Algebraic Lexicon Grammar

Giustino De Bueriis

Dept. of Political and Communication Sciences, University of Salerno

Via Ponte Don Melillo, Fisciano (SA), 84084, Italy.

Tel: 39-324-540-3366. E-mail: gdebueriis@unisa.it

Alberto Maria Langella (Corresponding author)

Dept. of Political and Communication Sciences, University of Salerno

Via Ponte Don Melillo, Fisciano (SA), 84084, Italy

Tel: 39-329-887-8544. E-mail: allangella@unisa.it

Received: May 11, 2019 Accepted: May 28, 2019 Published: June 17, 2019

doi:10.5296/ijl.v11i3.xxx URL: https://doi.org/10.5296/ijl.v11i3.xxx

Abstract

This article aims at showing an application of graph theory to the description of the syntactic relations between words in English. Graph theory is part of Network Science, a new and compelling branch of mathematics that has undergone huge development over the past 20 years. The linguistic theoretical background is the Lexicon Grammar (LG) that in turn is built on the ground of the Harrisian grammar for operators and arguments. Graph theory is particularly useful in order to show how sentences have underlying structures that can be visualized through the use of graphs, and whose properties can be measured and quantified with the typical mathematical tools used by the researchers in the field of Network Science. A graph is a set of nodes (also called 'vertices') and links (also called 'edges'); the links connect the nodes. Each sentence will be described as a graph, in which the words have to be seen as the nodes and the syntactic relations between the words are the links.

Keywods: Lexicon Grammar, Harrisian transformation, graph, network properties

1. Introduction

1.1 Introduction to our perspective

Language can be considered the evolutionary result of a long series of historical and cultural events and presents itself as a set of discrete objects (phonetical or written) which can be described from a mathematical point of view. Our guiding principle is the necessity of conceiving natural language like a natural object, which as all other natural objects from physics, biology and other natural sciences, can be approached with a mathematical method. The reason to see languages as natural objects is that the direction they take in the course of their evolution cannot be planned and has to be considered accidental. We use a pioneering and new branch of mathematics called Network Science to describe the syntax of natural language. In this article we apply the network science to the description of English syntax, but the mathematical tools used can be extended to the description of all other languages. In paragraph 2 we discuss the theoretical assumption of our Algebraic Lexicon Grammar (ALG), which is which structural features turn natural language into an object which can be treated with mathematical tools. In paragraph 3 we discuss the basic principles of Network Science, how we describe English sentences and how this new method can shed light on the syntactic general properties of natural language.

1.2 Relevant related literature

Since the beginning of linguistic research with formal methods on syntax, at least two main and different approaches started from the late 1950s: generative grammar and valency theory. Valency theory was explored by scholars such as Lucien Tesnière and Zellig Sabbetai Harris who devised formal and algebraic-like methods of investigation. These new methods were inspired by the concept of valency in chemistry: a verb for Tesnière or an operator for Harris have a valency requiring a certain number of words in order to produce an acceptable sentence from a syntactic point of view.

Language can be seen as a mathematical object with words requiring other words under the constraint of valency. Generative grammar, by contrast, produced the concept of rules of production necessary to generate a sentence starting by an initial non-terminal symbol S and arriving through intermediate states to the terminal elements of the vocabulary. A sentence is defined by the formal rules necessary to generate it. Our ALG is rooted in valency theory, following the footsteps of Lucien Tesnière and Zellig Sabbetai Harris. More recently, the Lexicon Grammar of Maurice Gross has proposed methods of investigation very close to the ones we have chosen to use for our ALG.

2. Theoretical Assumption

2.1 Our method

A natural language in its physical shape looks like (presents itself like) a set of strings of alphabetic characters (at least in languages which use this system of writing). In particular, each sentence of a language looks like (presents itself like) a linear sequence of strings empirically recognized by the speakers like a grammatically correct sentence; in a convenient pragmatic context, the linear sequence of strings might turn out to be a simple string or even a single character. This hypothetical sentence might even be meaningless, yet grammatically correct[1], but since one of the main features of all the natural languages is to be a

[1] Or viceversa a sentence could be absolutely ungrammatical and, in this case, not belonging to L. The grade of ungrammaticality is gradual and an ungrammatical sentence could still have a meaning for the purpose of communication, in other words it could be coherent in absence of cohesion, as pointed out by text linguistics.

communication system, every sentence belonging to it usually has a meaning. The basic assumption of this brief outline – where we sketch a descriptive algebraic model – is that a natural language syntax is a "natural phenomenon" and not a "cultural phenomenon": no cultural planning is at the root of its organization and functioning – even in a dynamic (evolutionary) perspective of language. It is a self-organized system[2], meaning that there is no external project whatsoever to organize or regulate its structure, features and properties.

After all – and this is one of the few points that this model shares with generativism – Chomsky's theory, rooted in the rationalistic premises of sixteenth and seventeenth century philosophy, assumes that the human mind has an innate competence of the linguistic universals; therefore if the mind possesses "naturally" these universal principles, this implies that their origin, auto-organization and particularly their functioning principles are not cultural, devised, elaborated from outside (singularly or collectively), but are dependent on a spontaneous auto-organization, which in turn has to obey "natural laws". These laws must be explained, even if in a way which is partially peculiar to each language under investigation.

The model proposed here is based upon graph theory. This theory is a relatively recent branch of mathematics which studies the various ways in which specific objects come to be connected according to a network of relations, and whose results are valid regardless of the specificity of the objects constituting the network.

When in some way a syntactic representation of a sentence belonging to a natural language is given, the main objective is to show the non-linear structure of an event (oral or written) which appears (to the hearing or to the sight) in a form essentially linear[3]: By and large, language presents itself to us in the form of strings of sounds (or letters). "Yet linguists of all persuasion have argued that there is evidence for more than meets the eye" (Kracht, 2007a, p. 47); nevertheless, "Sentences are not only sequences of words. There is more structure than meets the eye" (Kracht, 2007b, p. 86). This structure has traditionally, at least since L. Tesnière (1959) and N. Chomsky (1957) on, been presented through a stemma or a tree: "Sentences consist of words. These words are arranged into groups of varying size, called constituents. The structure of constituents is a *tree*." (Kracht 2007b, p. 86). Phrase constituents are sometimes describable with a tree (specifically, a "rooted tree", since it is a hierarchical structure) – which incidentally is a specific type of graph – but a sentence, as a whole, is not. The point of radical divergence between the model we are discussing and every other model describing a sentence with phrase structure trees is that in the latter there is a source node (expressed by S: sentence) of all the other nodes; in our model this node corresponds, by contrast, to the entire sentence, which is not a node, but a graph – which is, by default, a set of nodes and relations: to represent in the same way concepts and objects so

[2] In a system, the self-organization is the spontaneous emergence of a global coherence beyond local interactions; moreover, these types of systems often show even adaptivity, which is the capacity to adapt to environmental changes without losing its own basic organization.

