For a new edition of the Encyclopedia of Language and Linguistics.

# The History of Natural Language Processing and Machine Translation.

**Yorick Wilks**
**University of Sheffield.**

## ABSTRACT

The article surveys fifty years of work in computational language processing and machine translation, and suggests that a great number of the important ideas were present in the earliest days and hampered only back lack of computational power. Sections review the influence of linguistics proper on the computational area, as well as the influence of artificial intelligence and concerns from logic and knowledge representation. Later, corpora and machine readable dictionaries were made available, which in turn made possible the recent statistically-based empirical emphasis in the subject, a trend that began in machine translation under the influence of success in automatic speech processing. Finally, it is suggested that, despite these many influences on the field from outside, there is nonetheless a distinctive process-based computational linguistics and examples are suggested.

**Keywords:**

machine translation
information retrieval
information extraction
parsing
thesaurus
beliefs
syntactic structures
semantic representations
logic
statistics
question answering
summarization
psychology
word-sense
part-of-speech-tagging
sense and reference
performance
case grammar
agents
computational semantics

## 1 Introduction

A remarkable feature of the fifty-year history of natural language processing (NLP) by computer, alias **computational linguistics** (CL), is how much of what we now take for granted in terms of topics of interest was there at the very beginning; all the pioneers lacked were computers. In the Fifties and Sixties, King was arguing for statistical **machine translation**, Masterman for the power of a semantic **thesaurus**, Ceccato for conceptual codings (Ceccato, 1961), and **Yngve**, still working at the time of writing, had designed COMIT, a special programming language for NLP, and had refined his famous claim about the effect of a limitation on processing resources on permissible **syntactic structures** in a language (Yngve, 1960). The latter project brought him into direct conflict with **Chomsky** over the permissible ways of drawing syntactic tree structures, which can now be seen to have constituted a defining moment of schism in the history of NLP in its relationship to mainstream linguistics. It was the foundational schism, not healed until decades later when Gazdar became the first major linguist to embrace a computational strategy explicitly:

**Machine Translation** (MT) is the subject of a separate article and it will be described only indirectly here, but it must always remembered that it was the original task of NLP, and remains a principal one, still the application within which CL theories embodied in programs can be tested. There is now a wide range of other NLP tasks that researchers investigate and for which companies sell software solutions: **question answering**, **information extraction**, document **summarisation** etc. Thus, NLP does require a task: it is not in itself a program of scientific investigation, which is what CL normally claims to be, and that remains a significant difference between two very close terms.

It is also important to distinguish major tasks, like those just mentioned, from a wide range of tasks that are defined only in terms of linguistic theories, and whose outcomes can only be judged by experts, as opposed to naïve users of the results of the major tasks above. These non-major tasks include **word-sense disambiguation** (e.g. Yarowsky, 1995), **part-of-speech tagging**, **syntactic analysis**, parallel text alignment etc. CL is more associated with these tasks than with the very general tasks listed earlier, and they can be taken as ways of testing theories rather than producing useful artefacts.

Linguists are not the only scientists wishing to test theories of language functioning — there are also **psychologists** and **neurophysiologists** — and the dominant linguistic paradigm of the last half century, Chomsky's, has never believed that CL was the way to test linguistic theories. This dispute is over what constitutes the data of language study: it separates out very clearly NLP and CL on the one side, from linguistics proper on the other, where data is intimately connected with the intuitions of a speaker rather than with computable processes. Since 1990 emphasis has shifted to the use of corpora, of actual texts, than those imagined or written by linguists. Corpora are now normally gleaned from the web, and have become the canonical data of NLP and CL.

An element in the history of NLP/CL that cannot be overemphasised is the effect of hardware developments that have produced extraordinary increases in the storage and processing power available for experiments. This is obvious, and its effect on the field's development can be seen by considering the case of Sparck Jones' thesis (1966/1986) which was almost certainly the first work to apply statistical clustering techniques to

semantic issues and the first to make use of a large lexical resource, namely Roget's Thesaurus. Her statistical "clump" algorithms required the computation of large matrices that simply could not be fully computed with the tiny machines of 1964, with the result that this work's significance was not appreciated at the time, and it has been regularly rediscovered, usually without knowledge of the original, at regular intervals ever since.

The first piece of work to capture attention outside mainstream NLP was Winograd's SHRDLU thesis at MIT in 1971 (Winograd, 1971). One reason for the interest it aroused in the wider AI community was its choice of domain: the MIT Blocks World used for robotics and planning research, which consisted of blocks of different shapes that could be stacked, and were either real or simulated (simulated in Winograd's case) as well as a crane and a box for putting blocks in, all on a table top. It was a small world about which it was possible to know every fact. Winograd designed a dialogue program that discussed this world and manipulated it by responding to requests like PUT THE RED BLOCK ON THE GREEN BLOCK INTO THE BOX.

