Technology and Linguistics Research

Antonio Zampolli
Instituto di Linguistica Computationale, Pisa

1 Introductory remarks

Whereas in most of the humanities the computer is considered essentially as a powerful tool which can assist researchers in their traditional disciplinary activities, in linguistics the computational approach has given rise to a new discipline, computational linguistics. This seems to me the single major specific impact of the computer in the field of linguistics.

2 The formation of computational linguistics and its relationship to literary and linguistic computing

When[5] the use of electronic data processing techniques began to be directed to linguistics data at the end of the 1940s, two main lines of research were activated quite independently:

- Machine Translation (MT), and
- Lexical Text Analysis (LTA: production of indices, concordances, frequency counts, etc.).

While MT was promoted mainly in 'hard-science' departments, LTA was developed mainly in humanities departments. For this reason there was very little contact between the two.

Although at the beginning of the 1960s, there was some recognition of a possible reciprocal interest in topics such as text encoding systems for different alphabets, frequency-counts of linguistic elements in large corpora, and automated dictionaries, real cooperation was very rare, if

---

5. This paragraph is extracted, in part, from Calzolari and Zampolli, 1991.

not totally absent. After 1966 the two fields diverged still further. The year 1966 was particularly important for both lines of research, but for opposing reasons. The Prague International Conference 'Les Machines dans la Linguistique' ratified the international acceptance of LTA as an autonomous disciplinary field, and its extension to a broader area, which included new dimensions of processing (phonology, historical linguistics, dialectology, etc.), called Literary and Linguistic Computing (LLC).

Around the same period, the use of computers spread to other humanities disciplines. Joseph Raben founded the journal *Computers in the Humanities* in 1966. The computational processing of large texts has characterized from the very beginning most humanistic computer-assisted projects, so that some researchers recognize a subfield of 'computers in the humanities' (text processing for the humanities: TPH), which utilizes the traditional tools of LLC (indices, concordances, textual database access, etc.) for retrieving and analyzing factual information referred to in the text.

In the very same years in which LLC and TPH gained ground, the well-known ALPAC report (1966) brought to a sudden stop the majority of MT projects in the world, and marked the beginning of the so-called 'dark ages' of MT. Following, *de facto*, the recommendations of the ALPAC report, basic research on natural language processing slowly occupied the area characterized so far by MT activities, and emerged as a new disciplinary activity, computational linguistics (CL).

In spite of ALPAC recommendations for research on large-scale grammars, dictionaries, and corpora, CL focused mainly on the development of methods for the utilization of formal linguistic models in the analysis and generation of isolated sentences, in an almost exclusively monolingual framework, and at the grammatical level. It almost completely neglected the development of lexica, which were effectively restricted to small toy-lexicons of a few dozen words. A distorted (I believe) interpretation of the Chomskyan paradigm led to an almost complete lack of interest in corpora and quantitative data, which were attracting much attention in the LLC area partly due to projects for national historical dictionaries and for frequency dictionaries.

On the other hand, LLC and TPH also delayed in taking advantage of the know-how, methodology, and tools produced from the very beginning by MT in the field of automatic lexica. Not only had MT developed research on specialized hardware, storage, access techniques, inflectional and derivational morphological analysis, but certain projects had

already begun the collection of large sets of monolingual and bilingual lexical and terminological data. Very few exceptions can be reported in the LLC field, all primarily motivated by attempts to automate the lemmatization of texts for the production of lemmatized indices and concordances. The first experiments are related to Latin (CAAL, Gallarate, and LASLA, Liège).

For several years practically no relationship existed between LLC/TPH and CL. As local organizer of the 1973 Pisa COLING, I endeavoured to include in the call for papers, and to promote in the Conference, sections explicitly dedicated to topics which could delineate the areas of common interest. The attempt was successful in terms of joint participation, and it was probably not just by chance that Joan Smith presented there, at an international level, the newly founded ALLC (Smith, 1973). But in the 1970s a (so to speak) 'purist' approach characterized the general reflections of CL, which was searching for a definitional and a disciplinary identity, focusing on problems of computation and on the nature of the algorithmic procedures, rather than on the nature of the results and on linguistic, in particular textual, data. The variety of points of view is exemplified in the *Foreword* by Karlgren, and in the *Introduction* by Zampolli, to the *Proceedings* of COLING 1973 (Zampolli and Calzolari, 1973).

The development of CL, in the following years, was influenced by the interest in Natural Language Processing (NLP) shown by large sectors of Artificial Intelligence. Many efforts were directed towards the study of methods and tools for prototypes performing a 'deep understanding' of natural language, necessarily limited to restricted linguistic fragments and to 'miniature' pragmatic subdomains, thus enlarging the gap between CL and LLC/TPH activities.

In the LLC framework, the attention of a large part of the research community was captured by the new technological developments, and efforts were directed towards mastering new hardware and software facilities: the increasing variety of rich sets of characters, OCR, photocomposition, large database techniques, personal computers, new storage media, general purpose editors and wordprocessors, standardized concordance packages, etc.

Only in the last two years has a variety of contributing factors started to arouse the reciprocal interest of people working both in CL and LLC/TPH. Increasing contacts and exchanges, joint organization of

conferences or conference sections, and cooperative projects formulated at the international level are all external signs of this process.

This convergence is partly due to the activities of some Institutes, with programmes of research in both fields, and thus naturally operating to construct a bridge and to promote synergies. However, in my opinion, the key fact is that both fields now recognize that an important aspect of their development depends on the capability of processing, at least at some level of linguistic analysis, large quantities of 'real' texts of various types. It seems to me that these trends characterize the actual 'Zeitgeist', and will be examined in further detail in section 5.

3 Computational linguistics: methodologies and results

Since[6] the first machine translation projects of the early 1950s, considerable effort has been devoted to the study and development of methods for the analysis and generation of natural languages. The reasons are both theoretical and practical.

1. At the 'scientific level': testing grammars and rules proposed by linguistic theories; studying and developing formalisms for representing morphological, syntactic, semantic, pragmatic knowledge; accumulating and assessing large formalized descriptions of natural languages; constructing models of the psychological processing of language understanding and of language users; etc.

2. At the 'applicative level': to develop linguistic components for specific systems, oriented to practical industrial and commercial applications, which involve the automatic processing of natural languages (NLP).

3.1 Analysis

Most effort has been directed to processing individual sentences. We can say that the overall objective of sentence analysis, roughly speaking, is to determine what a 'sentence' means. In practice, this usually involves translating the natural language input into a language (e.g. formal logic) which can be 'interpreted' by computer programs. The analysis usually

---

6. This paragraph is extracted, in part, from a survey prepared by the author for the European Science Foundation, in cooperation with S. Hockey.

involves components at the morphological, syntactic, lexical and semantic levels, which can intervene in sequence, or can be activated in turn by a supervisor, or can work in parallel.

