A State of the art in Computational Linguistics

Giacomo Ferrari

Department of Human Studies - Università del Piemonte Orientale "A.Avogadro"
to appear in the Proceedings of the 17th International Congress of Linguists - Prague

1. The position of Computational Linguistics

Computational Linguistics has a long history, dating back to the Fifties, during which it developed a whole set of computational models and implementations, theories, methodologies and applications. It is difficult to give a sensible account of its present state without going back a little to the main steps through which this discipline evolved towards its present state. Since its origins, Computational Linguistics has been in an intermediate position between Computer Science and Artificial Intelligence, Linguistics and Cognitive Science, and Engineering. Computer Science itself shares its roots with Computational Linguistics; *parsing*, which is central for the design of compilers for programming languages (Aho and Ullmann 1977: 6), is also the building block of any natural language processing engine, and both are the realizations of the chomskian theory of formal languages (Chomsky 1957). The same theory, together with the corresponding computational model, has given a contribution to the general hypothesis of Artificial Intelligence, that human behaviours usually judged intelligent could be simulated in a computer in a principled way. Oversimplifying,

Artificial Intelligence aims at modelling a number of behaviours through three very general paradigms, theorem proving, problem solving and planning, and language understanding and production. The history of both disciplines is rich in intersections, especially between language processing and planning, as in SHRDLU (Winograd 1971) or, more recently, in ARGOT (Allen et al. 1982, Allen 1983), with all its practical and theoretical follow-ups; modern dialogue systems in all their forms and applications are derived from the dialogue model J.Allen designed for ARGOT.

The commitment to "simulation of behaviour", shared by Artificial Intelligence and and a relevant part of Computational Linguistics, makes them also share the effort for "cognitive modelling" of different human behaviours, including the use of language. This is probably one of the reasons why Linguistics appears in the set of sciences originally interested in the arising of the new discipline called Cognitive Science (www.cognitivesciencesociety.org).

Since the Seventies, when language technology reached a state of maturity such as to allow the realization of some applications, Engineering has been interested in some of the language processing techniques, and it appeared soon that the approach introduced by engineers was certainly less theoretically and cognitively interesting, but more effective in many ways. By now, we can say that while Computational Linguists were, and are, more interested in the correctness and plausibility of their models, Engineers were, and are, more interested in the usability of tools and techniques, even

at the cost of some “dirty” solutions. The history of Computational Linguistics in the last decades is much the history of the evolving relations between all these conjuring approaches.

2. Main Achievements

It has been for a long time a common place that Computational Linguistics, as Artificial Intelligence, has created a number of models, programmes, and prototypes which claimed a lot but did nothing at all. This is an extreme view that points out the absolutely theoretical character of what has been done, despite the strong claims of practical utility, often due to constraints imposed by sponsors and funding agencies. Nevertheless, it is just thanks to these attempts, often impractical, though theoretically "clean", that some of the fields of Computational Linguistics reached maturity and has been able to produce stable technologies, as a preparatory step to engineering. The following paragraphs will analyse the main areas in which some achievements have been reached.

2.1 Parsing and Parsing Technology

Parsing has been probably the field of Computational Linguistics where some results have been achieved since an early stage. The first parsers

which bore some relation to grammatical theory go back to the early Seventies. The common background knowledge that made possible such a development is Generative and Transformational Grammar, but its computational interpretation gives rise to e number of different models, with different both technical and theoretical impacts. The key problem to solve is to reach a logical (deep) structure of the sentence, such as to satisfy the constraint of being mapped onto some sort of semantic (executable) representation. Transformational grammar does not offer a direct solution to this problem, leaving a free space to several interpretations. The most orthodox of these interpretations (Friedman 1971) proposed a parser that, after having produced a surface tree, by means of a context-free grammar, tried to yield such a deep structure using the same set of transformations proposed by Chomsky, inverting input and output.

On the contrary, the most successful solution was the one offered by ATN (Woods 1970), as not only it was a good answer to the question of dealing in a single process the mapping of a surface sentence/structure onto the corresponding logical form, but gave also origin to a number of grammatical (Kaplan 1973, 1975; Kaplan and Bresnan 1982) and psycholinguistic (Kaplan 1972, Wanner and Maratsos 1978) debates. This was also the prototype of an entire family of natural language grammars for parsing, called Augmented Phrase Structure Grammars (APSGs, see Heidorn 1975); in general APSGs are ordinary context-free grammars,

augmented with instructions to store the already reached (sub)trees and to carry linguistic tests.

From a plain computational point of view, the way how rewriting rules are stored, accessed, and utilised identifies different groups of algorithms (Early 1970, Kuno1966, Kasami 1965 etc.), and this is the base of the parsing theory and practice in computer science.

From a natural language point of view, starting from these achievements it has been possible, with the time, to develop further (computational) theories (see below § 4.2.2.1), as well as grammar engineering practices (Bates 1978), or natural language parsing best practices (Tomita 1986, 1987).

Turning to the application side, when facing the problem of parsing real natural language, some inadequacies of axiomatic grammars appeared clearly. For instance, people rarely use sentences, in the proper sense, but rather utterances which may be elliptical or really "erroneous"; some solutions have been tried, falling in the category called "treatment of ill-formed input" (Kwasny and Sondheimer 1981). Also, human language consists of sentences/utterances connected with one another, and such connections are realised by some pro-forma, like pronouns that represent nouns, but also sentence fragments; this is the "reference resolution" problem, that gave rise to a number of techniques for treatment (Sidner 1981, 1983; Grosz 1977; Grosz and Sidner 1986; Grosz, Joshi, and Weinstein 1983). In general, however, most of the proposed solutions failed to become technological achievements.

From the point of view of the structure of algorithms, grammatical theories like GPSG (Gazdar et al. 1985) or LFG (Kaplan and Bresnan 1982) have lead to adopt a model in which semantic interpretation goes together with syntactic parsing. This idea of a direct correspondence between syntactic and semantic rules comes form the montagovian approach to logic grammars (Montague 1973), which relied on a categorial grammar. A first attempt to associate semantic rules to context-free and transformational grammars is due to B. Hall-Partee (1973), but GPSG, in its early formulation (Gazdar 1982) makes this association systematic. LFG does not impose any constraint on the nature of the semantic rules, but makes the assumption that such an association takes always place. Presently, the direct mapping of sentences onto a semantic structure is taken as a standard both in logic terms, making reference to the DRT model (Kamp 1981, Pinkal 1985, Kamp and Reyle 1993), and in practical applications, as embedded in (spoken) dialogue systems (see Dahl 2000).

Presently, all this bunch of knowledge, technical know-how, models, programmes is the core technology on which any natural language application is based.

