# Learning Analytics - Scientific Description and Heuristic Validation of Languages NLG

Francesco Saverio Tortoriello, Ritamaria Bucciarelli, Roberto Capone, Javier Enriquez, Marianna Greco, Giulia Savarese

Je-LKS
PEER REVIEWED PAPERS
LEARNING ANALYTICS: FOR A DIALOGUE BETWEEN TEACHING PRACTICES AND EDUCATIONAL RESEARCH
Vol. 15, n.3, 2019
ISSN: 1826-6223 | eISSN: 1971-8829
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal

# LEARNING ANALYTICS -SCIENTIFIC DESCRIPTION AND HEURISTIC VALIDATION OF LANGUAGES NLG

**Ritamaria Bucciarelli[1], Roberto Capone[2], Javier Enriquez[3], Marianna Greco[4], Giulia Savarese[2], Francesco Saverio Tortoriello[2]**

[1] University of Siena
[2] University of Salerno
[3] University of Spain
[4] MIUR
rbucciarelli@unisa.it 1; rcapone@unisa.it; 2 janjuen@alumni.upv.es;3 marianna.greco2@istruzione.it 4; gsavarese@unisa.it 5; fstortoriello@unisa.it

**Keywords**: Digital intelligence; Mind R2D2; Emotional filtering

The educator is a "Translator" ie manufacturer of algorithms for a teaching in the infosphere. The teacher who turns into a robotic mind perhaps of the type R2D2 a research droid, will be the emblem of our future. The work aims to validate the moments of transformation through which, over the centuries, the mathematical sciences, with the help of philosophy, have elevated the languages Natural Language Generation (NLG) to formal models. The starting hypothesis is to corroborate an epistemological statute, which entrusts mental processes with logical-mathematical reasoning following four models: Chomsky (1956), which, with descriptive grammar, marks a new model for the rewriting of languages; Gross (1975), which, with the relationship between linguistics, informatics and mathematics, generates a relation concerning a strongly transdisciplinary domain, in which linguistics

for citations:
**Bucciarelli R., Capone R., Eriquez J.J., Greco M., Savarese G., Tortoriello F.S.** (2019), *Learning Analytics -Scientific Description and Heuristic Validation of Languages NLG*, Journal of e-Learning and Knowledge Society, v.15, n.3, 251-261. ISSN: 1826-6223, e-ISSN:1971-8829
*DOI: 10.20368/1971-8829/1135040*

realizes models and procedures of the informatics type; Silberzstein's Nooj system (2015) for the elaboration, description and analysis of fixed INLG sentences. The focal part of the research is the comparison work that the team has carried out to validate the processing of languages according to the Transformational Analysis of Direct Transitive by M. Silberzstein and the lexicon-grammar; the probabilistic calculation, according to the Probabilistic latent semantic Analysis (Hoffmann, 1999) and the empirical method.

## 1 Introduction

This research focuses on mathematical models for the description of languages and on some new generation software for the construction of natural languages The research hypothesis conducted in 2013 at the chair of written Italian by Bucciarelli and the team of Balboni with the project by Ateneo Ca' Foscari in which researchers go to analyze the scientific aspects of specialized languages to emphasize the interest that the theme of the specificity of these languages arouses from the sociolinguistic and socio-semiotic. Therefore, the team relies on epistemological models of reference because they are determined by the will to provide certainties to lay basic empirical foundations to the research with :- Popper's theories that with the principle of falsifiability or possibility of confutation and defines an interpretation of science based on error and leads to elaborate new theories that prove to be fallacious, because so much more can be circumscribed the horizon of truth. In our opinion, if the calculation of the possibility leads us to a possible solution of the truth, we need to rely on a second model that gives certainties such as: -Learning analytics, the integrated techniques of analytical learning mediated by didactic research and applied to data mining in Cabena *et al.* (1998). that is, the set of techniques and methodologies that have as their object the extraction of useful information from large amounts of data through automatic methods using the filtering methodologies "FC. This means the transfer of the possibility to the heuristic certainty of the collection given there seems to be a right solution for the analysis and description of the lexicon. The authors then rely on hypotheses to be validated to models the irrefutable and therefore confront themselves with those who previously explored the same areas of research such as:

Chomsky's model (1964), which, with descriptive grammar, marks a new model for the rewriting of languages. He proposes algorithmic forms to explain linguistic facts and shifts from the lemma to the minimum sentence the centrality of the role of representation of the semantic unit of signification; Gross's model (1975), which, with the relationship between linguistics, informatics and mathematics, generates a relation concerning a strongly transdisciplinary domain, in which linguistics realizes computer models and procedures to refine, formalize its own data and its own methods and then proceed to a taxonomic

classification of the possible sentences in Italian through the lexicon-grammar L.G.L.I.; Silberzstein's Nooj system (2015) for the elaboration, description and analysis of fixed sentences, which introduces the concept of text constructor supported by large "neutral" linguistic resources (dictionaries, morphology, sentence structure and transformation grammars), which can be used both to analyze and to automatically generate INLG (2017). The focal part of the research is the production according to the Transformational Analysis of Direct Transitive by M. Silberzstein according to the lexicon-grammar of the transformation of N0 V N1 into finished automata with the calculation of the Probabilistic latent Semantic Analysis (Hofmann, 1999). Our research question is: is the linguistic text subject to mathematical laws? We will try to give an answer keeping in mind that a sentence can be manipulated through spontaneous or pre-established algorithms and an algorithm can be considered a finite logical sequence of operations that is subject to mathematical laws. Our idea is that language is as innate as number and man manipulates it according to an algorithmic sequence of mathematical laws. We will try to show how natural language is subject to mathematical but random laws, while fixed language is subject to pre-established mathematical laws and therefore predictable.

## 2 Reference model: Noam Chomsky transformational grammar (TGT)

The transformation of elaborated codes and methods is carried out in the theory of the generative transformative grammar by Chomsky, in Lightfoot (2002), to which some essential elements are already present in the work "Syntactic Structures", characterized by the search for innate structures of natural language, an distinctive element of man as an animal species, overcoming the conception of traditional linguistics centered on the study of the peculiarities of spoken languages. He states that to understand the functioning of a language is not enough to discover its structure, since it is not enough to describe the components and relationships between them, nor to analyze and classify them. The formal grammar, that is to say, the generative grammar is a set of rules that "specify" or "generate" recursively (that is, through a rewriting system) the well-formed formulas of a language. This definition includes a large number of different approaches to grammar. The term "generative grammar" is also widely used to refer to the school of linguistics in which this type of formal grammar plays a crucial role. In fact, it is in the formal languages that the Chomsky hierarchy finds in the theory of proof, the validation and elevation of languages to mathematical techniques. In fact, it is the branch of mathematical logic that considers demonstrations in turn as mathematical objects, facilitating their analysis with mathematical techniques, one of which is… *an algorithm that is a procedure that solves a given problem through a finite number of*

*elementary steps, clear and unambiguous, in a reasonable time.* Chomsky (1957) points out that the "creativity" governed by the rules for which the new sentences are "generated" constantly and, therefore, the linguistic capacity that each speaker possesses not only consists of a set of words, expressions and sentences, but which is also a set of defined rules and principles. In fact, mental grammar is a competence of the speaker, which allows him to compose and transform an infinite number of sentences, based on innate knowledge and the universal principles that regulate the creation of language. The deep structure represents the core of the semantic relationships of a sentence and is reflected through transformations in the structure of the surface (which closely follows the phonological form of the sentences) and, therefore, it is only the competence of the speaker to transform the sentence.

