CHAPTER 32

COMPUTATIONAL LINGUISTICS

GRAEME HIRST

32.1 Origins in Machine Translation

The field of computational linguistics (CL) has its origins in research on machine translation (MT) in the early days of computing in the 1940s. The founding moment of MT is traced to a widely circulated memorandum written in July 1949 by the American mathematician and science administrator Warren Weaver. Weaver's memorandum, which drew on ideas from a letter he had written to the cybernetician Norbert Wiener two years earlier, proposed the idea of automatic translation that would use the (assumed) common underlying logical structure of all languages and would resolve ambiguities in context. His proposal thus went beyond the idea of simple 'translations' based on lexical-substitution dictionaries and rudimentary analysis of accidence that Richard H. (Dick) Richens and Andrew Donald Booth had already begun experimenting with in England (Richens and Booth 1955,[1] Booth and Locke 1955).

The concept was greeted with much enthusiasm, scepticism, and research funding. Many MT research groups formed in the US, UK, USSR, and elsewhere in the 1950s; see Hutchins (1986) for a detailed list and a discussion of their technical work.[2] Given the early state of development of both software and hardware at the time, research on MT systems was inextricably bound up with research on more general problems in computer science and engineering, including the development of larger and faster physical storage media with more efficient data structures and indexes so that large

[1] Although published only in 1955, this paper was written in 1948 (Spärck Jones 2000).

[2] The early history of machine translation will not be recounted here in any detail, as it has been documented extensively elsewhere (Hutchins 1986, 2000a, Zarechnak 1979).

dictionaries could be stored.[3] While the work was truly interdisciplinary, bringing together researchers from fields as disparate as engineering, physics, cybernetics, and linguistics (Hutchins 2000b), much of it was also naïve with regard to the nature of language and translation. Thus in 1955, Booth and Locke could earnestly write, 'It is not possible to define a language in terms of its words alone. Its grammar must also be taken into account.'

Notable exceptions to this naïvety could be found in those groups that worked beyond MT on language processing more generally. The (University of) Cambridge Language Research Unit (CLRU), founded in 1954, was run by Margaret Masterman, a student of Wittgenstein whose broad interests and unorthodox (for the time) approaches led to the development of ideas, such as theories of synonymy and interlinguas for MT (Richens 1958; see also Spärck Jones 2000), that were years ahead of their time (Wilks 2000, 2003). The CLRU was also home to important early research on information retrieval by Karen Spärck Jones (see §32.3.1.3 below). Similarly, at the RAND Corporation (Santa Monica, Calif.), the MT group led by David G. Hays followed the principle that more linguistic research was necessary before MT could be possible, believing that 'the broader field of computational linguistics deserved general attention' (Hays 1967: 15). Indeed, it was Hays who in 1962 coined the term 'computational linguistics' (Kay 2000).[4] Presaging the approaches of two decades later (§32.3 below), the RAND research had a strong empirical component, which included the creation of syntactically annotated corpora of Russian texts for descriptive analysis and grammar development (Hays 1967).

MT research in the USSR appears to have begun only in 1954, after Aleksej A. Ljapunov, a leading Soviet cybernetician, read about an American project (Kulagina 2000); but by the end of the 1950s, the USSR had more MT researchers than all other countries combined (Harper 1961). What the USSR lacked was sufficient computers, and MT was largely regarded as a 'thought-experiment' (Mel'čuk 2000: 216); systems were tested by having people who did not know the target language perform translations by following flowcharts (Harper 1961). In such a situation, and in contrast to most US MT research, linguistic theory flourished, including the development of the meaning–text model of language (Mel'čuk and Žolkovskij 1970).[5]

[3] Booth (1955) estimated that a simple translating machine would require 6 physical components, including input and output magnetic tape units and a magnetic or optical drum for its dictionary and grammar, and would cost around $100,000 to build (about 50 times the price of an automobile). The relatively low cost was due to Booth's naïve assumption that a simple special-purpose letter-manipulating 'computer,' rather than a general-purpose central processing unit, would be sufficient.

[4] In 1960, Hays wrote, 'We are fretting under the MT label . . . [and] we petitioned for a change. Our new titles are *linguistic research* and *automatic language-data processing*. These phrases cover MT, but they allow scope for other applications and for basic research. Machine translation is no doubt the easiest form of automatic language-data processing, but it is probably one of the least important. We are taking the first steps toward a revolutionary change in methods of handling every kind of natural-language material' (published in Hays 1961: 24–5).

[5] For details of the theoretical approaches, see Rozencvejg (1974), a collection of English translations of papers on formal and theoretical linguistic aspects of Soviet MT published in the journal *Машинный*

By the early 1960s, it became clear that current approaches to MT were not successful. Critics included Hays, as noted above, and Yehoshua Bar-Hillel, who had worked in MT since 1951 (in fact as the first full-time researcher in the field: Booth and Locke 1955), but who in 1960 published an acerbic and highly critical survey of the field. Bar-Hillel argued that fully automatic high-quality translation—FAHQT, as he called it—would never be possible because ambiguity resolution often required encyclopedic world knowledge that a computer could never have; it was therefore necessary to accept either low-quality translation or the need for a human post-editor. In any case, he also argued, MT was not cost-effective compared to human translation.

The same conclusion was reached six years later by an interdisciplinary committee established by the US National Academy of Sciences. The Automatic Language Processing Advisory Committee was chaired by John R. Pierce;[6] Hays was one of its members, later claiming to have been a dissenter on the committee (Zarechnak 1979). The committee's report (1966), known generally as the ALPAC report, stated bluntly that 'there is no immediate or predictable prospect of useful machine translation' (p. 32), and that MT was not needed anyway. The report recommended that research funding should instead be directed to computational linguistics, both for its many other potential applications and for consequent better understanding of the nature of language itself.

The report attracted much furious criticism, claiming that it was too narrowly focused, as it concentrated almost solely on US government needs for translation from Russian (Hutchins 1996); that it severely underestimated the usefulness of low-quality translation (Friedrich Krollmann quoted in Josselson 1971[7]); that it contained serious factual errors and based its conclusions on out-of-date information (Titus 1967, Josselson 1971); and even that the committee had a 'hostile and vindictive attitude' and 'concealed' and 'willful[ly] omi[tted]' data and views inconsistent with its conclusions (Zbigniew L. Pankowicz quoted in Josselson 1971).[8] Nonetheless, funding for MT was cut drastically, and many projects were terminated—not just in the US but in Europe and the USSR (Titus 1967, Hutchins 1986, Mel'čuk 2000). In 1968, the young Association for Machine Translation and Computational Linguistics removed the words 'Machine Translation' from its name to avoid the taint (Hutchins 1986). MT research did not begin to revive until the mid-1970s (Hutchins 2000b); but by 1976, a seminar on MT sponsored by the US Foreign Broadcast Information Service could report cautious optimism on the feasibility of MT (Hays and Mathias 1976).