This system had many sophisticated features including an implementation of a Halliday grammar in a procedural language, PROGRAMMAR, that prefigured the language that was designed specifically for processing strings of symbols, such as sentences. It also had a method of forming up truth conditions in a form in LISP that could then be evaluated against the state of the blocks world. These conditions expressed the semantic content of an utterance and their value, when run, gave the denotation of the sentence, which might be the name of a block, or false if nothing satisfied them. This was an elegant and procedural implementation of the Fregean distinction of **sense and reference**. Like most systems of its time, it was not available for general testing and performed on only a handful of sentences, as was normal at that time.

SHRDLU's virtues and failings can be seen by contrasting it with a contemporary system from Stanford: Colby's PARRY dialogue system (Colby, 1973). This, also programmed in LISP, was made available over the then young Internet and tested by thousands of users, who often refused to believe they had not been typing to a human being. It simulated a paranoid patient in a Veterans' Hospital, and had all the interest and conversational skills that Weizenbaum's more famous but trivial ELIZA lacked. It was very robust, appeared to remember what was said to it and reacted badly when internal parameters called FEAR and ANGER became high. It did not repeat itself and appeared anxious to contribute to the conversation when subjects about which it was paranoid were touched on: horses, racing, Italians and the Mafia. It had no grammar, parsing or logic like SHRDLU, but only a very fast table of some six thousand patterns that were matched onto its input.

Contrasts between these two systems show issues that became more important later in NLP: widely available and robust systems versus toy ones; grammar parsing, which was cumbrous and rarely successful, versus surface pattern matching (later to be called **Information Extraction**); systems driven by world knowledge versus those which were not, like PARRY, and which essentially "knew" nothing, although it would have been a far better choice as a desert island companion than SHRDLU.

We began this historical essay by looking briefly at samples of important and prescient early work, and then showing two contrasting, slightly later, approaches, to the extraction of content, evaluation, representation, and the role of knowledge. We shall now consider

five types of system based on their own theoretical and methodological assumptions, and in this way try to get a picture of the range of influences that have been brought to bear on CL/NLP since the early Seventies.

## 2. Systems in relation to Linguistics.

Explicit links between CL/NLP and linguistics proper are neither as numerous nor as productive as one might imagine. We have already referred to the early schism between Yngve and Chomsky over the nature of tree representations and, more importantly over the role of procedures and processing resources in the computation of syntactic structure. It was Yngve's claim that such computation had to respect limits on storage capacity for intermediate structures, which he assumed corresponded to innate constraints on human processing of languages, such as George Miller's contemporary claim about the depth of human linguistic processing. Chomsky, on the other hand, assigned all such considerations to mere language **performance**.

There were in the Sixties a number of attempts to program Chomsky's **transformational grammars** so as to parse sentences: the largest and longest running was at IBM in New York. These were uniformly unsuccessful in that they parsed little or nothing beyond the sentences for which they had been designed, and even then produced a large number of readings between which it was impossible to choose. This last was the fate of virtually all syntactic analysers until the more recent statistical developments described below, including the original Harvard analyser of Kuno and Oettinger (1962), and the parsers based on the more sophisticated linguistic grammars of the Seventies and Eighties.

The last were linguistically motivated but designed explicitly as the basis for parsers, unlike linguistic grammars; the best known was GPSG from Gazdar and colleagues (Gazdar, 1982) which constituted a return to phrase structure, together with procedures for access to deeply nested constituents that owed nothing to transformations. Later came LFG (Lexical-functional Grammar) from Kaplan and Bresnan (1982), and FUG (Functional Unification Grammar) from Martin Kay (1984) which, like Winograd earlier, was inspired by **Halliday's** grammars (Halliday, 1976), as well as the unification logic paradigm for grammar processing that came in with the programming language Prolog.

These researchers shared with Chomsky, and linguists in general, the belief that the determination of syntactic structure was not only an end in itself, in that it was a self-sufficient task, but was also necessary for the determination of semantic structure. It was not until much later, and the development of techniques like **Information Extraction** that this link was questioned with large-scale experimental results. However, it was questioned very early by those in NLP who saw semantic structure as primary and substantially independent of syntactic structure as far as the determination of content was concerned; these researchers, such and Schank and Wilks in the Sixties and Seventies, drew some inspiration and support from the **case grammar** of Fillmore (Fillmore, 1968). He had argued, initially within the Chomskyan paradigm, that the case elements of a verb are crucial to sentence structure (e.g. **agents**, patients, recipients of actions), an approach which came to emphasise the semantic content of language more than its grammatical structure, since these case elements could appear under many grammatical forms. There have been hundreds of attempts to parse sentences computationally into case structure and **Fillmore** remains almost certainly the linguist with most explicit influence on NLP/CL as a whole.

