SUMMARY

Of the monograph

LANGUAGE, CONCEPT FORMATION
AND
CHILD LANGUAGE ACQUISITION

An information modeling approach

by
Velina Slavova

ISBN-978-619-245-221-6

Professor Marin Drinov Publishing House
of the Bulgarian Academy of Sciences

Préface

Nobody knows exactly how many languages there are in the world. One authoritative source that has collected data from all over the world, The Ethnologue, listed the total number of languages as 7,139 (in 2021). When the purpose is to examine language as a phenomenon, an attribute of all humans, one must consider the complete set of languages in a general way. Till now science is insufficient at explaining not only how language came about, but also why we have so many different languages and what the common basis of all of them is if any.

This monograph examines language as a system for exchanging information. The approach presented here links reasoning from the scientific fields of language, cognition, and information in a joint framework. The approach in this work looks at language from the level of a highly organized phenomenon to the level of essential detail. The modeling steps start from a general model of language as a cognitive phenomenon to reach an information processing model of concept formation. The aim is to trace previously unexplained interrelations between language and mental information units and to look for general principles valid at all levels of processing of the substance of meaning.

The work is structured in four parts, shortly presented below.

Part 1.

A model of message comprehension designated as processing of parallel information flows is presented using a Generalized Net. The approach involves reflecting and tracking information flows occurring in the cognitive process of extracting the meaning of a message. The presented model considers two information flows known as top-down and bottom-up. Their interaction imitates the process of accumulation of new meaningful information using feedback from information already available in memory. The processing in the Generalized Net is represented as information flows occurring in two separate "knowledge spaces" – linguistic knowledge and semantic knowledge. The analysis of the flows shows that the high-order cognitive processes have to use both linguistic and semantic knowledge. This leads to the hypothesis that information processing takes place in a "fused" space of linguistic and semantic knowledge. Tracing the network leads to the conclusion that in order to construct a correct mental image of a sentence, an operation *Semantic Merge* is required. This operation leads to obtaining aggregated semantic images of concepts following the grammatical rules involved in the message. This leads to presenting the mental semantic and grammatical operations as running within a single "information system". The approach is illustrated with examples of a grammatical rule which leads to an ambiguous semantic image. The tracking of examples using the Generalized Net suggests that the semantic images of the semantic entities involved in the message undergo a *Semantic Merge* and the result influences the solution.

The proposed semantic merge leads to investigating the rules of structuring a sentence at the syntactic level using a linguistic model based on the *Merge* operation defined in syntax. Analyzing the well-known Chomskian model in terms of "processing" structure leads to obtaining a covering Fibonacci tree connecting only the elements of the basic syntactic structure - Subject, Object and Recipient, known in linguistics as "arguments" of a sentence. The obtained argument-based Fibonacci tree take account of the operations *Merge* and *Type-shift* as defined in linguistics. The resulting formal representation is a height-constrained structure. Unlike the existing model of a syntactic tree, the Fibonacci tree satisfies both facts issuing from linguistics and the requirement for limited cognitive resource, known in cognitive science. The successive bottom-up merging of semantic entities along this tree, viewed as an operator of obtaining larger meaning unit, "produces" unambiguous semantic images of the Subject in the sentence in relation with the other arguments, thus building an clear-cut semantic image of a sentence. This Fibonacci "processing" tree is optimal in terms of both the number of merging steps and the number of entities involved in the construction of meaning. It is suggested that this tree reflects the principle of processing information units at a pre-linguistic level, that is – it reflects the principle of the "language of thought". At this point, the hypothesis is raised that the Fibonacci tree expresses the manner of constructing thoughts as larger information units.

To test this hypothesis, two experiments were conducted based on semantically ambiguous grammatical rules in the Bulgarian language. The results support the hypothesis that the mental interpretation of the basic syntactic roles takes place as a spontaneous "computation" according to the argument-based Fibonacci tree. According to the underlying linguistic model proposed by Soschen, the concepts of the entities involved as arguments in the "computation" undergo the same Merge operation at the bottom level of the processing. This gives rise to the hypothesis that concept formation obeys a similar process with similar optimality requirements. These considerations lead to further detailed investigation of the relationship between language and thought when examining current scientific results in multiple fields.