2.2. Interaction Models

Since the early Seventies, natural language has been viewed as the most natural way of interacting with computers, capable of breaking the barriers

of man-machine interaction. The notion of “naturality” has evolved through the years: at the very begging it seemed that the simple possibility to use English sentences was more natural than using some computer language, later it appeared that the “naturality” of natural language was defective if some other “tricks” of natural language are not adopted, like anaphora or paraphrase of an utterance, finally a real “blue-print” for graceful interaction has been dictated, which included a number of ingredients for “naturality” (Hayes and Reddy 1983). At the end, this notion of “naturality” has been overcome in the Eighties and Nineties by the introduction of easy graphical interfaces; nevertheless, natural language interfaces have opened a number of problems of interaction, some of which have a larger import than plain natural language itself.

The most relevant result is the introduction of "dialogue", i.e. the model of (natural language) interaction based on pragmatics and the inference of the user's intentions (Allen 1983). The concept underlying this approach is that human-human interaction is part of a more general “cooperative planned behaviour”. An agent has a plan to reach some objective, but, if she/he meets an obstacle, she/he can call for help; among the obstacles, a very serious one is the lack of the necessary information to carry the original plan, and the call for help can take the form of a request for information. Thus, the agent that responds to the request of information, often does not answer to the literal meaning of the question, but to the intentions of the calling agent she/he infers from the question itself. The plan, the planning

activity and the plan reconstruction are represented in the *planning* formalism, well known to Artificial Intelligence, while Searle's theory of *Speech Acts* gives the basis for the inference of the meaning of an utterance underlying its propositional content. This "marriage" between Speech Acts and Planning has probably been the most theoretically original result in Computational Linguistics, and it gave birth to a set of researches in the field of natural language interaction in terms of planned behaviour (Cohen and Perrault 1979, Allen and Perrault 1980, Allen and Litmann 1986, Carberry 1999), but with the spreading of Spoken Dialogue Systems (Peckham 1993, Smith and Hipp 1994), these models became of practical relevance and are again a paradigm of interaction. On the other hand, the assumption that any given speech act allows a specific set of inferences (as in Traum 1993), promoted the idea that there is a strict correlation between the rhetorical features of an utterance, thus referred to as dialogue act or discourse act, identifiable by a rhetorical label, and the set of inferences this utterance allows. By this conceptual bridge, Allen's theory of dialogue stands also at the origin of all the activities of pragmatic tagging of corpora (see §§ 4.2.1 and 4.2.4).

2.3. Morphology and Dictionaries

Until the Eighties, natural language processing programmes used "built-in" dictionaries of few thousands words. This has been seen for a long time as the only solution to the problem of POS-classifying words before parsing.

However, the small number of lexical items available and "searchable" was seen as a strong limitation to the power of those prototypes.

The problem of finding a word in a list (file, database) of large dimensions could be seen as a plain technical problem of search optimisation. Nevertheless, the idea of making a lexical system able to predict incoming forms motivated some research on the morphological approach to dictionary look-up since the Eighties (see Byrd 1983, Byrd et al. 1986). The typical morphology-based look-up system consists of a set of dictionaries of lexical segments (bases, prefixes, suffixes, endings) and a set of rules of word formation, consisting of two parts, composition of segments and phonetic modifications.

The (casual) appearance of highly inflected languages on the stage of computational treatment, pointed out the weakness of existing lexical analysis systems, mostly designed for English. It is not a case, probably, that the present standard for morphological and lexical analysis comes from a Finnish environment (Koskenniemi 1983). This is based on the above sketched framework, but the form of the (phonological) rules is particularly sophisticated, as it deals with complex morpho-phonetic phenomena.

2.4. The Treatment of corpora

The treatment of corpora goes as back to the origin of Computational Linguistics as Machine Translation. The first project of building a computational lexicon of the works of Saint Thomas (Busa 1980) dates to

the end of the Forties. From this project a number of large projects for the acquiring and the treatment of corpora for different languages was encouraged. Following this initiative, many collections of (literary) corpora have been started, such as the *Trésor de la Langue Française* started in 1971, and computerised since 1977 (see http://atilf.inalf.fr/tlfv3.htm) , the Brown Corpus since 1963 for American English (see http://www.hit.uib.no/icame/brown/bcm.html#bc3), or the COBUILD for British English, started by Collins around 1980 (see http://titania.cobuild.collins.co.uk/).

In the early stages, much attention was given to the acquisition and coding techniques (how to code all the existing characters and diacritics, how to represent textual references), to the context delimitation techniques (how to define a context, how long should it be etc.), to the output procedures (how to print out special characters or diacritics). Obviously, all these problems have been overcome with the introduction of mark-up languages, which allowed to define any subset of characters, while the introduction of high resolution screens, graphic interfaces, and laser printers made the representation of any kind of character easy. Thus the attention moved from technical problems to more substantial problems like the techniques for acquisition and for processing of linguistic data, and the standards for tagging and mark-up (see § 4.2.4).

3. From prototypes to engineering

One of the main objections moved to traditional Computational Linguistics was, and often still is, that it has produced a lot of theoretical work, but no practical application at all. Engineering is still a phase to be reached. The causes of this situation have been, in general, recognised to be a general lack of empirical knowledge and, going farther, the lack of automatically learning processes.

The first argument goes the following way: Theoretical "chomskian" linguistics deals with "competence", our programmes, on the contrary, deal with "performance" phenomena. This causes two inadequacies; (i) what we know about language is based on highly theoretical assumptions, but it has little or nothing to do with the linguistics of "real language", and (ii) even if our knowledge should be valuable, it would cover a number of phenomena largely smaller than those presented by "real language".

The second, more serious, argument runs as follows: no theory or programme will be able to model the whole linguistic knowledge necessary to both analyse any text that is presented and keep aligned with language innovation. The only system that can do that is the one that acquires documents and learns from them new linguistic structures.

Both trends have lead to a rivitalization of corpus linguistics. The main features of this activity are:

- the attention given to the standards of the collection of corpora, be they written texts or spoken samples. Linguists have always been very careful in setting the principles for the collection of their data, but the use in computational applications puts forward problems that never arose before, like the use of standard for document exchange
- the development of methods to prepare those materials, like annotating them
- the development of a number of techniques for the inference of linguistic features from the corpus. These are methods based on the ordinary automatic learning techniques, such as, for instance, Hidden Markov Models, statistical methods, or even neural networks (see § 4.2.3).

The emergence of these new trends does not imply that more traditional research objectives have been dropped; simply, priorities have been changed. Thus theoretical and formal studies are still carried with less effort than in the early stages of Computational Linguistics, while the study of techniques for the treatment of massive linguistic materials acquires a high priority.

4. Computational Linguistics, theoretical Linguistics and Corpora

Having examined the major trends of research and application, we can finally give an account of the present state of Computational Linguistics.