[3] Prosodical events anyway, does not violate the linear nature of the oral utterance: duration, intensity, height, pauses, variations of velocity, height, volume, accent, quantity, tone and syllables respect the linearity of the event. Nor other features of the orality not present – unless there is some clear convention o some artifice – in the written form deny it: intensity of pronunciation, intonation,.

radically different is – we think – a big mistake of principle, even if one has established a hierarchy between the nodes.

Generally, the analysis in constituents – from the American structuralism to L. Bloomfield and up to Chomsky's models – shows the relations between the parts with the whole to which they belong, but these models do not show the dependence relations existing between those same elements; and, moreover, the model in immediate constituents is not capable of expressing the difference of the roles between the dominant element and the dominated element (i.e. between the verb and the direct object). However, the models based upon the dependence (DGs) do not show the differences between the relations established by the verb with, respectively, the subject, the direct object or the prepositional object: each one of these relations is considered a dependence tout court.

2.2 Subsections

In paragraph 3 we deal with the way a sentence can be described through a graph, how, inside this graph, the syntactic relations can be expressed by adjacency matrices and the way a graph describing a sentence can grow. The description of a sentence with a graph and the related use of adjacency matrices are discussed in subparagraphs *3.1* and *3.2*. In subparagraph *3.3* we deal with how different classes of words contribute to the growing network. In subparagraphs *3.3.1*, *3.3.2*, and *3.3.3*, we discuss how, respectively, prepositions, conjunctions and articles contribute to the growing network.

3. Networks

Network Science, over the past twenty years, has found many mathematical patterns in real world objects, as well as in cultural phenomena: crystal patterns, Bose-Einstein condensation, the spreading of viruses, social networks, the World Wide Web (WWW), the Internet network, the cellular metabolism, the corporations and so forth. Scientists are uncovering interesting and common mathematical patterns behind objects which appear to be absolutely different from one another. These discoveries are producing a new science of reality, a unified approach to the study of disparate natural objects. Language syntax, according to the description given in this paper, seems to hide mathematical properties very similar to those hidden behind many other natural objects, objects that apparently seem to share no common features with syntax and language at large.

3.1 Graphs, links and adjacency matrices

A sentence is described by a graph $G = (V, E)$; V is a set of vertices and E is a set of edges connecting the nodes. Under each vertex there is the corresponding word and the number of the vertex is placed inside the vertex. The numeration of the vertices follows the linear order of the words inside the sentences. In graph theory the edges can be oriented or not oriented; in our model we use oriented edges expressed by arrows departing from a node and entering another node (if the outward and the inward vertex is the same, the edge is called a loop). The edges express the existence of a syntactic relation between two vertices and possess a vector. The vector is a succession of 6 binary digits (an n-tuple), and each binary digit refers to different morpho-syntactic properties of the syntactic relation. A vector is an *n*-tuple and has the following form:

$$\boldsymbol{w} = (\mathrm{D/I},\ \mathrm{S/P},\ \mathrm{M/F},\ \mathrm{T{+}M},\ \mathrm{P}(v))$$

Each of the 6 elements of the vector is expressed by a binary digit and each binary digit refers to different morpho-syntactic properties of the syntactic relation:

- ⦿ First component (D/I):
  - Value 1 = argument of the operator
  - Value 0 = non-argument of the operator

- ⦿ Second component (S/P):
  - Value 1 = singular/plural agreement
  - Value 0 = no singular/plural agreement

- ⦿ Third component (M/F):
  - Value 1 = masculine/feminine agreement
  - Value 0 = no masculine/feminine agreement

- ⦿ Fourth component (T+M):
  - Value 1 = tense or mood requirement
  - Value 0 = no tense or mood requirement

- ⦿ Fifth component (P(v))[4]:
  - Value 1-1 = certain event
  - Value 1-0 = possible event
  - Value 0-0 = impossible event

In our model, to each graph is associated an adjacency matrix, which is a square matrix of order *n*, and *n* is equal to the number of vertices of the graph ($n = |V|$). The adjacency matrix has the same number of rows and columns, and the value of whatever element a_{ij} is 1 if a link exists between the vertex *i* and the vertex *j* or 0 if there is no link between them. In ALG we have substituted to each 1 in the adjacecy matrices a number in base 10 which represents the weight of the link. In the next sub-paragraph we will show how the vectors binary digits are converted in numbers in base 10.

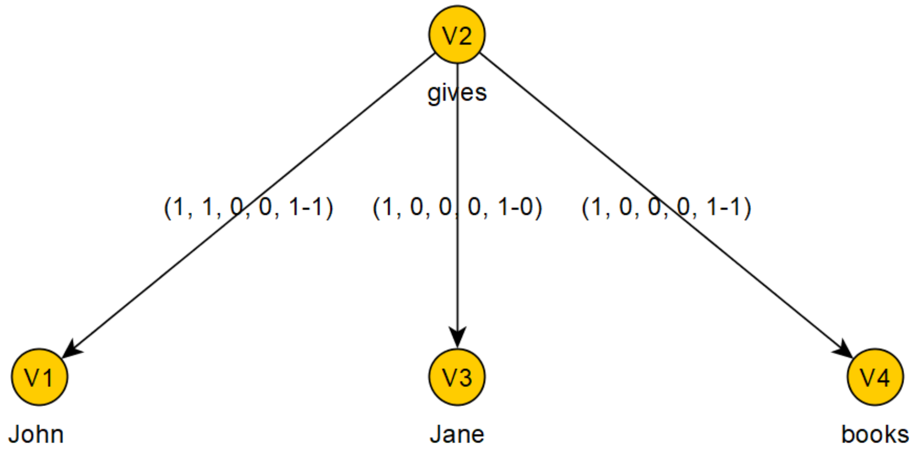
3.2 The Network of a Sentence

Algebraic Lexicon Grammar treats sentences as networks of elements. As mentioned above networks are graphs and the ones used here are weighted directed graphs. Indeed each sentence is described by a graph. Let’s consider the following sentence:

(1) *John gives Jane books*

[4] This represents the vertex probability according to the formal model proposed by A.N. Kolmogorov (1930-1987). It should be noted that the probability is not related to the entire network, but it is a local property, i.e., relating to a single vertex of the network.

We describe this sentence with the following graph:



Graph 1.