3.1.1 Morphological analysis

Morphological analyzers usually try to decompose graphical words into an invariant string (i.e. the part which remains constant in all the inflected forms of a lemma) and inflectional endings. The recognition of endings provides the description of the morphosyntactical properties of the word: part of speech, gender, number, tense, person, etc. The invariant string is normally used as an access key to the computational lexicon of the system, to obtain, from the relevant entry, the associated linguistic information on the syntactic and semantic properties of the word. Some analyzers try to recognize, in addition to the inflectional endings, also the affixes. In this case, the system tries to decompose the word into '(prefix) base (suffix)*, ending', and the access key to the lexicon is the 'base'. In a certain sense, we can say that those systems include a formal representation of the derivational morphology. The two level morphology model deserves special mention (Koskenniemi, 1983).

3.1.2 Computational lexicon

A computational lexicon is a collection of lexical entries, properly structured and stored in a form easily accessible by computer. Each entry consists of two parts:

1. The head: a specific string of characters which is matched to locate a particular lexical entry (see the previous section).

2. Information on the linguistic properties of the lexical entry. Typical examples are: the canonical form (lemma); part of speech; inflectional paradigm; number, form, and selectional restrictions of the arguments; semantic features; semantic relations with other entries (e.g. synonyms) or within a conceptual structure (e.g. taxonomy).

The form of the syntactic and semantic information stored in each entry usually depends on the requirements of the grammar and the semantic analyzer of the specific NLP system associated with the lexicon. When used in association with a concordance program, a computational lexicon can suggest the lemmatization of the words. Of course, if a word is homographic, two or more lemmata are proposed, and usually

the researcher will choose among them, with or without the help of the computer.

3.1.3 Syntactic component

The syntactic analysis component basically performs two functions.

The first function determines the syntactical structure of the input sentence. For example, it identifies the various phrases, their functions (subject, object, predicate), etc. This is more often done by assigning a tree-structure to the input. The analysis is performed by an algorithm which applies a set of formal syntactic rules (the 'grammar') to the sentence, starting from the information provided by the lexical look-up for each word.

The most used grammatical formalism has been the so-called 'Augmented Phrase Structure Grammar'. Several algorithms are now available, and their computational properties are well understood. Several difficult problems are instead associated with parsers based on transformational grammars. The Augmented Transition Network Model (ATN), introduced by Woods in 1970, has been established as one of the standard tools in computational linguistics. It is particularly suited to write small, efficient, syntactically oriented systems. CHART, (based on work by Kay, 1973, and Kaplan, 1973) is a very powerful data structure for parsing, providing a very general framework for representing input, output and intermediate results in all sorts of linguistic data processing.

In recent years, various new syntactic formalisms have emerged, both as a reaction to, and as a continuation of the Chomskyan paradigm, which have in common the utilization, in a principled way, of the unification mechanism: Lexical Functional Grammars (Bresnan, 1982), Generalized Phrase Structure Grammars (Gazdar *et al*, 1985), Head-driven phase structure Grammar (Pollard *et al*, 1987), etc.. It is important to note that these new trends pervade both theoretical and computational linguistics at the same time, and have emerged in contexts where linguists and computational linguists cooperate, almost without distinction. Another trend is that of restricting the role of syntax in favour of the lexical component.

The second function of syntactic analysis is to 'regularize' the syntactic structure. Various types of structures are mapped into a smaller number of simple canonical structures, thus simplifying subsequent processing. Those structures are often intended to represent the functional relationships among the various phrases within a sentence. The

verbal element is usually the focus around which the other phrases revolve.

3.1.4 Semantic component

The major aims of semantic components are:

1. To disambiguate ambiguous syntactic structures
2. To disambiguate homographic or polysemic words
3. To determine the 'general meaning of a sentence'.

The structure produced by the syntactic component is usually mapped into a formal language, which is designed to be unambiguous and to have simple rules for interpretation and inference. In practical systems, the 'meaning' of a sentence is, roughly speaking, what we want the system to do in response to our input: retrieve data, direct a robot, etc. In general, this means translating the natural language input into the formal language of the database retrieval system, of a robot command system, etc. Within the paradigm of formal logic, both propositional and predicate logic are used.

Selectional restrictions are often used for disambiguation. If one of the competing structures, produced by the syntax, violates a selectional restriction constraint, it is rejected as semantically anomalous. For example, a verb can be 'restricted' in the range of items it can accept as subjects, objects, etc. A structure is rejected if the proposed subject is not a member of the accepted class. Preference semantic (e.g. Wilks, 1975) analyzers do not reject structures. They merely 'prefer' some to others. In this way, for example, slightly non-standard subjects and objects can be allowed, so accepting 'non-literal' uses. Ambiguity is resolved by selecting the most preferred readings.

Significant generalizations can be made concerning how noun phrases are semantically related to the verbs and to the adjectives in a sentence. The most influential work for computational approaches has been case grammar (Fillmore, 1968) and its successors and modifications (Bruce, 1975).

Already during the '60s, work in cognitive psychology and artificial intelligence has developed a type of structure known as a semantic network for representing 'meaning'. These networks are graphs made up of nodes, which generally represent a word-meaning, and links between the nodes which reflect the relationships among the nodes (for an overview see Woods, 1975).

Various conceptual analyzers are based on a common semantic representation, called 'conceptual dependency' network (developed originally by R. Schank (1975)). Basically the action, normally referred to by the verb, is represented as a conceptual skeleton, consisting of 'primitive acts' (selected in a short list), which has a fixed number of 'slots' (e.g. the 'actor', the 'object', the 'direction': from-to), to be filled by items appearing in the input sentence. The analyzer tries to 'discover' compatible fillers for each slot.

Disambiguating and interpreting a sentence requires more than just linguistic knowledge. It also involves accessing knowledge of the world, general or domain-specific, and of the specific characteristics of the communicative context (dialogue, etc.). The distinction between linguistic and pragmatic knowledge is known to be very difficult. A great deal of effort is being directed at a cooperation between computational linguistics, artificial intelligence and cognitive science, to study appropriate methods for representing and using knowledge.

A knowledge representation, in its more general sense, is any framework in which information about language and the world can be stored and retrieved. While there is a wide range of knowledge representation formalism, all share common properties. Knowledge representation systems can be thought of as being made up of two distinct parts: the knowledge base (KB), which is the set of data structures that store the information, and the inference engine, which provides a set of operations on the knowledge base (Allen, 1987, p.315 ff.).

