Natural Language Processing and Biological Methods

Gemma Bel Enguix
Rovira i Virgili University, Spain

M. Dolores Jiménez López
Rovira i Virgili University, Spain

INTRODUCTION

During the 20th century, biology—especially molecular biology—has become a pilot science, so that many disciplines have formulated their theories under models taken from biology. Computer science has become almost a bio-inspired field thanks to the great development of natural computing and DNA computing.

From linguistics, interactions with biology have not been frequent during the 20th century. Nevertheless, because of the "linguistic" consideration of the genetic code, molecular biology has taken several models from formal language theory in order to explain the structure and working of DNA. Such attempts have been focused in the design of grammar-based approaches to define a combinatorics in protein and DNA sequences (Searls, 1993). Also linguistics of natural language has made some contributions in this field by means of Collado (1989), who applied generativist approaches to the analysis of the genetic code.

On the other hand, and only from theoretical interest a strictly, several attempts of establishing structural parallelisms between DNA sequences and verbal language have been performed (Jakobson, 1973, Marcus, 1998, Ji, 2002). However, there is a lack of theory on the attempt of explaining the structure of human language from the results of the semiosis of the genetic code. And this is probably the only arrow that remains incomplete in order to close the path between computer science, molecular biology, biosemiotics and linguistics.

Natural Language Processing (NLP) –a subfield of Artificial Intelligence that concerns the automated generation and understanding of natural languages— can take great advantage of the structural and "semantic" similarities between those codes. Specifically, taking the systemic code units and methods of combination of the genetic code, the methods of such entity can be translated to the study of natural language. Therefore, NLP could become another "bio-inspired" science, by means of theoretical computer science, that provides the theoretical tools and formalizations which are necessary for approaching such exchange of methodology.

In this way, we obtain a theoretical framework where biology, NLP and computer science exchange methods and interact, thanks to the semiotic parallelism between the genetic code and natural language.

BACKGROUND

Most current natural language approaches show several facts that somehow invite to the search of new formalisms to account in a simpler and more natural way for natural languages. Two main facts lead us to look for a more natural computational system to give a formal account of natural languages: a) natural language sentences cannot be placed in any of the families of the Chomsky hierarchy (Chomsky, 1956) in which current computational models are basically based, and b) rewriting methods used in a large number of natural language approaches seem to be not very adequate, from a cognitive perspective, to account for the processing of language.

Now, if to these we add (1) that languages that have been generated following a molecular computational model are placed in-between Context-Sensitive and Context-Free families; (2) that genetic model offers simpler alternatives to the rewriting rules; (3) and that genetics is a natural informational system as natural language is, we have the ideal scene to propose biological models in NLP.

The idea of using biological methods in the description and processing of natural languages is backed up by a long tradition of interchanging methods in biology and natural/formal language theory:

1. **Results and methods in the field of formal language theory have been applied to biology**:

(1) Pawlak (1965) dependency grammars as an approach in the study of protein formation; (2) transformational grammars for modeling gene regulations (Collado, 1989); (3) stochastic context-free grammars for modeling RNA (Sakakibara et al., 1994); (4) definite clause grammars and cut grammars to investigate gene structure and mutations and rearrangement in it (Searls, 1989); (5) tree-adjoining grammars for predicting RNA structure of biological data (Uemura et al., 1999).

2. **Natural languages as models for biology:** (1) Watson (1968) understanding of heredity as a form of communication; (2) Asimov (1968) idea that nucleotide bases are letters and they form an alphabet; (3) Jacob (1970) consideration that the sense of the genetic message is given by the combination of its signs in words and by the arrangement of words in phrases; (4) Jakobson (1970) ideas about taking the nucleotide bases as phonemes of the genetic code or about the binary oppositions in phonemes and in the nucleic code.
3. **Biological ideas in linguistics:** (1) the "tree model" proposed by Schleicher (1863); (2) the "wave model" due to Schmidt (1872); (3) the "geometric network model" proposed by Forster (1997); or (3) the naturalistic metaphor in Linguistics defended by Jakobson (1970, 1973).
4. **Using DNA as a support for computation** is the basic idea of Molecular Computing (Păun et al., 1998). Speculations about this possibility can be found in Feynman (1961), Bennett (1973) and Conrad (1995).

BIOLOGICAL METHODS IN NLP

Here, we present an overview of different bio-inspired methods that during the last years have been successfully applied to several NLP issues, from syntax to pragmatics. Those methods are taken mainly from computer science and are basically the following: *DNA computing, membrane computing and networks of evolutionary processors*.

DNA Computing

One of the most developed lines of research in natural computing is the named molecular computing, a model based on molecular biology, which arose mainly after Adleman (1994). An active area in molecular computing is DNA computing (Păun et al., 1998) inspired in the way that DNA perform operations to generate, replicate or change the configuration of the strings.

Application of molecular computing methods to natural language syntax gives rise to **molecular syntax** (Bel-Enguix & Jiménez-López, 2005a). Molecular syntax takes as a model two types of mechanisms used in biology in order to modify or generate DNA sequences: *mutations* and *splicing*. Mutations refer to changes performed in a linguistic string, being this a phrase, sentence or text. Splicing is a process carried out involving two or more linguistic sequences. It is a good framework for approaching syntax, both from the sentential or dialogical perspective.

Methods used by molecular syntax are based on basic genetic processes: *cut*, *paste*, *delete* and *move*. Combining these elementary rules most of the complex structures of natural language can be obtained, with a high degree of s*implicity*.

This approach is a test of the generative power of splicing for syntax. It seems, according to the results achieved, that splicing is quite powerful for generating, in a very simple way, most of the patterns of the traditional syntax. Moreover, the new perspectives and results it provides, could mean a transformation in the general perspective of syntax.

From here, we think that bio-NLP, applied in a methodological and clear way, is a powerful and simple model that can be very useful to a) formulate some systems capable of generating the larger part of structures of language, and b) define a formalization that can be implemented and may be able to describe and predict the behavior of natural language structures.

Membrane Computing

Membrane Systems (MS) (Păun, 2000) are models of computation inspired by some basic features of biological membranes. They can be viewed as a new paradigm in the field of natural computing based on the functioning of membranes inside the cell. MS can be used as generative, computing or decidability devices. This new computing model has several intrinsically