As we can see each edge has its own vector and each vector expresses the type of syntactic relation existing between the nodes linked. The edge linking the vertex V2 (*gives*) to the vertex V1 (*John*) has the weight expressed by the vector (1,1,0,0,1-1). The first value is 1, meaning that the subject *John* is required by the verb *gives* (*John* is an argument of *gives*). The second is 1, since the form of the verb *gives* requires a singular subject. The third and forth value are zero, indeed gender agreement is not required, as well as time or mode requirement. The last couple of values refers to the fact that the occurrence of *John* is certain since English, in this case, does not allow the dropping of the subject. According to Lexicon Grammar, the verb *to give* has three arguments: its valence is 3. The third argument, *books*, like the subject, is obligatory, meaning that its occurrence is certain, otherwise we would have a sentence with no meaning. Indeed the last couples of values of the vectors associated with the links connecting *gives* to *John*, and *books*, are expressed by the binary digits 1-1 (certain event). Indeed the following three sentences are not acceptable:

(2) **Gives Jane books*
(3) **John gives Jane*
(4) **John gives*

By contrast the link between the vertex V2 (*gives*) and the vertex V3 (*Jane*) is expressed by the binary digits 1-0, meaning possible event, thus Jane can be dropped and we still have an acceptable sentence:

(5) *John gives books*

In graph 1 the vector associated to the link between the vertex V2 (*gives*) and the vertex V1 (*John*) is (1,1,0,0,1-1) which can be seen as the binary number 110011. This binary number can be converted in a number in base 10; the number in base 10 is $2^5+2^4+2^1+2^0$ which is equal to 51. This number represents a measure of the complexity of the syntactic relation between the vertex V2 and the vertex V1 in graph 1. The vector associated to the link between V2 (*give)* and V3 (*Jane*) instead corresponds to the binary number 100010 which is converted into the number in base 10 $2^5+2^1=34$. The vector associated to to the link between V2 (*gives*) and V4 (*books*) is the binary number 100011, converted in the number in base 10 $2^5+2^1+2^0 = 35$. 35 between *gives* and *books* measures a stronger syntactic relation than the one expressed by 34 between *gives* and *Jane* (*Jane* can be dropped). Hence the weights of every link are converted in numbers in base 10. The following matrix is accociated with graph 1:

	V1 (*John*)	**V2** (*gives*)	**V3** (*Jane*)	**V4** (*books*)
V1 (*John*)	0	0	0	0
V2 (*gives*)	51	0	34	35
V3 (*Jane*)	0	0	0	0
V4 (*books*)	0	0	0	0

Adjacency Matrix 1

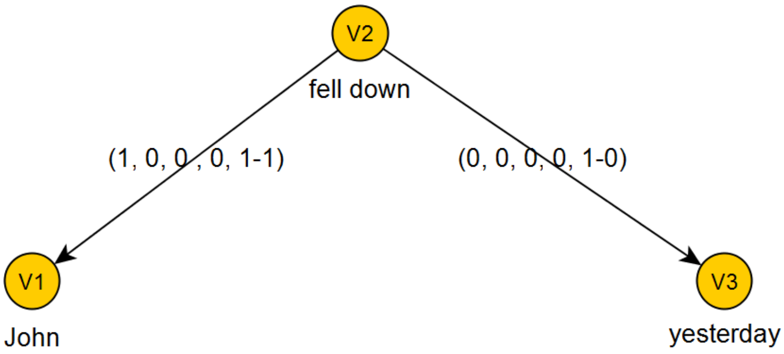
As shown above in the adjacency matrix 1, V2 (*gives*) has a stronger and more complex syntactic relation with V1 (*John*) than the syntactic relations that it has with V3 (*Jane*) and V4 (*books*): the number in base 10 associated to the link between V2 and V1 is 51 and the other two are equal to respectively 34 and 35 (in 51 we have also the number agreement which is missing in the other two syntactic relations).

The ALG distinguishes arguments and adjuncts[5]. Adjuncts can be individuated with the relative clause diagnostic:

(6) *John fell down* ***yesterday*** → *John fell down, which happened* ***yesterday*** (*yesterday* is an adjunct)

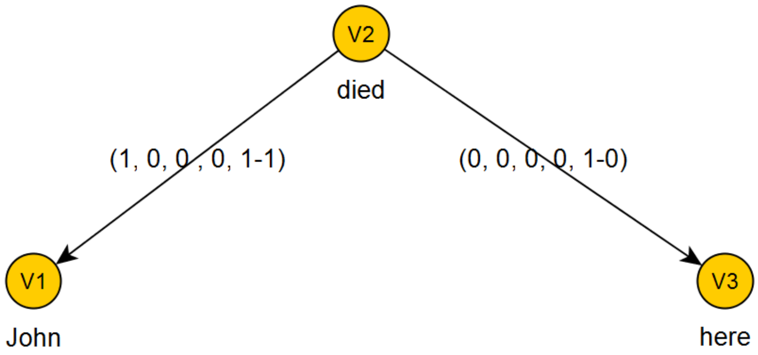
(7) *John died* ***here*** → *John died, which happened* ***here*** (*here* is an adjunct)

The following graphs are associated with the two previous sentences:



Graph 2.

[5] We are currently working on the possibility of taking into account semi-arguments.



Graph 3.

Graph 2 and graph 3 have the vector (1, 0, 0, 0, 1-1) between the verbs and the respective subjects. This vector has 1 as first value, meaning that the subjects are arguments of the verbs and the binary digits 1-1 as last value, meaning that the subjects are arguments that cannot be dropped (certain event). By contrast, the vector (0,0,0,0,1-0) between the verbs and the respective adjuncts has 0 as first value because *yesterday* and *here* are adjuncts, and the binary digits 1-0 as last value, meaning that the adjuncts can be dropped. The second, third and fourth value in the vectors are zero, because no number or gender agreements is required, as well as time or mode requirements. The following are the adjacency matrices of the two previous graphs:

	V1 (*John*)	**V2** (*fell down*)	**V3** (*yesterday*)
V1 (*John*)	0	0	0
V2 (*fell down*)	35	0	2
V3 (*yesterday*)	0	0	0

Adjacency Matrix 2

	V1 (*John*)	**V2** (*died*)	**V3** (*here*)
V1 (*John*)	0	0	0
V2 (*died*)	35	0	2
V3 (*here*)	0	0	0

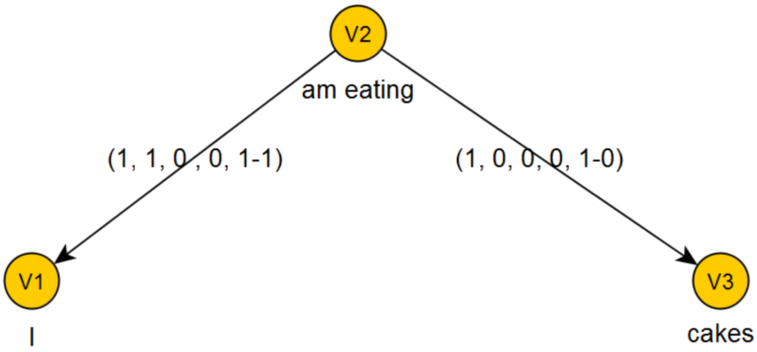
Adjacency Matrix 3

In both adjacency matrix 2 and adjacency matrix 3 the link between V2 (the predicates *died* and *fell down*) and V1 (the subjects *John* and *John*) has a weight equal to 35, while the link between V2 and V3 has a weight equal to 2. The numerical difference represents the different strength of the syntactic relations between the predicates and the subjects (35) on the one hand, and the predicates and the adjuncts (2) on the other hand.

