Korean Society for Biotechnology and Bioengineering Journal 32(1): 29-34 (2017)
http://dx.doi.org/10.7841/ksbbj.2017.32.1.29 ISSN 1225-7117 / eISSN 2288-8268

Research Paper

생물체의 정보소통전략에 대한 언어학적 접근

김수연[1], 오덕재[2]*

A Linguistic Approach to Communication Strategies of Biological Systems

Soo-Yeon Kim[1] and Duk Jae Oh[2]*

Received: 2 February 2017 / Accepted: 23 February 2017

Abstract: The completion of the Human Genome Project that identified all 3 billion base pairs in the human genome can be seen as a step towards understanding the relay of information and intention within an organism, or in other words, the language of life. The faculty of human language, key to differentiating humans from other animate species, works for conveying information to others by mapping meaning to sound based on syntactic structures. This resemblance between life and language has not gone unnoticed; the literature on RNA transcription and translation research regularly uses linguistic metaphors and the biolinguistic perspective of language has also been studied. By examining the biological characteristics of language and the linguistic characteristics of life, this study aims to identify key mechanisms shared between the two systems in order to promote a stronger connection between them. It furthers this goal by pointing out two general messages to which these mechanisms aim, productivity and accuracy, and discovers what lesson these messages give to a human society geared for sustainability.

Keywords: Cellese, Humanese, Generative grammar, Projection, Recursion, Filtering, Sustainability

1. INTRODUCTION

In both linguistics and biology, sporadic yet meaningful attempts have been made to address parallels between biological and linguistic phenomena, recognizing the need for linguistic analyses of biological processes [1-5] and for analysis of human language as an example of a natural organism [6-10]. In *The Language Instinct*, for instance, Pinker observes the similarity between language systems and genetic systems across the relationship between components and combinations, the methods of combination, and the systems of communication. “In a discrete combinatorial system like language, there can be an unlimited number of completely distinct combinations with an infinite range of properties. Another noteworthy discrete combinatorial system in the natural world is the genetic code in DNA, where four kinds of nucleotides are combined into 64 kinds of codons, and the codons can be strung into an unlimited number of different genes. Many biologists have capitalized on the close parallel between the principles of genetic combination and the principles of grammatical combination. In the technical language of genetics, sequences of DNA are said to contain “letters” and “punctuation”; may be “palindromic,” “meaningless” or “synonymous” are “transcribed” and ‘translated” and are even stored “libraries.”” [11].

A great deal of work has been done to find linguistic organization in macromolecular structures and biological sequences and to use the framework from modern linguistics to describe biological phenomena [12-14]. The existing literature on the striking resemblance between two disparate disciplines has

[1]세종대학교 영어영문학과
[1]Department of English Language and Literature, Sejong University, Seoul 05006, Korea

[2]세종대학교 바이오융합공학과
[2]Department of Integrative Bioscience and Biotechnology, Sejong University, Seoul 05006, Korea
Tel: +82-2-3408-3764, Fax: +82-2-3409-3764
e-mail: djoh@sejong.ac.kr

identified the need to research such resemblance, pointing out conceptual units or entities of linguistics that can be observed in biological phenomena. This study focuses on the fundamental characteristics, rather than matching entities, shared by these two systems: how both systems produce a virtually infinite variety of outputs and how information is processed under the requirement of accuracy to produce legitimate outputs in the systems. Based on these findings, this study extends the key features of effective communication that have led to the sustainability of the two systems across time and space to the other system, society. In a society where addressing issues of sustainability is becoming increasingly urgent, these systems, having respectively demonstrated remarkable strength, can be a resource in deriving strategies for its sustainability. This study makes no effort to be a comprehensive attempt to cover all relevant issues, but is rather an interdisciplinary attempt to seek implications for one of the most important questions of human existence in formulating strategies for sustainability through an examination of the language of life and the life of language.

2. ANALOGIES BETWEEN LIFE AND LANGUAGE

Approaches beyond those that are purely biological and chemical are valid and even necessary in order to understand biological information and information processing systems [1]. This section examines how linguistics, a field that analyzes signs, symbols, and their structure and meaning, can be employed as something more than just a metaphor to understand information processing systems in living organisms. The rich productivity of the biological system with very limited resources shows significant parallels to the way in which humans produce language from a finite set of resources to a virtually infinite array of discrete expressions. We also compare the modes in which information is conveyed in the biological process of transcription and translation with the way to project features of the head of a linguistic unit in individual languages to its maximal projection.