Syntactic and semantic structure can be linked in another way to procedures by considering the traditional issue of the **centre-embedding** of sentences in English. The rule

$$S \rightarrow aSb,$$

where ab is a sentence, is generally considered a rule of English, producing sentences such as "the cat the man bit died". The problem is that repeated applications of the rule rapidly produce sentences that are well-formed but incomprehensible, such as "The cat the man the dog chased bit died" and so on. Evidence suggests there may be resource limitations on repeated applications of rules, corresponding in some way to syntactic processing limitations in the human, which is no surprise within NLP but which has no place within linguistics.

However, the situation is more complex: DeRoeck and colleagues (DeRoeck et al., 1982) found the following perfectly comprehensible sentence: "Isn't it more likely that example sentences that people that you know produce are more likely to be accepted" which, give or take the "Isn't it true that", has the same depth of syntactic centre-embedding as the (incomprehensible) cat-dog sentence above. This seems to show that, even given some depth limitation on the comprehension of centre-embeddings, there may be another effect at work, namely that the sentence above is understood not by means of syntactic analysis at all but by some other, possibly more superficial, surface semantic coherence, which the cat-dog sentence fails to possess. This is precisely the sort of consideration that motivated the semantics-based understanding movement of the Sixties and Seventies.

## 3. Representation issues: logic, knowledge and semantics.

There is an extreme view of NLP, held by AI researchers for whom logic and knowledge representation are still its main technique, that, in Hewitt's words, "language is just a side-effect" (Hewitt, 1971). By that he meant that, since AI could be seen as knowledge-based processing then, if only we had a full computer-based representation of knowledge, that alone would effect the understanding of human language, a matter which then has no intrinsic interest on its own. Unsurprisingly, this view has little support in NLP/CL but it does capture a core AI view about the universal power of logic-based knowledge representation, a vision of some antiquity, going back at least to Carnap's Logische Aufbau der Welt, the logical structure of the world (Carnap, 1928).

The central AI vision (e.g. McCarthy & Hayes, 1969) is that some version of the **First Order Predicate Calculus** (FOPC), augmented by whatever mechanisms are necessary, will be found sufficient for this task of representing language and knowledge, a standard view since (McCarthy & Hayes, 1969). This position, and its parallel movement in linguistic **semantics** claim that logic can and should provide the underlying semantics of natural language, and it has had a profound and continuing effect on CL/NLP. In linguistics the view is usually ascribed first to **Lakoff's Generative Semantics** movement, in some ways a natural extension to **transformational grammar**, albeit never acknowledged by Chomsky, given the logical origins of that movement in **Carnap's rules of transformation** as part of what he called **logical syntax**. Its high point was **Montague's model theoretic semantics** (Montague, 1970) for English in the late Sixties, which aimed to formalise language semantics independently of Chomsky's theories.

Although these movements, in AI and linguistics, have many formal achievements in print, they have had little success in producing any general and usable program to translate English to formal logic, nor indeed any demonstration from psychology, that such a translation into logic would correspond to the human storage and manipulation of meaning. In more surface-oriented and recent movements like that of **Information Extraction**, a task driven largely by evaluation competitions run by the US agency DARPA, the translation of English to FOPC structures remains a goal, but no one has yet set realistic standards for its achievement.

Part of the problem that any such translation scheme raises is the following: logical structure is not a mere decoration but something designed to take part in proofs. There will undoubtedly be NLP applications that require logical inferences to be established between sentence representations but, if those are only part of an application (e.g. the consistency of times in an airline reservation system), it is not clear they have anything to do with the underlying meaning structure of natural language, and hence with CL/NLP proper. At this point, there are a number of possible routes to take: one can say (a) that logical inferences are intimately involved in the meaning of sentences, since to know their meanings is to be able to draw inferences, and logic is the best way to do that. A recent survey of such approaches in linguistics is in Pulman (2005). One can also say (b) that there can be meaning representation outside logic, and this can be found in linguistics back to the **semantic marker** theories of **Fodor and Katz** (1963), developed within the transformational paradigm, as well quite independently, in NLP as forms of **computational semantics**. There is also a more extreme position (c) that the predicates of logic, and formal systems generally, such as ontologies, only appear to be different from human language (often accentuated by writing their predicates in capital letters) but this is an illusion, and their terms are in fact the language words they appear to be, as prone to ambiguity and vagueness as other words; both sides of this are argued in (Nirenburg & Wilks, 2001).