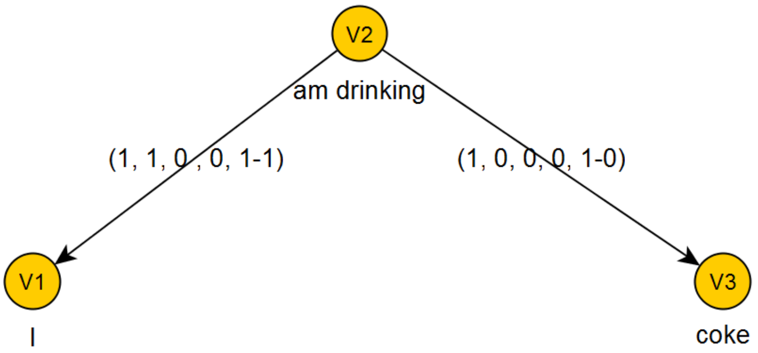
In many cases not only the adjuncts but even the arguments can be dropped:

(8) *I am eating cakes* → *I am eating*
(9) *I am drinking coke* → *I am drinking*

sentences described by the two following graphs:



Graph 4.



Graph 5.

The vector between *am eating* and *cakes* and between *am drinking* and *coke* is (1,0,0,0,1-0) and has 1 as first value (*cakes* is argument of *am eating* and *coke* is argument of *am drinking*) and 1-0 as last value (possible event: the arguments can be dropped). This vector differs by the vector referring to the relations between the verbs and the adjuncts in graph 2 and graph 3 because of the first value (0 in graph 2 and graph 3, and 1 in graph 4 and graph 5). *Here* and *yesterday*, respectively in graph 2 and 3, are not arguments and can be dropped, *cakes* and *coke*, respectively in graph 4 and 5, are arguments but can be dropped as well (in both cases the last couple of binary digits is 1-0, meaning possible occurrence).

Here I show the adjacency matrices referring to graph 4 and graph 5:

	V1 (*I*)	**V2** (*am eating*)	**V3** (*cakes*)
V1 (*I*)	0	0	0
V2 (*am eating*)	51	0	34
V3 (*cakes*)	0	0	0

Adjacency Matrix 4

	V1 (*I*)	**V2** (*am drinking*)	**V3** (*coke*)
V1 (*I*)	0	0	0
V2 (*am drinking*)	51	0	34
V3 (*coke*)	0	0	0

Adjacency Matrix 5

In adjacency matrix 4 and adjacency matrix 5 the links between the predicates and the object-arguments have weights equal to 34, indicating a strong syntactic link but inferior to the weight of the link between V2 and V1 (51), because of the fact that the object-arguments *cakes* and *coke* can be dropped and there is no number agreement. The value 34 indicates a stronger syntactic link than that between the predicates and the adjuncts in graph 2 and graph 3: 34 for the syntactic relations between the predicates and the arguments in graph 4 and graph 5, and 2 for the syntactic relations between the predicates and the adjuncts in graph 2 and graph 3.

Also to be noted is that the weight of the links between the vertices V2 and V1 in graph 4 and 5 is 51, which is higher (a stronger syntactic relation) than the links between the vertices V2 and V1 in graph 2 and 3 (which is equal to 35). Indeed these links in graph 4 and 5 have the second binary digit of the vector equal to 1 in comparison with the zero of the same links between the predicates and the subjects in graph 2 and 3 where no number agreement is required. This difference is the result of the first person of the present continuous in graph 4 and 5 respectively for the predicates *am eating* and *am drinking*, which requires as subjects the personal pronoun *I*. By contrast the preterite for the predicates *fell down* and *died* respectively in graph 2 and 3 does not require any particular number agreement by the nouns or pronouns selected as subject-arguments.

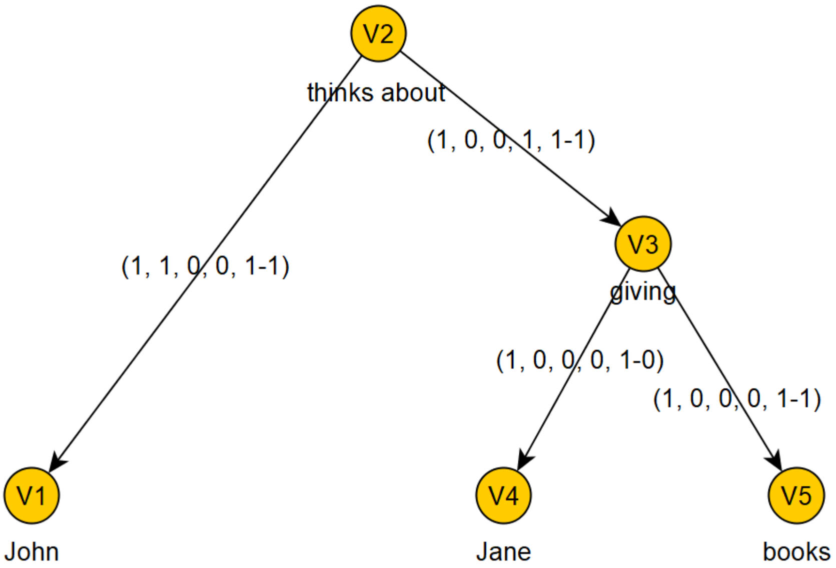
In graph 4 and graph 5 we have seen sentences whose object-arguments can be dropped and are required by the predicates through a link with weight 34, a weight inferior to the 35 expressing the weight of the link between the predicate and the object-arguments in graph 1. This difference is due to the fact that in graph 1 the occurrence of the object-arguments is obligatory in order to produce a meaningful sentence (the vector has 1-1 as last binary digits instead of 1-0 of the vectors in graph 4 and graph 5).

3.3 Growing Networks

Some elements are responsible for the growth of a network. The sentences of natural language are no exception (some words are responsible for it). In the following sentence

(10) *John thinks about giving Jane books*

thinks about brings *John* and *giving* into the sentence (respectively subject-argument and predicate-argument). *Giving* in turn is responsible for bringing into the sentence its own two arguments: *Jane* and *books*. The third argument of *giving* (*John*) is dropped because of its co-reference with the subject of the main clause. We can see that the main predicate *thinks about* has only outbound links while the other words in sentence 6 can have both inbound and outbound links as shown in graph 6 below:



Graph 6.

