Natural Language And The Genetic Code: From The Semiotic Analogy To Biolinguistics

Gemma Bel Enguix, M. Dolores Jiménez-López
Universitat Rovira i Virgili, Tarragona (Spain)

Abstract
With the discovery of the DNA structure (Watson and Crick, 1953), the idea of DNA as a linguistic code arose (Monod, 1970). Many researchers have considered DNA as a language, pointing out the semiotic parallelism between genetic code and natural language. This idea had been discussed, almost dismissed and somehow accepted. This paper does not claim that the genetic code is a linguistic structure, but it highlights several important semiotic analogies between DNA and verbal language. Genetic code and natural language share a number of units, structures and operations. The syntactic and semantic parallelisms between those codes should lead to a methodological exchange between biology, linguistics and semiotics. During the 20th century, 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. Biology and semiotics are two different sciences challenged by the same common goal of deciphering the codes of the nature. Linguistics could become another «bio-inspired» science by taking advantage of the structural and «semantic» similarities between the genetic code and natural language. Biological methods coming from computer science can be very useful in the field of linguistics, since they provide flexible and intuitive tools for describing natural languages. In this way, we obtain a theoretical framework where biology, linguistics and computer science exchange methods and interact, thanks to the semiotic parallelism between the genetic code a natural language. The influence of the semiotics of the genetic code in linguistics is parallel to the need of achieving an implementable formal description of natural language. In this paper we present an overview of different bio-inspired methods — from theoretical computer science — that during the last years have been successfully applied to several linguistics issues, from syntax to pragmatics.

Proceedings of the 10th World Congress of the International Association for Semiotic Studies (IASS/AIS)
Universidade da Coruña (España / Spain), 2012. ISBN: 978-84-9749-522-6 Pp. 771-780

1. INTRODUCTION

> The 1969 edition of the Encyclopaedia Britannica... has twenty-three alphabetically ordered volumes of articles that altogether contain about two hundred million letters. Most of our cells have pairs of each chromosome; the twenty-three pairs contain about six billion base pairs, so a single human genome is a text about three billion letters long. In each volume of an encyclopaedia the string of letters is organized into thousands of separate articles about discrete subjects. In the long string of DNA letters in a chromosome, there are thousands of stretches of letters — genes — that each address a particular topic: how to make a particular protein perhaps, or how to find another stretch of DNA. The index of the Britannica has about two hundred thousand entries. Altogether, the chromosomes of a person contain at least a hundred thousand genes [...]
>
> The topics in an encyclopaedia are ordered by their spelling, rather than their meaning, so that a reader can quickly find the right place without knowing much about any given topic. Sometimes, especially when the topic words themselves have a common origin, adjacent topics may be related by meaning. Similarly, the genes in each chromosome are present in a precise order that seems usually — but not always — to be arbitrary; when genes that do similar things are next to each other, they are likely to be related by common descent from a single gene. The orderliness of genes in each chromosome shows up during the division of one cell into two, when the DNA of each chromosome coils up on itself, giving each one a characteristic set of crosswise bands. Each band marks the presence of a few hundred genes; the pattern of bands on each chromosome is very regular from person to person.
>
> Pollack (1994: 20-21)

Robert Pollack describes the genetic code as a large text, following a parallelism that has been widely used from the discovery of the structure of DNA (Watson & Crick 1953). This paper develops this recurrent metaphor in the literature of codes and analyzes the possible impact of this analogy in the design of new methodologies for linguistics. This can be understood just as a game, but it can be also relevant from a computational and linguistic point of view. It seems that using natural models —very well studied by mathematics— for the description of natural entities, like natural language, can give some advantage for achieving simpler and easily implementable approaches. If human language can be better understood thanks to the collaboration of the genetic code, then a new relationship between biology and linguistics would be established, giving rise to a new sense of biolinguistics, as we will highlight in Section 4.

In Section 2, we explain a general correspondence between the syntactic units of the genetic code and natural language. Section 3 deals with some semantic problems that can be found in DNA/RNA interpretation. Finally, section 4 suggests some clues on how methodological interaction between molecular biology, linguistics and computer science can help these sciences to develop new paradigms of research.

2. SYNTACTIC ANALOGIES

Natural language and the genetic code seem to share a syntactical combinatorial structure. The purpose of this section is to establish a general analogy between the main units of both systems.