Part 2.

In the second part, a cognitive model of the language faculty is proposed. Based on a broad analysis of contemporary sources from cognitive science, brain science, brain imaging, language evolution, linguistics, language acquisition, etc., the main paths of reasoning in model building are presented. The neurobiological correspondence of a "concept" is defined in terms of measurable brain activity. Concepts are viewed in the context of known "grounded cognition", i.e., as arising from perception. The model proposed by Barsalou and widely accepted in cognitive psychology, according to which making sense of the world is done from the perspective of an actor in the environment, is used as a basic framework.

The analysis of the sources leads to the conclusion that information units are represented and organized as mental images reflecting the point of view of the operating biological system itself, i.e. from the first person. This is developed as a cognitive "Self-centered" model of the language faculty. The model considers sources of information as levels of cognitive functions, from perception and emotion to conceptualization, thought, and language. Based on the analysis of a variety of contemporary sources, this model assumes that the information necessary for the construction of the concept of Self is genetically and biologically embedded and that this primary concept determines both the way information about the world is internally represented and the construction of mental images in which the Self is the natural agent. This cognitive model is a generalization of the argument-based Fibonacci tree and the "Self-centered" way of creating information units. The model is supported with analysis of sentences from a large corpus of children's speech. Analysis of the data shows that children's utterances develop according to the distinct constructs of the argument-based Fibonacci tree, which gradually appear in children's utterances from 11 months to four years of age. The first use of each successive mental syntactic construction always occurs with the child speaker himself in the role of subject of the uttered phrase.

Part 3.

In the third part, an empirical approach is applied to study the process of concept formation based on statistical analysis of children's speech at the initial stage of speech production. The approach aims at studying the biological principles of internal representation of the surrounding world, since children do not know language at birth and innate mechanisms for classifying and organizing information support the process of concept formation during the child's interaction with the environment. The data sources are presented, as well as the procedure that led to the collection of a very large amount of data from free dialogues in English and French of children aged 9 months to 5 years. The statistical analysis begins with the detection of gender differences in noun use, where the noun categories of meaning of are inferred by applying the cognitive Self-centered model proposed in the previous part. This result suggests that the reasoning followed in the cognitive model leads to a meaningful and consistent result.

Statistical analysis of the acquisition processes of the two languages shows that when children's words are classified according to the part-of-speech systems existing in the two languages, the acquisition process of English and French shows almost no similarity. However, the applied factor analysis shows that, on the contrary, the two languages are acquired in an identical way, as is clear from the first principal component of the statistical space of part-of-speech usage across the acquisition time. This statistically high degree of similarity (0.905) is due to the universal cognitive basis of the language acquisition process and should be explained at some deeper level related to the formation of concepts and the "computational" complexity of concepts as information units. This requires concepts to be represented in classes that reflect both meaning and cognitive "computational" complexity.

Based on the cognitive Self-centered model, a way in which information is internally represented in classes of meaning is proposed. The proposed set of classes is based on reasoning about the different types of information needed to manage an autonomous system that can and should act in its environment. The classes obtained in this way coincide almost completely with the known theory of semantic bootstrapping proposed by Pinker in cognitive linguistics. Statistical analysis shows that when the parts of speech in the two languages are considered as belonging to the proposed classes of meaning, the acquisition processes of the two languages by children are very similar. However, for two classes - Relationships and Circumstances - the statistical analysis shows the presence of different proportions of use by the two groups of children from a certain age onwards. It is assumed that the meaning expressed by children of the same age is equally complex and that the differences are due to the peculiarities of the languages. This suggests that there are other factors that explain the almost complete correspondence of the first principal components of the two statistical spaces, which principal components reflect the unified cognitive basis for the acquisition of the two languages. As is well known, in the studied period of child development cognitive abilities grow and the complexity of solved cognitive tasks increases as a consequence of the increase in cognitive resources. This makes it necessary to analyze the process of concept formation in terms of its "computational" complexity.