We have said that some words are responsible for the growth of the network: *talks about* brings into the sentence the subject *John* and the predicate *giving*. *Giving* in turn is responsible for making the network grow by linking to itself other two arguments: *Jane* and *books* (the third potential argument, *John*, as said above has been dropped). The principles governing the network growth in the graphs of ALG have been very well pointed out by Albert-Lázló Barabási, an Hungarian mathematician and one of the leading figures today in Network Science:

“Putting the pieces of the puzzle together, we find that real networks are governed by two laws: *growth* and *preferential attachment*. Each network starts from a small nucleus and expands with the addition of new nodes. Then these new nodes, when deciding where to link, prefer the nodes that have more links...The model is very simple, as growth and preferential attachment naturally lead to an algorithm defined by two straightforward rules:

A. *Growth*: For each given period of time we add a new node to the network. This step underscores the fact that networks are assembled one node at a time.
B. *Preferential attachment*: We assume that each new node connects to the existing nodes...The probability that it will choose a given node is proportional to the number of links the chosen node has. That is, given the choice between two nodes, one with twice as many links as the other, it is twice as likely that the new node will connect to the more connected node."[6]

According to Barabási the network growth happens adding one node at a time to the network, just as we add linearly one word at a time to the sentences, and when deciding where to attach themselves, the new nodes choose for the nodes with more links. We have seen, for example in graph 6, that the nodes with more links are the two predicates (*thinks* and *giving*); so they are the nodes responsible for the network growth.

The principle of the *preferential attachment* translates the linguistic concept of valence in the mathematical language of Network Science: the predicate-argument structure is ruled by the predicate, and in order to have a sentence its valence has to be satisfied. The principle of the preferential attachment has been very well pointed out by Guido Caldarelli and Michele Catanzaro, two experts of Network Science:

"The Barabási-Albert model shows that a bottom-up mechanism of growth can generate heterogeneity, without imposing any top-down blueprint...the network is the outcome of the iteration of an individual local choice: preferring more connected nodes to less connected ones. The model uses probability to allow for individual deviations from this behaviour: some nodes can decide to connect to low-degree nodes. However, the general tendency sets the outcome. As a further confirmation, one can check that the heterogeneity of the network disappears if it is grown without the preferential attachment rule. Indeed, new nodes connect to old ones at random, in such a way that the degree of old nodes (the degree is the total number of inbound and outbound links (A/N)) does not influence their capacity to attract new links; the outcome is a homogeneous network in which every node ends up having more or less the same degree."[7]

The previous quotation points out well the property of the real-world networks to grow according to the heterogeneity produced by the preferential attachment: the nodes which seem to attract new links with a more than random probability are in our model first and foremost the predicates, and other elements that we will discuss later in this article.

The following is the adjacency matrix for graph 6:

	V1 (*John*)	**V2** (*thinks about*)	**V3** (*giving*)	**V4** (*Jane*)	**V5** (*books*)
V1 (*John*)	0	0	0	0	0
V2 (*thinks about*)	51	0	39	0	0

[6] Barabasi, Albert-Lazlo, *Linked*, p. 86.
[7] Caldarelli, Guido, Catanzaro, Michele, *Networks*, p. 71.

V3 (*giving*)	0	0	0	34	35
V4 (*Jane*)	0	0	0	0	0
V5 (*books*)	0	0	0	0	0

Adjacency Matrix 6

The adjacency matrix 6 shows a strong syntactic relation between the main predicate *talks about* and the subject *John* (equal to 51). Besides, the first value of the vector linking V2 to V1 (see graph 6 above) tells us the fact that *John* is an argument of *talks about*; the vector has also the second value equal to 1, because of the number agreement between the predicate and the subject of the main clause. By contrast, the syntactic relations between the subordinate predicate *giving* and its two arguments *Jane* and *books* is less strong, because these syntactic relations lack the number agreement: their value is 35 (see adjacency matrix 6).

We have said that the graphs used in the ALG model are weighted directed graphs. For this type of graphs, a certain number of measures can be made. For example the number of inbound and outbound links for each node. In Network Science the number of inbound links of a node is called degree-in (abbreviated as deg-in), the number of outbound links is called degree-out (abbreviated deg-out), and the total number of degrees-in plus degrees-out is called degree-tot (abbreviated deg-tot):

	Deg-in	**Deg-out**	**Deg-tot**
V1 (*John*)	1	0	1
V2 (*thinks about*)	0	2	2
V3 (*giving*)	1	2	3
V4 (*Jane*)	1	0	1
V5 (*books*)	1	0	1

Degrees of links for graph 6

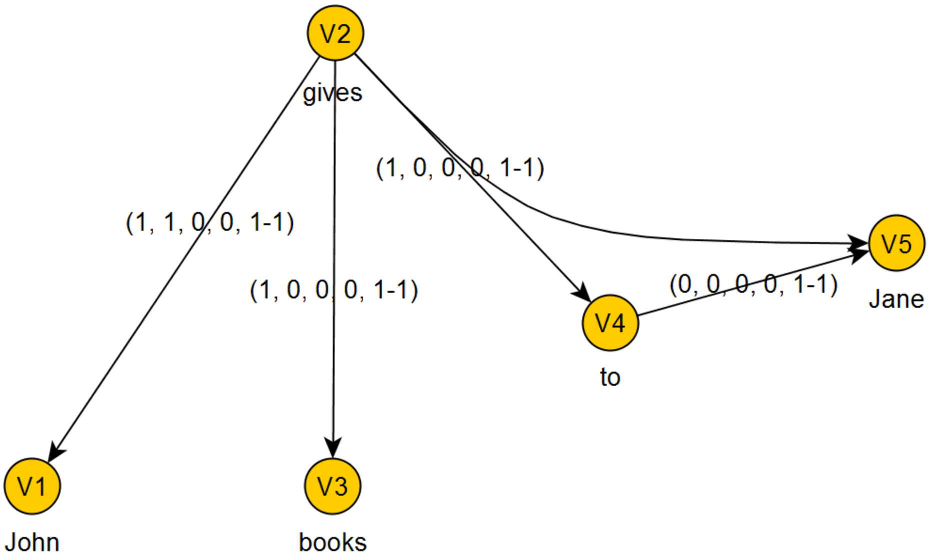
We can see that the three arguments *John* (of the predicate *talks about*), *Jane* and *books* (of the predicate *giving*) have only one inbound link, therefore their deg-tot is 1. By contrast, the two predicates play a key role in the sentence: *thinks* scores a deg-tot equal to 2 and *giving* scores a deg-tot equal to 3. We can distinguish the main operator because it is the only node with only outbound links (its deg-tot is equal to its deg-out).

3.3.1 The Role of the Prepositions in the Growing Network

It should be noticed that prepositions can play a key role in networks growing. In order to evaluate this, let's consider the following sentence:

(10) *John gives books to Jane*

The previous sentence carries the same meaning of sentence 1 but instead of the double object it displays a direct object (*books*) plus a prepositional phrase (*to Jane*). Sentence 7 can be described by the following graph:



Graph 7.