KL-ONE is a language which is used extensively for research into problems of understanding natural language and knowledge representation (Schmolze and Brachmann, 1982). It is based on the representation of general concepts (classes of individuals), individual concepts (instantiations of general concepts), and of the relations between concepts. A general concept is part of conceptual taxonomy, in which it inherits the properties of the superordinates. An appropriate formalism allows the description of the 'internal structure' of a concept: e.g. the parts of the concept (which are in turn concepts) and their functional relations within the structure of the concept. Appropriate mechanisms use the knowledge to make inferences.

3.1.5 Discourse component

Much more is known about the processing of individual sentences than about the determination of discourse structure, despite the fact that the

resolution of ambiguities in individual sentences in certain cases (e.g. pronouns) presupposes the ability 'to understand' the connections between sentences.

There have been a number of efforts to define conditions which must be satisfied for a text to be coherent. This work, carried on mainly in the framework of text linguistics, has shown that text analysis will depend critically on the ability to organize relevant world knowledge and make substantial inferencing. Much of the recent research focuses on methods of organizing knowledge for processing new data.

Several approaches use 'frames'. A 'prototype frame' describes, with a set of labelled 'slots', properties, constituents and participants of a class of objects or static situations. When a frame is activated, the parts of the frame specify what kind of information needs to be found in the discourse for a situation to be understable (Minsky, 1975).

'Scripts' are designed to capture the typical knowledge of speakers about a stereotyped sequence of events. A script enables the storage of an outline of a certain type of 'episode', providing the capability of predicting activities, actors, and objects which can be assumed to be present, even if they are not referred to in the input, thus allowing inferences, disambiguations, and connections (Schank, Abelson, 1977).

'Plans' are used to analyze descriptions of a novel sequence of events. A plan consists of a goal, alternative sequences of actions for achieving the goal, and preconditions to apply a given sequence. 'To construct a casual chain for a novel sequence, a means-end analysis must be performed; that is, we must try "to understand" how later events in a text act to further previously stated goals'. (Grishmann, 1986i; Wilensky, 1981).

3.2 Generation

The task of generating language has been undertaken mainly for two kinds of purposes:

1. To test grammars proposed by theoretical linguists. Because of the complex interactions possible, it is desirable to use the computer to verify that a proposed set of rules actually works. The Friedman (1971) Transformational Grammar Tester generated sentences in accordance with a proposed transformational grammar, so that linguists could verify that their grammar did, in fact, generate only grammatical sentences.

2. To generate natural language answers for the user of a given computational system, in cases where a pre-compiled list of fixed messages is not sufficient. User acceptance often requires the generation of complex sentences, and even multi-sentence texts. The generation consists of translating a 'meaning' representation into natural language. The typical process goes from a logical representation, through a deep structure, to the sentence.

It is not generally agreed whether a generation system could be obtained by simply reversing the order of the application of the components and the rules of the analysis. In practice, the problems which arise in generation are often different. The difference between collocations ('idioms of encoding') and fixed phrases ('idioms of decoding') is an example.

Generation has been given less effort than analysis. The specific problems of generation have been under-estimated. One of the reasons is that, whereas an analyzer should be able to accept and recognize many paraphrases of the same information or command from the user, it will suffice to generate only one of these forms. Furthermore, analysis has to deal with the ambiguities present in the texts.

4 Computational linguistics and linguistics

4.1 Grammatical formalisms

Grammar is the field in which interactions have been more intense between linguistics and CL, despite the different purposes of the two disciplines.[7]

Much of the work that led to the development of new grammatical formalisms grew out of an attempt to overcome the difficulties of applying the formalism of transformational grammars in parsing.

This work has been performed from different starting points, but

---

7. Whereas in generative linguistics the emphasis was mainly on the formalisms by which linguistic descriptions can be specified (both in the form of constructive rules which define the range of possible structures and in the form of constraints on the possible allowable structures), and in claims about the nature of language implicit in that formalism, computational linguistics has pursued the goal of specifying a theory to such a level of detail and completeness that computer programs based on it can be written which analyze and generate natural languages.

several approaches make substantial use of *features*, and *functions*, and are influenced by the notion of *cases*.

The creation of *Lexical Functional Grammar* (Bresnan, 1982) is perhaps the best well-known example of a methodological and theoretical new development due to the interaction of linguists and computational linguists.[8] It is an attempt to solve problems that arise in transformational and in ATN grammars by using *additive descriptions*.

Other types of grammars, designed in the CL context, include for example *definite clause grammar* (Colmerauer, 1978), *slot grammar* (McCord, 1980), *functional unification grammar* (Kay, 1985), *head-driven phrase structure grammar* (Pollard and Sag, 1987), etc.

4.2 The generative and the computational paradigms

During the second half of the 1970s, a computational paradigm (the term is used here in the sense of Kuhn, 1970) was proposed for linguistic research, as opposed to, or at least considerably different from the generative paradigm, in those days dominant in theoretical linguistics.

The computational paradigm views the language as a communicative process based on knowledge. The task of the linguist should consist in understanding the organization of this process and the structure of knowledge.

> Our metaphor is that of computation, as we understand it from our experience with *stored program digital computers*. The computer shares with the human mind the ability to manipulate symbols and carry out complex processes that include making decisions on the basis of stored knowledge. Unlike the human mind, the computer's workings are completely open to inspection and study, and we can experiment by building programs and knowledge bases to our specifications. Theoretical concepts of *program* and *data* can form the basis for building precise computational models of mental processing. (Winograd, 1983).

In this approach, CL is nearly identified with theoretical linguistics *tout-court*.

---

8. As an example of the interactions between CL and linguistics, I think it would be interesting to trace the history of CL in the Bay area, which also represents a substantial part of the history of CL.

> Computational linguists are not simply linguists who have found ways of avoiding some of the labour that their trade would otherwise force upon them by consigning it to a machine. There are linguists who have found, or who hope to find, something in the metaphors and theories of computing that reflects in a fundamental way on the human linguistic faculty. (Kay, 1982).

One should look to computers for fundamental insight into human language because computers are the only devices we have to embody a notion of abstract symbolic processing.

> Computers are, as Alan Newell is fond of saying, the only semiotic engines we have. It is to them that we must look for the parts out of which convincing psychological models of linguistic performance will be built. (Kay, 1982).

In this way, the basic goals of linguistics largely coincide with those of psycholinguistics, and the differences reside in the experimental tools. Computational linguistics *and* linguistics are considered as a part of cognitive science.

An intense debate on the differences between the linguistic paradigm, as represented by the generative-transformational school, and the computational linguistic paradigm appeared in a series of papers in *Cognition* in 1976-77 (Dresher and Hornstein, 1976, 1977a, 1977b; Schank and Abelson, 1977; Winograd, 1977).