перевод и прикладная лингвистика (Machine Translation and Applied Linguistics) in the period 1957–70. For details of Soviet MT systems of the period, see Hutchins (1986).

[6] Pierce was also known as a fierce critic of automatic speech recognition in the 1960s; see Church (2012).

[7] Josselson misspells the name as 'Krollman.'

[8] Winfred P. Lehmann, who was leader of MT research at the University of Texas, Austin, later wrote, 'I gave a presentation [to the committee] on 18 March 1964. . . . The effort was pointless. It's still difficult to forget the sneer on Pierce's face' (Lehmann 2000: 161).

32.2 Symbolic Methods, Artificial Intelligence, and Natural Language Understanding

32.2.1 Automatic Language Understanding as Artificial Intelligence

In 1950, in a seminal paper defending the concept of artificial intelligence (AI), Alan Turing proposed that if a computer had the ability to use language and converse knowledgeably about the world as well as any human, we should regard it as intelligent—a criterion that became known as the Turing test. The development of computational methods of natural language understanding, or NLU, thus became one part of the problem of building intelligent machines.

Nonetheless, little early work in AI addressed NLU. Only two papers in Feigenbaum and Feldman's (1963) representative collection concerned language; both involved largely unprincipled syntactic analysis of highly restricted English, but distinguished themselves from research in machine translation by their goal of complete semantic analysis, albeit within a trivial domain—kinship relations (Lindsay 1963) and questions about baseball statistics (Green et al. 1963). This work set the stage for computational linguistics, as it disentangled itself from machine translation, to be viewed as the facet of the then-glamorous field of AI, whose primary goal was human-like language understanding. Given this goal, CL became more open to influence from linguistics and psycholinguistics.

32.2.2 Early End-to-end Systems

Early NLU was advanced by three important developments, all based on linguistic theory and purely symbolic, non-quantitative processing. The first was Procedural Semantics, a method developed by William (Bill) Woods (1968) for deriving an executable database query from a parse tree that represented a natural language query on the database. The second, to produce the necessary parse tree, was the Augmented Transition Network (ATN) parser, also developed by Woods (1970) (see also Woods 2010); ATN parsers had the power of a transformational grammar. The two coupled together could act as a natural language query system for a database, and in 1970, Woods and his colleagues were commissioned by NASA to build a 'natural language understanding system' (Woods 1973: 441) that would answer geologists' questions about samples brought back from the recent lunar missions, such as 'What is the average concentration of aluminum in high alkali rocks?' Although the system

that was built (Woods 1973) never actually entered regular use by geologists, it was an influential proof-of-concept for subsequent research.

The third important development was also an end-to-end system billed as natural language understanding. Terry Winograd's 1972 SHRDLU system was based on early Hallidayan systemic grammar[9] and a procedural approach to semantics in which meanings were represented as executable procedures in a deductive problem-solving language named PLANNER. SHRDLU operated in the (simulated) world of a robot arm that could be ordered to manipulate a set of blocks on a table: 'Find a block which is taller than the one you are holding and put it into the box.' Just as Woods's systems could understand only utterances about the contents of their databases, and their replies were data, not sentences, SHRDLU relied on having perfect knowledge of its tiny world, and its responses were actions of the robot arm. Nonetheless, it was regarded as a tour de force, even as an indication that full NLU was close at hand.

32.2.3 The Representation of Knowledge as a Central Issue

The success of these systems focused attention on the need for research in artificial intelligence on computational representations of knowledge of the world—not just for language understanding but more generally for reasoning about and acting in the world. Moreover, a knowledgeable language understanding system would be able to read books and newspapers, and thereby gain more knowledge—learning by reading. It was necessary only to build a system just smart enough to get started.[10] Thus, an implicitly reader-based view of meaning was taken, in which systems would interpret new text in light of their present knowledge (Hirst 2007).

Early research by Marvin Minsky (1975) and by Eugene Charniak (1976, 1978) on the representation of knowledge emphasized recursive template-like slot-and-filler structures called 'frames.' Frames represented stereotypical or commonsense knowledge, could be arranged into default inheritance hierarchies, and could be used to make deductive and inductive inferences to answer questions about information that was not explicit in a text; for example, *Janet needed some money; she got her piggybank and started to shake it. Why did she shake the piggybank?* (Charniak 1973). Frames could be set to be invoked or triggered by certain keywords or concepts, or by the addition of new facts to the system. (In this respect, frames generalized Charniak's earlier (1973) approach in which inferences were made by snippets of knowledge characterized as 'demons' that were triggered by keywords or by prior inferences.) In many ways, frames resembled Fillmore's (1968) case structures in linguistics, although an explication of the relationship between the two came somewhat later (Charniak 1981).

[9] See Ch. 21 above.

[10] In fact, the so-called knowledge acquisition bottleneck remains a major problem in AI and CL. Contemporary CL includes the topics of 'knowledge acquisition from text' and 'learning by reading'; see §32.3.2 below.

A competing approach, known as conceptual dependency (CD), was developed by Roger Schank and his colleagues at Stanford University and subsequently at Yale University (Schank 1973, 1975). CD decomposed verbs and represented them as structures built from a small set of semantic primitives; a sentence was then represented as an instantiation of one or more such structures. Larger CD structures, known as scripts, represented stereotypical knowledge about situations and events (Schank and Abelson 1977); they were not dissimilar to frames but put a much greater emphasis on temporal sequence. Schank and his students built systems that could interpret paragraph-length stories, make inferences about the situations they described, and answer questions about them.

Although he had a Ph.D in linguistics, Schank deprecated linguistics and rejected the reality of syntax ('we have never been convinced of the need for grammars at all' (Schank 1975: 12)). The surface-form analysers in the systems that he and his students built created semantic structures directly, with no intervening syntactic representation. They were based wholly on heuristics and procedures invoked in response to the occurrence of particular words. (In this regard, the approach is similar to Charniak's described above, but at the surface level, not just in deeper understanding.)