2.1. Productivity with Unlimited Expressive Power

Shortly after Watson and Crick's discovery of the DNA double helix in 1953, a revolutionary proposal was made in the field of linguistics which assumed an innate language ability endowed uniquely to human [15-18].[1] Generative grammar, led by Noam Chomsky, is based on the observation that a seemingly infinite variety of language expressions is produced by finite resources from finite experience. For this school of thought, language is viewed as a natural object analogous to a body organ. Langauge cannot be taught to children, any more than children are taught to grow hair. There can be no conscious learning procedures for our first language just as we do not consciously study how to walk. One of the most important analogy points between biology and linguistics from a generative grammarian's perspective comes from the fact that both human languages and biological objects are hierarchical, generative, recursive, and virtually limitless with respect to their scope of expression [8].

As language is an infinitely productive system, a human being can produce and understand sentences he/she has never heard before. There are two crucial points to be noted here with respect to the infinity of language: finite resources and finite experience. Humans have an innate language ability hard-wired into our brains by our genes. This ability, the so-called 'Universal Grammar', enables humans to master languages despite paucity of experience. Despite their limited exposure to only a small variety of sentences, children reach a stage where they can produce and understand a virtually infinite number of sentences with immense variety thanks to the existence of the biological capacity of humans, or 'the faculty of language,'[2] which allows them to readily master any human language without explicit instruction [8].

The next point to discuss with respect to the limitless expressive power of language is concerned with finite resources. A particular human language consists of words from a lexicon and computational operations (i.e., rules such as Merge, Copy) to construct expressions (i.e., legitimate output or grammatical sentences). As illustrated in Fig. 1, a computational system exists in the faculty of human language which takes words

[1] There have been many different phases in the development of generative grammar. From 1960-1980, it was called Standard Theory, and then Government and Binding Theory was the main theme of generative grammarians through to the mid-nineties. Since 1995, the theory of generative grammar has been called the Minimalist program. It claims there are three factors that enter into the growth and development of language [18]: Genetic Endowment, External data, and Principles that are not specific to the organ under investigation and may be organism-independent. In this study, we adhere to the on-going philosophy of generative grammar – i.e., existence of Universal Grammar (UG) as the innate ability of human beings and infinite outputs from finite resources in human language.

[2] The faculty of language is further classified into two categories [8]: the faculty of language in a broad sense (FLB) and the faculty of language in a narrow sense (FLN). FLB is combined with the 'sensory-motor' system and 'conceptual-international' system. FLN, a component of FLB, is the abstract linguistic system whose main component is a computational system. This computational system (syntax) generates internal representations of language and also performs a mapping function for these internal representations to the conceptual-intentional interface (semantics) and to the sensory-motor interface (phonology).

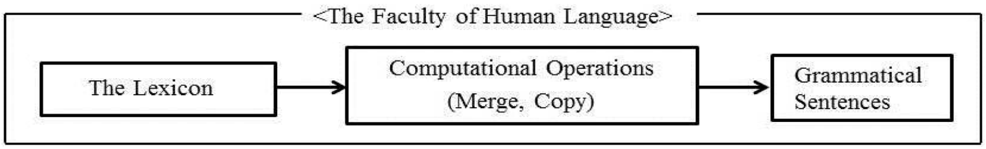

Fig. 1. Schematization of syntactic processes of human language based on generative grammarians' perspective: a computational component takes words from the lexicon and starts building structures by employing computational operations (Merge, Copy) to produce bigger units such as phrases, clauses and sentences.