The link between vertex 2 and the couple of vertices V4 and V5 is a hyperlink[8]. The following are the adjacency matrix and the degree of links for graph 7:

	V1 (*John*)	**V2** (*gives*)	**V3** (*books*)	**V4** (*to*)	**(V4,V5)** (*to Jane*)	**V5** (*Jane*)
V1 (*John*)	0	0	0	0	0	0
V2 (*gives*)	51	0	0	0	35	0
V3 (*books*)	0	0	0	0	0	0
V4 (*to*)	0	0	0	0	0	3
V5 (*Jane*)	0	0	0	0	0	0

Adjacency Matrix 7

	Deg-in	**Deg-out**	**Deg-tot**
V1 (*John*)	1	0	1
V2 (*gives*)	0	3	3
V3 (*books*)	1	0	1
V4 (*to*)	0	1	1
V4, V5 (*to Jane*)	1	0	1
V6 (*Jane*)	1	0	1

[8] A hyperlink is every outbound link on at least two vertices.

Degree of links for graph 7

In graph 7 we see that the nodes receiving an inbound hyperlink behave like a unique node, more precisely as an ordered pair (V4, V5), but as a source of an outbound link they have to be treated separately; indeed the preposition *to* has an outbound link on the node V5 (*Jane*). It should be noticed that the preposition (*to*), although not adding any relevant information to the meaning carried by the sentence, plays a strategic role in networks growing anyway. Its deg-tot is 1 as single node plus 1 as member of the ordered pair scoring a total of 2 (see degree of links for graph 7 below), only lower than the number scored by the predicate *gives*. Furthermore prepositions always have outbound links, feature that enables them to make the network grow. The values of the links between the predicate *gives* and the three arguments *John*, *books* and *Jane* in the previous adjacency matrix 7 are identical to those of adjacency matrix 1 (51, 35 and 35). Here the novelty is represented by the presence of the preposition *to*, which is responsible for the substitution of the double object (*John gives Jane books*) with a direct object plus a prepositional phrase (*John gives books to Jane*). It should be remembered that the high value of the outbound link of V2 (*gives*) on V1 (*John*) which is equal to 51, is the result of the existence of the number agreement (the second binary digit of the vector is equal to 1).

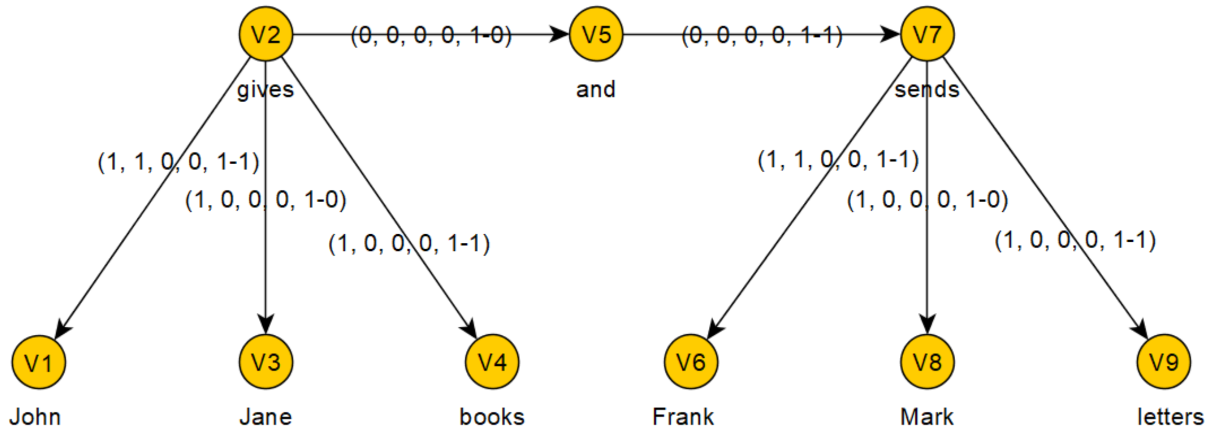
The pivotal role of the prepositions in the sentences seems to be a general feature of the English syntax just as with the syntax of other languages we have studied[9] with this same method.

3.3.2 The Role of the Conjunctions in the Growing Network

The conjunctions also display outbound links, and are definitely responsible for the network growing in the sentences. For example, in order to show how the conjunctions can be responsible for the network growing, we analyze the following sentence:

(11) *John gives Jane books and Frank sends Mark letters*

which is described in ALG by the following graph:



[9] We have tested the ALG on Italian and Latin.

Graph 8.

In graph 8 we can see how *gives*, which has a valence equal to 3, brings *John*, *Jane* and *books* into the sentence, and *sends*, which also has a valence equal to 3, brings *Frank*, *Mark* and *letters* into the sentence. The conjunction *and* is responsible for linking one each other the two predicates *gives* and *sends*. The two clauses display the typical double direct object. The following tables are the adjacency matrix and the degree of links for graph 8:

	V1 (*John*)	**V2** (*gives*)	**V3** (*Jane*)	**V4** (*books*)	**V5** (*and*)	**V6** (*Frank*)	**V7** (*sends*)	**V8** (*Mark*)	**V9** (*letters*)
V1(*John*)	0	0	0	0	0	0	0	0	0
V2(*gives*)	51	0	34	35	2	0	0	0	0
V3(*Jane*)	0	0	0	0	0	0	0	0	0
V4(*books*)	0	0	0	0	0	0	0	0	0
V5(*and*)	0	0	0	0	0	0	3	0	0
V6(*Frank*)	0	0	0	0	0	0	0	0	0
V7(*sends*)	0	0	0	0	0	51	0	34	35
V8(*Mark*)	0	0	0	0	0	0	0	0	0
V9(*letters*)	0	0	0	0	0	0	0	0	0

Adjacency Matrix 8

	Deg-in	**Deg-out**	**Deg-tot**
V1 (*John*)	1	0	1
V2 (*gives*)	0	3	3
V3 (*Jane*)	1	0	1
V4 (*books*)	1	0	1
V5 (*and*)	1	1	2
V6 (*Frank*)	1	0	1
V7 (*sends*)	0	3	3
V8 (*Mark*)	1	0	1
V9 (*letters*)	1	0	1

Degree of links for graph 8

There is no need to comment the syntactic relations inside the two clauses in graph 8 because they are identical to the syntactic relations already discussed for sentence 1. What instead is interesting here is the key role of the conjunction *and* in graph 8. The conjunction *and* is linked with an inbound link to the predicate *gives* of the first clause and with an outbound link to the predicate *sends* of the second clause. The inbound link is expressed by the vector