The first relationship was introduced by Jakobson (1973: 52-53). The linguist suggests a first interpretation that correlates the elements of the genetic code and verbal language in the following way:

- *Nucleotides - phonemes/letters*
- *Codons - words*
- *Lexicon - 64 codons*

The relationship between *nucleotides* and *letters* (or phonemes) is motivated by the fact that nucleotides are units that cannot work isolated, but in combination with others, like phonemes.

Assuming that nucleotides are minimal non-meaningful units of the genetic code, and phonemes are minimal non-meaningful units of the verbal language, then the next step consists in looking for a class of units with meaning obtained from the combination of the smaller non-meaningful ones. These units are *codons*, groups of three nucleotides that code for amino acids.

Jakobson (1973) points out the relationship between *codons* and *words*, and Marcus (1995) relates *codons* and *morphemes*. Both analogies can be useful.

Jakobson also suggests a classification of codons, distinguishing between those which have «lexical» meaning («signification propre») and those which have «grammatical» meaning. This is a non-developed but quite interesting idea. We know there are several codons that only codify proteins for regulation. This means that the goal of some genes is to regulate the transcription of other genes. These ones could be considered parts of the genetic code with an internal or grammatical meaning. However, some other codons (and genes) are the ones that are finally expressed. These could be said to have an external or lexical meaning.

Going on with Jakobson's analogy, the Russian linguist postulated that the *genetic lexicon* consisted of 64 words. In fact, these 64 «words» of DNA only refer to codons. Since codons are words of three letters (nucleotides), and there are four different nucleotides (thymine, guanine, cytosine, adenine), the number of codons is $V_4^3=64$. Despite their small number, 64, many of them are synonymous because there exist only twenty-one meanings, the number of existing amino acids.

Concerning the major syntactic units, Jakobson (1973: 53) considers *cistrons* and *operons*. Since the word *cistron* means essentially the same as *gene*, both units, genes and operons, can be taken for the analogy between major units.

Therefore, we have to explore the possible relationship between *operons/genes* and *sentences* — the major syntactic units. This is an approach already suggested by several biologists. Collado (1989) shows how sentences can be related to syntax of the genetic code. The whole work of this author is an attempt to establish a generative mechanism based on linguistics to explain the regulation. It seems, according to these studies, that the operon is the genetic concept that better fits the sentence.

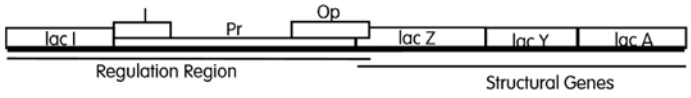

FIGURE 1: OPERON LAC

Operons (Jacob & Monod 1963) are considered as «units of transcriptive activity [...] coordinated by a genetic element». There are two essential elements in the operons: the *promoter* (regulating area) and the area of *structural genes*. Collado et al (1998), following the above description, consider two main parts in an operon (TU: Transcriptional Unit): Pr'' (regulating unit) and S' (structural area). And from here, they find a grammar:

$$\begin{aligned} TU= &\quad Pr'' \; S' \\ Pr''= &\quad Pr' \; S' \\ S'= &\quad S^1 \; S^2 \; S^n \end{aligned}$$

The grammar generates the tree in Figure 2.

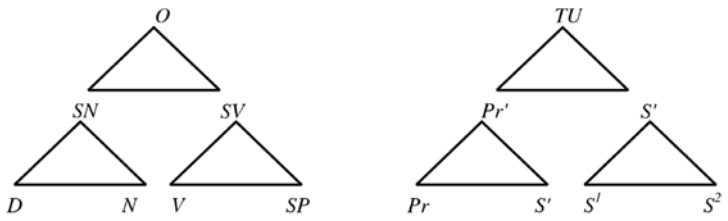

FIGURE 2: SENTENCE AND OPERON GENERATION TREES

Finally, there are in the genetic code some complex structures TU - c = TU TU that are parallel to the complex sentences. Representation of such complexity can be seen in Figure 3.