Many other semantic representations were developed, almost all of which could be considered variations on the basic case-role-like theme of slot-and-filler. They included Yorick Wilks's (1975, 1978) 'preference semantics,' which contained mechanisms for using selectional restrictions to choose word senses and for backing off to non-literal meanings when selectional restrictions could not be satisfied, and John Sowa's (1984) 'conceptual graphs.' Often, the representations were motivated by research in cognitive psychology and were presented as psychological models, as evidenced by the subtitles of Schank and Abelson's (1977) and Sowa's (1984) books: *An Enquiry into Human Knowledge Structure* and *Information Processing in Mind and Machine*; copying how people do it was thought to be a good strategy for artificial intelligence. Indeed, many researchers in AI and CL regarded their work as part of the new discipline of cognitive science.

The systems that were built on all these theories typically had tiny vocabularies and knowledge bases whose domain was a minute sliver of the world. (CD researchers were frequently teased for their focus on the twin domains of interpersonal violence and eating in restaurants.) They were evaluated merely by demonstration on a handful of examples, with the tacit implication that if they could 'handle' a few representative examples, then all they needed in order to become useful was a realistically large knowledge base, which could be developed with some extra time and effort.[11] But almost without exception, these systems were equivalent to first-order predicate or propositional logic, and hence inherently inadequate as representations of the full expressivity of natural language. Concurrent research in formal semantics, such as the work of Richard Montague (1974b) and his successors (see §23.4.2.3 above), did not go unnoticed (e.g., Friedman et al. 1978b, Hobbs and Rosenschein 1977), but the higher-order and intensional logics that they required were computationally infeasible (Friedman et al. 1978a).

[11] See n. 10.

32.2.4 Parsing

By the early 1980s, the failure of syntax-free and syntax-lite approaches had become obvious (e.g. Lytinen 1985; for discussion, see Hirst 1987: 2–3), and those who rejected them in the first place had not been idle. In the 1960s, efficient algorithms for syntactic analysis were a topic of much interest. Much of this work took its cue from research in the analysis of formal languages such as programming languages. The Cocke–Kasami–Younger (or CKY or CYK) algorithm for parsing context-free languages in a time bounded above by the cube of the length of the input string was discovered independently by its three eponyms.[12] Earley's (1970) and Valiant's (1975) algorithms further improved efficiency, and laid the foundation for 'chart parsing,' in which a data structure known as a chart is used to store partially complete analyses of constituents for possible reuse in alternative analyses, greatly speeding up the process.

The belief of the time, following Chomsky (1957a), that context-free grammars were insufficient for natural languages led other researchers to develop more powerful approaches such as the transformation-based ATN parser (§32.2.2 above). Aravind Joshi's tree-adjoining grammars (TAGs), introduced by Joshi et al. (1975) and developed over a number of years by Joshi and his students (Abeillé and Rambow 2000), are mildly context-sensitive, and moreover were shown to be weakly equivalent (generating the same class of string languages) to three other independently developed formalisms of contemporaneous interest (Vijay-Shanker and Weir 1994). TAGs are unusual in that the primitives are not strings but trees, with distinguished root and foot nodes that enable operations of substitution and adjunction. TAGs were pioneers in the *lexicalization* of grammars to facilitate parsing—that is, associating each lexical item with one of a finite set of structures that may be composed by a set of operations (Schabes et al. 1988); lexicalization became particularly important later in the development of probabilistic parsers (§32.3.2 below).

Parsers for a number of other formalisms were also developed in this period, including lexical-functional grammars (Kaplan and Bresnan 1982), head-driven phrase structure grammars (HPSG) (Pollard and Sag 1987, 1994) (see §§8.3.1 and 21.1 above), and combinatory categorial grammars (Pareschi and Steedman 1987, Vijay-Shanker and Weir 1990). Many of these parsers were based on the operation of unification, which combines two partial descriptions of linguistic objects insofar as they are compatible. Unification permits the description of the grammar of a language to be largely declarative rather than process-based. It was developed independently (including the same choice of name) by both Martin Kay (1979) in the context of functional grammars and Alain Colmerauer and colleagues (e.g. Colmerauer 1978) in the context of applying logic and computational theorem-proving to parsing and machine

[12] John Cocke developed the algorithm in its earliest form in 1960 (Kay 2000), but did not publish it; it was first reported by Hays (1962). Younger's work appeared in 1967. Kasami's work was first reported in a 1966 technical report (cited by Younger 1967 in a last-minute footnote) and was published as Kasami and Torii (1969).

translation (see also Kay 1992, 2005, Colmerauer and Roussel 1993). Unification and theorem-proving were the basis of the logic-programming language Prolog, in which 'definite clause grammars' for natural languages could conveniently be written, with the programming language itself providing much of the operation of parsing; parsing thus was viewed as deduction (Pereira and Warren 1980, 1983; Pereira and Shieber 1987).

Most parsing algorithms, including those based on unification, are non-deterministic—i.e. the process must sometimes make a guess at what the next step in its analysis is, and, if subsequently stymied because the guess was wrong, must back up, throw away its work, and try a different guess. But in trying a different guess, the algorithm might recreate some of the work it did for the earlier guess, as parts of the analysis will be the same for both; this is a motivation for chart parsing (see above). A contrasting approach to parsing was inspired by research in psycholinguistics; people analyse a sentence even as it is being heard or read, and rarely if ever change their initial analysis. Mitchell (Mitch) Marcus (1980) developed the idea of deterministic parsing, in which syntactic structures, once built, could not be discarded or modified; and when making a structural decision at a point in a sentence, the parser could look only a certain distance further to the right. The definition of distance in structural terms modelled the fact that short so-called garden-path sentences such as *The horse raced past the barn fell* (Bever 1970) can lead people (and Marcus's parser) into unrecoverable errors of analysis, whereas much longer sentences ordinarily present no problems.

A parser by itself cannot decide which of the competing syntactic analyses of a sentence is the one intended in the context of utterance, even if it has an a priori preference for some structures over others. The syntactic disambiguation problem was taken to be one of semantics and, again, knowledge of the world, often interwoven with the problem of disambiguating polysemous and homonymous words (Hirst 1987).

The interest (and publications and sharing of software) of researchers in automatic syntactic analysis fell disproportionately on the largely language-independent process of parsing itself, rather than the development of computational grammars for the particular languages to be analysed. Published grammars were rarely larger than toy examples, and practical, broad-coverage grammars tended to be guarded as proprietary. One exception is the HPSG-based English Resource Grammar (Copestake and Flickinger 2000), which is freely available.