(0,0,0,0,1-0); the first value is zero because *and* is not required by the valence of *gives*, whose three arguments are *John*, *Jane* and *books*. The second and third value are zero as well because no number or gender agreement hold for the syntactic relation between the predicate *gives* and the conjunction *and*. The forth value is zero as well, because no time or mode is required by *gives* for *and*. The outbound link by *and* on the second predicate *sends* is expressed by the vector (0,0,0,0,1-1). The vector is identical to the vector of the inbound link now discussed except for the last binary digit, which is equal to 1 instead of zero. This means that the probability of the occurrence of a predicate after the conjunction *and* is certain or obligatory in order to produce a grammatical sentence. In fact the following sentence is not grammatical:

(11a) **John gives Jane books and*

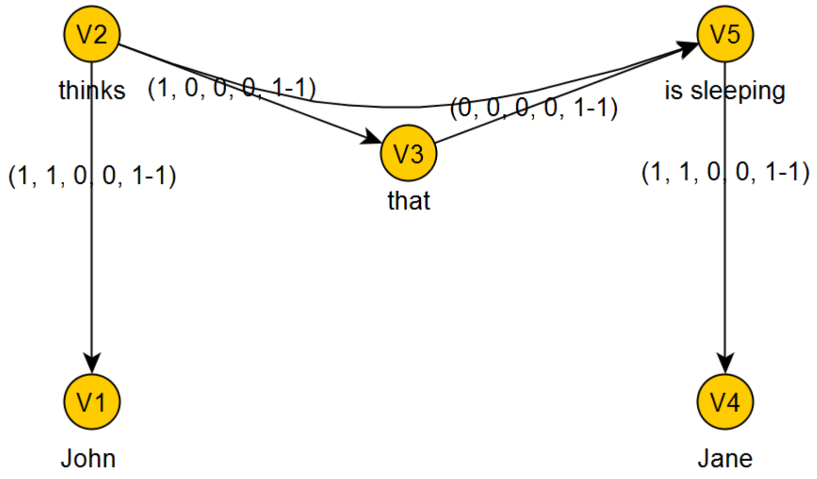
Thus the conjunction has to be followed by a predicate in the structural order (in this case the predicate *sends*).

As regards the role of the conjunctions in the network growing of sentence 11 we need to take a look at the degree of links for graph 8. It has a deg-tot equal to 2, with an inbound link and an outbound link, a property that, as already pointed out for the preposition *to* in graph 7, is responsible for the network growing, because allows some nodes (words) of a graph (a sentence) to bring other nodes into it: in the case of graph 8, the conjunction brings into the sentence the predicate *sends* that in turn brings into the sentence its three arguments. The conjunction *and* is the only node to possess an outbound link except the two predicates *gives* and *sends*; the six arguments in graph 8 only have inbound links. So both the preposition *to* in graph 7 and the conjunction *and* in graph 8 share the same property of being responsible for the network growing: to always have inbound and outbound links.

In the previous graph the coordinating conjunction *and* has linked together two independent clauses. The conjunction *that* instead is responsible for making a predicate realize its valence as it is necessary for bringing into the sentence a second predicate with the role of argument:

(12) *John thinks that Jane is sleeping*

In sentence 9 *thinks* has to be considered as the main predicate with an elementary argument *John* as subject, and a predicate-argument *is sleeping* which in turn brings into the sentence its own argument *Jane*. The previous sentence is described by the following graph:



Graph 9.

to which the following adjacency matrix and degree of links are associated:

	V1 (*John*)	**V2** (*thinks*)	**V3** (*that*)	**V4** (*Jane*)	**(V3,V5)** (*that /is sleeping*)	**V5** (*is sleeping*)
V1 (*John*)	0	0	0	0	0	0
V2 (*thinks*)	51	0	0	0	35	0
V3 (*that*)	0	0	0	0	0	3
V4 (*Jane*)	0	0	0	0	0	0
V5 (*is sleping*)	51	0	0	0	0	0

Adjacency Matrix 9

	Deg-in	**Deg-out**	**Deg-tot**
V1 (*John*)	1	0	1
V2 (*thinks*)	0	2	2
V3 (*that*)	0	1	1
V4 (*Jane*)	0	1	1
V3, V5 (*that/ is sleeping*)	1	0	1
V5 (*is sleeping*)	1	1	2

Degree of links for graph 9

The conjunction in the previous graph behaves like the preposition *to* in graph 7, having one inbound link as member of the ordered pair *that*/*is sleeping* that receives an hyperlink by the main predicate *thinks*, and an outbound link as single vertex on the predicate-argument *is sleeping*. Hence, the conjunction *that*, having an inbound and an outbound link, shares with the predicate-arguments and the prepositions the property of making the network grow. The vector associated with the outbound hyperlink of the main predicate *thinks* on the ordered pair *that*/*is sleeping* is (1, 0, 0, 0, 1-1): the first value means that the ordered pair is an argument of the predicate *thinks*, while the following two zeros mean that no number or gender agreement is required; the fourth value is zero again because there is no mode or time requirement. The last couple of binary digits is 1-1, because the occurrence of the ordered pair consisting of the conjunction *that* plus the argument-predicate *is sleeping* is necessary in order to produce a meaningful sentence. Both the main predicate *thinks* and the predicate-argument in graph 9 have outbound links respectively on *John* and *Jane* whose weights are 51 (value already discussed earlier in this article for the outbound links from predicates to noun-arguments).

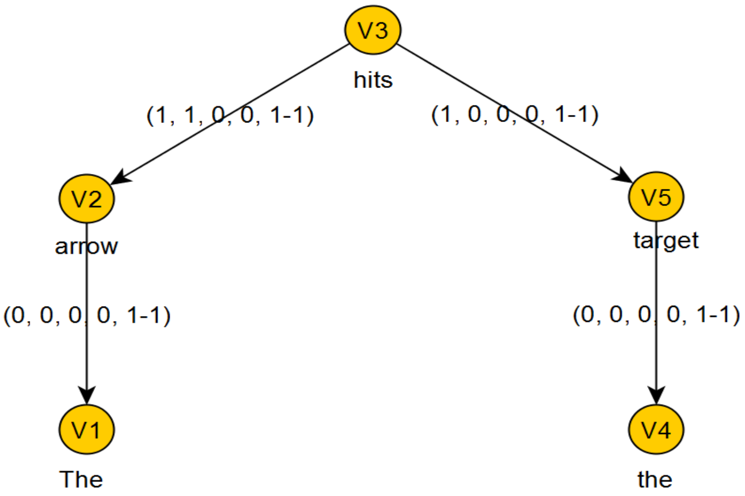
The outbound link from the conjunction *that* to the predicate-argument *is sleeping* has the weight expressed by the vector (0, 0, 0, 0, 1-1): the first binary digit means that *is sleeping* is not an argument of *that*; the second and the third zero mean that no number or gender agreement holds between the two vertices; the fourth zero means that there is no mode or time requirement.