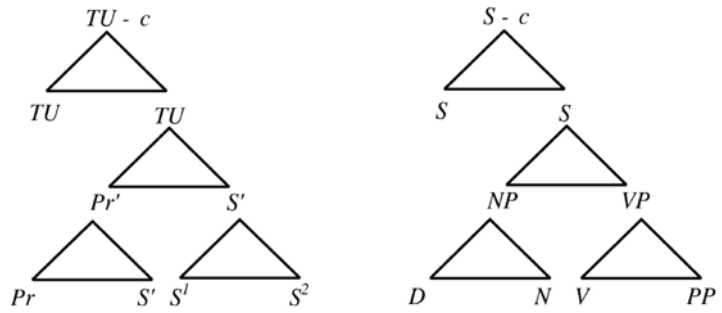

FIGURE 3: COMPLEX SENTENCE AND OPERON

The last analogy can be given by regulons, or networks of regulation in the genetic code. Following Neidhardt (1996), «*a regulon is defined as a network of operons with a common regulator, usually a protein repressor or activator that recognizes a particular sequence common to the controlling regions of these operons*». This author asserts that «*a given regulon might be induced by more than one environmental condition, a given environmental condition might induce more than one regulon, and certain stimuli might induce only a subset of the operons*

of a regulon», and later that «*multiple regulators might permit subroutines of the same regulon to be activated independently, in accord with either the strength of a signal or the presence of different signals*».

Therefore, we should think that the relations among the operons that act within a regulon might be multiple, and it is the same author who gives an example of the regulatory system of nitrogen described as a «bicyclic cascade» (Figure 4).

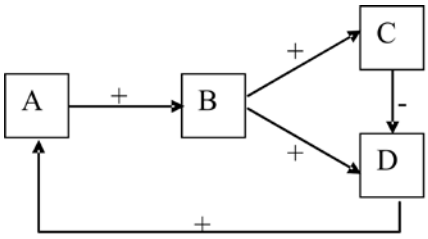

Figure 4: Regulatory System of Nitrogen

In the example, the functions of connection among the operons within the regulon are reduced to the meanings *on* and *off*. These concepts can be expanded introducing semantic connections to large sequences. This is very similar to the concept of texts in that, more than a row of sentences, they are a whole related with different connectors. In respect to these particles, the sense of the text can vary.

The identification regulon/text is completely new, and we suggest a deeper study of regulatory proteins in order to get a more comprehensive knowledge of topicalization in macro-structures.

The mapping between linguistic macrostructures and the structure of regulons is a totally new idea that implies the assumption of pragmatic and contextual categories in the description of genetic code and the possible formalization of influences inside a text in discourse linguistics.

Summing up this first part of the work, we suggest the analogy of syntactic units of genetic code and verbal language that can be seen in Figure 5.

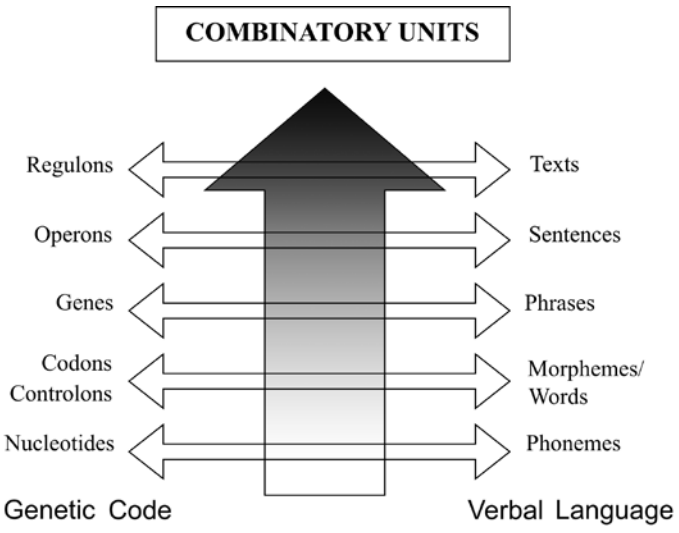

Figure 5: Syntactic Units

3. SEMANTIC ANALOGIES

In this section, we study three «semantic» phenomena that can be found in the genetic code: *synonymy, polysemy, ambiguity*. The reasons of having such phenomena are not clear, but it seems that they can be traces of evolution from ancient configurations of DNA. A synchronic explanation could refer to the protection of the information of the code in front of frequent mutations.

3.1 Synonymy

Synonymy is an ultra represented phenomenon within the genetic code. The clearest and best-known example is found in the structural area: the same two codons can represent the same amino acid. Therefore, the 64 known codons only correspond to 21 meanings. However, biologists do not talk about synonymy but redundancy or degeneracy.

3.2 Polysemy

Polysemy is a linguistic-like phenomenon that plays an important role within the genetic language. It refers to the different meaning that some sequences can have depending on where they are located. The TATAAT sequence can be a good example of this. If it is placed in the regulating area, it means «beginning». If it is in a structural gene, TAT AAT are interpreted as codons, and the sequence has to be read as «tyrosine» «asparagine».