32.2.5 Discourse and Dialogue

Reflecting the assumption from AI that people would converse with intelligent machines in order to instruct them or to seek information or advice from them, determining the intent of a human interlocutor became an important theme in CL in the mid-1970s. Thus, in this work, a speaker-based view of meaning was taken (Hirst 2007); language understanding was construed as recognizing the goals and plans that underlaid a person's utterances to the machine. For example, a person who asks *Is there a coffee shop around*

here? probably has the goal of buying coffee; a *yes/no* answer or the directions to a coffee shop known to be closed would not be appropriate responses. This work drew on research in the philosophy of language that viewed utterances as actions in the world (Austin 1962, Searle 1969) (see §26.8–9) and the meaning of an utterance as the speaker's intention in uttering it (Grice 1957). More generally, studies of linguistic pragmatics, including indirect speech acts (Searle 1975) (see §26.9.1), presuppositions (Wilson 1975) (see Chapter 23 and §26.4.2 above), and implicatures and maxims of conversation (Grice 1975) (§26.5), became very influential in computational linguistics.

To recognize an interlocutor's goals and plans from their utterances in order to construct an appropriate response, even though the utterances might be overly terse, or contain misconceptions, or only indirectly indicate what is wanted, requires reasoning about the interlocutor's beliefs, including their beliefs about one's own beliefs. James Allen, Philip Cohen, and Raymond Perrault developed systems that used plan recognition techniques to reason from first principles about the underlying intent of an utterance (Allen and Perrault 1980), and to construct appropriate utterances seeking action from others (Cohen and Perrault 1979), given initial beliefs about the interlocutor's beliefs. These ideas were further developed by Sandra (Sandee) Carberry (1990), who explicated the need for cooperative dialogue systems to construct and maintain, as the conversation progressed, a model of the user—the system's interlocutor—that includes not only inferences about their beliefs and intentions but anything else that might be relevant to the system's understanding and response.

The resolution of anaphora and of definite reference in general was another problem of conversation and discourse (Hirst 1981). Algorithms were developed to determine what elements of a text or conversation were available for reference at any particular point (Hobbs 1978, Sidner 1978). The problem of definite reference to implicitly invoked entities—for example, referring to *the wheels* after mention of *a car*—again pointed to the need for knowledge of the world. Research on coreference resolution converged with that on understanding intentions in discourse in an important paper by Barbara Grosz and Candace (Candy) Sidner (1986) that presented a general computational model of discourse segmentation taking into account both the intentions of the speaker and the focus or attentional state that the speaker of a segment associates with that segment. Within this model, the widely influential centering theory (Grosz et al. 1995)[13] developed rules that sought to relate focus of attention to the speaker's choice of referring expression and the consequent degree of coherence of the discourse; under the assumption that the speaker seeks to maximize coherence, these rules constrain the choice of antecedents in anaphor resolution.

Nonetheless, all this research was based on typewritten input from the human. While it was assumed that in due course, people would speak to computers, not type text to them, apart from an early project that brought computational linguists together with researchers working in signal processing and related fields (Woods et al. 1976; Wolf and Woods 1977),

[13] Although it was not published until 1995, earlier manuscript versions of this paper had circulated since 1983 and had been widely cited.

speech recognition itself remained a largely separate research field dominated by engineers and uninformed by linguistics and CL (see also §32.3.2 below).

32.2.6 Machine Translation

Despite the ALPAC report (§32.1 above), some work on machine translation did continue, mostly outside the United States. The TAUM-MÉTÉO system for the translation of weather forecasts from English into French (Chandioux and Guéraud 1981) entered daily use at the Canadian Meteorological Centre in 1977, translating millions of words each year (Thouin 1982, Hutchins 1986). In the European Community (as it then was), all twelve member states sponsored EUROTRA, a large and ambitious MT project for all nine EC languages that was firmly based on syntactic and semantic theory (Steiner et al. 1988).

32.3 Empirical Computational Linguistics and Natural Language Processing

In the early 1990s computational linguistics underwent a revolution, becoming far more quantitative and data-oriented, and separating itself from artificial intelligence. While there are obvious parallels in the rapid development of corpus linguistics in the 1990s (see Chapter 33 below)—both, after all, involve computers and corpora—computational linguistics and corpus linguistics remained largely separate research fields with different motivations and methods of analysis.

32.3.1 The Rise of Empiricism

When it came, the transformation of computational linguistics from a rationalist enterprise inspired by artificial intelligence and armchair linguistics to an empiricist undertaking based on corpora and statistics was extremely rapid. In just a few years, the early 1990s, the character of research papers in the field changed radically: the number of papers on symbolic topics such as semantics and plan-based dialogue declined greatly, while those on empirical topics such as text classification and statistically based parsing (§32.3.2 below) showed a corresponding increase. This is depicted in Fig. 32.1, reproduced from the work of Hall et al. (2008), who documented these changes.[14] Of course, symbolic and deep formal approaches did not disappear

[14] Interestingly, Hall et al. used the methods of empirical computational linguistics to study the history of empirical computational linguistics. Because of a commitment to open-access publishing by the Association for Computational Linguistics, a corpus of all its conference and journal publications