This will run in two directions, a description of the major application areas, in order to understand their impact on the research trends, and a brief account of the most popular research objectives and methodologies.

4.1 Application Areas

The studies in Computational Linguistics have always been affected in some way by the nature of prospective applications. Thus, when human-computer interaction in natural language was in the focus of attention, sentence processing and dialogue models were extensively studied. It is, therefore, very important to have a notion of what are the presently preferred application domains.

4.1.1 Spoken Dialogue Systems

The study of dialogue structures and dialogue modelling has been one of the research themes since the early Eighties, and different paradigms have been developed, most of which dealt with typed input and output. By the end of the Nineties some vocal services have been activated, especially in Europe, in which vocal input is interpreted by a signal processor and a parser, while the answer, also in vocal form, is planned by a dialogue manager module according to a number of contextual parameters. There are some systems that can be considered a sort of milestones, such as SUNDIAL (Peckham

1993), Dialogos (Albesano et al. 1997), or the DUKE University system (Bierman et al. 1993).

These systems have large application in transaction situations, like trip information, train, flight, or hotel reservations. The main features of these systems are a robust speech recognition component, a simple parser that performs semantic interpretation of the input, a dialogue manager that acquires the interpretation of the input and, according to a number of parameters, decides whether it is necessary to ask the user to complete the parameters of her/his query, or, if they are sufficient, the query can be processed by the back-end component (data-base). The reactions of the dialogue module are translated into sentences and hence into speech utterances by a speech synthesis module.

Spoken dialogue systems integrate technological developments of early dialogue systems (see § 2.2), based on a simplified theory of speech acts, then dialogue acts, and corpus techniques, largely used to infer dialogic rules. These applications are mostly responsible for the creation of dialogue corpora, annotated in terms of dialogue or discourse acts (see http://www.sigdial.org for a relatively exhaustive list of such initiatives).

4.1.2 Multilingualism

Machine translation has been one of the first, if not the first application domain of computational linguistics. This has been experimented in several

areas, like military message translation, technical translation, institutional document translation. Translation has always been intended to be from one language to the other, from Russian to English and vice-versa at the beginning, from Japanese to English and vice-versa in the large Japanese projects of the Seventies and the Eighties, or from French to English and vice-versa in Canada. The European Community started, in the Eighties, the first project many-languages-to-many-languages, Eurotra (Raw et al. 1988). Later, it is worth mentioning the large project Verbmobil (Kay et al. 1994), which couples machine translation techniques with spoken dialogue techniques.

The spreading use of the Internet proposes a new dimension of the translation problem. The scenario is the search for some information, or some documents, using a generic search engine; once key-words have been introduced in one language, they can be translated into other languages by means of bilingual dictionaries, and the search can be launched in the target languages. This limits the search to those documents into whose languages the query words have been translated, thus missing possible other documents in other languages. To overcome this limitations, the idea has been proposed that any document and any query is translated into a sort of (universal) interlanguage (Universal Networking Language, also UNL, see http://www.ias.unu.edu/research/unl.cfm); the match is carried onto expressions of this interlanguage, rather than onto one-to-one translations.

4.1.3 Document classification and retrieval

Studies in classification and retrieval of documents start in the field of information retrieval, which has been for years separated from any kind of natural language processing. Many statistical techniques have been tried to establish degrees of relevance of words in a text (see Salton and Buckley 1988), or degrees of relations between couples or triples of words (collocations; see e.g. Sinclair 1987, Sekine et al. 1992, Dagan et al. 1999). More recently, Computational Linguistics tried to propose contributions coming from linguistic analysis. On one hand, the grouping of key-words by concepts has been proposed, using the WordNet approach (see Mandala et al. 1998), on the other hand some (parsing) techniques has been developed to recognise and use phrases as queries (the Cristal LRE Project, see http://www.ilc.pi.cnr.it/Mac/cristal.html).

Document classification and retrieval has several applications. Web browsers are generally based on such a technique, but also more local services like dedicated portals, *pull* information systems, technical information systems and any other service that relies on document management.

4.2 Research Areas

In order to develop the above mentioned applications, a number of basic techniques and theoretical issues is now being developed. We will sketch, in

the following few pages, the main lines of both theoretical and technical developments of research in Computational Linguistics.

4.2.1 Linguistic Resources

By linguistic resources all kinds of repositories where some linguistic piece of knowledge is contained in an explicit or implicit way are meant. The study of the techniques for acquisition, storage, manipulation of information, standards, access, and use of these resources has become an almost autonomous branch of Computational Linguistics, with its own characteristics and a large community of practitioners, as is shown by the LREConference (Linguistic Resources and Evaluation Conference, see http://www.lrec-conf.org/), a biennial event that assembles several hundreds of researches. This branch will be dealt with by N. Calzolari, elsewhere in this volume. However I will briefly treat here those subjects that somehow are closer to traditional Computational Linguistics or interact with it.

Although linguistic resources may fall into different categories, the most common ones are dictionaries and corpora.

Dictionaries may take various forms. The most common dictionaries are the general coded dictionaries, where POS and other morpho-syntactic codes are associated to lexical items. There are also a number of special language dictionaries in specific (technical) fields. A very peculiar one is WordNet. This is a sort of "conceptual" dictionary, where words are presented in a

structure that accounts for links like synonymy, hyperonymy etc. This dictionary comes from an initiative taken by the psychologist G.Miller in the early Eighties (Miller 1985); the general idea was to present lexical material in a more concept-oriented way. More recently it has been taken to represent real conceptual relations, and it has been tested in many programmes of document classification and document retrieval to group key-words in “ideas”. With this objective in mind, some projects have been started, and some finished, to build WordNets for other languages or multilingual. This particular usage of WordNet has stimulated a number of researches aimed at analysing single word classes, proposing methodologies to carry word disambiguation, or statistically based inferences of different type (see for all http://engr.smu.edu/~rada/wnb/).

The other large family of resources are “corpora”. A “corpus” is a collection of language samples stored in specific formats and annotated in such a way as to distinguish different linguistic units. The utility of such large collections of texts is that they offer the empirical base for the inference of different linguistic regularities that may form a rule or a norm. The ground level of such an inferential function is the simple distribution of words; this is the basis of a large enterprise like COBUILD (http://titania.cobuild.collins.co.uk/), a corpus in continuous evolution, where texts are stored in a rough format and words and word sequences can be searched through the whole material. However, the most spread opinion is that texts should be “annotated” or “tagged”. This means that linguistic

units belonging to some level of analysis should be assigned labels that classify them. Thus, at the word level, any word is isolated and added the POS-tag, i.e. the label indicating its Part Of Speech and other morphological properties. Higher levels are also possible, like coreference or, more commonly, the pragmatic level, i.e. the discourse segments.