Under (a) above, one should note the highly original work of Perrault and colleagues at Toronto in the late Seventies (Perrault et al., 1980) who were the first group to compute over beliefs represented in FOPC so as to assign **Speech Acts** to utterances in a dialogue system. Speech Acts are a notion drawn from **Searle's** work in philosophy which has become the central concept in computational **pragmatics**, one that might enable a system to distinguish a request for information from an apparent question that is really a command, such as "Can you close the door?" The Toronto system was designed as a railway advisory system for passengers, and made use of limited logical reasoning to establish e.g. that the system knew when a given train arrived, and the passenger knew it did, so the question "Do you know when the next train from Montreal arrives" would not be, as it might appear, about the system's own knowledge.

Under (b) above, one can indicate the NLP tradition of the Seventies and Eighties of conceptual/semantic codings of meaning (already mentioned in the last section) by means of a language of primitive elements and the drawing of (non-logical) inferences from structures based on them. The best known of such systems were Schank's Conceptual Dependency system (1975) and Wilks (Wilks & Fass, 1992) Preference Semantics system; both were implemented in interlingual MT systems, and a range of other applications. Schank's system was based on a set of 14 primitive verbs and Wilks' on a set of about 80 primitives of various types. Schank asserted firmly that his primitives

were not English words, in spite of similarities of appearance (e.g. with INGEST), whereas Wilks argued there could be many sets of primitives and that they were no more than privileged words, as in dictionary definitions (see part 4 below). Wilks' notion of "preference" became well known: that verbs and adjectives have preferred agents, objects etc. and that knowledge of these default preferences is the major method of ambiguity resolution. Such preferences were later computed statistically when NLP became larger scale and more empirical (see part 5 below). Schank later developed larger scale structures called **scripts** that became highly influential as a way of capturing the overall meaning of texts and dialogues.

There are strong analogies between this strand of NLP work and contemporary work in linguistics, particularly with Fillmore and Lakoff, but there was at that time little or no direct contact between researchers in NLP and linguistics proper. That is one of the most striking changes over the last twenty years, and the simplest explanation is distance from Chomsky's distaste for all things computational, and the realization by linguists, at least since the work of Gazdar, that computational methods could be central for them. In spite of this distance there were undoubtedly influences across the divide: no one can see the semantic structures of **Jackendoff** (1983), involving structured sequences of primitives like:

CAUSE GO LIQUID TO IN MOUTHOF

as representing "drink", without feeling their similarity to the earlier NLP structures mentioned above.

## 4. Corpora, resources and dictionaries.

In the Sixties, Masterman (1957) and Sparck Jones (1966/1986) had made use of Roget's **thesaurus**, punched onto IBM cards, as a device for word sense disambiguation and semantic primitive derivation, respectively, even though they could not do serious computations on them with the computers then available. After that, large scale linguistic computation was found only in MT, and in the era of the influence of AI methods in CL/NLP the vocabularies of working systems were found to average about 35, which gave rise to the term "toy systems" to refer to most of the systems described above.

But there were movements to bring together substantial **corpora** of texts for experiments, although these were driven largely from the humanities and in the interests of stylistic studies and statistical measures of word use and distribution. The **best known** of these was the Brown-Oslo-Bergen corpus of English (Francis & Kucera, 1964), but the British National Corpus (http://www.natcorp.ox.ac.uk/index.html) was constructed explicitly with the needs of NLP in mind, and the University of Lancaster team, under Geoffrey Leech, played a key role in its construction. This group had already created the first effective piece of corpora-based statistical NLP, the part-of-speech tagger CLAWS4 (Garside, 1987).

At very much the same time, in the early Eighties, interest arose in the value to NLP, not only of text corpora, but specifically of the texts that are dictionaries, both monolingual and bilingual. Bran Boguraev in Cambridge was one of the first researchers (since very early work on Webster's Third Dictionary at Systems Development Corporation in the Sixties (Olney, 1968)) to seek to make use of dictionary via its electronic printing tape, in this case of the Longmans Dictionary of Contemporary English, a dictionary specifically

designed for foreign learners of the language. This had definitions with restricted syntax drawn from a vocabulary of only 2000 words.

In the Eighties there was a great deal of activity devoted to extracting computational meaning on a large scale from such machine-readable dictionaries (see Wilks et al., 1996): it seemed a sensible way to overcome the "toy system" problem, and after all dictionaries contained meaning, did they not, so why not seek it there? Substantial and useful semantic databases were constructed automatically from LDOCE and a range of other dictionaries, again usually dictionaries for learners of English since they expressed themselves more explicitly than traditional dictionaries for scholars and the broadly educated. Hierarchical ontologies were constructed automatically, and these databases of definitions remain, along with thesauri, a component database for many major systems for resolving word sense ambiguity.