## 3 Language environments the lexicon grammar an elementary calculation

During a decade of experimental work carried out in the Department of Communication Sciences of the University of Salerno in collaboration with other research centers and, in particular, with the "Laboratoire d'Autornatique et Linguistique (CNRS - Paris 7)", new methods for linguistic investigation have been developed. Research has been carried out based essentially on the construction of syntactic lexicons that, taking advantage of the opportunities offered by computerized data processing, point to a description, as exhaustive and formal as possible, of a specific language. The research is part of the project "Lexicon grammar of the Italian language (LGLI)". The theoretical reference model is represented by the "Operator-argument Grammar " (Harris, 1964).A rigorously analytical approach has been derived in which, despite the centrality of the syntax and the scientific nature of the rules of transformation, the grammar of a language should no longer be interpreted as an abstract model, but be investigated based on concrete statements. The activity focused on the deepening of methods for linguistic research and was directed, for the interested parties, to identify the modalities of curricular applications for a modern glottodidactics (Ibrahim *et al.*, 2003). If we would like to proceed with a taxonomic classification of the possible sentences in spoken Italian, it would be appropriate to clarify the importance of the verb in the sentence through the method of research and experimentation of L.G.L.I. (Elia *et al.*, 1981) On the basis of these premises, to describe a language from a lexical-grammatical point of view, we will have to do so. Research on sentence structures involves a lexicon-grammatical classification of verbs and controls the real possibilities of aggregations with nominal forms. According to the theories of Harris and Chomsky, when studying the combinatorial possibilities of sentences, they are considered "free" sentences that have a wide possibility of changing lexical

entries within the N position (productivity of class N.) A second characteristic of simple sentences is characterized by the co-occurrence of a class of compatible operators and verbs. The third, of idiomatic sentences that are also called fixed sentences. Therefore, by operating a syntactic classification of Italian verbs we will have the following results of identification of the sentence, as well as the following syntactic mechanisms:

- Handling of conversion and replancements
- A taxonomic classification
- DB categorization

The basic structure "SB" is represented by sentences that present one or more arguments, with a greater presence of direct and inferior complements of prepositions.

- Catullus wants Lesbia = $N_0V\ N_1$
- Catullus hates Lesbia = $N_0\ V\ N_1$

The classification operation is not simple because there are more lexical-grammatical entries than a single word, since the lexical system is rich and "irregular" in the creation of constellations due to the meanings that can be multiple: Max hates Maria:- Max is hateful with-Maria;- Max has hatred with Mary;- Max has in hatred with Mary. For a new grammar and a new positional calculation, and a new code like: Completive sentences have been defined as simple ones, because verbs have a semantic content that is not clearly defined and the sentence is completed when the first verb is completed with the other effect:

- $N_0$ V Ch S (43)
- $N_0$ V The fact Ch S(43)

It is a simple calculation for the production of these complete and direct sentences introduced by the phrase the fact Ch:

- $N_0$ V Che F cong = Enea checks that everything is in order
- $N_0$ V F o se F = Enea checks whether Max has told the truth

In the lexical-grammatical classification of the Italian Elia *et al.* (1984), the class No V Ch F has a remarkable presence of emotional verbs. These verbs are 440, of which 298 are inserted in the class (43) Elia (1984). Verbs that are included in the class (43) have a homogeneous behaviour: they present a human subject (active), except for someone who is not active [-human] (Elia, 1984, p.16). In the following tables the occurrences and the computational probabilistic calculation of the verbs of will are explained, as we would define them, properties of the class (43) and among these it is opportune to include

the extension of N0 V Ch F(43) (love, hate) :

| | |
|---|---|
| $N_0$ =: Nhum | $N_1 = V^0\Omega$ |
| $N_0$ =: Nnr | $N_1$ = di $V^0\Omega$ |
| $N_0$ =: il fatto Ch F | $N_1 = N_{1\,ch}\Omega$ |
| $N_0$ =: $V^1\Omega$ | $N_1$ = se Fo se F ciò |
| | |
| $N_0$ di $V_1$ Ω | Ppv = Io |
| $N_0$ V | $N_1$ = Nhum |
| $N_0$VN$_1$ contro $N_1$ | $N_1$ = N-hum |
| $N_0$V (presso con) Nhum | $N_1$ = il fatto ChF |
| | $N_1$ di Nhum |
| $N_0$ =: Che F | $N_1$ da $N_1$ |
| $N_0$ =: $V^0\Omega$ | $N_1$ dal fatto ChF |
| $N_0$ =: Aux $V^0\Omega$ | ChF a $N_1$hum |
| $N_0$ di $V^0\Omega$ | Passivo |
| $N_0$ =: di Aux $V^0\Omega$ | |
| Neg/interrog = Fcong | $N_1$ =: $N_1$Agg$^1$ |
| Imp = Fcong | $N_1$ =: Agg$^1$ChF |
| $N_0$ =: Che Fcong | $N_0$V (essere cong) |
| $N_0$ =: Fcong | $N_1$ hum per Aux $V^1\Omega$ |

