Chapter VII
Natural Language Parsing:
Perspectives from Contemporary Biolinguistics

Pauli Brattico
University of Jyväskylä, Finland

Mikko Määttä
University of Helsinki, Finland

ABSTRACT

Automatic natural language processing captures a lion's share of the attention in open information management. In one way or another, many applications have to deal with natural language input. In this chapter the authors investigate the problem of natural language parsing from the perspective of biolinguistics. They argue that the human mind succeeds in the parsing task without the help of language-specific rules of parsing and language-specific rules of grammar. Instead, there is a universal parser incorporating a universal grammar. The main argument comes from language acquisition: Children cannot learn language specific parsing rules by rule induction due to the complexity of unconstrained inductive learning. They suggest that the universal parser presents a manageable solution to the problem of automatic natural language processing when compared with parsers tinkered for specific purposes. A model for a completely language independent parser is presented, taking a recent minimalist theory as a starting point.

INTRODUCTION

Natural language parsing constitutes a skill that has been the target of automatization since the emergence of modern digital computers. Open information management environments are no exception. Whatever the intended application, the undertaking is nontrivial due to several features of natural languages, such

as ambiguities and phonologically empty elements. A specialized scientific literature and parsing technology has thus emerged to tackle various aspects of this problem (Aho & Ullman, 1972). Some parsing algorithms supply parse trees for any context-free language; others work with a more restricted set of grammars or language fragments and could be conceived as special purpose algorithms or instances of "shallow" (quick and dirty) parsing (Abney, 1996; Argamon et al., 1998; Ramshaw & Marcus, 1995).

In this chapter we adopt a supplementary perspective to the problem of parsing, that of biolinguistics. Biolinguistics takes the human ability to use language to be a specialized faculty of the brain. The aim of this chapter is to explore the implications of this view for the study of (automatic) natural language parsing. More specifically, we argue for the existence of a completely language independent and universal (innately given) parser that is part of the language faculty and describe some of the properties this universal parser ought to have. The core of the parser is based on the minimalist theory of grammar (Chomsky, 1995), while we suggest that the problem of computational complexity is dealt with an unsupervised extraction of templates for partial parse trees corresponding to skill automatization.

BACKGROUND

In linguistics, as in the field of parsing technologies mentioned above, there are basically two different perspectives one can assume. First, one can concentrate on the detailed description of a specific language. There are around 6000 languages spoken around the world, each with its own intricate rules of construction, vocabulary, and stylistic rules (Comrie, 2001; Greenberg, 1963). Each individual grammar can be further dissolved into several interacting levels, such as semantics, syntax, morphology, morphosyntax, phonology and phonetics. It is thus possible to develop specialized grammatical systems, specialized scientific techniques, and nomenclature for the description of different languages and their subcomponents. The first attempts in this direction were made already two thousand years ago, as in the case of Panini's grammar for Sanskrit. Moreover, such descriptions can achieve considerable precision due to the fact that the different levels of human language consist of smaller units that are put together according to well-defined, combinatorial rules.

Some 50 years ago linguists studying natural language grammars began to pursue a different track. Instead of developing new technologies and methods for the description of individual languages and their subcomponents, they studied the method 2-4-year old children use in order to break the code of their own native language. It is well-known that this happens effortlessly, without much linguistic input or cognitive sophistication, and in effect in a couple of years (Chomsky, 1969; Graffi, 2001; Marcus, 1993; Moro, 2008; Pinker, 1994). When the problem is set this way, it becomes feasible to find out the cognitive representational apparatus that the child uses when acquiring and using her own language(s).

One could entertain the possibility that the child acquires her native language merely by paying sufficient attention and drawing inferences. It is obvious that since languages differ from each other, some learning must be going on (Yang, 2002). Under this scheme, the child uses a primitive representational system sufficiently rich to express the rules of possible languages and then tries to formulate the rules of the target language based on the linguistic and extra-linguistic input. Since it is known that children learn languages without being explicitly taught, rule formulation must be based on an inductive logic, or rule induction. Rule induction allows the child to generalize beyond what she has perceived in the past. It is sometimes assumed that rule induction is implemented by general cognitive skills, such as pattern recognition, statistical inference or analogical reasoning (Bates & Elman, 1996; Deacon, 1997;