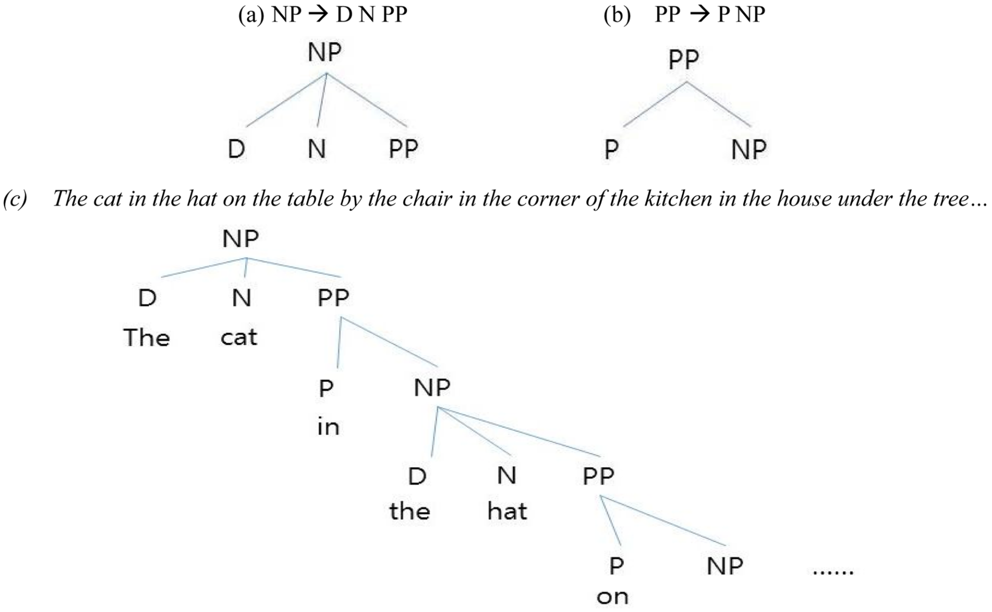
(a) NP → D N PP
NP
D
N
PP
(b) PP → P NP
PP
P
NP
(c) The cat in the hat on the table by the chair in the corner of the kitchen in the house under the tree…
NP
D
The
N
cat
PP
P
in
NP
D
the
N
hat
PP
P
on
NP
……

Fig. 2. Example of Recursion: if (a) and (b) are applied recursively, virtually endless output can be produced as in illustrated in (c).

from the lexicon and starts building structures by employing syntactic operations to produce bigger units such as phrases, clauses, and sentences. From its inception, Chomskyan generative grammar has assumed that the most elementary property of language lies in its discrete infinity, consisting of objects organized in a hierarchical way. Human language's recursion property based on the hierarchical structure is a fundamental characteristic of the human language faculty which distinguishes humans from other organisms [8]. For example, consider the two PS (phrase structure) rules in Fig. 2: (a) states that NP consists of D, N and PP (NP → D N PP), where NP stands for noun phrase, D for determiner, N for noun and PP for prepositional phrase; in (b), PP consists of P and NP (PP → P NP), where P stands for preposition. If we apply these two PS rules in a recursive fashion in such a way that an output of (a) becomes an input to (b), then this PP can again become an input to (a) producing another NP. These two rules can repeat endlessly as illustrated in (c) in Fig. 2. This recursion property accounts at least partially for the infinite nature of human language. There are no limits on what we can talk about. There is neither a longest sentence nor a non-arbitrary upper bound to sentence length [8].[3] In sum, the fact that a seemingly infinite variety of language expressions is produced by finite resources from finite experience is a primary feature of human language. This natural genetic productivity is absent in inanimate nature and therefore represents a core capability of life.

In biological systems, there are very limited biological resources to keep and convey the variety of genetic information. Only four kinds of nucleotides in DNA or RNA are used to store the genetic information for the sequence of proteins composed of 20 kinds of amino acids. When considering, as an example, the method to assign a small protein with 50 amino acids (in general, a protein has 50~2,000 amino acids), 20^{50} (equals 1.1 $\times 10^{65}$) different proteins can be addressed by the combination

[3] In this sense, language is analogous to natural numbers [8].