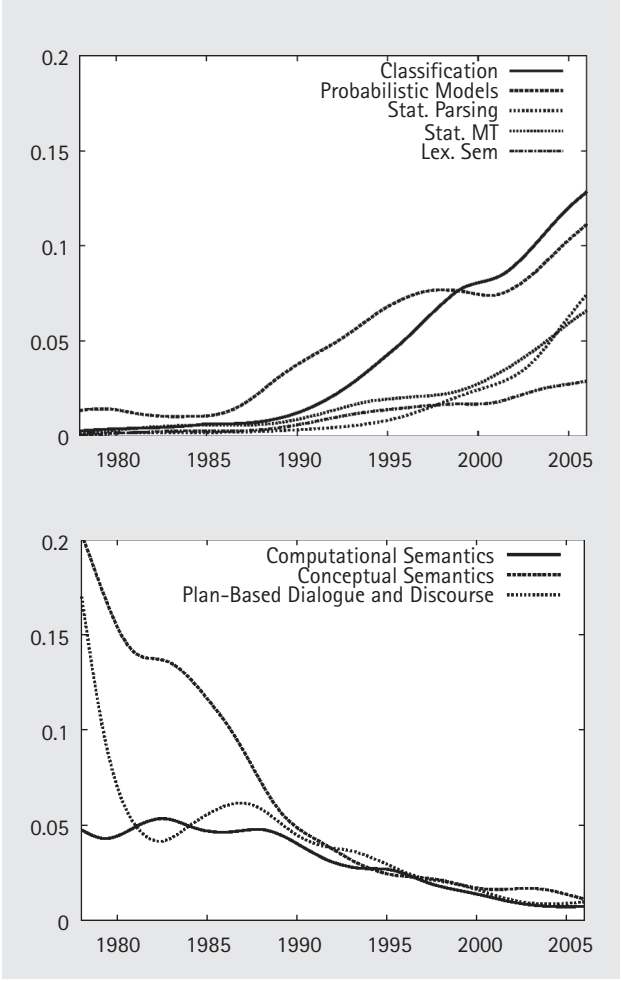

FIG. 32.1 Topics that became more prevalent (top panel) and less prevalent (bottom panel) in research papers in computational linguistics with the rise of empiricism. The *y*-axis is the probability that any given paper in a particular year included a particular topic. (From Hall et al. 2008)

overnight. Many researchers remained sceptical of empirical methods, and even those who saw their value did not immediately abandon their existing research. Nonetheless, as Fig. 32.1 shows, the displacement of the older approaches was quite rapid.

This new view of computational linguistics took language as a stochastic or probabilistic system. In this view, language understanding and its component tasks should not be viewed as processes deemed to have succeeded if they have returned the official

from 1965 onwards, 14,000 papers in total, was freely available for analysis. Treating the papers as bags of words, Hall et al. used topic models and latent Dirichlet allocation (Blei et al. 2003) (see §32.3.2) to cluster and classify them.

correct answer and to have failed if they have done otherwise; nor is a system that fails on some particular input refuted as if it were a mathematical theorem refuted by a counterexample. Rather, a system should use probabilistic methods to return what it believes to be the most likely answer; the more often a system is correct, the better it is, but perfection should not be expected and an imperfect system may still be very useful in practice (§32.3.2 below). Thus, the view is essentially Firthian (see this volume §§20.3.2, and 33.8) and explicitly anti-Chomskyan, repudiating Chomsky's famous (1957a) arguments against statistical models of language (see §33.3).[15] It has its roots in information-theoretic views of language and noisy-channel models of communication that had influenced Weaver (§32.1 above) but had themselves fallen out of favor for many years as a result of Chomsky's anti-statistical arguments (Liberman 2010). Its rise was the result of influential research within CL, advances in related areas of research and in computing, the increasing availability of online text, and other factors to be discussed below.

32.3.1.1 *Influential Research Within CL*

Perhaps the single most influential early piece of research in empirical CL was the IBM statistical model of machine translation. Frederick Jelinek and his research group at IBM Thomas J. Watson Research Center had used an information-theoretic noisy-channel model in the 1970s–80s to develop systems for speech recognition that were very successful for their day, and in 1987 began to apply the same methods to machine translation (Jelinek 2009). The result was a model of statistical machine translation based on correspondences derived from a parallel 'bitext'—a corpus that contains both a text and its translation into another language (Brown et al. 1990). The demonstration system was trained on 40,000 sentence pairs of English and the corresponding French (800,000 words of each language) from the bilingual proceedings of the Canadian Parliament (Hansard). The translations resulting from this method were at least as good as those of the contemporaneous systems that used syntax and semantics, and this was unsettling to many researchers who had worked long and hard on those systems. It seemed both impossible and unfair that translation could be accomplished by a system that had no knowledge of language[16] and had been constructed by researchers with no knowledge of linguistics.[17] The relative merits of rationalist and

[15] Nonetheless, Penn (2012) argues that Chomsky's views on syntax and language played a very significant role in laying the philosophical foundations of statistical parsing (see §32.3.2), which 'has obediently received a view of grammar that has been very carefully circumscribed to exclude any built-in symbolic apparatus for disambiguation, as well as most of the world knowledge or reasoning that might be useful for doing so.'

[16] In fact, the system had plenty of knowledge of language: it knew about correspondences between French and English and their probabilities, and latent within that was some knowledge of syntax and semantics. But its critics construed the absence of overt syntactic and semantic knowledge as no knowledge of language at all.

[17] The members of the IBM research team flaunted their ignorance of linguistics as if to taunt other researchers. Fred Jelinek is famously quoted as saying: 'Every time I fire a linguist from our project, the performance of our system gets better' (or words to that effect; Jurafsky and Martin 2009: 83). At the 1992

empiricist approaches were furiously debated at the 1992 TMI machine translation conference in Montreal.[18]

Kenneth Church and his colleagues at Bell Laboratories (later AT&T Research), especially William (Bill) Gale, were also early advocates for empirical CL. In particular, their papers on statistical analysis of lexical collocations (Church and Hanks 1990, Church et al. 1991) demonstrated the large amount of usable knowledge that can be derived from the word *n*-grams of a sufficiently large corpus. Church and colleagues also published influential work on part-of-speech tagging and on matching up ('alignment') of words and sentences in bitexts for training statistical machine translation systems (Church 1988, Gale and Church 1993).

A third influential researcher was Eugene Charniak, who had been one of the pioneers of knowledge-based approaches to NLU in the early 1970s (§32.2.3 above) and who was a co-author of two well-known textbooks on artificial intelligence. In 1993 Charniak published the first textbook on statistical methods for natural language processing, noting in the preface that it represented his own personal conversion from an artificial intelligence-based approach that 'is not going anywhere fast' (p. xvii–xviii).[19] The availability of this concise and easy introduction to help researchers master the new methods was certainly a factor in their rapid acceptance. By 1999, Charniak's small book had been superseded by Manning and Schütze's large introductory textbook, which could be used as the basis for a complete university course, and whose very existence indicated that statistical methods were now mainstream.[20]

32.3.1.2 *Advances in Computing*

Three advances in computing were instrumental in the rise of empiricism. The first, almost trivially, was the continuing exponential growth in the power of computers themselves; large corpora could be processed in a reasonable amount of time. The second was the parallel development of ever cheaper, ever denser storage media for files, which made the online storage of large corpora feasible. Moreover, the advent of the CD-ROM in the late 1980s brought a qualitative change, making the distribution of large corpora cheap and easy; suddenly, 650 Mb of data could be easily duplicated,

TMI conference (see n. 18), members of the research team presented a paper (Brown et al. 1992) showing that the results of the statistical model could be greatly improved by adding simple linguistic knowledge (such as basic morphological analysis), which they had gone to the library to read about in a book; they characterized this as a great advance, much to the frustration of other researchers present whose systems already included such elements and who regarded them as obvious.

