittgenstein would put it, able to go on. This is most certainly a derivation of something clearly semantic "from nothing" but word data, the problem being the system does not know what the name of the class is!

A second notion is that of ontologies, forms of knowledge representation that have now become the standard way of looking at formalised knowledge in a wide range of AI and web applications: they contain technical and everyday information about set inclusion and membership as well as functional, causal etc. information about sets and objects, and they or may not (see Brewster et al., 2004) have a strong underlying logical structure. The problem about such structures has always been, as with other forms of knowledge discussed here, that they are traditionally written down by human intuition. So what are we to make of the meanings of the terms they contain: are they or causal in meaning and can we gather anything from looking at their place in an ordered ontological hierarchy?

This is a straightforwardly Wittgensteinian question and the only proper answer is his own: namely that we cannot tell any term's meaning by looking at it, only by seeing it deployed in use. It is a corollary of that view, often advanced in this paper, that all such terms are terms in language, the language they appear to be in (usually) English, and that is

so no matter how much their designers protest to the contrary. This is an issue discussed in detail in (Nirenburg and Wilks, 2000).

Ontologies, then, pose something of the problem here that logic does, or formal features in linguistics (such as Fodor & Katz' semantic markers, q.v.): they are claimed to be formal objects, kept apart from language and its vagaries, and with only the meanings assigned to them by scientists. But this isolation cannot in fact be maintained, and a more reasonable position is that ontologies will have justifiable meanings when they can be linked directly to language corpora, chiefly by being built automatically from them. An example of such a current project is ABRAXAS (Brewster et al. 2004), one of a number of projects that claims to do exactly that.

Let us now turn to the issue of enlarging language models because this will lead us to another language derived object that may help bridge the gap between languages for the expression of knowledge and models derived from usage. I want to argue that it may now be possible, using much more of the whole web, to produce far larger models of a language and to come closer to the full language model that will be needed for tasks like complete annotation and automatically generated ontologies. These results are only suggestive and not complete (see Guthrie et al., 2006), yet but they do seem to make the data for a language much less sparse and without loss by means of skip-grams. What follows is a very brief description of the kind of results coming from the EPCRC REVEAL project at Sheffield, which takes a 1.5 billion word corpus from the web and ask how much of a test corpus is covered by the trigrams of that large training corpus, both as regular trigrams and as 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. The 1.5 billion word training corpus gives a 67%+ coverage by trigrams of 1000 word test texts in English.

Suppose, as a way of extending the training corpus, we consider skipgrams, and take:

Chelsea celebrate Premiership success.

the normal tri-grams that contains (of contiguous three-word sequences) will be:

Chelsea celebrate Premiership

celebrate Premiership success

But one-skip tri-grams (allowing a one word gap in the trigram) will be:

Chelsea celebrate success

Chelsea Premiership success

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

We obtained 74% coverage with 4skiptrigrams over test texts. This suggests, by extrapolation, that it would need 7.5x10*11 words to give 100% trigram coverage. Our corpus giving 74% 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 indexes), so the corpus needed for complete coverage would be about seven times the full English world wide web in 2003, which is presumably somewhat closer to today's English web, and now certainly within the realm of realistic search

These corpora are so vast they cannot possibly offer a model of how humans process meaning, so any 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 more likely (perhaps with the aid of a largescale fast surface parser) to correspond to text items representing related Agent-Action-Object triples in very large numbers. This is a old dream going back at least to (Wilks, 1965) where they were presented as potentially Wittgensteinian "forms of fact", later revived by Greffenstette as the concept a "massive lexicon" one now beginning to be available as inventories of derived surface facts at ISI (Pantel and Hovy, 2005) and elsewhere. If one is asked what function they have in language interpretation, one could just say, as in (Wilks 1978) that, if asked what "my car drinks gasoline" means, one could just consult one's huge inventory of facts and ask what cars normally do with gasoline (i.e. use it cause an engine to cause the car to travel)—and from that infer that as the the meaning of "drink" in the last example. This gives the flavour of how Grefenstette envisaged his "vast lexicon" could be deployed to interpret language without assumptions or a priori structures.