Both paradigms recognize, as a basic task, the study of the structure of the knowledge processed by an individual who uses a language and, as a basic principle, the hypothesis that this knowledge can be understood as formal rules concerning the structure of symbols (Winograd, 1983i). But major differences exist.

The generative paradigm recognizes two aspects in language: *competence*, an abstract characterization of a speaker's knowledge; and *performance*, the processes that actually determine what a speaker says or how an utterance is understood in a particular context. But, as a matter of fact, the study of performance is practically ignored. The structure of a person's linguistic competence is characterized independently of any process by which it is manifested. 'The performance component is seen as theoretically secondary to the independent specification of competence' (Winograd, 1983). Furthermore, most researchers have adopted the 'autonomy of syntax hypothesis': there would be relatively

independent bodies of phenomena that can be characterized by syntactic rules without considering other aspects of language or thought.

In the computational paradigm, instead, the organization of the processes of comprehension and production play a central role. As a consequence, particular attention is given to the interaction between linguistic and non-linguistic knowledge, and to how linguistic acts fit into a larger context of action and knowledge.

But the debate between the two paradigms has progressively cooled down for several reasons. It has been recognized that the present state of knowledge about natural language processing is so preliminary that the attempt to build a cognitively correct model (i.e. a computational analogue of the human processing mechanism in language production and comprehension) is not feasible.

'Before researchers can begin a project to build such a model, there would have to be simultaneous major advances in both computational linguistics and the experimental techniques used by psycholinguists' (Allen, 1987i). The current goal is, instead, 'a comprehensive, computational theory of language understanding and production that is well-defined and linguistically motivated' (Allen, 1987). Constructing such a computational theory would be a first preparatory step in producing a cognitively correct theory.

Even if for the moment the ambition of constructing cognitive models has been postponed, the production of this computational theory seems to be a very long-term research programme, and CL needs to explore the entire process of language understanding and generation.

A major problem is to overcome the present phase in which isolated attempts lead to the continuous appearance of new, rapidly abandoned proposals. It is necessary to move to a stage in which it will be possible to build on the outcome of previous research.

This seems to me also a necessary condition to draw linguists' attention to the specific problems which CL was the first to raise: for example, the construction of knowledge representation formalisms which can support semantic analyses of sentences; modelling of reasoning processes that account for the way in which context, both textual and extralinguistic, affects the interpretation of sentences; generalizations on the meaning differentiation of polysemous words; discourse structures of various text-types (narrative, dialogue); strategies for semantic interpretation, etc.

In this way we shall perhaps come closer to making it possible, as

forecast by Charles Fillmore in 1977 'for workers in linguistic semantics, cognitive psychology, and artificial intelligence—and maybe even language philosophy—to talk to each other using more or less the same language, and thinking about more or less the same problems'.

A new paradigm, *connectionism*, seems to call for the attention of CL. But it seems to me too early to try to evaluate the possible impact of connectionism, through the mediation of CL, on linguistics, even if the first signs of interest are already appearing, in particular in psycholinguistics.

5 Current trends and possible developments

5.1 CL and language industries

The need to consolidate and progressively build on partial achievements has become even more urgent in the light of the increasing interest of national and international public authorities and private companies for the technical, economic, and social potentialities of the field of the so-called 'language industries' (LI).[9]

This expression, coined on the occasion of a Congress sponsored by the Council of Europe in Tours in February 1986, is used to indicate activities based on computational systems, oriented to practical industrial and commercial applications, which contain, as an essential part, natural language processing components. Examples of typical applications include, within the domain of speech technology: access control, command and control to data entry, driver stations, document creation, telephone enquiries, transaction processing by telephone, database enquiry, environmental control, voice messaging, announcement systems, augmented communication for handicapped people, etc. For written texts, we can quote: spelling checkers, computer-assisted lexicography and terminology, natural language interfaces, machine translation, information retrieval, computer-assisted language learning and teaching, computer-assisted consultation of reference works, translator workstations, etc.

A set of different factors and conditions are requiring today the

---

9. From the early times of machine translation, the perspective applications of linguistic theories have gained financial support for linguistics, in particular in the USA, where the selection of particular lines of research for support has been strongly influenced by intended computer applications.

promotion and development of LI. The key is, in my opinion, the advent of the so-called 'information society'. The global dimension of the economy conceived as a worldwide system, together with the technological development of telecommunications systems, entails a growing information flow. The principal information vehicles are still natural languages, for both the production and the storing tasks. Furthermore, the major part of the information in natural language is nowadays produced directly through computer use, and recorded on machine-readable media: wordprocessors, office automation, electronic mail, photocomposition, databases, etc. Various countries are considering the possibility of progressively recording entire libraries in machine-readable form.

This situation puts an obvious pressure for the creation of new products and services for the various economic activities primarily involved in information handling. The following passage of Makoto Nagao (1989) seems particularly relevant:

> Computers are a fusion with and unification of communications technology at both the hardware and the software levels, and computer systems will undoubtedly enter every corner of future society. When that day arrives, the most important technology will be specifically concerned with neither hardware nor software, but with what I have been advocating for many years: 'information-ware'. In other words, the central problem will regard the ways in which the information signals sent by human beings will be mechanically processed, transmitted, stored, and then recalled in a form which can be interpreted by other human beings. The essence of informationware is therefore how information can be efficiently stored in a computer and activated in response to the various demands of its users. Information can in fact take different forms, including writing, speech and visual images, but objectively, the most accurate means for transmitting and receiving information is writing. For this reason, of the various aspects of informationware, linguistic information and its processing technique will be the primary technology at the heart of the information society. Such technology might be called 'language engineering', and the industry which it will span will be the 'language industry'.

A central aspect of the language industry is multilingualism. Only an 'elite' minority in the world can operate today in a foreign language,

without sacrificing its performance (Perschke, 1988). Furthermore, the conservation of national languages, a principle adopted from the beginning, for example, by the EC, is an important condition for the preservation of national cultural identities.

The need for monolingual and multilingual natural language processing systems, to be used in products for information handling in the LI framework, is uncontroversial. Some studies are carried out in order to narrow down and focus the most urgent tasks and targets, identifying the principal sectors of activities and their economic dimension.

It seems urgent to evaluate the present state of the art in linguistic research and engineering, and the possibilities of large-scale development, in particular:—which products can be created on the basis of existing technologies;—which applications can be envisaged at short and medium terms;—which are the priority areas and tasks for linguistic basic and applied research;—which can be an appropriate research and development strategy;—by which measures, at the organizational level, the public authorities and professional scientific associations can stimulate progress in the field.