Without the distinction between «regulating area» and «structural gene», the sequence would be ambiguous and the genetic code would be dangerous.

Therefore, it could be said that there is a phenomenon of polysemy sensitive to the site in biology, what gives to the context a very important role in molecular biology. That could be expressed in the following «formula»:

DNA fragment meaning = nucleotide sequence + context

Which is somehow an extrapolation to what happens to a phrase and its function within the verbal language:

Phrase meaning = internal structure + context

3.3 Ambiguity

The genetic code also implies ambiguity at two levels: one of them refers to the relation between a gene and a protein and is very specific; the other one refers to the relation between the genotype and the phenotype and is very global.

3.3.1 From the gene to the protein and from the protein to the gene

This type of ambiguity is explained by Pollack as follows:

> The polysemy interjected by editing adds to an ambiguity already brought on by redundancy in the genetic code: two or more DNA sequences may in fact encode the same protein by using redundant codons to order the assembly of the same amino acid sequence. This ambi-

> guity is one way: looking at a protein's amino acid sequence, it is not possible to predict the DNA sequence that encoded it. In contrast, every messenger sequence is unambiguous: it leads to only one protein. A one-way ambiguity of this sort is called a degeneracy, and we say that the genetic code is degenerate.
> Pollack (1994: 83-84)

This author bases his theory on the concept of synonymy or degeneracy, which we have already explained in order to show the existence of ambiguity. However, he warns about the necessity of observing the message from two different points of view:

1. Given a gene, the protein that synthesizes is not ambiguous. Every DNA sequence codes for only one sequence of amino acids.

2. Given a protein, the DNA sequence that has synthesized it cannot be inferred. Here is the ambiguity.

Synonymy, inversely seen, gives many input sequences for an amino acid string. For example, given a sequence Arg | Ser | Pro | Ile | Met | Val | Val | Glu | Lys, which combination of nucleotides could have been its origin? It is easy to check it knowing the codons that produce every amino acid:

Arg	is the meaning of CGU, CGC, CGA, CGG, AGA, AGG,
Ser	is the meaning of UCU, UCC, UCA, UCG, AGU, AGC,
Pro	is the meaning of CCU, CCC, CCA, CAG,
Ile	is the meaning of AUU, AUC, AUA,
Met	is the meaning of AUG,
Val	is the meaning of GUU, GUC, GUA, GUG,
Glu	is the meaning of GAA, GAG,
Lys	is the meaning of AAA, AAG.

Therefore, given a sequence like this one, with only 8 amino acids, we get almost 7.000 possible combinations that could have produced it!

3.3.2 From the genotype to the phenotype and from the phenotype to the genotype

The same ambiguity can be found in a more general way, going from the genotype to the phenotype:

> A crucial ambiguity allows different genotypes to generate the same phenotype. This ambiguity arises whenever either one or two copies of an active allele produce the same phenotype: a person with the phenotype of type A blood may be carrying either one A and one O allele or two A alleles. Usually, an allele that has no effect on the phenotype unless it is present in both copies is inactive. For example, while the A and B alleles are responsible for two slightly different chemicals on the surface of a red blood cell, the O allele is silent and inactive, neither adding nor subtracting anything. In this silence it offers no barrier to the function of a second allele of either the A or the B type, so genotypes of AO

> and BO generate phenotypes indistinguishable form AA and BB, respectively. An active allele producing the same phenotype, whether present on one chromosome or both — like the one for blood type A-- is called dominant.
> Pollack (1994: 42)

The problem that this text introduces is based on the notion of *allele*. Two alleles are two different copies of the same gene. Each person has two alleles of each one of the genes, one is maternal and the other is paternal, but many times one of them remains aside: this is what we call a *recessive gene*. The example used by Pollack refers to blood. Allele O is recessive. If a person has an O copy and an A copy, his/her phenotype and his/her type of blood will be A. If he/she has OO will be O, and if he/she has AA will be A. The problem is the following: for an O phenotype, the genotype is obviously OO. But a phenotype A can have two possible genotypes, AA or AO. Therefore, the A phenotype is ambiguous.

The same happens with many more examples like that one observed by Mendel with a pea's peel (Mendel 1865). A rough pea corresponds to a rough phenotype. However, a smooth peel may correspond to a rough smooth genotype or smooth smooth genotype because «rough peel» is a recessive gene.