Table 1. Comparison between human and cell languages [1]

	Human Language	Cell Language
1. Alphabet (L)	Letters	4 Nucleotides (or 20 amino acids)
2. Lexicon (W)	Words	Structural genes (or polypeptides)
3. Sentences (S)	Strings of words	Sets of genes expressed coordinately in space and time under the control of spatiotemporal genes
4. Grammar (G)	Rules of sentence formation	Laws of chemistry and physics of nucleic acids that determine the folding patterns of DNA according to nucleotide sequences and microenvironmental conditions. Only a small subset of grammatically folded (hence *syntactically* correct) chromatin structures is selected by evolution and hence carry genetic (i.e., *semantic) information.*
5. Phonetics (P)	Physiologic structures and processes underlying phonation, audition, and interpretation	Conformational dynamics of DNA that enables the expression of genetic information through input of free energy via protein binding and/or ATP-dependent super coiling of DNA
6. Semantics (M)	Meaning of words and sentences	Gene-directed cell processes driven by conformons and intracellular dissipative structures (IDSs)

of 150 (3×50) nucleotides made of only 4 kinds (A, G, T, and C). Furthermore, since there are multiple different codons that match with the same single amino acid, more flexibility is given to the usage of nucleotides. This way of expressing limitless products with extremely limited resources is an immensely productive process. Additionally, not only the sequences of proteins but the post-translational modifications of proteins including protein cleavages, S-S bonds, glycosylation, and the 2- or 3-dimensional structures of proteins endow more flexibility to a cell to make more proteins correctly by using limited resources. As lexical items and grammar rules can be used multiple times in producing sentences, the same gene product may be used multiple times in different cells at different times [6].

2.2. Accuracy in Information Conveying

Since the discovery of the double helix in the 1950s, the genetic information carrying system has been described using borrowed terminology from linguistics. According to one view [1, 2], "cellese," the language of cells that is used to record, express, and induce genetic progress within cells, exhibits more than coincidental similarity to humanese; there is a systematic pattern that can be detected from the similarity of the two languages. For this view, letters in human language can be compared to 4 nucleotides (or 20 amino acids) in cellese, words to structural genes (or polypeptides), and strings of words to sets of genes as illustrated in Table 1.[4] The application of the grammatical rules of human languages to produce legitimate outputs works in the same fashion as the transcription and translation of the genetic information in cells. Yet, questions arise when we consider what the outputs of grammatical rule applications are. This needs to be compared to a somewhat recent biological perspective which argues that the largest parts of the genome do not code for proteins but serve as regulatory elements [19]. Theoretically, infinite sentences that we, humans, can produce are not the combination of rules/constraints but their realization employing lexical items. In this respect, we need to distinguish rules that are used to convey genetic information and outputs of the application of the rules [4].[5] This study considers proteins, not sets of genes, as outputs of information conveying rule applications in organisms [4].[6]

Let us consider the Central Dogma. 'Transcribing genetic information from DNA to RNA' and 'translating it from RNA to proteins in a systematic way' are clear indications of parallels between human and cell languages. In human language, infor-

[4] This approach later expands its analogy to proteinese whose function is to make a judgment [2]. Cellese consists of 5 sub-languages: DNese (DNA language), RNese (RNA language), proteinese (Protein language), metabolese (Metabolite language), and intercellese (Intercellular language).

[5] Modifying Ji's analogy [1], somewhat different parallels in the hierarchical level has been made [4]: folded polypeptides behave like phrases in humanese and protein networks are considered as human speech, which is defined as an ordered set of utterances forming a coherent unit in communication. This reasoning is based on their size; as the size of the set of utterances varies, so does the size of proteins.

[6] As we have seen, applying linguistics to analyze biological sequences and information carrying systems, such as those related to DNA, RNA, and amino acids, begins with looking at the structure of DNA using linguistic concepts. Previous work has agreed on the first step of analogy between linguistics and biology where nucleotides, which are the design feature that forms DNA, follow the same pattern of the basic unit of human language. There have been debates on which units of human language nucleotides are analogous to, i.e., distinctive features, phonemes, morphemes, or words. In this study, we do not go into the detailed matching between biological units and linguistic units. Rather, we focus more on the phenomena where the pattern of nucleotide sequence in DNA demonstrates that the structure and function of DNA has a systemic resemblance to the sentence structure of language, namely its syntactic process.