In this framework, one of the priority needs, recognized by several researchers in various countries, is a description of natural language, in a form which is suitable for computer use, performed as far as possible exhaustively, at least for those linguistic aspects which can be treated at the present state-of-the-art linguistics and natural language processing. Such extended descriptions are considered the bases for the construction of 'robust' components capable of dealing with the various types of large real texts which are the typical objects of a wide range of LI applications already possible or foreseeable at short and medium term.

These descriptions concern, first of all, grammars and lexica, and can take the form of repositories of grammatical and lexical knowledge bases. Large corpora of textual material in the form of textual databases are considered essential sources of information.

5.2 The impact of current needs on research priorities and directions

I shall now consider how the need for robust NLP components and for the consolidation of linguistic knowledge description are interrelated, and which directions are likely to be taken in CL and linguistics research, in order to try to satisfy those needs.

5.2.1 Robust NLP components

The recognition of lexical units and syntactic structures is needed by all language industry applications. But the humanistic disciplines will also benefit from robust analyzers, capable of dealing, at some syntactic level, with large quantities and varieties of texts.

Linguistics. Different linguistic schools assign various theoretical status to (spoken and written) texts: results of performance acts; samples of statistical populations; instantiation of 'la parole'; etc.[10]

Linguists typically interact with texts to construct inventories of linguistic units, to examine syntagmatic behaviour, to discover personal, social, temporal, genre variations. Frequency of use in texts is considered to be the result of voluntary and/or unconscious choices within the range of alternatives offered by the linguistic system, and their variations are connected in various ways to research on performance mechanisms, stylistic habits, and sociolinguistic trends (Halliday, 1990). In compiling reference works (lexica, grammars), linguists collect evidence and examples.

The interaction with texts is obviously more useful if the access keys to the texts include not only strings of textual characters (occurrence and co-occurrence of graphical forms or parts of forms), but also the formal representation of units, structures, and relations, categories identified at various linguistic levels. Until now, computational research on textual corpora has generally been performed only on graphical forms. This is due to the considerable time and high cost required to manually perform the linguistic analysis and to encode its results. The (at least partial) automatization of the analysis operations is a necessary condition if we wish to exploit adequately the growing wealth of textual data which is progressively available in machine-readable form. The structures recognized by the parser can be either 'annotated' (i.e. explicitly represented) in the texts, for subsequent retrieval and access, or directly 'computed' each time by the parser, in performance of specific requests.[11]

---

10. See Zampolli, 1975.
11. A debate is going on at present on the relative merits of the two approaches. See the discussions at the January 1990 SALT Club Workshop on Corpus Linguistics in Oxford.

Text-oriented disciplines (philosophy, stylistics, literary research, etc.). These disciplines will also benefit from the possibility of retrieving, in the texts, explicitly represented linguistic units, possibly in connection with conceptual units, structures, and relations. So-called 'content-analysis' has, from the very beginning, associated information derived from dictionary look-up with words in the texts.

Textual database access systems, produced in CL, now make it possible to search texts for the occurrence or co-occurrence of families of semantically and/or conceptually related 'terms', interactively defined by the researcher, and considered as 'indicators' of themes, motives and, in general, compositive modules. The researcher can also invoke the assistance of structured knowledge sources such as, for example, lexical knowledge bases, in which semantic/conceptual relations among lexical units are explicitly represented: conceptual taxonomies, synonyms, antonyms, etc.[12]

Humanistic disciplines which consider texts as sources of factual information. Linguistic tools can enhance the capabilities of information retrieval systems on large quantities of full texts. Historical and legal researches are obvious examples of disciplines which require information retrieval systems aimed at identifying extralinguistic entities, and their relations referred to in the text. Very often, given the historical distribution of the source texts, the neutralization of diachronic linguistic variants is requested.[13]

The recognition of morphological variations, synonyms, paraphrases, anaphoric references, etc., can help to reduce the 'silence' in the retrieval processes. The solution of lexical and structural ambiguities will help to reduce the 'noise'. The 'conceptual' lexicon can function as a 'thesaurus' of the common core language (Bindi and Calzolari, 1990).

5.2.2 Features of robust components

The use of computers for analyzing texts could be very useful, and even permit new types of research and applications, both in the humanities and in the language industries, even if:

---

12. See as an example the DBT full text database system developed at the ILC in Pisa (Picchi, 1988), and its interaction with lexical knowledge bases (Calzolari and Picchi, 1988).
13. See the SIL system developed in Pisa by Bozzi and Cappelli (1987), which is now connected to the CLIO system.

1. the analysis has not been carried out completely successfully: for example, if some parts of the sentence are not fully analyzed, or if the parser is unable to reach the final level of the structure, but stops with partial, not fully connected, substructures;
2. the computer is requested, not to perform the analysis in a fully automatic way, but only to assist the human operator in performing his tasks.

We need to develop and test 'robust parsers', i.e. parsers capable of:

1. Processing the variety of phenomena occurring in real texts (e.g. repetitions, ellipses, 'agrammaticalities', abbreviated styles), and using large grammars, covering extended subsets of a language.

   For various historical reasons in the last decades, theoretical linguists have relied mainly on introspection and on native speaker intuition. In an effort to evaluate competing syntactic theories, their work has focused on the theoretical properties of their models, concentrating on the explanation of peculiar linguistic phenomena. As a consequence, a large amount of CL research in practice tends to revolve around little toy subsets of artificially constructed linguistic forms. Only a few hope that such systems may be expanded and linked together until they cover the entire language. The grossly unrepresentative nature of such examples is evident. Therefore, their systems fail as soon as they are exposed to genuinely unselected, authentic input. Only the analysis of corpora, constructed in such a way as to represent a realistic variety of text-types, and pragmatic and communicative contexts, can give appropriate insight into the real concrete usages of language, which often elude the attention of theoretical linguists (Garside, Leech and Sampson, 1987).

2. Continuing to work even if they do not reach the intended level of analysis, providing the results of eventual lower level analysis, and presenting unresolved ambiguity in an economical way. Alternatively, they can call for human assistance.

3. Making use also of statistical knowledge, derived from corpus analysis. When natural language is used in specific domains or communicative contexts (sublanguages: e.g. maintenance manuals, weather reports, technical articles for a specific field), it may be restricted in lexical, syntactic, semantic, and discourse properties.

Sometimes it includes peculiar features absent in the general language. In particular, semantic constraints can be enumerated in more detail, and are more strictly respected in the textual use, so that we can expect a significant contribution to the solution of syntactic and lexical ambiguities. It has also been shown that the frequency distribution, in the texts, of specific linguistic units can be related in a characteristic way to specific sublanguages or text-types.