As it has been mentioned before, while POS-tags belong to a more or less stable tradition, discourse-act tagging has been originated and strongly influenced by previous studies on dialogue models. Thus, DAMSL (Allen and Core 1997) tagging system is the result of a long involvement in dialogue modelling of University of Rochester, as well as the RSTool by D.Marcu (2000) is a derivation of the studies carried at ISI in the domain of text generation, not to mention the MapTask (Carletta et al. 1996) initiative, or the Verbmobil corpus (Alexandersson et al. 1997).

Linguistic resources are used mostly to develop learning techniques to be applied to different fields of natural language processing (see § 4.2.3).

4.2.2 Parsing sentences and documents

Parsing, keeps being an important component of any process having natural language as object. It has evolved in two main directions, quite opposite one to the other. On one side a substantial evolution of grammar formalisms have produced heavy modifications also in the parsing techniques; on the other side, the need for more realistic and engineered solutions led to the use

of some sort of mass-parsing techniques, based on substantial simplifications of traditional parsing.

4.2.2.1 New formalisms and new algorithms

Since the introduction of APSGs, it appeared that a neater parsing process should rely on both a structural (syntactic trees) and a functional (labels) component, which could manage agreement tests (in terms of identity of feature values), "functional" transformations, like passive (in terms of inversion of SUBJECT-OBJECT labels) and other parsing situations. This awareness somehow percolated to theoretical linguistics, where a number of double-sided theories arouse, like Lexical-Functional Grammar (LFG-Kaplan and Bresnan 1982), Generalised Phrase Structure Grammar (GPSG - Gazdar et al.1985), Head-driven Phrase Structure Grammar (HPSG - Pollard and Sag 1994), or plain functional computational models like Functional Unification Grammar (FUG – Kay 1979, 1984, 1985). From a procedural point of view all these theories, FUG excepted, more or less agreed in identifying two main blocks, a structural process, which remains close to traditional parsing, and a functional process, which takes care of computing the values of higher nodes labels as a function of lower nodes labels values (in terms of computing new values or raising lower values). LFG proposed an algebraic system of meta-equations, GPSG proposed a set of constraints to the admissible values, while FUG completely eliminated the structural

computation, trying to use a single algorithm, called *functional unification*, which directly produced a single functional representation of any sentence.

It is not necessary, here, to discuss the viability of each of these solutions; the functional computation has in itself many difficulties, and some attempts to solve them have been done (see Schieber 1986). All the grammars of this group, and the related algorithms, have been classified as feature-based grammars, feature unification algorithms, unification algorithms.

More recently, a general logic for the treatment of feature unification has been developed (Carpenter 1992), and some of the related algorithms have been implemented (see for instance Kepser 1994).

The solution offered by feature-based grammars and typed-feature logic seemed, for a while a sort of definite solution for every parsing problem, but despite its formal clarity and neatness it appeared soon inefficient for the needs that were emerging from the market of natural language products.

4.2.2.2 Mass-parsing

In fact, the idea of producing a complete and deep analysis of sentences was loosing appeal, as the more promising applications of the end of the Nineties became text analysis, text classification, and any other domain where large quantities of text were to be analysed. In addition, it turned out that for the purpose of identifying some key-words or key expressions, no deep and accurate analysis was needed.

Thus the idea emerged that a complete parsing was not necessary, at least for "real" purposes, but some sort of "surface" analysis was sufficient. This job is successfully carried by "shallow parsers", whenever a syntactic analysis is necessary, or even "chunkers" where just the identification of some words is sufficient.

Shallow parsers are traditional parsers to which some complexity restriction has been added, especially in those areas where a deep analysis generates ambiguity with no interpretation benefit corresponding to it. An example for all could be the PP-attachment problem. By PP-attachment the ambiguity in the interpretation of Prepositional Phrases if meant; in the sentence

John saw the man in the park with a telescope

many interpretations are possible according to whether "with a telescope" is attached to "park" (a park where a telescope is located), to "man" (a man walking with a telescope under his arm), or to "John" (John saw the man by a telescope). This gives rise to an explosion of ambiguities, which cannot be solved unless one knows practically what happened. But, despite these difficulties, if one has to classify the sentence, or build a query to some search engine, or even translate it into another language, the attachment of the PP does not make any difference. The conclusion is that a deep syntactic analysis introduces more information than actually needed. It is enough to identify the PPs without trying to attach them one to the other, and any solution will hold.

In an extreme view, it could be possible just to delimit constituents on the basis of some separators, like articles, prepositions or conjunctions (see Abney 1991). In this case, no structure, even partial, is built, but sentences are simply divided into segments; segments are called "chunks" and this approach is known as "chunking".

Actually there is no clear cut difference between "shallow parsing" and "chunking" as they simply lay in different locations of a continuous line. By "chunking", a text is divided into chunks that may belong to the same phrase (but this is to be verified) while "shallow parsing" implies in any case the attribution of some shallow structure, as it is the case in those resources known as Tree-banks (see http://www.cis.upenn.edu/~treebank/). It is usually believed that this preliminary processing of a text (in both cases) is an intermediate step towards deeper parsing, if it is necessary.

Thus syntactic processing of texts fall into two categories, according to whether hand-crafted patterns or machine learning techniques (see § 4.2.3) are used. By the first method, the analysis still relies on some sort of rules, be they relaxed enough to allow chunking; by the second method it is assumed that parsing of large texts becomes feasible without much work of grammar writing, but simply by inferring new linguistic constructions.

4.2.2.3 Finite State Transducers Technology

A finite state transducer is an extension of a specific family of automata, essentially it is a finite state automaton that works on two (or more) tapes.

The most common way to think about transducers is as a kind of ``translating machine", that reads from one of the tapes and write onto the other. Transducers can, however, be used in other modes as well: in the generation mode transducers write on both tapes and in the recognition mode they read from both tapes. Furthermore, the direction of translation can be turned around: i.e. the expression *a:b* can not only be read as ``read *a* from the first tape and write *b* onto the second tape", but also as ``read *b* from the second tape and write *a* onto the first tape". In procedural terms,

it consists of a set of states and transitions; each transition is allowed by a specified transition symbol, and associates to it a translation symbol, including the null one.

For may years Finite State Automata, and, therefore, Finite State Transducers, have been considered an efficient tool for the treatment of natural language, but less powerful and accurate than other formalisms as Context-Free grammars. We are not interested, here, into a computational evaluation of FSTs, but it is a fact that in more recent times they found a large employ in different areas, such as morphological analysis (see Clemenceau and Roche 1993, Karttunen et al. 1992, Silberztein 1993), following the work by Koskenniemi (1983), lexical recognition (see Pereira et al. 1994), speech processing (see Laporte 1993 and 1996), and even syntactic analysis (Pereira and Wright 1996).