3.3.3 The Role of Articles in the Growing Network

If we consider the topological behaviour of the articles we see that they have inbound but no outbound links. Hence they are not responsible for the network growing: when a sentence comes to the articles is from a topological point of view at a dead end. Let’s see how articles behave in the following sentence:

(13) *The arrow hits the target*

This sentence is described in our ALG with the following graph:



Graph 10.

to which the following adjacency matrix and degree of links are associated:

	V1 (*The*)	**V2** (*arrow*)	**V3** (*hits*)	**V4** (*the*)	**V5** (*target*)

V1 (*The*)	0	0	0	0	0
V2 (*arrow*)	3	0	0	0	0
V3 (*hits*)	0	51	0	0	35
V4 (*the*)	0	0	0	0	0
V5 (*target*)	0	0	0	3	0

Adjacency matrix 10

	Deg-in	**Deg-out**	**Deg-tot**
V1 (*The*)	1	0	1
V2 (*arrow*)	1	1	2
V3 (*hits*)	1	1	2
V4 (*the*)	1	0	1
V5 (*target*)	1	1	2

Degree of links for graph 10

Adjacency matrix 10 displays again an unbalanced syntactic relation between the predicate and its two arguments. The edge between the predicate V3 (*hits*) and the subject-argument V2 (*arrow*) has a vector associated with it with a value of 51, while the edge between V3 and the object-argument V5 has a vector associated with it with a value of 35. In the first case, there is the number agreement which is missing in the second case. We have seen this typical difference before in this article with these two values always expressing it: 51 (predicate-subject relation) compared to 35 (predicate-object). As regards the two articles we can see how they have one inbound link each, with a vector whose value is 3. The vector associated with the link between the first argument *arrow* (vertex V2) and the article *the* selected by it (vertex V1) is (0, 0, 0, 0, 1-1). The first four binary digits are 0, because according to the Harrisian principles, the article *the* is not considered an argument of the noun *arrow*, and no number or gender agreement nor time or mode requirements are necessary. The last couple of binary digits is 1-1, meaning that the occurrence of the article *the* is necessary in order to generate a grammatical sentence. In fact the following sentence is ungrammatical:

(13a) **Arrow hits the target*

What pointed out for the syntactic relation between *arrow* and *the* applies for the syntactic relation between *target* and *the* too.

Thus the links between the nouns and the articles in sentence 13 have very low weights (3), and this, according to our model, seems to be a general topological feature of English[10]. If we take a look at the degrees of the links for graph 10 we notice the noun-arguments *arrow* and

[10] The same seems to hold for the other languages that we have studied applying the ALG.

target have one inbound and one outbound link each, scoring a total degree equal to 2. This local property allows them to make the network grow. The predicate *hits* has two outbound links, and scores a total number of links equal to 2, just as its two arguments. The difference between the predicate and its two arguments is that the former on the one hand possesses strong ties with high weight linking it to its arguments (51 and 31), on the other hand it possesses only outbound links, a property always associated with the main predicates of the sentences.

3. Conclusions

We hope to have shown in this article how Network Science can be a very useful analytical tool in order to visualize the hidden patterns lying behind natural language syntax. While putting linguistics at the core of what may be, without any doubt, called a scientific revolution, ALG is able to measure and quantify the syntactic relations between the words of natural language. Here we have applied ALG to the syntax of English, but, as it seems to be the case, it could be applied to the syntax of any language (in an earlier article we applied it to the syntax of Italian). This possibility lays down the theoretical foundations for a new and promising way of treating and analyzing natural language syntax. At the same time, it enables linguists to produce interesting comparisons between the syntaxes of different languages on a mathematical ground.

Even more interesting seems to be the possibility of treating natural language as a natural object whose mathematical structure has developed naturally without any top-down blueprint (this seems to be the case for a great number of other real world networks). When formalized in this way, the syntax of a language appears like any other real world network, and as such can be compared to other networks, regardless of the nature of the single elements that represent the building blocks of the networks. Such comparison allows us to have a thoughtful insight in the mathematical structures of language and to see if some general principles are hidden behind the transitory form of so many real world networks.

References

Barabási, A. L. (2002). *Linked: the new science of networks*. New York: Basic Books.

Berge, C. (1989). *Hypergraphs*. New York: Elsevier.

Bollobábas, B. (1979). *Graph Theory: An introductory course*. New York: Springer-Verlag.

Bretto, A. (2013). *Hypergraph theory. An introduction*. Cham: Springer.

Buchanan, M. (2003). *Nexus: small worlds and the groundbreaking science of networks*. New York: W. W. Norton & Company.

Caldarelli, G., & Catanzaro M. (2012). *Networks: a very short introduction*. Gosport: Oxford University Press.

Chartrand, G. (1977). *Introductory Graph Theory*. New York: Dover Publications.

Chomsky, N. (1957). *Syntactic structures*. Berlin: Mouton & Co.

Erdős, P., & Rényi, A. (1959). On random graphs. I, *Publicationes mathematicae debrecen* 6 (pp. 290-297). Debecren, Institute of Mathematics, University of Debecren.

Erdős, P., & Rényi, A. (1960). On the evolution of random graphs, *Publication of the mathematical institute of the hungarian academy of sciences* (pp. 17-61). Berlin: Springer.

Gross, M. (1975). *Méthodes en syntaxe*. Paris: Hermann.

Harary, F., Norman R. Z., & Cartwright, D. (1965). *Structural models: an introduction to the theory of directed graphs*. New York: Wiley.

Harris, Z. S. (1957). Co-occurrence and transformation in linguistic structure, *Language* vol. 33 (pp. 283-340). Washington D.C.: Linguistic Society of America. https://doi.org/10.2307/411155_

Harris, Z. S. (1982). *A grammar of English on mathematical principles*. NewYork: Wiley-Interscience.

Hudson, R. (1984). *Word grammar*, Oxford: Blackwell.

Jakobson, R. (1971). Quest for the essence of language, *Selected writing II. word and language* (345-359). The Hague: Mouton. https://doi.org/10.1515/9783110873269.345

Kracht, M. (2007). The emergence of syntactic structure. *Linguistics and philosophy, 30*(1), (pp. 47-95). Berlin: Springer. https://doi.org/10.1007/s10988-006-9011-5

Mel'čuk, I. A. (1988). *Dependency syntax: theory and practice*. New York: State University of New York Press.

Sgall, P., Hajičová, E., & Panevová, J. (1986). *The meaning of the sentence in its pragmatic aspects*. Dordtrecht: Reidel Publishing Company.

Strogatz, S. (2004). *Sync: how order emerges from chaos in the universe, nature, and daily life*. New York: Hachette Books.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Editions Klincksieck.

Watts, D. (2004). *Six degrees of separation: the science of a connected age*. W. W. Norton & Company.

Copyrights