Some of the recent NLP systems, which are most successful in terms of concrete applications for LI, rely strongly on the use of probabilities which are established by observing the frequencies in language corpora. As an example, we can quote, in particular, text to speech, speech recognition, optical character recognition, spelling checkers, linguistic critiquing and, more recently, practically-oriented speech-connected machine-translation prototypes.

4. Having access to large computational lexica. To analyze real texts a computer must recognize thousands of words. The majority of NLP systems have so far concentrated their efforts on grammar development. A recent poll has shown that the average lexicon in NLP projects includes only a few dozen words. Furthermore, each new project usually starts the construction of its lexicon from scratch.

5.3 Reusable linguistic knowledge sources

Robust NLP components thus require that CL creates large grammars, lexica, and textual corpora.

The construction of lexica, grammars, and corpora of adequate size and coverage is a very difficult, expensive, time-consuming task. Therefore, various disciplines (linguistics, CL, AI), are today considering the problem of the re-usability of linguistic knowledge.

5.3.1 Lexica

NLP is based on information about words: what they are, how they sound, how they connect, what they mean.

> There is good evidence for the power of [a] dictionary-intensive approach to NLP. Although clever algorithms are also necessary, the quality of [a] broad-coverage NLP program depends mainly on the number of words that it knows about, and the amount that it knows about each one. (Libermann, 1990, personal communication).

Broad-coverage programs, once they exist, serve as important tools in further research; dictionaries themselves, once constructed, serve as data for other research.

Theorists of many persuasions are converging on one form or another of 'lexicalized grammars', in which most syntactic and semantic information is part of the representation of particular words.

The problem of re-usability in computational lexica and in linguistic resources in general has two complementary aspects:

1. It is important to re-use existing data, in particular traditional dictionaries, which now often exist in machine-readable form for photocomposition. Computer assisted procedures can extract not only information which is explicitly formulated, but also information implicitly embedded in the dictionary. For example, a variety of semantic and conceptual relations can be extracted by the definitions: taxonomy, typical subject, 'set of', 'used for', etc. (Calzolari, 1988).

2. New large lexical databases must be multifunctional, i.e. must be re-usable in a variety of applications, to avoid duplication of effort.

   The investigation of the feasibility of standardization will concern, first of all, a description of lexical units at various levels of linguistic information and representation formalism. Some levels of description (orthography, phonetics, phonology) are obviously less controversial than others.

   Other levels present specific problems.

   At the syntagmatic level, recent work in traditional and computational lexicography, but also, among others, in meaning- text-theory, has emphasized the necessity for a description of collocational possibilities of individual lemmas. Knowledge about possible collocations is still very limited. Besides investigating typologies of collocation phenomena, and to what degree a formalization of collocational description is feasible, it appears urgent to design methods and tools to identify and collect data from textual corpora (Bindi and Calzolari, 1990).

   The most crucial issue for a description of the syntactic properties of lexical entries is to find a representation which is sufficiently abstract to serve as input to different syntactic theories. By theories we do not only intend explicit grammatical frameworks, but also theories implicitly encoded in the program of a parser, generator, etc.

Although the descriptive goals are similar in most cases, differences can be found with respect to:

- the descriptive vocabulary used in different frameworks and the distinctions made;
- the amount of syntactic information that is encoded in the lexicon;
- the purpose which a certain application is to fulfil.

The central problem is to 'investigate in how far compromises or abstractions can be found that will provide at least fundamental syntactic information in the lexicon which can be augmented with theory specific information' (Rohrer, 1990). This problem is obviously directly connected with that of the re-usability of grammatical knowledge (see section 5.3.2).

The current situation of semantic theories constitutes an additional obstacle to the creation of standard semantic descriptions. Here, the very provisional and limited development of the various semantic approaches makes the comparison, the evaluation, and the abstraction process extremely difficult. In a certain sense, we are still at the stage of evaluating whether any approach has so far reached a sufficient degree of maturity for application to the semantic description of a large lexicon. A very promising approach, which, however, demands further research, seems to be that of lexical semantics, whose aim is also to interrelate regularities at the syntactical and semantic level. Lexical semantics (Pustejovsky, 1989) can be considered as an example of a theoretical development promoted by strict cooperation of linguists and computational linguists, in a certain sense similar to the LFG case mentioned above (section 4).[14]

The problem of including in the lexicon domain-specific world knowledge is certainly the most difficult. Only limited actions seem to be foreseeable in the immediate future such as, for example, to evaluate descriptive devices used in existing dictionaries in order to capture linguistic variation according to different pragmatic factors,

---

14. It seems that 'we have reached an interesting turning point in research, where linguistic studies can be informed by computational tools for lexicology as well as an appreciation of the computational complexity of large lexical databases' (Pustejovsky, 1989).

and to consider the role of these descriptions for preferential mechanisms in NL applications. Possibilities of standardizing the description of pragmatic stratification of the technical and scientific vocabulary in machine dictionaries for special purposes must be considered in connection with the inclusion of technology. (Rohrer, 1990).

Another interesting issue to be investigated in the near future is the feasibility of linking different monolingual dictionaries to yield a multilingual dictionary, possibly in the framework of a common underlying conceptual representation (Boguraev *et al.*, 1988).

3. Several types of partners must be involved, to contribute specific know-how and resources, and to represent various kinds of needs: linguists, computational linguists, humanists, industries, publishing houses. With the help of a carefully structured lexical database, it is possible to provide more coherent and informative dictionaries for human users. The consultation of both monolingual and bilingual electronic dictionaries for everyday use (CD-ROM, wordprocessors, etc.) can be greatly enhanced. For example, browsing techniques in a well-structured lexical database allow better access to any point of the dictionary. Appropriately structured dictionary definitions can be semi-automatically processed to generate a thesaurus of the general language. Links between monolingual and bilingual dictionaries allow searches for translation equivalents using families of conceptually interrelated words.[15]

5.3.2 Re-usability of grammatical knowledge: the polytheoretical issue

Most NLP systems are heavily based on grammatical components. Grammatical knowledge is represented in a formal language, which is usually very system-specific. As a consequence, the linguistic knowledge put into a component is not re-usable in other systems. Even the

---

15. Through appropriate semantic procedures, several types of semantic information implicitly embedded in the traditional lexicographical definitions can be extracted from machine-readable dictionaries and structured in a Lexical Knowledge Base (LKB). In this way, a specific word can be connected to other words through a network of various semantic-conceptual relations: synonymic, antinomic, taxonomic, etc. An LKB can be associated to a textual database (TDB), to increase the retrieval capabilities of the TDB user. Usually, the user of a TDB can search in the texts the occurrences (or co-occurrences) of word forms explicitly specified on the keyboard. If an LKB is available, the system can search not only the word form specified by the user, but also all the word forms connected to it in the LKB, both through morphological (inflectional variants) and conceptual-semantic links.

same project cannot easily modify or replace the formal representation language, without the risk of losing the results of an often huge bulk of work. A new project, adopting a different formalism, cannot re-use the grammatical knowledge embodied in another project, and the grammar rule writers must start from scratch. As a side effect, corpora annotated according to a given formalism cannot be easily exploited by other researchers. Following the example of very recent efforts in the field of lexical knowledge bases, the problem of the re-usability of grammatical knowledge is about to be faced.