The appeal of this approach, started with the "two-level morphology" by K.Koskenniemi, and spread over other areas, is that it is possible to map

whatever onto whatever else. Thus, phonetic lattices can be mapped onto tentative words, idioms can also be mapped onto words, morphologically complex words onto their morphemes (and related interpretations), strings of words onto phrases. This allows an unprecedented flexibility, without missing computational tractability and neatness. On the contrary, it seems that from a computational point of view FSTs are more efficient than CF grammars, and related algorithms. For these reasons FST technology is often regarded as the most promising technology for massive natural language applications.

4.2.3 Acquisition and Learning

Another relevant drawback for the building of efficient and robust natural language processing systems is the lack of flexibility and adaptivity of linguistic knowledge embedded in these kinds of system. This is commonly attributed to the fact that it takes long time and a considerable effort to build the linguistic knowledge sources, like dictionaries or grammars, but they remain unchanged through time, thus they may become obsolete in short; also they are not easily adapted to new kinds of applications. In addition, any single new linguistic item, be it a word unknown to the dictionary, a sentence type unforeseen by the grammar, or an unknown semantic frame, causes a Natural Language Processing System to stop or crash (see Armstrong-Warwick 1993). The answer to this situation seems to be the

building of systems able to infer and learn new lexical items or linguistic regularities. In addition, experiences carried in the field of speech recognition showed that learning approaches, especially based on statistic methods, were highly successful (Stolcke 1997, Jelinek, 1998). Also, the collection of large corpora, started in early stages of Computational Linguistics, has produced a huge amount of data that can be used to induce new patterns or train learning systems.

Learning techniques have been mainly applied to morphologic and syntactic analysis (Charniak 1997, Brent 1999, Daelemans et al. 1999, Haruno et al. 1999, Ratnaparkhi 1999), semantic disambiguation (Ng & Zelle 1997, Dagan et al. 1999), discourse processing and information extraction (Cardie 1997, Beeferman et al. 1999, Soderland 1999), and machine translation (Knight 1997).

Natural language Learning studies have been carried mostly without strict contacts with traditional Machine Learning; nevertheless some of the traditional techniques have much in common with this branch (Brill and Mooney 1997). The basic techniques are those based on statistics: a probabilistic model, such as Hidden Markov Models (HMM) or Probabilistic Context Free Grammars (PCFGs) can be roughly represented in terms of a finite state machine or a production system. Statistical data derived from a corpus can be attached to such models in order to serve as training for them. The corpus can be pre-annotated or not; in the first case

training consists just in counting the already marked linguistic phenomena, in the second case some estimation techniques shall be used.

Another important method is symbolic learning, which consist in automatic acquisition of rules; this has always a statistical base, but the model is a rule-based rather than a probabilistic one. A variety of this method consists in remembering past examples and make decision on the base of similarity to those examples.

A third method relies on the neural networks, a computational model of neural cells (Rumelhart and McClelland 1986) that learns from examples.

It is evident, in all three cases, the relevant role played by a *training corpus*, i.e. a set of data that are used by the learning mechanism to infer, in any of the sketched ways, linguistic regularities. Training can be *supervised*, if the data have been somehow manipulated before training, or *unsupervised* if the data have not been interpreted in advance.

Many research efforts are deployed in this area, although stable technological results are still far. However, the assumption that the only possibility of improving our natural language systems and reaching a level of engineering close to real applications makes of this research field, together with the strictly related corpus based (computational) linguistics, the domain to which the majority of efforts is devoted.

4.2.4 Text and dialogue studies

As it has been anticipated above, discourse and dialogue studies have evolved in two different directions. On one side, many of the functional or speech acts based theories have been turned into sets of pragmatic tags, as a preparation for the analysis of corpora. On the other side some of the traditional issues have been tackled with new methodologies.

A commonly accepted result is that discourse has an internal structure, however it is recognised and marked. It is also accepted that the discourse structure can be exploited in different natural language processing applications, such as summarisation, machine translation, natural language generation. Thus, research has been carried to produce systems that automatically chunk discourse into segments and assign them labels according to some theory of discourse acts, the *discourse parsers*. Many of these attempts have been based on the recognition of the internal structure of discourse (see for instance (see Kurohashi & Nagao 1994, Asher Lascarides 1993), but they failed to produce reliable results, especially with real texts. Also logical models, like Discourse Representation Theory (Kamp 1981, Kamp and Reyle 1993), could not be applied to large texts, despite their formal elegance and semantic depth. Better results have been attained by using incremental algorithms (see Polanyi 1988, Cristea and Webber 1997).

But the most successful approach to discourse parsing is considered, presently, Marcu's algorithm (see Marcu 2000). This is based on the

recognition of cue phrases and discourse markers and the use of RST (Mann and Thomson 1988). Segments fall into three levels, *sentence*, *paragraph*, and *section*, and the rhetorical relations holding between them are hypothesised on the basis of the corpus analysis of the cue-phrases. Then the trees, for each level of granularity, are built according to RST. The final discourse trees are built by merging the trees corresponding to each level.

A general issue strictly connected to discourse processing is discourse coherence. This is a strong theoretical issue, also connected to discourse parsing, as discourse segments are characterised and identified by coherence relations. In this field two basic approaches have been developed, Centering Theory (Grosz, Joshi, Weinstein 1995) and Veins Theory (Cristea et al. 1998). Both conjure to improve discourse parsability, also providing methods to solve anaphora and other forms of co-reference.

Approaches to discourse and dialogue do not appear systematic, but rather a collection of solutions for different issues. The original systematic and coherent studies have shown that the phenomenon is much more complex than one could think of, and several different aspects are being studied now. However, it is to be seen as a high priority topic, as the large amount of (textual) documents available in the Internet demands for intelligent, flexible and effective programmes for text retrieval, text summarisation, and content extraction.

5. Recommendations for linguists

Computational Linguistics has often been divided in two extreme positions. On one hand the community of researchers involved in lexicography and computational treatment of texts have developed tools for the treatment of large amounts of linguistic data, as well as linguistic studies based mostly on statistics. This was true in the Sixties and early Seventies, when text processing was based on punched cards, the software definition of special characters and diacritics, and printing chains for those special characters, and is true now, when texts are marked-up and standardised by the use of exchange languages like SGML or XML.

The results of this trend of studies tend to be probabilistic.

On the other hand, the community of researchers involved in syntactic analysis, semantic interpretation, pragmatics, dialogue modelling, and discourse planning tend to look for regularities that can be expressed in rules that can generate corresponding behaviours. Sometimes these rules have been offered by theoretical linguistics, as for syntactic parsing, or logic, as for semantic interpretation, sometimes have been derived by the interpretation of some empirical material, assisted by some general theories as it has been the case for dialogue modelling.