But such dictionaries were not a panacea that cured the problem of meaning, and it became clear that dictionaries themselves require substantial implicit knowledge to be of computational use, knowledge both of the world and of the primitive vocabulary contained in their definitions.

Brief mention here should be made of systematic annotation codings—the automatic attachment of tags representing linguistic information to the words of a text—which began, again, in the humanities with the language SGML for marking up corpora. This type of annotation has now become a huge range of annotations in differing modalities, the best known of which are HTML and XML the annotations underlying the World Wide Web. A curious effect of all this has been to bring programs, once thought of as quite disjoint from texts, into the space of objects that are themselves annotated texts, which is an unexpected new universality for linguistics, taken broadly.

Another, quite independent, source of annotated corpus resources were tree banks, of which the Penn Tree Bank (Marcus, 1993) is the best known: a corpus syntactically structured by hand, with the syntactic structure being added to the text as annotations, indicating structure and not merely categories. One effect of the wide use of the Penn Tree Bank for experiments was to enshrine the texts used for it, in particular sections of the Wall Street Journal, as ueber-corpora, used so much and so often that some believed their particular features had distorted NLP research.

In the recent past much energy and discussion was put into the selecting and "balancing" corpora—so much dialogue, so many novels and memoranda etc.---- but this activity is becoming irrelevant because of the growing use of very large parts of the world wide web itself as a corpus that can be annotated. The so-called Semantic Web project (Berners-Lee et al., 2001) has as one of its aims the annotation of the whole web-as-a-corpus, so that machines as well as humans can read its content. This is a project that envisages such annotations as reaching further than traditional linguistic annotations, of say syntactic or semantic type, right up to annotating logical structure. This goal brings the project back to the traditional AI one of automatically translating the whole of human language into logic. The value of this translation, even if possible, has yet to be shown in practice.

## 5. Statistical and quantitative methods in NLP

This movement is the most difficult to survey in brief, largely because we are still within it at the time of writing (see Manning & Schuetze, 1999). In the Sixties, Gilbert King

predicted that MT could be done by statistical methods, on the ground of the well-known 50% redundancy of characters and words in Western languages, though it is not easy to see why the second justified the first. Later, and as we saw earlier, Sparck Jones pioneered what were essentially IR methods to produce semantic classifications, intended ultimately for use in MT.

We noted earlier that the first clear example of modern statistical NLP was the work by Leech and his colleagues on the CLAWS4 part-of-speech tagger in the late Seventies. At the time, few could see the interest of assigning part-of-speech categories to text words. Yet now, only two decades later, almost all text processing work starts with a part-of-speech assignment phase, since this is now believed (even at about 98% accuracy, the usual level achieved) to simplify all subsequent linguistic processes, by filtering out a large range of possibilities that used to overtax syntactic analysers. The undoubted success of such methods, showed that analysis decisions previously believed to need "high level" syntactic or semantic information, could in fact be taken at a low level by methods like n-gram statistics over sequences of words.

The greatest impetus for statistical NLP, however, came from work on the MT, research program of Jelinek (Brown et al., 1990) and his group at IBM who were applying methods that had been successful in automatic speech recognition (ASR) to what had been considered a purely symbolic (linguistic or AI) problem. Jelinek began asking what was the phenomenon to be modelled—answer, translation—and then seeking examples of that human skill to apply machine learning to. The most obvious case was parallel corpora: texts expressing the same meaning in more than one language. These were widely available and he took the Canadian Hansard texts in English and French.

We can see already some of the major forms machine learning (ML) in NLP can take: in the CLAWS4 work the phenomenon (part-of-speech tagging) had been annotated onto the text by humans and the ML algorithms were then set to recapitulate those, in the sense of being able to tag new, unseen, texts at some acceptable level of accuracy.

This is called supervised ML; in the Jelinek work, on the other hand, although the targets to be learned are given, namely the translations in the parallel texts, the training material had not been produced specifically for this task, but consisted of naturally occurring texts, albeit produced by people. Many would call this weakly supervised ML. In the work of Sparck Jones, described earlier, however, the clusters found were not set up in advance, and this is normally called unsupervised ML.

The Jelinek work produced an accuracy level of about 50% of sentences translated correctly, a remarkable fact given that the system had no linguistic knowledge of any kind. When applied to new, unseen, texts it failed to beat the traditional, hand-coded, MT system SYSTRAN, which had not been trained for specific kinds of text.