However, in the meantime, a new use for structures of this general type has appeared, namely the subject-relation-object triples that are to carry basic knowledge at the bottom level of the Semantic Web (REF), the proposed structure intended to encapsulate human knowledge, based on the world wide web we now have, but annotated in a form to display something of a text's meaning so that computers can use the web themselves. This is too large a vision to discuss here, but one last historical association may be worth making.

Long ago, Bar Hillel (1964) attacked the very possibility of machine translation (MT) on the ground that the kinds of interpretation that translators made required knowledge of vast numbers of facts about the world, and machine translation would therefore need them too. So, you cannot interpret (and so translate) "carbon and sodium chloride" unless you know whether or not there is such a thing as carbon chloride, and so know the inner structure of that phrase (i.e. as carbon+sodium chloride OR carbon chloride + sodium chloride---it is of course the first in this

universe). Bar Hillel went on to argue that machines could not have such extensive knowledge of the facts of the world, and so MT was demonstrably impossible.

It was from exactly that point that AI set out on its long journey to develop mechanisms for representing all the facts in the world (of which the CyC project, (ibid.), is the longest running example.) All this was done in a practical spirit, of course, with no thought or memory of Wittgenstein's declaration that the world was the totality of facts (ibid.), and what if anything, that could possibly mean. It was all practical, energetic computation rather than philosophical thinking, but still, in some sense, fell under Longuet-Higgins' declaration (after Clausewitz) that AI was the pursuit of metaphysics by other means. It is surely interesting that empirically based NLP/CL has now brought back concepts like the derivation of a totality of facts, not painfully hand-constructed as in CyC, but extracted perhaps by relative simple means from the vast resources of the world's corpora.

Margaret Masterman and the search for a Wittgensteinian theory of language processing.

It would not be proper, in a paper with this title, to ignore the contribution of Margaret Masterman, since one could say the goal of her life's research was just such a notion of computational linguistics. She had been a student of Wittgenstein at the time of the Blue Book (1958), and later founded the Cambridge Language Research Unit, which for many years in the 1960s and 1970s did fundamental work on language processing.

There is no doubt that Masterman wanted her theories of language (see Wilks, 2006) to lead to some such goal, one that sought the special nature of the coherence that holds language use together, a coherence not captured as yet by conventional logic or linguistics. Such a goal would also be one that drew natural language and metaphysics together in a way undreamed of by linguistic philosophers, and one in which the

solution to problems of language would have profound consequences for the understanding of the world and mind itself. And in that last, of course, she differed profoundly from Wittgenstein himself, who believed that that consequence could only be the insight that there were no solutions to such problems, even in principle.

It is also a goal that some would consider self-contradictory, in that any formalism that was proposed to cover the infinite extensibility of natural language would, almost by definition, be inadequate by Wittgenstein's own criteria, and in just the way she considered Chomsky's theories inadequate and his notion of generativity and creativity a trivial parody.

The solution for her lay in a theory that in some way allowed for extensibility of word sense, and also justified ab initio the creation of primitives. This is a paradox, of course, and no one can see how to break out of it at the moment: if initially there were humans with no language at all, not even a primitive or reduced language, then how can their language when it emerges be represented (in the mind or anywhere else) other than by itself. It was this that drove Fodor (1975) to the highly implausible, but logically impeccable, claim that there is a language of thought predating real languages, and containing not primitives but concepts as fully formed as “telephone”, on the ground that concepts cannot be built from or expressed by combinations of primitive concepts, and so must always be as wholes in any language of thought. This is, of course, the joke of a very clever man, but it is unclear what the alternatives can be, nor, more specifically, what an evolutionary computational theory of language can be.

It is this very issue that the later wave of theories labelled “connectionist” (e.g. Sejnowski and Rosenberg, 1986) sought to tackle: how underlying classifiers can emerge spontaneously from data by using no more than association and classification algorithms. Masterman would have sympathised with its anti-logicism, but would have found its statistical basis only thin mathematics, and would have not been sympathetic to its anti-symbolic disposition.

It is easier to set down what insights Masterman would have wanted to see captured within a Wittgensteinian linguistics than to show what such a theory is in terms of structures and principles. It would include that same ambiguous attitude that Wittgenstein himself had towards language and its relation to logic: that logic is magnificent, but no guide to language. If anything, the reverse is the case, and logic and reasoning itself can only be understood as a scholarly product of language-users: language itself is always primary.