One approach is to construct a grammatical knowledge representation that in some sense is more abstract than the system-specific representation, and can act as a 'neutral' basis from which various system-specific representations can be derived. The derivation will be fully automatic in the ideal case, but even if it were only interactive, this would already be a great improvement. The knowledge represented would not be entirely lost when a project would replace its system-specific grammar formalism; and there would be one formalism which could be shared, by linguists working in different environments, to communicate about descriptive questions (H. van Riemsdijk and L. des Tombe, 1990, personal communication).

Much intellectual effort will certainly be necessary to:

- compare in detail the various formalisms and grammatical theories;
- define and test a language for the abstract grammatical knowledge representation;
- define the interfaces necessary to semi-automatically generate grammars for some specific applications.

The theoretical relevance and implication of a polytheoretical approach to the description of linguistic knowledge at various levels is very controversial. Some linguists consider the goal of a 'polytheoretical' description unfeasible. Others consider it irrelevant if not counterproductive. Essentially linguistic theories should rely on explicative power. The description of real large subsets of natural languages is considered as a secondary task. On the contrary, other linguists assign, beside a practical value, also a theoretical interest to the polytheoretical approach.

A large, progressively incrementable, structured and formally explicit collection of information for different languages, not committed to a

specific theory, is considered essential for the progress of research both at the monolingual and comparative level, as a source of data to be accounted for by the theories, and as a testbed for their evaluation. Some linguists also believe that the effort will show that most of the differences between linguistic theories could be considered as 'dialectal-social variations', which could be reduced and neutralized in order to allow the linguists to deal with more substantial problems. Ongoing projects (e.g. the Dutch project EUROGRAMMAR, and the ESF 'LANGUAGE TYPOLOGY'), aimed at compiling large encyclopedias on (comparative) grammars of various languages, will find these grammatical knowledge repositories a very useful tool, through which the data can be conveniently stored and accessed.

5.3.3 Textual reference corpora

Carefully constructed, large, written and spoken corpora are essential knowledge sources for:

1. Extensive description of the concrete use of language in real texts.
2. Identification of particular properties of linguistic units (various meanings, collocations, etc: cf. lexicographic practice).
3. Identification and characterization of sublanguages.
4. Studying frequencies and deriving probabilities.

National (British National Corpus, UK) and international (DCI, USA; NERC, Europe) corpora are being promoted. Humanists have a rich tradition, and precious data collections and know-how on corpus collection and processing.

It is necessary for linguistics, CL, AI, humanities, industries, publishing houses, to cooperate in developing:

- Methods to design the composition of corpora
- Standards for text representation and analysis annotation (cf the Text Encoding Initiative)[16]
- Tools for (semi)automatic linguistic analysis of large corpora

---

16. The Text Encoding Initiative, sponsored by the Association for Computers and the Humanities, the Association for Computational Linguistics and the Association for Computational Linguistics is preparing guidelines for the encoding of and the interchange of machine-readable texts using the Standard Generalized Markup Language.

- Methods for identification and characterization of sublanguages
- Adequate statistical models
- Methods for knowledge extraction from corpora
- Methods for collection and handling of spoken corpora
- Tools to reuse existing lexicographic, literary, and humanistic corpora
- Guidelines for dealing with (text) copyright problems

6 Concluding remarks: CL and other humanistic disciplines

Methods which have been developed in CL are not yet very commonly used in humanistic disciplines which use computers for text processing. There are many areas where CL methodology could be applied more, and a need for convergence is emerging between areas which, for various reasons, had very little contact in the past.

As examples of the consequence of the lack of communication, consider:

1. The know-how developed by humanistic disciplines in text collection, text representation, corpus linguistics, quantitative linguistics, text processing, style and sublanguages, lexicographic analysis, has been practically ignored by CL.

2. The majority of humanistic computing applications, frequency counts, concordance production, interactive text browsing, pattern recognition, information retrieval, etc., usually operate only on graphical (strings of textual characters), not on linguistic or conceptual units.

Researchers in the various lines must recognize that there is a need to exchange data and know-how, and cooperate to develop:

- Large reusable linguistic knowledge repositories (textual corpora, lexical databases, grammars)
- Robust taggers and analyzers
- Standards for linguistic data
- Specialized workstations, including intelligent browsing tools, for access to texts and dictionaries

- Computer-assisted procedures for expanding and integrating, through the interaction with reference sources, the linguistic and factual knowledge of the researcher interrogating the text.

They can find partners, resources, and support from the so-called language industries field. Multilingualism is a central aspect of language industries. The conservation of natural languages is an important factor for the preservation of national cultural identities. 'Informatization' has been indicated as a key element for the conservation of the vehicular function of a language. B. Quemada (1990, personal communication) has drawn an analogy between the introduction of printing and the informatization of languages. Languages which have not been involved with printing, have become dialects or have disappeared. The same thing could happen to languages that will not be 'informatized'.

References

Allen, J. (1987) **Natural Language Understanding,** (i) p. 2. Menlo Park, California: Benjamin Cumming.

ALPAC (1966) National Research Council. Automatic Language Processing Advisory Committee. **Language and Machine-Computers in Translation and Linguistics.** Washington, DC: National Academy of Sciences, National Research Council.

Bindi, R. and Calzolari, N. (1990) Statistical analysis of a large textual Italian corpus in search of lexical information. In **Proceedings of the EURALEX 1990 Conference** (Malaga).

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

Boguraev, B., Briscoe, T., Calzolari, N., Cater, A., Meijs, W., and Zampolli, A. (1988) **Acquisition of Lexical Knowledge for Natural Language Processing Systems (ACQUILEX).** Technical Annex, ESPRIT Basic Research Action No. 3030. Unpublished.

Bozzi, A. and Cappelli, G. (1987) The Latin lexical database and problems of standardization in the analysis of Latin texts. In **Data Networks for the Historical Disciplines,** edited by F. Hausmann, pp. 28-45. Graz: Leykam-Verlag.

Bresnan, J. (1982) Editor. **The Mental Representation of Grammatical Relations.** Cambridge, Mass.: MIT Press.