This dichotomy somehow interact with another opposition, sketched at the beginning of this article, between a highly "prototypical" approach and extreme low level "engineering" solutions.

Till the early Nineties, researchers created highly sophisticated prototypes, where modules often integrated complex theoretical formal models. Prototypes were costly, in terms of required know-how, but it was often claimed that easy portability from one application to the other (say form a database to another database) was a source of revenue to cover those expenses.

In fact this turned out to be partly false. Portability was also costly and difficult, and, in addition, users seemed not to be happy with the few commercial products that were proposed in the early Eighties in the field of natural language interfaces. When the engineering of some products started, some of the strongest assumptions were torn apart. Modularity of early prototypes was changed into the view that modules should have been interchangeable and the notion of *interoperability* became fashionable, meaning that complex systems should be built by gluing together "off-the-shelf" pieces. This view destroyed the assumption that "the more complex a problem is, the more theoretically sophisticated the solution", and, somehow, demotivated advanced research, in favour of less sophisticated empirical studies.

Extreme views have never helped research and produced real advances. It is abvious that no engineering view could have been even possible, without those very principled prototypes. We are now able to assemble "off-the-shelf" commercial modules because, at a given time point, people designed theoretically motivated objects. The mistake was, probably, to try to stand

on the side of applications before being mature for them. With an *a posteriori* evaluation we should admit that Computational Linguistics *is* a theoretical, high-risk branch of research participating into the cycle of information technology and services. Engineering of products cannot be done by computational linguists; Computational Linguistics must produce advanced research, know-how, but not get involved in production. To make a parallel, the fact that bridges and buildings stand because of some principles of physics does not purport the involvement of physicians in ordinary projects; nevertheless physicians have the duty to keep studying the principles of statics.

Talking of syntactic approaches, we observe that in many cases the move towards empirical studies is uncontrolled, because linguistics has no interpretation mechanisms. Computational Linguistics gave important results in syntactic and morphological analysis because it could rely on robust theoretical linguistic models like Chomsky's grammar(s) and the theory of automata. This is not true in other fields like, for example, discourse and dialogue modelling.

Discourse and dialogue have been studied by linguists according to different paradigms, such as discourse analysis (Sinclair and Coulthard 1975), conversationalism (Sack and Schegloff 1973), or sociolinguistics (Tannen 1984). The speech acts/planning approach and all its developments arouse in the area of Computational Linguistics with no link to linguistics; thus the empirical studies necessary for a better modelling have been often carried

with a weak linguistic view. This is even worst in the field of corpus-based linguistics, where often tag-sets are chosen with the objective of getting a "pre-theoretical" or "theoreticaly neutral" classification of facts.

In the field of semantics, Computational Linguistics turned to logic because linguistics offered absolutely nothing for what concerns the meaning of sentences. Thus, the conclusion is that Linguistics is in debt here, and should take advantage of the stimuli coming from CL to build new theories which take into account computational modelling.

Again, Computational Linguistics is not an applicative domain, but a part of theoretical linguistics in its own rights; it provides an extra view on language, besides historical linguistics, structural linguistics, generative linguistics, cognitive linguistics, etc. This is the reason why we should think of integrating Computational Linguistics approach and stimuli into linguistics, rather than just promoting co-operation between Computational and Theoretical Linguistics.

References

Abney, S. 1991. "Parsing By Chunks." In *Principle-Based Parsing: Computation and Psycholinguistics*, R. Berwick , S. Abney, and C. Tenny (eds.), 257-278. Dordrecht: Kluwer Academic Publishers.

Aho, J. A. V. and Ullman D. 1977. *Principles of Compiler Design*. Addison-Wesley.

Albesano, D., Baggia, P., Danieli, M., Gemello, R., Gerbino, E., Rullent, C. 1997. "Dialogos: A Robust System for Human-Machine Spoken Dialogue on the Telephone." In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. Munich.

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N., Schmitz, B., and Siegel, M. 1997. *Dialogue Acts in VERBMOBIL-2 Second Edition*, Verbmobil Report 226.

Allen, J.F., Frisch, A.M., Litman, D.J. 1982. "ARGOT: The Rochester Dialogue System." In *AAAI- Proceedings of the Conference*: 66-70

Allen, J.F. 1983. "Recognizing Intentions from Natural Language Utterances." In *Computational Models of Discourse*, M. Brady and R.C. Berwick (eds.), 107-166. Cambridge, MA: MIT Press.

Allen, J.F. and Perrault, C.R. 1980. "Analyzing intention in utterances." In *Artificial Intelligence*, 15: 143-178.

Allen, J.F. and Core, M. 1997. "Coding Dialogs with the DAMSL Annotation Scheme." In *Working notes of the AAAI Fall 1997 Symposium on Communicative Action in Humans and Machines*.

Allen, J.F. and Litman D.J. 1986. "Plans, Goals, and Language." In *IEEE Proceedings: Special issue on Natural Language Processing*, G. Ferrari (ed.): 939-947.

Armstrong-Warwick, S. 1993. "Preface". In *Computational Linguistics* 19(1): iii–iv.

Asher, N. and Lascarides, A. 1993. "Temporal interpretation, discourse relations and common sense entailment". In *Linguistics and Philosphy*, 16(5): 437-493.

Bates, M. 1978. "Theory and practice of augmented transition networks." In *Natural Language Communication with Computers*, L.Bolc (ed). Springer.

Beeferman, D., Berger, A., Lafferty J. 1999. "Statistical Models for Text Segmentation." In *Machine Learning: Special Issue on Natural Language Learning* , 34 (1): 177-210.

Biermann, A., Guinn, C., Hipp, D.R., and Smith, R. 1993. "Efficient Collaborative Discourse: A Theory and its Implementation". In *ARPA Workshop on Human Language Technology*, Princeton, NJ.

Brent, M.R. 1999. "An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery." In *Machine Learning: Special Issue on Natural Language Learning* , 34 (1): 71-105.

Brill, E. and Mooney, R.J. 1997. "An overview of Empirical Natural Language Processing." In *AI Magazine*, 18(4): 13-24.

Busa, R.S.J. (ed.) 1980. *S.Thomae Aquinatis Opera Omnia ut sunt in Indice Tomistico additis 61 scriptis ex aliis medii aevi auctoribus curante Roberto Busa S.J.* Stuttgart: Frommann-Holzboog.

Byrd, R. 1983. “Word formation in Natural Language Processing systems”. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*.

Byrd, R., Klavans, J.L., Aronoff, M., Anshen, F. 1986. “Computer methods for morphological analysis.” In *24th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, Columbia University: 120-127.