Jelinek's CANDIDE system was a benchmark in that it suggested there were limits to purely ASR-derived statistical methods applied to a linguistic task like MT, and he himself began a program for the derivation of linguistic structures (lexicons, grammars etc.) by those same statistical ML methods, in an attempt to raise the levels CANDIDE's success, and in doing so he set in motion a movement throughout NLP to learn traditional NLP/CL structures at every linguistic level by those methods.

There are now far too many such applications to cite here: ML methods have been applied to the alignment of texts, syntactic analysis, semantic tagging, word-sense

disambiguation (Yarowsky, 1995), speech act assignment, and even dialogue management. In the case of some of these traditional tasks, the nature of the task has changed with the evaluation and scoring regimes that have come along with the paradigm shift. For example, it was conventional to say, only a few years ago, that syntactic parsers had failed, at least for languages like English, and that there simply was no parser that could be relied on to produce a correct parse for an unseen English sentence, or at least not one that could be reliably picked out, by probabilities or other ordering, from a forest of alternatives. However, now that statistically-based parsers learn over tree banks and are scored by the number of brackets they can correctly insert, and the appropriate phrase structure annotations they can assign, the issue is merely quantitative and it is no longer considered essential that a "full parse" (i.e. to the S symbol) is produced. Charniak currently produces the best figures (2001) for doing this.

There is a general perception that statistical, or corpus-driven (alias empirical), linguistics has resulted in a shift to surface considerations in language: e.g. the shallower syntactic structures just mentioned that have allowed syntactic analysis to become more useful in linguistic processing, because more successful and reliable. One could also point to the success of the independent task **Information Extraction** (IE, Gaizauskas and Wilks, 1997) which consists, in broad terms, in extracting fact-like structures from texts on a large scale for practical purposes, e.g. all those in IBM promoted in the 1990s, extracted from public source newspapers. At the 95+% level that is the norm of acceptability in empirical linguistics, IE has become an established technology and this has been achieved largely by surface pattern matching, rather than by syntactic analysis and the use of knowledge structures, although those have played a role in some successful systems.

However, many of the more recent successes of empirical linguistics, again based on ML over corpora, have been in areas normally considered semantic or "less superficial" in nature, such as word-sense disambiguation and the annotation of dialogue utterances with their dialogue or speech acts, indicating their function in the overall dialogue.

It may well be that raising the currently low figure for tagging dialogue acts (80%) to an acceptable level does require more complex structures to be modelled, as was shown to be the case in Jelinek's approach to MT e.g. the modelling of dialogue managers and agent belief systems, but it is proving much harder to model and evaluate these independently than was the case for components of an MT system.

## 6. Computational Linguistics as an independent paradigm?

In conclusion, let us consider briefly to what extent CL/NLP is an independent paradigm (see Cole, 1996), rather than being just a sub-division of Linguistics (or even AI). It is certainly the case that a small number of linguists have had disproportionate and continuing influence on the development of CL/NLP: Halliday and Fillmore's work continues to appear in computational paradigms, and Halliday's influence on Kay's Functional Unificational Grammar is clear. Chomsky, by contrast, has had little influence in CL since the unsuccessful attempts in the Sixties to program Transformational Grammars.

It is also clear that much of the development described in this article can be traced to the influence on CL/NLP of some combination of the following movements:

1. linguistics itself,
2. logic and knowledge representation in AI,
3. statistical methods: speech research, neural net/connectionists, the evaluation community and Information Retrieval,
4. lexicographers and corpus experts.

But there is another strand of influence, one hard to describe, but coming directly from computation itself, namely procedurally-based theories: those in which the procedures are essential, and not merely the programming of rules constraining some domain. Some elements of NLP constitute a kind of core NLP, definitive of the subject. In such a list one could include:

i). Winograd's procedural expressions of grammar, truth conditions and the movement content of verbs;

ii). Marcus' syntactic parser (1980), which put a resource bound on searching structures in an attempt to capture the notion of "garden path sentences";

iii). Charniak's (1983) attempt to limit searches of semantic nets by means of a finite resource, on the assumption that correct results are defined partly by resources available.

iv). Wilks' Preference semantics (Wilks & Fass, 1992), an attempt to define the best semantic structure for an utterance as the maximally coherent one in terms of satisfied preferences;

v). Woods' display of grammars as a path tracking procedures (ATNs) augmented by recursive pushdown stacks and registers (Woods et al., 1974).

vi). A number of authors, including Gazdar (Evans & Gazdar, 1996) and Pustejovsky (1996), who attempted to define appropriate dictionary entries by some level of maximal compression of information.

vii). Waltz and Pollack's (1985) connectionist model of word-sense choice in terms of affinity and repulsion.

viii). Much of the work of Yngve referred to at the beginning of the paper, especially the notion of limiting syntactic depth.

ix). Hirst (1990) and others who have attempted to define semantic structure as one progressively revealed and specified by incoming information.

x). Grosz's definition of the accessibility of **discourse** constituents with a network where partitions are progressively closed off (Grosz & Sidner, 1986).