[18] Formally, the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Montreal, 25–7 June 1992.

[19] In his speech accepting a Lifetime Achievement Award from the Association for Computational Linguistics in June 2011, Charniak said that he now regarded the research of the first half of his career, up to 1992, as not worth looking at.

[20] Nonetheless, the almost simultaneous appearance of another popular textbook, Jurafsky and Martin (2000), which put greater emphasis on symbolic approaches, served to retain these methods in university courses. Some instructors used both books, or divided their offerings into a 'Manning and Schütze course' and a 'Jurafsky and Martin course.'

stored on an extremely cheap medium, and mailed as an ordinary first-class letter, obviating the need for duplicating and shipping heavy reels or cartridges of magnetic tape. The third advance was the increasingly broad availability of the Internet,[21] as it was by then becoming known, permitting files of data of a reasonable size to be transferred directly over the network from machine to machine with the File Transfer Protocol (FTP).

32.3.1.3 *Advances in Related Fields*

Advances in the neighbouring research fields of information retrieval and machine learning were important factors in the development of empirical CL.

Early research in document and information retrieval had little to do with the contemporaneous research in computational linguistics (although a few key researchers, most notably Karen Spärck Jones, were involved in both areas). Rather, early systems for information retrieval were based on Boolean queries against bibliographic fields and keywords that were assigned to documents by human indexers (Liddy 2003). As it became feasible to store complete documents, and hence search the complete text, document representations were developed that sought to eliminate the need for a human indexer and a controlled indexing vocabulary. By far the most influential representation was Gerard Salton's vector-space model (Salton et al. 1975, Dubin 2004), in which a document is represented as a vector in a n-dimensional vector space, where n is the number of words in the vocabulary and the ith component of the vector is a weighted count of the number of times the ith word of the vocabulary appears in the document; the similarity between two documents, or between a document and a retrieval request similarly represented, is then measured as the cosine of the angle between the vectors. The vector-space model of text became the basis for much research in CL in which a text is treated as a bag of words or other features without order or structure and can thus be represented as a vector. Moreover, because many problems in empirical CL can be construed as retrieval problems (some linguistic objects are selected from a larger set and possibly are rank-ordered), two metrics used for evaluation in information retrieval, precision and recall, became commonly used for quantitative evaluation in empirical CL.

Empirical CL drew heavily on methods of machine learning, and became a consumer of research in that field in much the same way that symbolic CL was a consumer of research in knowledge representation (§32.2.3 above). The 1960s–80s had been a fertile period in machine learning, producing a number of methods and algorithms that subsequently became mainstays of CL, such as hidden Markov models, the Viterbi algorithm (Viterbi 1967, Forney 1973; see also Viterbi 2006), the EM (expectation-maximization) algorithm, and Bayesian networks (Mitchell 1997). Nonetheless, CL remained cool towards machine learning methods that were based on neural networks

[21] By 'Internet' here, I mean the Internet itself as a physical network, and not the World Wide Web, implemented by the Hypertext Transfer Protocol (HTTP) on this network, which arose a few years later (see §32.3.1.4) and which is often referred to metonymously as 'the Internet' by lay persons.

and spreading activation, which had some limited uptake in CL in the 1980s but which did not prosper, perhaps because Minsky and Papert's 1969 critique of the related notion of perceptrons (Rosenblatt 1958) led to general scepticism (Olazaran 1996), and perhaps because it was never clear how they could be used for complex linguistic tasks.

32.3.1.4 *Availability of Online Text*

By definition, empirical CL requires data—corpora of text in a machine-readable form. Previously, the largest online corpus was the laboriously built Brown Corpus of Present-Day Edited American English (Francis and Kučera 1964) (see §33.4 below), which at only one million words across many genres, though it had seemed large in 1964, was far too small for most work in empirical CL. However, the development of computerized typography and word processing in the 1980s allowed the relatively easy construction of large online corpora as a side effect; any text typeset this way would persist in a machine-readable form and could be compiled into a corpus, although much work might be needed to regularize the format of the data. (It was from such a source that Jelinek and colleagues obtained the data for their Canadian Hansard corpus.) Optical character recognition was also used for digitizing printed material.

However, it remained difficult in practice for researchers to get their hands on sufficient data; copyright in the text itself and proprietary rights in the compilation were among the barriers. In 1989, the Association for Computational Linguistics began its Data Collection Initiative (ACL/DCI) with the goal of at-cost distribution of corpora for non-profit scientific research, using appropriate licensing agreements to reassure rights holders (Liberman 1989). This in turn led to the establishment in 1992 of the Linguistic Data Consortium (LDC), housed at the University of Pennsylvania and led by Mark Liberman, for the collection and distribution of text and speech corpora, lexicons, and other linguistic resources in many languages.

A particularly influential early corpus obtained by the ACL/DCI was the *Wall Street Journal* corpus donated by Dow Jones Inc., 30 million words published in that newspaper in the years 1987–9, which became part of the LDC's first distribution in 1993. The corpus grew as later years of the newspaper were added. It was (and still is) used extensively in CL research. Parts of the *WSJ* corpus, along with the Brown Corpus, were selected as the textual basis of the Penn Treebank (Marcus et al. 1993) to be semi-automatically annotated with parts of speech and so-called 'skeletal' parse trees; subsequently 50,000 *WSJ* sentences (one million words) received more-detailed hand parsing for the Penn Treebank (Marcus et al. 1994), and this became the primary data for research in statistical parsing. The dominance of the *WSJ* corpus was such that CL was sometimes accused of studying not English but the language of English news reporting and of the *Wall Street Journal* in particular; it was not unusual for a system, especially a parser, that was trained on this corpus to perform well on new text from the *Wall Street Journal* but markedly less well on any other genre of English or even on the *New York Times*. The larger and more balanced (but rather more expensive) British National Corpus (Leech 1993, Burnard 1995, Aston and Burnard 1998; see also §33.5 below) was often used as an alternative.