Carberry, S. 1999. “A Process Model for Recognizing Communicative Acts and Modeling Negotiation Subdialogues.” In *Computational Linguistics*, 25 (1): 1-53.

Cardie, C. 1997. "Empirical methods in information extraction". In *AI Magazine*, 18 (4): 65-79.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. 1996. *HCRC dialogue structure coding manual*. Edinburgh: Technical Report HCRC/TR-82.

Carpenter, R. 1992. *The Logic of Typed Feature Structures.* Cambridge, MA: Cambridge University Press.

Charniak, E. 1993. *Statistical Language Learning*. Cambridge, MA: MIT Press.

Charniak, E. 1997. "Statistical techniques for natural language parsing". *AI Magazine*, 18(4): 33-43.

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

Clemenceau, D. and Roche, E. 1993. "Enhancing a large scale dictionary with a two-level system." In *EACL-93 Proceedings of the conference:* 465.

Cohen, P.R. and Perrault, C.R. 1979. “Elements of a Plan-based Theory of Speech Acts.” In *Cognitive Science*, 6: 177-212.

Cristea, D., Ide, N., Romary, L. 1998. "Veins Theory: Model of global Discourse Cohesion and Coherence." In *Proceedings of Coling/ACL, Montreal*: 281-285.

Cristea, D. and Webber, B. 1997. "Expectations in incremental discourse processing." In *Proceedings of ACL/EACL-97*, Madrid: 88-95.

Daelemans, W., van den Bosch, A., Zavrel, J. 1999. "Forgetting Exceptions is Harmful in Language Learning." In *Machine Learning: Special Issue on Natural Language Learning* , 34 (1): 11_43.

Dagan, I., Lee, L., and Pereira, F. 1999. "Similarity-based models of word cooccurrence probabilities." In *Machine Learning*, 34(1): 43-69.

Dahl, D.A. (ed.) 2000. *Natural Language Semantics Markup Language for the Speech Interface Framework*, http://www.w3.org/TR/2000/WD-nl-spec-20001120

Earley, J. C. 1970. "An efficient context-free parsing algorithm." In *Communications of the ACM* 13(2):94-102.

Friedman, J. 1971. *A Computer Model of Transformational Grammar*. Elsevier.

Gazdar, G. 1982. "Phrase structure grammar." In *The Nature of Syntactic Representation*, P.Jacobson and G.K.Pullum (eds.), Dordrecht, Reidel:131-186.

Gazdar, G., Klein, E.H., Pullum, G.K., and Sag, I.A. 1985. *Generalized Phrase Structure Grammar*. Oxford: Blackwell, and Cambridge, MA: Harvard University.

Grosz, B.J. 1977. “The representation and use of focus in a system for understandings dialogs.” In *Proceedings of IJCAI*: 67-76.

Grosz, B.J., Joshi, A.K., and Weinstein, S. 1983. "Providing a unified account of definite noun phrases in Discourse." In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*: 44-50.

Grosz, B.J., Joshi, A.K. and Weinstein, S. 1995. "Centering: A framework for modeling the local coherence of discourse." In *Computational Linguistics*, 12(2): 203-225.

Grosz, B.J. and Sidner, C.L. 1986. "Attention, Intention and the structure of Discourse." In *Computational Linguistics* 12(3): 175-204.

Hall Partee, B. 1973. "Some transformational extensions of Montague Grammar." In *Journal of Philosophical Logic* (2): 509-534.

Haruno, M., Shirai, S., Ooyama, Y., Aizawa, H. 1999. "Using Decision Trees to Construct a Practical Parser." In *COLING-ACL 1998*: 505-511.

Hayes, P.J. and Reddy, R. 1983. "Steps Toward Graceful Interaction in Spoken and Written Man-Machine Communication." In *International Journal of Man-Machine Studies* 19(3): 231-284.

Heidorn, G.E. 1975. "Augmented Phrase Structure Grammars." In *Theoretical Issues in Natural Language Processing*, B.L. Nash-Webber and R.C. Schank (eds.), Assoc. for Computational Linguistics.

Jelinek, F. 1998. *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.

Kamp, H. 1981. "A theory of Truth and Semantic Representation". In *Fromal Methods in the Study of Language*, J. Groendijk, J. Janssen, and M. Stokhof (eds.), 277-322. Dordrecht: Foris.

Kamp, H. and Reyle, U. 1993. *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Studies in Linguistics and Philosophy 42*. Dordrecht, Kluwer, 1993.

Kaplan, R. 1972. "Augmented Transition Networks as Psychological Models of Sentence Comprehension." In *Artificial Intelligence* 3: 77-100.

Kaplan, R. 1973. "A general syntactic processor." In *Natural Language Processing*, R.Rustin (ed.), 193-241. New York: Algorithmic Press.

Kaplan, R. 1975. "On Process Models for Sentence Analysis." In *Explorations in Cognition*, D.A.Norman and D.R.Rumelhart (eds.), 117-134. San Francisco: Freeman.

Kaplan, R. and Bresnan, J. 1982. "Lexical-functional Grammar: A formal system for grammatical representation." In *The mental representation of grammatical relations*, J.Bresnan (ed.), 173-281. Cambridge, MA: MIT Press.

Karttunen, L., Kaplan, R.M., and Zaenen, A. 1992. "Two-level morphology with composition." In *Proceedings of the 14th International Conference on Computational Linguistics (COLING 92)*: 141-148.

Kasami, T. 1965. *An efficient recognition and syntax algorithm for context-free languages.* Technical Report AF-CRL-65-758, Air Force Cambridge Research Laboratory, Bedford, Massachusetts.

Kay, M. 1979. “Functional Grammar.” In *Proceedings of the 5th Annual Meeting of the Berkeley Linguistic Society*, 142-158. Berkeley:

Kay, M. 1984. “Functional Unification Grammar: A Formalism for Machine Translation.” In *Proceedings of Coling 84*, 75-78. Stanford.

Kay, M. 1985. “Parsing in Functional Unification Grammar.” In *Natural Language Parsing*, D. Dowty, L. Karttunen, A. Zwicky (eds.), 251-278. Cambridge, MA: Cambridge University Press.

Kay, M., Gawron, J.M., Norvig, P. 1994. *Verbmobil. A Translation System for Face-to-Face Dialog*. Chicago: Chicago University Press (CSLI Lecture Notes n. 33).

Kepser, S. 1994. *A Satisfiability Algorithm for a Typed Feature Logic*. m.a. thesis. Arbeitsberichte des SFB 340, Nr. 60.

Knight, K. 1997. "Automating knowledge acquisition for machine translation." In *AI Magazine*, 18(4): 81-96.