Bruce, B. (1975) Case Systems for Natural Language. **Artificial Intelligence**, 6, 327-360.

Calzolari, N. (1988) The dictionary and the thesaurus can be combined. In **Relational Models of the Lexicon**, edited by M. Evans, pp. 75-96. Cambridge: Cambridge University Press.

Calzolari, N. and Picchi, E. (1988) Acquisition of semantic information from an on-line dictionary. In **Proceedings of the 12th COLING** (Budapest, 1988), pp.87-92.

Calzolari, N. and Zampolli, A. (1991) Lexical databases and textual corpora: a trend of convergence between computational linguistics and literary and linguistic computing. In **Research in Humanities Computing**, edited by S. Hockey and N. Ide, pp. 273- 307. Oxford: Oxford University Press.

Colmerauer, A. (1978) Metamorphosis Grammar. Natural Language Communication with Computers. In **Lecture Notes in Computer Science**, vol.63, edited by Leonard Bolc. Berlin: Springer.

Dresher, B.E. and Hornstein, N. (1976) On some supposed contributions of artificial intelligence to the scientific study of language. **Cognition**, 4, 321-398.

Dresher, B.E. and Hornstein, N. (1977a) Response to Schank and Vilensky. **Cognition**, 5, 147-150.

Dresher, B.E. and Hornstein, N. (1977b) Reply to Winograd. **Cognition**, 5, 377-392.

Fillmore, C. (1968) A case for Case. In **Universals in Linguistic Theory**, edited by E. Bach and R.T. Harms. New York: Rinehart Winston.

Fillmore, C. (1977) Scenes-and-frames semantics. In **Linguistic Structures Processing**, edited by A. Zampolli, pp.55-82. Amsterdam: North-Holland.

Friedman, J. (1971) **A Computer Model of Transformational Grammar**. New York: Elsevier.

Garside, R., Leech, G. and Sampson, G. (1987) Editors. **The Computational Analysis of English**. London and New York: Longman.

Gazdar, G., Klein, E., Pullum, G.K., and Sag, I.A. (1985) **Generalized Phrase Structure Grammar**. Cambridge, Mass.: Harvard University Press.

Grishmann, R. (1986) **Computational Linguistics**, (i) p.146. Cambridge: Cambridge University Press.

Halliday, M. (1990) Invited lecture at the 1990 AILA World Congress, Thessaloniki (not yet published).

Kaplan, R. (1973) A general syntactic processor. In **Natural Language Processing**, edited by R. Rustin, pp. 193-241. New York: Algorithmics Press.

Karlgren, H. (1973) Foreword. In **Computational and Mathematical Linguistics**, edited by A.Zampolli and N. Calzolari, pp. xiii-xiv. Firenzi: Olschki.

Kay, M. (1973) The MIND system. In **Natural Language Processing**, edited by R. Rustin, pp. 155-188. New York: Algorithmics Press.

Kay, M. (1982) Grammatico-semantic analysis. **The Prague Bulletin of Mathematical Linguistics**, 39, 110-113.

Kay, M. (1985) Parsing in Functional Unification Grammar. In **Natural Language Parsing**, edited by D.R. Dowty, L. Karttunen and A. Zwicky, pp.251-278. New York: Cambridge University Press.

Koskenniemi, K. (1983) **Two-Level Morphology: a general computational model for word-form recognition and production**, University of Helsinki, Dept. of General Linguistics, Publication no.11, Helsinki.

Kuhn, T. (1970) **The Structure of Scientific Revolution**, 2nd edn. Chicago: University of Chicago Press.

McCord, M. (1980) Slot Grammars. **American Journal of Computational Linguistics**, 6, 1, 31-43.

Minsky, M. (1975) A Framework for Representing Knowledge. In **Psychology of Computer Vision**, edited by P.H. Winston. McGraw-Hill.

Nagao, M. (1989) **Machine Translation—How Far Can It Go?** Oxford: Oxford University Press.

Perschke, S. (1988) **Hearing on the language industry in the European community. Questions put to the participants.** (Background paper for discussion).

Picchi, E. (1988) **D.B.T. A Full-Text Data Base System: Methods and Tools for Searching Full-Text Data.** Pisa:ILC. Internal report.

Pollard, C., and Sag, I.A. (1987) **Information-based Syntax and Semantics**, Vol.I. Stanford: CLSI.

Pustejovsky, J. (1989) Current issues in computational semantics. In **Proceedings of the 4th Conference of the European Chapter of the ACL** (Manchester, 1989) pp. xvii-xxv.

Rohrer, C. (1990) **EUROTRA-7, Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerized Applications.** Presented to the EEC. Unpublished.

Schank, R. (1975) **Conceptual Information Processing**, Amsterdam: North-Holland.

Schank, R. and Abelson, R. (1977) **Scripts, Plans, Goals and Understanding.** Hillsdale, NJ: Erlbaum.

Schmolze, J.G. and Brachmann, R.J. (1982) **Proceedings of the 1981 KL-ONE Workshop**, Fairchild Technical Report no. 618.

Smith, J. (1973) Ideals versus practicalities in linguistic data processing. In **Computational and Mathematical Linguistics**, edited by A.Zampolli and N. Calzolari, pp. 895-898. Firenzi: Olschki.

Wilensky, R., (1981) PAM and MICRO-PAM. In **Inside Computer Understanding**, edited by R. Schank and C. Riesbeck, pp.136-96. Hillsdale, New Jersey: Laurence Erlbaum Associates.

Wilks, Y. (1975) An Intelligent Analyzer and Understander of English. **Comm. ACM**, 18, 5, 264-74.

Winograd, T. (1977) On some contested suppositions of generative linguistics about the scientific study of language. **Cognition**, 5, 151-179.

Winograd, T. (1983) **Language as a Cognitive Process.** Syntax, (i) p.20. Reading, Massachusetts: Addison-Wesley.

Woods, W. (1970) Transition network grammars for natural language analysis. **CAMC** 13:10, 591-606.

Woods, W. (1975) What's in a Link: Foundations for semantic networks. In **Representation and Understanding**, edited by D.Bobrow and A. Collins. New York: Academic Press.

Zampolli, A. (1973) Introduction. In **Computational and Mathematical Linguistics**, edited by A.Zampolli and N. Calzolari, pp. xix-xxviii. Firenzi: Olschki.

Zampolli, A. (1975) **Problemi di linguistica applicata computazionale**. Pisa: CNUCE-CNR.

Zampolli, A. and Calzolari, N. (1973) Editors. **Computational and Mathematical Linguistics**. Firenzi: Olschki.

Zampolli, A. and Hockey, S. (1990). Unpublished Memorandum on Computing in the Humanities, presented to the European Science Foundation.