Fig. 1 - A Descriptive table and inclusion in classroom (Elia, 1984)

## 4 Linguistic environments NooJ: Probabilistic latent semantic analysis"

Starting from the description of the linguistic environment Nooj we propose the transformation of Silberzstein's sentence into a probabilistic calculation. [… *it is true that the probabilistic calculation must be elaborated in the laboratory, but man unconsciously produces involuntary calculations in the manipulative reproduction of some textual techniques, or for advertising and market needs. As indicated in this analysis (Silberztein 2016)]* NooJ allows linguists to formalize various types of linguistic description: orthography and spelling, lexicons for simple words, multiword units and frozen expressions, inflectional and derivational morphology, local, structural and transformational syntax. One important characteristic of NooJ is that all the linguistic descriptions are reversible, i.e. they can be used both by a parser (to recognize sentences) as well as a generator (to produce sentences). (Silberztein 2011, 2016) show how, by combining a parser and a generator and applying them to a syntactic grammar, we can build a system that:

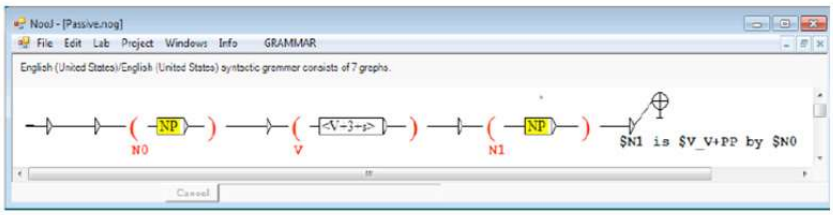


Fig. 2 - A descriptive table in formal logic (Silberzstein, 2011)

In the sentence *Joe loves Lea* = $N_0$ *V* $N_1$ the three variables, in which the acronyms were used, were used:-*variable = $NO = Joe's acronym;* -variable = $V = acronym of loves; variable = $N1 = acronym of Lea. Outgoing acronyms, second ALU: Plays the string: *$N1 is $V_(V_V+PP) of $NO* which equals Lea is loved by Joe:- *$NO cat. the word Lea; $V op. supp. (is); $V op.optional choice (love, lover etc.) $N1 cat. the word Joe;* The author shows how in Silberztein (2016), any serious attempt to describe a significant part of a language will involve the creation of a large number of elementary transformations. "Probabilistic latent semantic analysis" (PLSA), also known as probabilistic "Latent semantic indexing" (PLSI, especially in information retrieval circles) is a statistical technique for the analysis of two-mode and co-occurrence data. The purpose of the "EM algorithm" (Hoffmann, 1999) is to increase, and possibly maximize, the probability of the parameters of a probabilistic model M with respect to a set of data, results of a stochastic process that involves an unknown process, thus indicating with the current $\Theta^0$ parameters of the model. The objective is, therefore, to obtain a new set of parameters $\Theta$ such that: The standard procedure for maximum likelihood estimation in latent variable models is the Expectation Maximization (EM) algorithm. EM alternates two coupled steps: (i) an expectation (E) step where posterior probabilities are computed for the latent variables, (ii) an maximization (M) step, where parameters are updated. Standard calculations yield the E-step equation: parameters $\Theta$ such that: $\log P(s|\theta, M - \log P(s\theta^0, M) > \_0$

By introducing the hidden variables, we will have:]

$$P(s|\theta, M) = \frac{P(s, \pi|\theta, M)}{P(\pi|s, \theta, M)}$$

So, moving on to logarithms:

$$\log P(s|\theta, M) = \log P(s, \pi|\theta, M) - \log P(\pi|s, \theta, M)$$

Multiplying the current parameters by the probability distribution of the hidden variable M, $P(\pi|s, \theta^0, M)$ and adding up all the values that the hidden variable can take is obtained:

$$\log P(s\theta, M) = \Sigma_{\Sigma} P(\pi|s, \theta^0, M).(\log P(s, \pi|\theta, M) - \log P(\pi|s, \theta, M)$$