Her language-centredness led her to retain a firm belief in a linguistic level of meaning and representation: she shared with all linguists the belief that language understanding could not be reduced, as some artificial intelligence researchers assume, to the representation of knowledge in general, and independent of representational formalisms (a contradiction in terms, of course), and with no special status being accorded to language itself. Indeed, she would have turned the tables on them, as on the logicians, and said that their knowledge representation schemes were based in turn on natural languages, whether they knew it or not.

On the notion of a unified Cognitive Science, I think her attitude would have been quite different from those who tend to seek the basis of it all in psychology or, ultimately, in brain research. Chomskyans have tended to put their money on the latter, perhaps because the final results (and hence the possible refutations of merely linguistic theories) look so far off. Masterman had little time for psychology, considering it largely a restatement of the obvious, and would I think have argued for a metaphysically-rather than psychologically-orientated Cognitive Science. Language and Metaphysics were, for her, closely intertwined and only they, together, tell us about the nature of mind, reasoning and, ultimately, the world

Conclusion

Earlier in the paper, we touched on the concept of the Semantic Web (SW), and the process its construction requires of what one call giving meaning progressively to the "upper level" concepts in ontologies, as used in language processing and scientific knowledge structures. These upper level concepts are still written down by intuition, which may have validity in scientific area if done by experts---who else can write a map of biology?----but is as much at risk as all the knowledge structures in AI if not grounded in something firmer. I want to argue, in conclusion, that the future SW may offer the best place to see the core of a Wittgensteinian computational linguistics coming into being, as a way of grounding high-level concepts, such as the primitives at the tops of ontologies, in real usage of the sort we see in the web-as-corpus.

What I think we are seeing in the SW is a growing together of these upper conceptual levels based on the name spaces and concept triples derived from texts by Information Extraction, a successful shallow technology for extracting items and facts that now rests wholly on the success of automated annotation, and which has now been successfully extended to the automatic induction of ontologies. My belief is that the top and bottom levels will grow together and that interpretation or meaning will "trickle up" from the lower levels to the higher: this is the only way I can imagine the higher conceptual labels being justified on an empirical base. It is a process reminiscent of the concept of "semantic ascent" pioneered by Braithwaite in the 1950s as a description of the way in which interpretation "trickled up" scientific theories from observables like cloud-chamber tracks to unobservables like neutrinos. I cannot see any other route from the distributional analysis on which the revolution in language processing rests and the interpretation of serious concepts. It is also a process reminiscent of Kant's great dictum synthesising Rationalism and Empiricism: "Concepts without percepts are empty; percepts without concepts are blind."

I would like argue that the SW is a development of great importance to AI as a whole, even though we still dispute about what it means, and how it can come into being. Many seem to believe that it means Good

Old Fashioned AI is back in a new form, a rebranding of the old tasks of logic, inference, agents and knowledge representation. It is true that core AI tasks have come to something of an impasse: we do not see them marketed in products much after fifty years. But a key feature of SW, I assume, is that its delivery must be gradual, coming into being at points on the World Wide Web (WWW), possibly starting with the modelling of biology and medicine. I cannot imagine how it could start somewhere completely new, and without being piggy-backed in on the WWW, yet it will be much more than those same texts "annotated with their meanings", as some would put it.

The key possibility I think the SW offers to traditional AI is to deliver some of its value in a depleted form initially, by trading representational expressiveness for tractability, as some have put it. The model here could be search technology and machine translation on the WWW (or even speech technology): each is available now in forms that are not perfect but we cannot imagine living without them. This may all seem obvious, but machine translation has only recently crossed the border from impossible (or failed) to commonplace. It is far better for a field to be thought useful, if a little dim at times, than impossible or failed. It will be important that web services using the Semantic Web are chosen so as not to be crucial, merely a nuisance if they fail. My own current interests are in lifelong personal agents, or Companions, conversationalists as well as. Agents, where it should not matter if they are sometimes wrong or misleading, any more than it does for people, as long as we have alternative ways of checking information.