One could continue with this list, but it might not be especially revealing, and it would certainly not include all or most of NLP/CL. The act of making it does seek, however, to raise the question of whether there is some distinctive core of CL/NLP that captures human language behaviour, as well as machine behaviour, by some set of procedures based on information compression and the minimisation of effort, a component in several theories on that list. All science is information compression, in a wide sense, and it is certainly plausible that the brain, and any other language machine, will have available distinctive procedures to do this, as opposed to the brute force methods of statistics which are implausible as models of human language processing. About this last, Chomsky was probably right.

Finally, it is not possible to understand the history of NLP/CL over the last half-century without seeing the crucial role of funders in it, and particularly the US Defense Department, which created MT from nothing in the US and, through DARPA and ARPA, has continued to shape the field there, and to some extent world-wide. In recent years, it has been the DARPA evaluation competitions, open to all, that created Information Extraction and then the whole empirical linguistics movement we are still participating in. Whether all this effort defended anybody or anything is, of course, another question.

## References


Berners-Lee, T., Hendler, J. & Lassila, O. (2001) The semantic web. *Scientific American*.

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 Linguistics 16:2*: 79-85.

Carnap, R. (1928) *Der Logische Aufbau der Welt*. Berlin: Weltkreis.

Ceccato, S. (1961) Operational linguistics and translation. In S.Ceccato (Ed.) *Linguistic analysis and programming for mechanical translation*. New York: Gordon & Breach, pp. 117-129.

Charniak, E. (1983) Passing markers: A theory of contextual influence in language comprehension. *Cognitive Science, 7*, pp. 171-190.

Charniak, E. (2001) Immediate-head parsing for language models. *In Proc. of the 39th Annual Meeting of the Association for Computational Linguistics*.

Colby, K.M. (1973) Simulation of belief systems. In Schank & Colby, (eds.) *Computer Models of Thought and Language*, (San Francisco, CA: W.H. Freeman)

Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A. & Zue, V. (eds.) (1996) *Survey of the State of the Art in Human Language Technology.* MIT Press, Cambridge, MA.

Cooper, R.P. (1996) Head-Driven Phrase Structure Grammar. In Brown & Miller (eds.) *Concise Encyclopedia of Syntactic Theories 191-196*. Oxford: Pergamon, pp. 152-179.

De Roeck, A. et al. (1982) A myth about centre-embedding. *Lingua 58*. 327-40.

Evans, R. & Gazdar, G. (1996) DATR: A language for lexical knowledge representation. *Computational Linguistics, 22.2*, pp. 167-216

Fillmore, C. (1968) The Case for case. *In Universals in Linguistic Theory*. Bach & Harms (eds.), pp. 1-90, New York: Holt, Rinehart and Winston.

Fillmore, C. (1977) The case for case reopened. In Cole and Sadock, (eds.) *Syntax and semantics 8: Grammatical relations*. NY: Academic Press, pp. 59-81.

Francis, W. & Kucera, H. (1964) A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Providence, Rhode Island Department of Linguistics Brown University.

Fodor, J.A. & Katz, J. (1963) The structure of a semantic theory. *LANGUAGE*, pp. 170-210.

Gaizauskas, R. & Wilks, Y. (1997) Information Extraction: beyond document retrieval. *Journal of Documentation*.

Garside, R. (1987) The CLAWS word-tagging system. In Garside, Leech & Sampson (eds.), *The Computational Analysis of English*. London and New York: Longman.

Gazdar, G. (1982) Phrase structure grammar. In Jacobson & Pullum (eds.) *The Nature of Syntactic Representation*. Dordrecht: Reidel, pp. 131-186. Reprinted in Kulas, Fetzer & Rankin (eds.) *Philosophy, Language, and Artificial Intelligence*. Dordrecht: Kluwer, pp. 163-218, 1988.

Grosz, J.B. & Sidner, C. (1986) Attention, intentions and the structure of discourse. *Computational Linguistics, 12(3)*, 175-204.

Halliday, M.A.K. (1976) Halliday: System and function in language. *Selected papers*, Kress (ed.). London: Oxford University Press.

Hewitt, C. (1971) Procedural Semantics. In Rustin (ed.) *Natural Language Processing Courant Computer Science Symposium 8*. Algorithmics Press, pp. 180-198.