An auxiliary function is defined $Q(\Theta \mid \Theta^0)$ as the expectation value of the logarithm of the joint probability of s and p on the possible values of the hidden variable:

$$Q(\theta|\theta^0) = \Sigma_{\pi} P(\pi|s, \theta^0, M).\log P(s, \pi|\theta, M)$$

The expression to be made maximum becomes:

$$\log P(s|\theta, M) - \log P(s|\theta^0, M) =$$
$$= Q(\theta|\theta^0) - Q(\theta^0|\theta^0) - \Sigma_\pi P(\pi|s, \theta^0, M).log = \frac{P(\pi|s,\theta,M)}{P(\pi|s,\theta^0,M)}$$

The third term of the second member of this equality is the relative entropy of the distributions $P(\pi||s,\theta,M)P(\pi|s,\theta,M)$ which, as seen in the previous section, is always positive. It follows that

$$\log P(s|\theta, M) - logP(s|\theta^0, M) > -Q(\theta|\theta^0|\theta^0)$$

This inequality is the core of the EM algorithm. In fact, if we can calculate a set of parameters $\Theta^0$ that makes the difference of the auxiliary functions positive; this will increase the probability of the model with respect to the data. In particular, the objective is to find the values that $\Theta^{MAX}$ maximize this difference, that

$$\theta^{MAX} = argmax_\theta\ Q(\theta|\theta^0)$$

The EM algorithm is therefore composed of two steps

- Calculation of the expectation value $Q(\Theta \mid \Theta^0)$ starting from the parameters of the current model
- Maximization of $Q(\Theta \mid \Theta^0)$ in the variables $\Theta$ in the variables

From an initial hypothesis about the parameters of the model, these two steps are applied iteratively until convergence is reached when the updating of the parameters no longer increases the probability. The algorithm does not guarantee the achievement of the maximum global probability, but only its increase with each subsequent application and the convergence to a local maximum. In addition, sometimes it is not possible for this to carry out the maximization stage exactly, or at least not in an efficient and computationally economic way From grammar to the description of an automaton: Once you have obtained a context-free grammar, it is easy From grammar to the relative non- deterministic automaton as S: =aS|aB; B::b|B = we will have finished robot

## 5 Finite grammar and infinite languages

Viewed from this perspective, we take the view that we might draw an analogy between Universal Grammar Model and Second Language Learning. Universal Grammar might be able to enable students to map or link the structure of a foreign language that will last forever, even if students do not study this

second language any more. Later, if we wanted to reach the mastery of any second language, we would have to go over it by practicing its language skills. That is to say, learning a second language might be considered as a gradual change from declarative to procedural knowledge. In order to achieve this, students may use some strategies, which begin as declarative knowledge that can become proceduralized with practice (procedural knowledge).Then, how could this be explained in a more detailed way?Anderson 1983, 1985 (cit. in O'Malley & Chamot, 1990) defines declarative knowledge "as the knowledge about the facts and things we know and stored in terms of units of meaning that can be represented by propositional networks requiring a schema."The principal value of schemata is that they facilitate making inferences about concepts; consequently, in learning, the new information is linked to prior knowledge stored in memory in the form of knowledge frameworks or schemata. Here, in our view, the principles and parameters model of Universal Grammar plays an important role by building up a mental dictionary in the students' mind. For example, according to the Oxford Advanced American Dictionary: *Entrust / In 'trust / verb / [VN] Entrust A (to B) / Entrust B with A* to make someone responsible for doing something or taking care of someone. As procedural knowledge is concerned, as well, Anderson (1983, 1985) defines it "as the things that we know how to do and includes mental activities such as language production skills (writing, speaking), and language comprehension skills (reading, listening). In line with the previous example, this is the result:Entrust A to B**.** *He entrusted the task to his nephew.* Entrust B with A. *He entrusted his nephew with the task.*Here, we believe that the principles and parameters of Universal Grammar also plays an important role by enabling students to know not only the dictionary meaning of words or pronunciation, but also how they are used and behave in sentences as well as the creation of the ability to interact with other people.In other words, Chomsky (1964) distinguishes between syntactic and lexical components, on the one hand, and between deep structure and superficial structure of the syntax, on the other. Based on these assumptions, then, as Noawak *et al.* (2002: 612) indicate, a grammar is a finite list of rules specifying a language: Subsequently, as Noawak *et al.* (2002: 612) specify "there is a correspondence between languages, grammars and machines. 'Regular' languages are generated by finite-state grammars, which are equivalent to finite-state automata. Finite-state automata have a start, a finite number of intermediate states and a finish." Progressing in the exposed sense, Chambers *et al.* (2004) also bring to light certain aspects relating this topic when they state that experts and researchers in the field of Information and Communications Technologies (ICT) and language learning are increasingly emphasizing that, once a new form of technology has become available, the starting point of research projects should not be the innovation