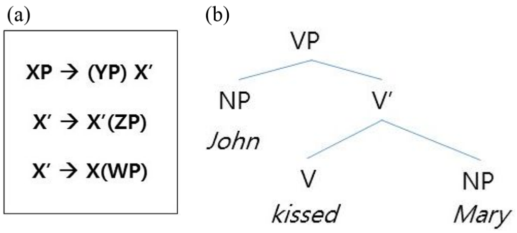

Fig. 3. X'-Theory: (a) represents the basic patterns of X'-Theory that is employed for constructing phrases in generative grammar. X, Y, and Z in (a) are arbitrary variables that stand for heads such as N (oun), V(erb), A(djective) etc. XP, YP, WP, and ZP are maximal projections whose heads are X, Y, W, and Z respectively. (b) is one example of phrase structures where X is V. Features of the head V is projected to its maximal projection, VP: the internal argument (object, *Mary*) of the head V (*kiss*) is realized as its sister in its maximal projection (VP) and the external argument (subject, *John*) of V is in the specifier positon in VP. The syntactic requirements of the head V (*kiss*) are satisfied in VP.

Table 2. Analogy of information conveying modules

	Human Language	Cell Language
Lexicon	Words (Heads)	20 amino acids (Nucleotide)
Computational operations	Copy, Merge	Transcribe, Translate, Elongation
Outputs	Sentences	Proteins
Filters	Linguistic Constraints	Splicing, Cleavage, Bonds

mation of the head is projected to its phrase and becomes a base to build a sentence as genetic information of DNA is transcribed to RNA. For instance, features of a head V (verb) are projected to a VP (verb phrase) and need to be fully realized in VP. Features of a head V include the number of arguments it requires, the structural Case it checks, and the semantic requirements for its arguments as represented in Fig. 3.[7] In other words, as the promoter regions of DNA share the same role and structure, the head of the phrase structure of language shares its features with its projections; a promoter initiates transcription of a particular gene, and a property of a head is projected to its structure.

Considering amino acids in cellese as a counterpart of words in humanese, this study assumes the overview of the analogy between humanese and cellese in Table 2 and discloses the mechanism that both share when they function properly. At each phase in Table 2, information processing is the most important factor where productivity of information and its accuracy sustain the dynamism of human language and biological phenomena. This high degree of commonality between biological and linguistic information processing mechanisms highlights the role of communication in these phenomena.

3. MESSAGES TO A SUSTAINABLE SOCIETY

This study notes that human language and biological phenomena both achieved sustainability by active communication within each system. It identifies two commonalities that enable the sustainability of a system. Defining these two commonalities as an essence of sustainable life, this section discusses how they can be extended to human societies.

The first commonality that gives a message to a sustainable society is "productivity" based on information recycling. Human language can create an infinite number of sentences based on a finite number of grammar rules due to the property of recursion [8]. In fact, recycling in the base system, which enables productive communication among modules, distinguishes human language from other communication systems found in animal behavior. This property plays a key role in biological phenomena as well, in the recycling of information within DNA. RNA, the messenger of important genetic information, does not get immediately destroyed after it serves its purpose of expressing a protein. It is retained for a certain period to be reused as a mold for information to create the same protein. The productivity of two systems, based on the role of recursion in language and recycling in life, can find applications in building a strategy for a sustainable society.

The second commonality is the accuracy strategy in each module of organizations. The central dogma of cell biology explains how genetic information produces materials required for life. When an RNA becomes mistranslated and the cell synthesizes an incorrect protein as a result, the cell's protease immediately breaks down the protein. This monitoring function of a cell is also found in the transcription process; when an RNA does not accurately contain the genetic information of the DNA or when an external factor such as a viral infection causes the cell to produce RNAs regardless of what the cell's DNA codes for, RNA-degrading enzymes break down the RNA to filter out wrong information. The role of promoters and enhancers, or transcription factors in RNA transcription and the steps in RNA processing such as splicing also follow the same mechanism as the processes of filtering in language production. In gene level processing, the accuracy of conveying genetic information is based on the universal physical/chemical rule of thumb, thermo-

[7] Case refers to the morphology that is associated with grammatical relations representing how an NP is functioning in the sentence syntactically. For instance, the Nominative Case is found with subjects, and the Accusative Case, found with objects, informally speaking.

dynamics. Paring two nucleotides is very specific (A-G and T-C or U-C), satisfying thermodynamically the most stable binding. This prevents and filters the mismatch of two nucleotides that often causes lethal results in DNA replication, in RNA transcription, and in the translation process in production of the final output, proteins. Human language can also be defined as an information conveying system in which basic modules are combined and transformed to create meaning and sound. In the process of computing human language, no rules would work unless the process satisfies the given constraints. If trials violating constraints occur, they crash immediately and will not be able to reach its spell-out stage as a legitimate output. This sort of filtering mechanism ensures that incorrect mapping from meaning to sound is filtered, and that optimal balance is reached between meaning and sound. The accuracy strategy in these two distinct systems investigated in this study can be applied to human societies for information balancing to formulate a sustainable model.