The development of the World Wide Web from about 1995 onwards was another major boost, as it was quickly realized that the Web was itself a corpus and that corpora of many different genres, from literature to reviews of consumer products, could be easily created from its contents (Kilgarriff and Grefenstette 2003). Moreover, network bandwidth and the size of data storage media increased rapidly, enabling ever larger corpora. By 2005, multi-billion-word corpora were not unusual, and were motivated by the results of Banko and Brill (2001), who showed that, at least in some situations, the choice of algorithm for a system is a less important factor in the quality of its performance than the sheer volume of data on which it is trained.

32.3.1.5 *Quantitative Evaluation and Competitive Shared Tasks*

A backlash against the hand-waving proof-by-demo system evaluations that were common in symbolic CL research (§32.2.3 above) began in the later 1980s with a move towards more rigorous and quantitative evaluations. This was facilitated by the development of shared tasks and competitive evaluations, usually in conjunction with a workshop or conference. In a shared task competition, research groups are given a very specific problem, along with sample data and possibly some other relevant resources. The tasks are typically well-defined instantiations of unsolved problems—for example, finding antecedents of certain kinds of anaphors in certain kinds of texts—that are central to larger problems, so that breakthroughs or major improvements would have a large effect. Each group develops its own methods and software for the task, which is then run on data previously unseen by the competitors; at the subsequent event, quantitative results are released and each competitor discusses their system and its methods. Even after the competition, the datasets involved may continue to be used by researchers as training data and as test data to compare new systems with those of the original competition.

The first shared tasks in CL were associated with the Message Understanding Conference (MUC) series, beginning in 1989;[22] competing teams had to fill pre-defined templates with information extracted from military messages or, later, short news items (Grishman and Sundheim 1996). The Text Retrieval Conference (TREC), focusing on tasks related to document and information retrieval, followed in 1992 (Voorhees and Harman 2005). Participation in the shared tasks of these conferences was a requirement of some US government research-funding programmes (Crystal 1993), and poor performers could have their funding discontinued. However, many subsequent shared tasks and research challenges were not associated with funding but were voluntary and open to any interested researcher. Well-chosen shared tasks have spurred innovation in CL, and their necessary emphasis on quantitative evaluation and comparability of different systems and methods was another factor in the rise of empirical CL.

[22] The Message Understanding Conferences began in 1987, but it was only for the second MUC in 1989, that the essential characteristics of the shared task were developed (Grishman and Sundheim 1996). MUC ran until 1998.

32.3.2 The Development of Empiricism and Cross-disciplinary Influences

While some early empirical research in CL, as noted above, made a virtue of the complete absence of linguistic theory, this was not true in general. Rather, what was rejected was the idea that deep analysis, especially formal semantics, and complex representations of knowledge of the world are a necessary or desirable part of the solution of any practical problem in language processing. The new view was that many useful things can be done with surface form plus some linguistic analysis that could be shallow or deeper, depending on the problem. The syntactic analysis of a sentence that might be necessary could range from none at all, to part-of-speech tagging (e.g. Brill 1995, Ratnaparkhi 1996), to chunking into phrases ('partial parsing') (Abney 1991), to a complete parse tree. And the analysis itself could be carried out by a data-driven statistical process; great advances were made in the 1990s on probabilistic part-of-speech tagging and on parsing with probabilistic grammars derived from treebanks (§32.3.1.4 above)—in particular, lexicalized probabilistic grammars that thereby encode lexically conditioned structural preferences (Collins 1997, 2003). Interest in parsing with dependency grammars also increased markedly as a result of the development of a constraint-based approach to the problem (Maruyama 1990) and probabilistic parsing models (Eisner 1996) for these grammars (Kudo and Matsumoto 2000, Kübler et al. 2009). The data-driven constituent-based parsers of English by Charniak (2000) and Collins (1997, 2003) and the data-driven dependency-based MaltParser by Nivre et al. (2007), made available free to other researchers, were widely used.

Nor did interest in semantics and world-knowledge disappear. Rather, the new focus was on lexical semantics and the knowledge implicit in word meaning and word distribution. A particularly important development was the creation of WordNet (Fellbaum 1998) by George Miller and his colleagues at Princeton University. WordNet is a database of words and the lexical relations between them that gives primacy to word sense rather than word form; different senses of a word form are listed separately, and each separate sense of the form is grouped together in a so-called synset with synonymous senses of other word forms. Because synsets are connected by relations such as hyponymy, meronymy, and antonymy, WordNet encodes some taxonomic knowledge of the world. WordNets have now been developed or are under development for more than 50 languages.[23]

At the semantic levels above the lexical, considerable attention was given to the development of methods for semantic role labelling, i.e. determining the relationships asserted between the entities mentioned in a sentence by determining their thematic or case roles (Palmer et al. 2010). Learning by reading was revived as a topic of minor interest (Hovy 2006, Forbus et al. 2007), but it was also recognized that even if full text understanding could not be achieved, much knowledge could be systematically gleaned from texts for

[23] The Global Wordnet Association tabulates wordnets that comply with the standard design; see www.globalwordnet.org.

inclusion in knowledge-based resources and for use in tasks such as question answering. For example, from the text fragment *dachshunds and other small dogs* it can be learned that a dachshund is a small dog (Hearst 1998); from *bringing back her washed clothes* it can be learned that clothes may be washed (Schubert and Tong 2003).

Advances in machine learning continued to influence CL. In particular, support-vector machines[24] (SVMs), introduced by Cortes and Vapnik (1995), were rapidly adopted as a preferred method for classifying texts or other linguistic objects (Joachims 2002). Many applications of NLP can usefully be viewed as problems of text classification. For example, news articles gathered from a variety of sources may be grouped by each distinct topic or event; consumers' online reviews of a book or film may be classified as favourable or unfavourable, a task known as sentiment analysis (Pang and Lee 2008). Similarly, word sense disambiguation may be viewed as the classification of each occurrence of a word according to its set of possible senses. Two complementary developments in machine learning developed the idea of finding in a document the topics that are 'latent' in its words. Latent semantic analysis (Deerwester et al. 1990, Manning et al. 2008) reduces the number of dimensions in the vector-space representation of a set of documents, thereby implicitly bringing together documents on related topics, even if they happen to use different (synonymous or closely related) words for some concepts, while separating those on different topics that happen to use the same (homonymous) words. This idea was developed further in topic models and latent Dirichlet allocation (Blei et al. 2003, Blei 2012), in which a text is modelled as a Dirichlet distribution over a choice of topics and a topic is modelled as a distribution over a choice of words.