Part 4.

In the fourth part, an information processing model of concept formation is proposed. The reasoning is based on basic facts in neuroscience. The important proposal is that information units are represented as the result of comparing signals from two sources, those coming from the biological system, including memory, and those coming from the environment. Represented in this way, the process of creating information units fits the conditions of a theorem proved by Horibe concerning Fibonacci trees and their optimality. This modeling step leads to a representation of the process of creating information units by a Fibonacci tree, mathematically proven as a minimum cost structure (expressed as the weights of all paths to the root) with maximum entropy expressed by the leaves. Applied in the context of the model, this tree describes the minimum cost of biological and energy resources to obtain the maximum amount of classified information when it has "received" the maximum entropy input.

The proposed classes of conceptual meanings are mapped to the levels of this tree according to their abstractness. The resulting Fibonacci tree representation provides a way to evaluate the complexity of the classes in a manner analogous to computational complexity.

The proposed "cognitive" complexity of each class expresses the effort of forming an information unit as the number of operations of linking signals arising from the environment with those arising from the biological system.

The complexity of the different classes is thus substituted into an additive expression representing the overall cognitive complexity of what is expressed in the children's dialogues. When data from the two languages are substituted, this expression for the two languages evolves identically over time. This indicates that the observed acquisition process is determined by increasing cognitive resources and both English and French speaking children increase their cognitive resources in the same way. Differences in the use of the classes of meaning are due to differences in the linguistic expression of mental images of the same complexity. The Monte Carlo method confirms that the high correlation of 0.933 found for the total "complexity of concepts" used in the two languages results precisely from the application of the model.

The statistical measure of the complexity of the expressed utterance (sentence) also evolves over time in an identical way in both English and French when applying the proposed model of concept complexity. This in turn implies that the complexity of the communicated "ideas" is the same for both languages and that it is determined by the cognitive resources available. The acquisition of both English and French, at the level of both concept complexity and idea complexity, develops equally, gradually within the available cognitive resources. The differences arise because of the different linguistic means used to express the same thing - the thought to be conveyed.

Given that the resulting models, reflecting both the level of syntactic structure and the level of concept formation, are described by the same structure, a Fibonacci tree, it can be assumed that the creation of information units and the formation of larger information units obey similar principles that reflect resource optimization.

General scientific contribution (author’s assessment)

This work brings together key theories and results from the fields of linguistics, cognition, and psychology to build a model of how information is acquired and organized in biological systems. Analogous to computational complexity, a measure of the cognitive complexity of classes of concepts with different levels of abstractness is proposed. The model, based on results in mathematics, relates acquired and classified information to principles of optimal use of biological and energy resources. Empirical verification of the model through statistical analysis of children's speech leads to a representation of the acquisition of English and French as the same process. This empirical confirmation links children's linguistic expression presented in terms of processing complexity to their increasing cognitive abilities. To the best of my knowledge, there are no approaches to date that combine in a single model the emergence and classification of information, optimality principles, and language acquisition in order to account for language production as correlated with and constrained by cognitive abilities.

To the attention of the scientific jury were presented reviews of the monograph by:

Krassimir Atanassov
Professor in Mathematics, Institute of Biophysics and Biomedical Engineering,
Bulgarian Academy of Sciences,
Sofia, Bulgaria

Encho Gerganov
Professor in Psychology and Psycholinguistics, Department of Psychology and Cognitive Sciences,
New Bulgarian University,
Sofia, Bulgaria

Vassil Vassilev
Professor in Artificial Intelligence and Intelligent Systems, School of Computing and Digital Media
London Metropolitan University,
London, the United Kingdom

Vitalii Velychko
Professor in Information Technologies, Institute of Cybernetics,
National Academy of Sciences,
Kyiv, Ukraine

Irina Blinnikova
Professor in Cognitive Psychology, Faculty of Psychology
Lomonosov Moscow State University,
Moscow, Russia

Note - The review by prof. Blinnikova arrived after the monograph was submitted for printing.