itself but rather its role in the language learning process. Nevertheless, as stated by, Popper's epistemology and its demarcation criterion of science that makes the scientific nature of theories coincide closely with their falsifiability, is the N. Chomsky model, which with the descriptive grammar marks the transition of a heuristic model of formal grammar to the language rewriting. It proposes algorithmic forms to explain linguistic facts and shifts the centrality of the representation role of the semantic unity of signification from the headword to the minimal sentence, to describe and analyze. It follows the reference model Nooj M. Silbersztein for the description, analysis, production of fixed sentences, paraphrasing of sentences, and is specified on the facts of automatic data processing. Like this manner, this paper aims to integrate research and practice in this emerging field for further research and development in these areas.

## Conclusion

In this validation process the team has tried to make sense of this research,elaborating a working hypothesis, built on scientific bases, proposing choices of models, technologies in use, empirically valid theories. The validation technique presented is: -dissertation on natural languages; analysis and description of the language according to the lexicon-grammar; transformation into online languages and data collection and description of a fixed sentence. The hypothesis ends with a heuristic certainty on the calculation of quantum emotions. The answers we give are the suguenti: The human mind will govern the robotic mind with infallible tools, but will it be able to transmit real emotions? The research continues.

# REFERENCES

Chomsky, N. (1964), *Aspects of the Theory of Syntax*. Massachusetts Inst of Tech Cambridge Research Lab Of Electronics.

Elia, A., Martinelli, M., & d'Agostino, E. (1981), *Lessico e strutture sintattiche: introduzione alla sintassi del verbo italiano*. Liguori, Napoli.

Chambers, R. D. (2004). *Fluorine in organic chemistry*. CRC press.

Elia, A. (1984), *Le verbe italien: les complétives dans les phrases à un complément*. Schena; Nizet.

Gross, M. (1986), *Grammaire transformationnelle du français, Syntaxe du verbe*. Cantilène, Paris.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading research quarterly*, 233-253.

Harris, Z. S. (1964), *Transformations in Linguistic Structure*. Proceedings of the

American Philosophical Society 108:5, pp. 418-122.

Hoffmann, T., (1999), "Proceedings of the Fifteenth conference on Uncertainty…"

Ibrahim, A. H. (2003), *Le cadre du lexique-grammaire*. Linx. Revuedeslinguistes de l'Université Paris X Nanterre, (48), 101-122).

Landi E. Bucciarelli R. Landi A. (2000), *Dalla grammatica al testo poetico: lezioni di linguistica*. Loffredo editore Napoli

O'malley, J. M., O'Malley, M. J., Chamot, A. U., & O'Malley, J. M. (1990). *Learning strategies in second language acquisition*. Cambridge university press.

Page, K. M., & Nowak, M. A. (2002). Unifying evolutionary dynamics. *Journal of theoretical biology*, *219*(1), 93-98.

Page, K. M., & Nowak, M. A. (2002). Unifying evolutionary dynamics. *Journal of theoretical biology*, *219*(1), 93-98.

Popper, K. R. (1963), *Conjectures and Refutations*, Routledge and Kegan Paul, London, trad. it. *Congetture e confutazioni*, Il Mulino, Bologna 1972.

Silberztein, M. (2015). *Joe loves lea: transformationa lanalysis of direct transitive sentences*. In International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ (pp. 55-65). Springer, Cham.

Veronesi, C. (2007). *Popper filosofo della matematica*. Collana Pristem / Storia,monografia n. 18,Centro PristemEleusi, Università Commerciale Luigi Bocconi,Milano 2007, pagine 93.