Hirst, G. (1990) Mixed-depth representations for natural language text. *AAAI Spring Symposium on Text-Based Intelligent Systems*, Stanford, March, 25-29.

Jackendoff, R. (1983) *Semantics and Cognition.* MIT Press, Cambridge, MA.

Kaplan, R.M. & Bresnan, J. (1982) Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan (ed.) *The Mental Representation of Grammatical Relations*, pp. 173-281. Cambridge, MA: MIT Press.

Kay, M. (1984) Functional Unification Grammar: a formalism for machine translation. *In Proceedings of the 22nd conference on Association for Computational Linguistics*, pp. 75-78. Stanford, California.

King, G.W. (1961/2003) Stochastic Methods of Mechanical Translation. In Nirenburg, Somers & Wilks (eds.) 2003. *Readings in Machine Translation*. Chapter 4, pp. 45-51. MIT Press, Cambridge, MA.

Kuno, S., & Oettinger, A. (1962) Multiple-Path Syntactic Analyzer. *In Proceedings of IFIP Congress '62,* Munich, pp. 1143-1162.

Manning, C.D. & Schuetze, H. (1999) *Foundations of statistical natural language processing*, MIT Press, Cambridge, MA.

Marcus, M. (1980) *A Theory of Syntactic Recognition for Natural Language.* MIT Press.

Marcus, M. (1993) Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics, Vol.19*, pp. 87-105.

Masterman, M. (1957) The Thesaurus in Syntax and Semantics. *Mechanical Translation, 4(1-2)*, pp. 35-43.

McCarthy, J. & Hayes, P. (1969) Some philosophical problems from the standpoint of Artificial Intelligence. *In Machine Intelligence 4*. Edinburgh Univ. Press, Edinburgh.

Montague, R. (1970) English as a Formal Language. In Visentini, B. et al (eds.) *Linguaggi nella Societa e nella Tecnica. Edizioni di Comunita*. Milan, Italy, pp. 98-119.

Nirenburg, S. & Wilks, Y. (2001) What's in a symbol: ontology, representation and language. *Journal of Experimental and Theoretical Artificial Intelligence (JETAI).*

Olney, J., Revard, C. & Ziff, P. (1968) Some monsters in Noah's Ark. *Research memorandum SP-2698.* Systems Development Corp., Santa Monica, CA

Perrault, R., Cohen, P. &Allen, A. (1980) A plan-based analysis of indirect speech acts. *Computational Linguistics, Vol. 6, Issue 3-4*, pp. 167-182.

Pulman, S.G. (2005) Lexical decomposition: for and against, In Tait (ed.) *Charting a New Course: Natural Language Processing and Information Retrieval*, Cambridge, Cambridge Univ Press.

Pustejovsky, J. (1996) *The Generative Lexicon*. MIT Press.

Schank, R. (1975) *Conceptual information processing*. North Holland, Amsterdam.

Sparck Jones, K. (1966/1986) *Synonymy and semantic classification*. Edinburgh University Press, Edinburgh.

Varile, N. & Zampolli, A. (eds.) (1997) S*urvey of the state of the art in human language technology.* Cambridge: Cambridge University Press.

Waltz, D.L. & Pollack, J. (1985) Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation, *Cognitive Science 9,1*. January-March, pp. 57-84.

Wilks, Y. & Fass, D. (1992) Preference semantics: a family history. *In Computing and Mathematics with Applications, Vol. 23, No. 2*.

Wilks, Y., Slator, B. & Guthrie, L. (1996) *Electric words: dictionaries, computers and meanings*. Cambridge, MA: MIT Press.

Winograd, T. (1971) *Understanding natural language*. MIT Press, Cambridge, MA.

Woods, W., Kaplan, R. & Nash-Webber, B. (1974) The lunar sciences natural language information system, *Final Report 2378*, Bolt, Beranek & Newman, Inc., Cambridge, MA.

Yarowsky, D. (1995) Unsupervised word-sense disambiguation rivalling supervised methods. *In Proc. ACL'95*.

Yngve, V H. (1960) A model and an hypothesis for language structure. *In Proc. of the American Philosophical Society, Vol. 104, No. 5*, pp. 444-466.

## Mini bio

Yorick Wilks is the founder of the Sheffield NLP Research group, the largest in England, and Director of Sheffield's Institute of Language, Speech and Hearing. He has published numerous articles and seven books in that area of Artificial Intelligence. He is also a member of the EPSRC College of Computing and the UK Computing Research Council. He is a Fellow of the European and American Association for Artificial Intelligence, on advisory committees for the EU, the National Science Foundation, and on the boards of some fifteen AI-related journals.