This view of the future of the SW is personal and partial; many do not see the need to justify the meanings of logical predicates or ontological terms now than they did when they set out in AI and representation in the Sixties. But the history of the CyC project is a good demonstration, if one were needed, of why that cannot be a foundation for AI in the long term. There is a related view, also current in the SW, that meanings will be saved or preserved by trusted data bases of objects (URIs), referential items in the world, rather in the way digit strings "ground" personal

phone numbers in a data base. But this way out will not protect knowledge structures from the changes and vagueness of real words in use by human beings. Putnam considered this problem in the Sixties and declared that scientists should therefore be the ultimate "guardians of meaning". As long as they knew what "heavy water" really meant, it did not matter whether the public knew and perhaps better if they did not. But people call heavy water "water" because it is, because it is indistinguishable from water, otherwise it would have been called "deuterium dioxide". We, the people, are the guardians of meaning and "getting meaning into the machine", probably via the SW, should entail doing it our way, and what could be more in the spirit of Wittgenstein than that?

References

Bar Hillel, Y., 1964. Language and Information. Reading, MA: Addison Wesley.

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The Semantic Web, Scientific American, May 2001, p. 29-37.

Braithwaite, R., 1956. Scientific Explanation, Cambridge UP, Cambridge.

Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y., 2004. Data-driven Ontology Evaluation. In Proc. of 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal.

Briscoe, T., Copestake, A., De Pavia, V., 1991, Default inheritance in unification-based approaches to the lexicon, Technical report, Cambridge University Computer Laboratory

Brown, C.H., 1974. Wittgensteinian Linguistics. The Hague: Mouton & Co.

Brown, P.F., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R.L., Roossin, P., 1990. A Statistical Approach to Machine Translation. Computational Linguistics16:2: 79-85.

Buitelaar, P., 1997, A lexicon for underspecified semantic tagging, In Proceedings of the ACL-Siglex Workshop on Tagging Text with Lexical Semantics, Washington, D.C.

Carnap, R., 1936. Logische Syntax der Sprache. English translation 1937, The Logical Syntax of Language. Kegan Paul, London.

Chomsky, N., 1985. Aspects of the Theory of Syntax. Cambridge: The MIT Press.

Chomsky, N., 1957. Syntactic Structures, Mouton: The Hague.

Church, K., Gale, W., Hanks, P., Hindle, D., 1989. Parsing, word associations and typical predicate-argument relations, Proceedings of the workshop on Speech and Natural Language, October 15-18, Cape Cod, Massachusetts.

Dagan, I., Alon I., Word sense disambiguation using a second language monolingual corpus, Computational Linguistics, 1994, Vol. 20(4), pp. 563-596.

Dunning, T., 1993. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics. Computational Linguistics. 19.

Fillmore, The Case for Case. 1968. In Bach and Harms (Ed.): Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston, 1-88.

Fillmore, C., 1976. Frame semantics and the nature of language, In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech. Volume 280: 20-32.

Goedel, K., 1986. Ueber formal unentscheidbare S Saetze der Principia Mathematica und verwandter Systeme. In Solomon Feferman, (Rd.) Kurt Goedel: Collected Works, volume 1, pages 144–195. Oxford University Press, German text, parallel English translation.

Grefenstette, G., 2002. Lecture, Sheffield University.

Grefenstette, G., 2004. The scale of the multilingual web. Lecture at Search Engine Meeting, The Hague, Netherlands, April 2004.

Greibach, S A., The Unsolvability of the Recognition of Linear Context-Free Languages,
October 1966, Journal of the ACM (JACM), Vol 13 (4)

Guthrie, D., Allison, B., Liu, W., Guthrie, L., Wilks, Y., 2006. A Closer Look at Skip-gram Modelling. In Proc. Fifth International Conference on Language, Resources and Evaluation (LREC'06), pp. 1222-1225.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., Wilks, Y., 2006. A closer look at skip-gram modelling. In Proc. LREC'06, Genoa, Italy.

Hirst, G., "Context as a spurious concept." Proceedings, Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February 2000, p273—287 InformationExtraction: http://en.wikipedia.org/wiki/Information_extraction

Katz, J.J., Fodor, J., 1963. The structure of a semantic theory, Language.

Kilgarriff, A., Grefenstette, G., 2003. Introduction to the Special Issue on Web as Corpus. International Journal of Corpus Linguistics 6 (1).

Krotov, A., Gaizauskas, R. and Wilks, Y., 2001. Acquiring a stochastic context-free grammar from the Penn Treebank. In Proceedings of Third Conference on the Cognitive Science of Natural Language Processing.