Despite its influences from research in information retrieval and in speech recognition, computational linguistics remained largely distinct from these other fields. Some US funding agencies recognized the overlap of interests and the potential for greater synthesis and practical applications in a convergence of the fields, and strongly pushed for this from their grantees in the three fields, holding workshops, first in 1993–4 and then again in 2001, to bring the fields together under the name Human Language Technologies (HLT). From 2003, this name was adopted as a subtitle for the North American conferences of the Association for Computational Linguistics, but the move had only limited success; while a greater number of researchers now do work that overlaps more than one of the fields, they still remain largely distinct research communities.

32.3.3 Applied Natural Language Processing

As computational linguistics further developed its empirical orientation, it became more concerned with practical and commercial applications of natural language

[24] Support-vector machines are not physical hardware but machines in the abstract sense. They are a class of algorithms for classifying points in a multidimensional vector space—e.g. documents represented by counts of words or other features—into two or more categories by deriving from training data the hyperplane that separates the classes with the least amount of error.

processing, and for good reason—component methods and resources had become mature and robust enough that useful applications were now in view, including many based on text classification, such as sentiment analysis (§32.3.2 above). In addition, funding agencies, especially in the US, emphasized the development of methods for finding, synthesizing, and succinctly presenting information from large document collections, including the World Wide Web itself. This included question-answering (see §32.4 below)—finding the right few words among billions, or summarizing the text of one or many documents (Maybury 2004).

The flagship application, however, remained automatic language translation. Machine translation systems had by this time been available for a number of years, often styled as assistants to professional translators rather than as end-user products. But in 1997 they became broadly available to the public over the Web when SYSTRAN's translator became the basis of AltaVista's Babel Fish service (Yang and Lange 1998), and they are now routinely sold as end-user software. SYSTRAN's system has evolved from being based purely on linguistic rules and dictionaries (Yang and Lange 1998) to a hybrid system incorporating statistical methods (Dugast et al. 2008). In 2006–7, Google began offering a competing service, Google Translate, based on purely statistical methods (Halevy et al. 2009, Koehn 2010).[25] Although the quality of their translations was often poor, these systems have improved demonstrably over time, and their extensive use vindicates the opinions of the critics of ALPAC who asserted that a low-quality translation may nonetheless have high utility.

32.4 Computational Linguistics Today

Kenneth Church (2004, 2012) has suggested that research in CL oscillates in twenty-year cycles between rationalism and empiricism: empiricist in the 1950s–60s, rationalist in the 1970s–80s, and empiricist again in the 1990s–2000s. This suggests that the field is due for a revival of rationalism in the 2010s as the limitations of present empirical methods are felt.[26] Certainly, a frequent criticism of the field in the last decade is that there is a lack of innovation—that too many papers merely report minor incremental work based on the innovations of the 1990s, following a sterile paradigm of annotate–learn–evaluate, and that the last decade of progress has been due to larger corpora and

[25] In an echo of Frederick Jelinek's comment about the deleterious effect of linguists on machine translation (see n. 17), Peter Norvig, Google's director of research, claimed in 2007 that Google's experiments had found that their translation system performed less well when explicit syntactic knowledge was incorporated. Nonetheless, subsequent research at Google (e.g. Xu et al. 2009, DeNero and Uszkoreit 2011) has investigated the use of syntactic components, which may themselves be derived from bitexts, within statistical MT.

[26] Church (2004) hypothesizes that the 20-year components of the cycle are due to students rebelling against their teachers. Church urges that students be taught not only current methods in CL but those of the previous cycle so that when they rebel, they do not simply repeat the mistakes of their teachers' teachers.

more powerful computers more than to any new conceptual breakthroughs. CL is also criticized for having abandoned its connections to theoretical linguistics and psychology (Reiter 2007, Spärck Jones 2007, Wintner 2009, Krahmer 2010), and for focusing too much on NLP and its applications rather than advancing the development of computational models of language and human language processing.

But an indication of the achievements of CL and NLP came in February 2011 when the Watson question-answering system, developed by a team at the IBM T. J. Watson Research Center, competed against two human champions on the American television game show *Jeopardy!*[27] and won (Kroeker 2011). Watson was based on IBM's DeepQA architecture for question-answering, which combines many different algorithms for analysing questions, finding possible answers in its knowledge sources, and selecting and computing its degree of confidence in a final answer (Ferrucci et al. 2010). When playing *Jeopardy!*, Watson was not connected to the Internet and relied solely on a large set of pre-analysed reference sources, such as encyclopedias, newswire articles, and additional material that it had earlier extracted from the Web (Ferrucci et al. 2010, Kroeker 2011).

The technologies developed for question-answering systems such as DeepQA have clear commercial applications in many professional and social domains for the kind of information needs which cannot be satisfied just by keyword searches that return lists of whole documents or passages of text through which the user then has to sift. A fortiori, Hirst (2008) has suggested that a likely direction for these applications will be a confluence of several streams of research in NLP and information retrieval to build systems that will construct a complete answer to a user's question by selecting and summarizing relevant information from many sources, and that, moreover, do so by 'considering each document or passage from the point of view of the user's question (and anything else known about the user)' (p. 7); that is, they act as an agent for the user, in effect taking a reader-based view of the meaning of the text. Some other applications, conversely, will take a writer-based or intention-based view of the text; this will include not just sentiment analysis (§32.3.2 above) but intelligence-gathering, in the broadest sense—systems that aim to determine the opinions, beliefs, and plans of the writer. Machine translation, by definition, attempts to preserve a writer's intent across languages, and as new research on statistical methods in MT supplements them with a semantic sensibility, it too may take on a more explicitly writer-based view of text and meaning (Hirst 2008).[28]

[27] In *Jeopardy!*, contestants are given an answer and its context, and must supply a corresponding question. For example, given the answer *Washington* and the context *Leaders of history*, a correct response could be *Who was the first American president?*, whereas in the context *National capitals*, a correct response could be *What is the capital of the United States?* The problems are generally more challenging, both in language and in the knowledge they require, than this example suggests; many involve wordplay (e.g. requiring a response that rhymes), and many implicitly require the solution of two separate problems (such as *The most northerly country with which the US does not have diplomatic relations*) (Ferrucci et al. 2010).

[28] For comments on, suggestions for, and discussion of this chapter, I am grateful to Ken Church, Joakim Nivre, Gerald Penn, and Nadia Talent. This work was supported financially by the Natural Sciences and Engineering Research Council of Canada.