Grammar and NLP

Inasmuch as computational models include mathematical linguistic theory and stategies for the use of this theory, Grammar plays a role in computational linguistics. Principle-based parsers (Berwick l991, and related works) include the axiomatization of our knowledge of language, through sets of rules or principles, and control strategies representing our use of this knowledge. Interestingly, different strategies (left to right, ascending or descending, deterministic or non-deterministic, sequential or parallel) are used to associate one or more structures with a linguistic expression as input. The research in this area aims at reducing the possible analyses that can be generated by the parser to the one or the very few that correspond to the actual structure of the linguistic expressions under analysis.

In the history of the field several grammatical notions have been indirectly embedded in parsers. This is the case, for example, in the treatment of subjacency in Marcus's deterministic parser (1980) and in the treatment of c-command in Berwick and Weinberg's parser (1984). Currently, several researchers are developing more sophisticated parsers based on the satisfaction of feature-based constraints Stabler (1997), Harkema (2000), Niyoqi and Berwick (2002), Di Sciullo and Fong (2000, 2001). The computational models are not of the 'generate and filter' incorporating GB grammars (Chomsky, 1981) but rather of the check and generate type, incorporating Minimalists grammars (Chomsky 1995).

The importance of grammar in Natural language processing is evident, given the central role played by configurational relations, such as asymmetrical c-command, in principle-based parsers. However, the current problem in computational linguistics is to formulate a model that can process linguistic expressions efficiently and quickly. It is likely that the use of a simplified model, such as the Minimalist or asymmetry-based models, based on the generation and recovery of local grammatical relations, constitutes the first step towards the resolution of over-generation and speed.

The question arises whether the use of shallow and partial NLP techniques in heavily linguistically difficult tasks such as the one posed by summarization, question answering, and topic identification, will bring to fore more accurate results than robust parsing. In partial parsing and partial semantic interpretation, not all syntactic and semantic relations will be detected and encoded appropriately. Partial semantic interpretation can be carried out in order to execute anaphora resolution, as anaphora is based on a limited subset of grammatical elements and relations. A related question is whether purely statistical methods are sufficiently reliable to perform, without the use of syntactic parsing and semantic interpretation, tasks such as topic identification, summarization and translation tools.

This collection of papers brings together contributions in different areas of natural language processing. Some contributions are purely theoretical, while others aim at the resolution of more concrete problems. However, the contributions converge in one point: the centrality of grammar in natural language processing. They are briefly summarized in

the following paragraphs.

In 'Morpho-Syntactic Parsing', Anna Maria Di Sciullo and Sandiway Fong describe an implemented bottom-up parser for a theory of morphological selection in the Asymmetry framework (Di Sciullo, l995, 2001). Core lexical properties of derivational affixes, generally encoded in terms of subcategorization frames, are articulated in terms of asymmetrical relations. The selection of affixes is encoded in a uniform specifier-head-complement configuration, and predictions can be made with respect to composition and linking relation. Thus, the so-called lexical gaps fall out from the theory. Starting from a review of the underlying Asymmetry framework, they consider the computational implications of three different implementations. In particular, we examine the effect on bottom-up parsing from varying the specifier-head-complement order. Furthermore, computational motivation for the logical separation of overt and covert affixation is provided.

In "A Minimalist Implementation of Verb Subcategorization", Sourabh Niyoqi and Robert Berwick challenge the traditional accounts of verb subcategorization, from the classic work of Fillmore on requiring either a considerable number of syntactic rules to account for diverse sentence constructions, including cross-language variation, or else complex *linking rules* mapping the thematic roles of semantic event templates with possible syntactic forms. In this paper, we exhibit a third approach:we implement, via an explicit parser and lexicon, the *incorporation theory* of Hale and Keyser (1993, 1998) to systematically cover most patterns in *English Verb Classes and Alternations* (Levin l993), typically using only 1 or 2 lexical entries per verb to subsume a large number of syntactic constructions *and* also most information typically contained in semantic event templates, and. further, replacing the notion of “thematic roles"’ with precise structural configurations. The implemented parser uses the *merge* and *move* operations formalized by Stabler (1997) in the minimalist framework of Chomsky (2001).

In "Top-Down Recognition of Minimalist Languages", Henk Harkema describes a top-down recognition method for languages generated by Minimalist Grammars. Minimalist Grammars are formal grammars that incorporate certain aspects of current transformational linguistic theories: phrases are derived by applying structure building functions to lexical items and intermediate structures, and the applicability of these functions is determined by the syntactic features of the structures involved. The recognition method presented in this paper reduces phrase structures to simple expressions that encode the behavior of these structures with regard to the structure building functions of the grammar.

In Relative Clause Attachment and Anaphora: Conflicts in Grammar and Parser Architectures, Rodolfo Delmonte, is concerned with the use of shallow and partial NLP techniques in heavily linguistically demanding tasks such as the one posed by summarization. This approach should not be taken as an alternative way of coping with the same problems by means of a complete system of text understanding and summarization, but as a proposal in line with current NLP research in unrestricted texts that assumes that partial processing can be more suitable and nonetheless useful for better

satisfaction of certain requirements. In particular, morphological analysis is a prerequisite in order to better cope with Out of Vocabulary Words(OOW) by means of guessing techniques based on morphological rules; statistical processing is then assumed to be useful for tagging disambiguation. He concentrates his attention to parsing, and in particular to problems related to challenges coming from attachment of structurally ambiguous constituents such as prepositional phrases and relative clauses.