From all this, one draws the conclusion that in studying the phenotype of a person there are thousands of genotypes that may correspond to him/her because many pairs of different genes give equal results.

4. DISCUSSION: A NEW INTERDISCIPLINARY FRAMEWORK OF RESEARCH

This paper presents some structural and systematic parallelisms between the genetic code and verbal language with the goal of promoting the application of similar methods of research in molecular biology and linguistics, giving rise to a new meaning for the word *biolinguistics*: a science approaching linguistics with methods imported from biology and formalized by computer science. The main research interest of this new biolinguistics is *linguistics*, the main model comes from *biology* and *computer science* is the way to transfer biological concepts to formal linguistics and, at the same time, it provides a new methodology of research.

Indeed, we want to propose a new field of research, in a shared space between the boundaries of these three areas that seem to have a crucial impact in the future of the society: linguistics, biology and computer science.

Molecular biology has been the best and more increasingly developed area in the last 50 years, and it seems that most of the future goals of the humanity have something to do with the deciphering of DNA and genetic engineering. In fact, the publication of Darwin's work *The origin of the species*, in 1859, was the first step towards an integrative view of language as a biological construct. Currently, biology — especially molecular biology — has become a pilot science, so that many disciplines have formulated their theories under models taken from biology.

However, linguistics has still the challenge to understand how natural language is acquired, produced and processed, in a multi-lingual society that needs some tools for linguistic

interaction, automatic translation and human-computer interfaces. Up to now, linguistics has not been able to solve these challenges, partly, because of the fail in the models adopted. Indeed, it has been proofed that natural language does not fit in the classical Chomskyan hierarchy of languages the base of mathematical models for linguistics. Moreover, the metaphor of the mind as a computer seems to be almost exhausted. Finally, psychological models have a lack of formalization that prevents them to be conveniently implemented. Surprisingly, and despite the fact that genetic code has been considered a code from the beginning, linguistics has not attempt to construct a new paradigm taking advantage of the great developments in molecular biology. Nevertheless, it is an especially tempting idea, because of the interesting similarities between natural language and the genetic code in different levels.

Probably, the use of natural methods would increase the probabilities of success in the description of natural language, and the use of already formalized by theoretical computer science natural methods would help the necessary implementation of the theory. Natural Language Processing (NLP) can take great advantage of the structural and «semantic» similarities between these 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, a theoretical framework can be drawn, where biology, NLP and computer science exchange methods and interact. And, for linguistics, the bio-inspired computational paradigm can be a powerful tool for explaining language capacity from two different perspectives: *synchrony* —or the view of language as a complex system— and *diachrony* —or language understood as an evolutionary system.

REFERENCES

Collado-Vides, Julio (1989): «A transformation-grammar approach to the study of regulation of gene expression», *Journal of Theoretical Biology,* 136: 403-425.

Collado-Vides, Julio; Gutiérrez-Rios, Rosa María and Bel-Enguix, Gemma (1998): «Networks on transcriptional regulation encoded in a grammatical model», *BioSystems,* 47: 103-118.

Darwin, Charles (1859): *The origin of species.* Available at http://www.infidels.org/library/historical/charles_darwin/origin_of_species/.

Jacob, François and Monod, Jacques (1963): «Genetic repression, allosteric inhibition and cellular differentiation», in Locke, ed., *Cytodifferentiation and Macromolecular Synthesis.* New York: Academic Press, pp. 30-64.

Jakobson, Roman (1973): *Essais de linguistique générale. 2. Rapports Internes et Externes du Language.* Paris: Les Éditions de Minuit.

Marcus, Solomon (1995): *Language, logic, cognition and communication. A semiotic, computational and historical approach,* GRLMC Reports, 1, Tarragona.

Mendel, Gregor (1865): «Experiments in plant hybridization». Available at http://www.netspace.org/MendelWeb.

Monod, Jacques (1970): *Le hasard et la nécessité.* Paris: Éditions du Seuil.

Neidhardt, Frederick C. (1996): «Multigene systems and regulons», in Neidhardt, F., Curtiss, R.I., Gross, C.A., Ingraham, J.L., Riley, M., eds., Escherichia Coli and Salmonella Typhimurium. *Cellular and Molecular Biology, vol. I.* Washington: American Society for Microbiology, pp. 1313-1317.

Pollack, Robert (1994): *Signs of life. The language and meanings of DNA.* London: Penguin Books.

Watson, James and Crick, Francis (1953): «A structure for deoxyribose nucleic acid», *Nature,* 171: 737-738.