Koskenniemi, K. 1983. *Two-level Morphology: A general computational Model for Word-form Recognition and Production.* Publ.#11, Helsingin Yliopisto.

Kuno, S. 1966. "The augmented predictive analyzer for context-free languages; its relative efficiency." In *Communication of the ACM* 9,11: 810-823.

Kurohashi, S. and Nagao, M. 1994. "Automatic detection of discourse structure by checking surface information in sentences." In *Proceedings of the 15th International Conference on Computational Linguistics (Coling94),* 1123-1127. Kyoto.

Kwasny, S. and Sondheimer, N. 1981. "Relaxation techniques for parsing grammatically ill-formed input in natural language understanding systems." In *American Journal of Computational Linguistics*, 7(2):99-108.

Laporte, E. 1993. *Phonétique et transducteurs*, Technical report, Université Paris 7.

Laporte, E. "Rational transductions for phonetic conversion and phonology". In *MERL-TR-96-30.*

Mandala, R., Takenobu, T., and Hozumi, T. 1998. "The use of WordNet in information retrieval." In *Usage of WordNet in Natural Language Processing Systems, COLING/ACL-98 Workshop, Montreal, August 16, 1998 S*. Harabagiu (ed.): 31-37.

Mann, W.C. and Thompson, S.A. 1988. "Rhetorical structure theory: A theory of text organization". In *Text*, 8(3): 243-281.

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA:The MIT Press, 2000.

Miller, G. A. 1985. ``WordNet: a dictionary browser." In *Proceedings of the First International Conference on Information in Data*, University of Waterloo, Waterloo.

Montague R. 1973. "The proper Treatment of Quantification In Ordinary English." In *Formal Philosophy. Selected Papers of Richard Montague*, R.Thomason (ed.), Yale University Press.

Ng, H. T., & Zelle, J. 1997. "Corpus-based approaches to semantic interpretation in natural language processing". In *AI Magazine*, 18(4): 45-64.

Peckham, J. 1993. "A new generation of spoken dialogue systems: results and lessons from the SUNDIAL project." In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, 33-42.

Pereira, F.C.N., Riley, M., and Sproat, R.W. 1994. "Weighted rational transducers and their application to human language processing". In *ARPA Workshop on Human Language Technology*, Morgan Kaufmann.

Pereira, F.C.N. and Wright, R. 1996. "Finite state approximation of phrase structure grammars." In *MERL-TR-96-30.*

Pinkal, M. 1985. "Situationssemantic und Diskursrepräsentationstheorie: Einordnung und Anwendungsaspekte." In *8th German Workshop on Artificial Intelligence, Wingst/Stade, Germany, October 8-12, 1984, Proceedings*, 397-407. Springer.

Polanyi, L. 1988. "A formal model of the structure of discourse." In *Journal of Pragmatics*, 12: 601-638.

Pollard, C. and Sag, I. 1994. *Head-driven Phrase Structure Grammar*. Chicago: Chicago University Press (CSLI Pubblications).

Ramshaw, L.A. and Marcus, M.P. 1995. "Text Chunking Using Transformation-Based Learning." In *Proceedings of the Third ACL Workshop on Very Large Corpora*, Cambridge MA.

Ratnaparkhi, A. 1999. "Learning to Parse Natural Language with Maximum Entropy Models." In *Machine Learning: Special Issue on Natural Language Learning* , 34 (1): 151-175.

Raw, A., Vandecapelle, B. and Van Eynde, F. 1988. "Eurotra: an overview". In *Interface. Journal of Applied Linguistics* 3/1: 5-32.

Rumelhart, D. E. and McClelland,, J. L. (eds.) 1986. *Parallel Distributed Processing: Explorations in the Mi-crostructure of Cognition, Volumes 1 and 2*. Cambridge, MA: MIT Press.

Sacks, H. and Schegloff, E.A. 1973. "Opening Up Closings." *Semiotica*, VIII, 4 : 289-327.

Salton, G. and Buckley, C. 1988. "Term weighting approaches in automatic text retrieval." In *Information Processing and Management*, 24(5): 513-523.

Schieber, S. M. 1986, *An introduction to unification-based approaches to grammar*. Chicago: Chicago University Press (CSLI Publication).

Sekine, S., Carroll, J.J., Ananiadou, S., Tsuji J. 1992. "Automatic Learning for Semantic Collocation." In *Proceedings of the 3d Conference on Applied Natural Language Processing*, ACL:104-110.

Sidner, C.L. 1981. "Focusing for interpretation of pronouns." In *American Journal of Computational Linguistics*, 7(4): 217-231.

Sidner, C.L. 1983. "Focusing in the comprehension of definite anaphora." In *Computational Models of Discourse*, M.Brady and R.C.Berwick (eds.), 267-330. Cambridge, MA: MIT Press.

Silberztein, M. 1993. *Dictionnaires Electroniques et Analyse Lexicale du Français - Le Système INTEX*. Masson.

Sinclair, J.M. 1987. "Collocation: A progress report." In *Language Topics: Essays in Honour of Michael Halliday, volume II*, R. Steele and T. Threadgold (eds.), 319-31. Philadelphia: John Benjamins.

Sinclair, J. and Coulthard, M. 1975. *Towards an analysis of Discourse*. Oxford: Oxford University Press.

Smith, R. and Hipp, R.D. 1994. *Spoken Natural Language Dialog Systems: A Practical Approach*. Oxford University Press.

Soderland, S. 1999. "Learning Information Extraction Rules for Semi-Structured and Free Text." In *Machine Learning: Special Issue on Natural Language Learning* , 34 (1): 233-272.

Stolcke, A. 1997. "Linguistic knowledge and empirical methods in speech recognition." In *AI Magazine*, 18(4): 25-31.

Tannen, D. 1984. *Conversational Style: Analyzing Talk among Friends*. Ablex.

Tomita, M. 1986. *Efficient Parsing for Natural Language: a Fast Algorithm for Practical Systems*. Boston: Kluwer.

Tomita, M. 1987. “An efficient augmented context-free parsing algorithm.” In *Computational Linguistics*, 13(1-2):31-46.

Traum, D.R. 1993. "Rhetorical Relations, Action and Intentionality in Conversation." In *Proceedings ACL SIG Workshop on Intentionality and Structure in Discourse Relations*, 132-135.

Wanner, E. and Maratsos, H. 1978."An ATN approach to comprehension." In *Linguistic Theory and Psychological Reality* M. Halle, J. Bresnan, and G. Miller (eds.), 119-161. Cambridge, MA: MIT Press.

Winograd, T. 1971. *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*. PhD Thesis MAC TR-84

Woods, W.A. 1970. "Transition Network Grammars for Natural Language Analysis". In *Communications of ACM* 13: 591-606.