In "A prototype for a computational analysis of Modern Greek Compounds" Angela Ralli and Eleni Galiotou, point out to the fact that compounds are usually excluded by attempts to develop a computational processor of Modem Greek (MG) morphology. The reason for such an exclusion relies on the difficulties posed by their internal structure. Among other things, MG compounds display the following characteristics that are not easy to deal with in a linguistically-sound computational manner. They belong to two types of internal constituency: [stem stem] and [stem word]. Each type is related to certain peculiarities with respect to their inflectional pattern as well as to their stress pattern. MG compounds are usually right-headed, but the issue of headedness becomes problematic in a considerable number of exocentric compounds as well as in compounds displaying no fixed order between the internal constituents. They propose to analyze MG compounds by using the formalism of KIMM0, a well-known tool of morphological analysis. KIMMO is adapted to the requirements of MG morphology.

In "FipsC : A Syntax-Based System for Terminology Extractions", Jean- Philippe Goldman and Eric Wehrli describe FipsCo, a system developed to extract collocations and multi-word expressions from large text corpora, using a powerful syntactic parser. We will show that syntactically analyzed corpora constitute a more reliable basis for MWEs extraction than raw (or shallow parsed) texts. Preliminary results supporting this view are presented based on both French and English corpora.

In "Approaches for Learning Constraint Dependency Grammar from Corpora", M. P. Harper, W. Wang, and C. M. White evaluate two methods of learning constraint dependency grammars from corpora: one uses the sentences directly and the other uses sub grammar expanded sentences. Leaning curves and test set parsing results show that grammars generated directly from sentences have a low degree of parse ambiguity but at a cost of a slow learning rate and less grammar generality. Augmenting these sentences with sub grammars dramatically improves the grammar learning rate and generality with very little increase in parse ambiguity.

In "Parsing with constraint graphs: a flexible representation for robust parsing" Philippe Blache presents an approach for natural language processing relying on the idea that linguistic information can be represented by means of graphs. One of the interests of this approach is that a linguistic structure does not necessarily cover the entire input: in case of ill-formed sentences, the syntactic structure can for example be constituted by a set of sub-graphs. We propose in this paper a formalism, called Property Grammars, representing the linguistic information by means of constraints represented as graphs. This approach is very flexible and reusable for several applications such as shallow

parsing(which consists in using a subset of graphs) or corpus annotation (annotating a text with graph labels). More importantly, it provides a framework for integrating different levels of linguistic information.

In "Heuristic Syllabification and Statistical Syllable-Based Modeling for Speech-Input Topic Identification", Pierre Ouellet and Pierre Dumouchel describe a heuristic syllabification method and the use of a statistical syllable *n*-gram language model for discriminating between a closed set of topics. The syllabification method works by assigning costs to consonant clusters and then splitting the clusters where the cost in minimized. They apply the syllabification on a pronunciation dictionary which maps words to phone sequences; the result is then used for producing an inventory of syllables and to transform word- or phone-level transcriptions into syllable-level transcriptions. In topic identification (TID), each topic is represented by a backoff syllable *n*-gram statistical model trained on a sample of syllabified texts deemed representative of the topic. Identifying the topic of an unknown utterance is performed by transcribing it into syllables using a speech recognizer; the transcription is then evaluated (scored) with respect to each topic model, and the topic for which the score is highest is deemed the most likely. The training texts for topic models can to produced manually, or automatically using speech recognition. We evaluate the accuracy of the syllabification by referring to the Celex lexicon, which includes syllabifications; topic identification performance is evaluated on a subset of the Switchboard corpus.

In "Experiments with a Probabilistic Translation Assistant: would Statistical Grammar help?", Philippe Langlais presents the latest version of TRANSTYPE, a prototype which implements a novel approach to interactive machine translation; namely Target Text Mediated Interactive Machine Translation. He first gives an overview of the core system TRANSTYPE relies on. Then, he summarizes the results of an *in-situ* evaluation. Finally, he discusses the potential benefits that could be gained by integrating a probabilistic grammar in this approach.

In "Hybrid Approaches in Machine Translation: From Craft to Linguistic Engineering", Claude Coulombe, . . . claim that in order to attain the status of a true technology, MT must move beyond mere demonstration of pure theories and clever tinkering to linguistic engineering. On the one hand, MT-applied linguistic engineering must abandon craft, its toy-systems that will never become true applications due to insufficient theoretical foundations. On the other hand, linguistic engineering must also avoid approaches that mainly look to reproduce how the human mind operates for the same reasons that airplanes do not fly like birds. Linguistic engineering is characterized by its eclecticism, rejecting the infamous theoretical “schools” and theories that are supported by ideologies. MT-applied linguistic engineering systematically and incrementally builds a set of tools, methods and know-how based on proven theoretical results. Linguistic engineering integrates the results of various theories and paradigms into an essentially hybrid approach such as the use of large-coverage symbolic parsers, integrated semantic knowledge that is partially symbolic and partially based on statistical models and, finally, the user's interactive participation.

The papers assembled in this collection came out from the First Conference of the Federation on Natural Language Processing, held at the Université du Québec à Montréal in October 200l. Many thanks to the members of the Natural Language Processing project for their help in the organization of this conference. Their names are listed in the project homepage www.unites.uqam.ca/graln I am grateful to Valorisation-Recherche Québec for the financial support to the Natural Language Processing project, no 2200-006, out of which this first conference came through.