4. CONCLUDING REMARKS

Acknowledging similarities between the human language system and central dogma of biological systems, this research investigated specific structural properties that these two systems share, discovering the key message for maintaining their sustainability. Communication appears as key for their sustainability, whose main function is conveying information among various modules of each system productively and accurately. Living entities on Earth, including humans have been able to survive for several billion years based on well-organized information conveying systems, maintaining their internal homeostasis, reproduction, growth and development. Thanks to these communication systems, they have been able to cope with the dynamic external environment around them. Genetic information is processed in ways that show remarkable similarity to the linguistics concepts of feature projection, recursion, and condition-based filtering; thus, the tools of analysis for generative grammar, which analyze the way in which humans produce language, can be compared to the tools of analysis for genetic information processing. This study aims to go beyond accomplishments of the existing literature in identifying broad similarities between the two fields and suggests that accuracy of processing from the filtering mechanism and dynamic productivity from the recycling mechanism demonstrated in these two systems should be applied to other aspects of human existence, in particular for sustainability of society.

Acknowledgements

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014 S1A5A2A01015440).

REFERENCES

1. Ji, S. (1999) The linguistics of DNA: words, sentences, grammar, phonetics, and semantics. *Ann. NY Acad. Sciences* 870: 411-417.
2. Ji, S. (2006) The proteome as a molecular language ('Proteinese'). Poster presentation in *DIMACS Workshop on Sequence, Structure and Systems Approaches to Predict Protein Function.* May 3-5. Rutgers University, USA.
3. Raible, W. (2001) Linguistics and genetics: systematic parallels. *Language Typology and Language Universals. An International Handbook*. de Gruyter, Berlin, Germany.
4. Victorri, B. (2007) Analogy between language and biology: A functional approach. *Cogn. Process* 8: 11-19.
5. Witzany, G. (2014) Biological self-organization. *Int. J. Signs and Semiotic Systems* 3: 1-11.
6. Steels, L. (2004) Analogies between genome and language evolution. pp.200-206. In: J. Pollack (ed.) *Artificial Life IX*, MIT Press, Cambridge, MA, USA.
7. Fisher, S. E. (2016) Evolution of language: lessons from the genome. *Psychonomic Bulletin & Review*. Advance online publication. doi: 10.3758/s13423-016-1112-8.
8. Hauser, M. D., N. Chomsky, and W. T. Fitch. (2002) The faculty of language: what is it, who has it, and how did it evolve? *Science* 298: 1569-1579.
9. Jackendoff, R. (2002) *Foundations of Language*. Oxford Univ. Press, NY, USA.
10. Lenneberg, E. H. (1967) *Biological Foundations of Language*. Wiley, NY, USA.
11. Pinker, S. (1994) *The Language Instinct*. Morrow, NY, USA.
12. Searls, D. (1997) Linguistic approaches to biological sequences. *Comput. Appl. Biosci.* 13: 333-344.
13. Searls, D. (2002) The language of genes. *Nature* 420: 211-217.
14. Chiang, D., A. Joshi, and D. Searls. (2006) Grammatical representations of macromolecular structure. *J. Comput. Biology* 13: 1077-1100.
15. Chomsky, N. (1957) *Syntactic Structures*. Mouton, The Hague.
16. Chomsky, N. (1965) *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, USA.
17. Chomsky, N. (1975) *Reflections on Language*. Pantheon, NY, USA.
18. Chomsky, N. (1995) *The Minimalist Program*. MIT Press, Cambridge, MA, USA.
19. Mercer, T., and J. Mattick (2013). Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* 20: 300-307.