Lenat, D., Guha, R.V., 1990. Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley.

Malinowski, B., 1923. The problem of meaning in primitive languages. In: C.K. Ogden & I.A. Richards (Eds.), The meaning of meaning, pp. 296-346. London: Routledge & Kegan Paul.

Masterman, M., 2006. Language, Cohesion and Form: selected papers, (Ed.) Y. Wilks, Cambridge UP, Cambridge.

McCarthy, J., Hayes, P.J., 1969. Some philosophical problems from the point of view of Artificial intelligence. In Machine Intelligence 4, (Eds.) Michie and Meltzer, Edinburgh, Edinburgh UP.

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

Moore, R.K., 2007. Spoken language processing: Piecing together the puzzle, Speech Communication, 49.

Moon, R., 2007. Sinclair, lexicography, and the Cobuild Project: The application of theory. International Journal of Corpus Linguistics, Volume 12, Number 2.

Needham, R.M., 2003. Unpublished MS, personal communication from K. Sparck Jones.

Nirenburg S., Raskin, V., Ten choices for lexical semantics, Technical report, Computing Research Lab, Las Cruces, NM, 1996, MCCS-96-304.

Nirenburg, S., Wilks, Y., 2000. What's in symbol. In Journal of Theoretical and Experimental AI (JETAI).

Pustejovsky, J., Anick, P., Automatically acquiring conceptual patterns without an annotated corpus, In Proceedings of the Third Workshop on Very Large Corpora, 1988.

Pustejovsky, J., 1995, The Generative Lexicon. MIT.

Putnam, H., 1970. Is Semantics Possible? Metaphilosophy 1: p187-201.

Quine, W.V.O., 1960. Word and Object, Cambridge, Cambridge UP.

Riloff, E., Shoen, J., Automatically aquiring conceptual patterns without an annotated corpus, In Proceedings of the Third Workshop on Very Large Corpora, 1995.

Ritchie R. W., On the Generative Power of Transformational Grammars, *Information Sciences* 6 (1973): 49-83. Cited In "Citation Classics" In Current Contents, Social & Behavioral Sciences 19 (1987): 15, and in Current Contents, Arts & Humanities 9 (1987): 15.

Sampson, G., 2004. The 'Language Instinct' Debate, Continuum.

Sejnowski, T., Rosenberg, C.,1987. Parallel networks that learn to pronounce English text. Complex Systems, 1: 145-168.

Singhal, A., 2001. Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4).p 35-43.

Sparck Jones, K., 2003. Shallow data: deep theories. Procd. ECIR@003, Springer Verlag, Berlin.

Spärck Jones, K., 1986, Synonymy and semantic classification, Doctoral

dissertation, University of Cambridge, 1964, Edinburgh: Edinburgh University Press.
Surowiecki, J., 2004, The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations Little, Brown

Veronis, J., 1993. Sense tagging, does it make sense? http://citeseer.ist.psu.edu/685898.html

Waltz, D.L., Pollack, J.B., 1985. Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation.
Cognitive Science 9(1): 51-74/

Wittgenstein, L., 1958. Philosophische Untersuchungen, Philosophical Investigations, 2nd ed. Oxford: Basil Blackwell.

Wittgenstein, L., 1978. Remarks on the Foundations of Mathematics, rev. edn, ed. G. H. von Wright, R. Rhees, and G. M. Anscombe, trans. G. E. M. Anscombe, Cambridge, MA: MIT Press.

Wittgenstein, L., 1958. The Blue and Brown Books, Oxford: Basil Blackwell.

Wittgenstein, L., 1961. Tractatus Logico-philosophicus, Routledge: London.

Wilks, Y., 1971. Decidability and Natural Language, Mind LXXX.

Wilks, Y., 1990, Form and content in semantics, Synthese, Vol.92.

Wilks, Y., 1968. Preference Semantics, In The formal semantics of natural language (Ed.) E. Keenan, Cambridge, Cambridge UP (1975)

Wilks, Y., 1964. Text Searching with Templates. Cambridge Language Research Unit Memo, ML.156.

Wilks, Y., 2005. The History of Natural Language Processing and Machine Translation. In Encyclopedia of Language and Linguistics, Kluwer: Amsterdam.

Winograd, T., 1971, Understanding Natural Language, MIT Press: Cambridge, MA.