Advanced Mathematical Models & Applications
Vol.8, No.3, 2023, pp.437-451

Jomard Publishing

NEW UNIVERSAL LABELING STRATEGY FOR MEANING REPRESENTATION IN NLP

D. Monte-Serrat[1*], **E. Ruiz**[1], **C. Cattani**[3]

[1]Department of Computing and Mathematics, FFCLRP, University of Sao Paulo (USP), Brazil
[2]Department of Engineering, UNITUS, University of Tuscia, Italy

Abstract. This article shows that the unambiguous meaning representation of machine language must be inspired by human cognition to find a universal label for Natural Language Processing resources. We show, that the meaning representation is linked to the quality and not to the quantity of data, since the architecture of human symbolization, taken as a model, incorporates contextual elements in its dynamics. We argue that language, when conceived under this dynamic concept, must be analyzed in terms of sequence-by-sequence transformations, focusing on the structural bases of the meaning construction that: i) offers relative certainty for different possible responses; ii) provides more accurate and reliable input data; and iii) avoids the curse of dimensionality. We conclude by suggesting that the construction of databases should indicate the upper hierarchical structure based on the lower structures, referring to the string-to-meaning derivation in the fundamental tree of the language. This strategy, applied to meaning representation, frees the operator from dictating a specific sequence for each application. Thus, the construction of meaning in natural language processing becomes more contextualized and unambiguous. Based on the dynamic structure of the language, therefore, it is possible to develop semantic databases in any spoken language.

Keywords: Meaning representation. Natural Language Processing. Language Dynamic Concept.
AMS Subject Classification: 68T50 Natural Language Processing.
Corresponding author: Dioneia, Motta Monte-Serrat, Faculty of Philosophy, Sciences and Letters of University of Sao Paulo at Ribeirao Preto, Brazil, Tel.: +55 16 3315-0429, e-mail: *dserrat@unaepr.br*
Received: 9 July 2023; Revised: 4 August 2023; Accepted: 28 October 2023; Published: 20 December 2023.

1 Introduction

Professionals and researchers working in the field of language processing, whether in linguistics or computer science, are used to gaps between theory and practical implementation of tools. This article aims to fill these gaps by proposing that language, whether human or computational, involves a universal structure that builds understanding and cognition.

Taking functioning of human cognition as a source of inspiration for the language of Artificial Intelligence, AI, it is possible to conceive that language and cognition, human or artificial, are connected. The representation of meaning is not treated by consensus between semanticists and philosophers of language (Pitt, 2020). Also, theories and theoretical concepts about language vary substantially. It is necessary not to lose the thread of these theories. There is, in language, be it natural or representational, something in common in the building of meaning: Both contain a sort of representation.

Representation, linguistically speaking, is a hotly debated subject (Pitt, 2020), and the fact that it overlaps natural and artificial language (Monte-Serrat & Cattani, 2021; 2021a) requires

How to cite (APA): Monte-Serrat, D., Ruiz, E., & Cattani, C. (2023). New universal labeling strategy for meaning representation in NLP. *Advanced Mathematical Models & Applications*, *8*(3), 437-451.

more analysis and careful thought to give it a linguistic treatment with more attention than has been the case so far. We believe that this is a fundamental element to guide some conceptions related to meaning representation in computational language, to constitute a strategy that reinforces the quantitative analysis as the basis of formal sciences.

Generic linguistic principles applied to the intelligent system in the task of building a simple and readable database of sentences can be the foundation for them to be extended to databases of any language, since they are principles closely linked to a structural feature of the language: the symbolic system.

Once one works with language, symbols (gestures, letters, numbers etc.) are manipulated. The symbolic system makes the intermediation between the real world and the individual (human cognition) (Voloshinov, 1986; Wallon, 1949), so that the latter can interpret, giving meaning to the environment. Symbolic systems have then, as a principle, manipulation: Since they make intermediation, they depend on 'someone' to represent, process and act on information.

On the other hand, the expression 'abstract meaning representation' suggests an abstraction, better saying, presupposes a lack of manipulation in the process of meaning formation for language computer programming. Abstraction or representation consist of theoretical knowledge. Knowledge, in turn, results from a process that incorporates elements of contextual reality and elements from linguistics to cognition (Copstead-Kirkhorn et al., 2013; Terlow, 2020). This process sets up a mechanism that organizes cognition in a dynamic balance between functional segregation and integration of subparts in a network (Friston, 1994; Tononi et al., 1994; Sporns, 2013). Very little is yet known about how the human mind represents things in the outside world through symbolic systems. Dealing with language means dealing with symbols in a context that prints values to the symbolic system to avoid ambiguities (Monte-Serrat & Cattani, 2021; 2021a). It is at this point that the abstract meaning representation (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014) detaches from the context of the in vivo language, in order to establish previously categorized contexts, in specific sequences dictated by the operator for the in silico language, focusing on the symbolization architecture.

Grammatical content detached from contextual content may mislead the reader into (Maia & Santos, 2018) error. The British linguist Halliday claims that language results from a potential network of meanings constructed from a context. Halliday, (1961; 1995) explains that the abstract categories of grammar give coherence to language through "relationships" between systems, but language still needs a context of semantic choice (Halliday, 1961), of a sociological and dynamic event so that the meaning is completed and does not give rise to ambiguity.

Language cannot be understood only as a set of all grammatical sentences or as something essentially linked to formal logic and its dichotomies (Halliday, 1985; 1995). Language should not be seen just as systemic, but as systemic and functional (Halliday, 2003), considering the interaction of human beings with the environment.

A text is defined (Halliday, 2006, p.4) as "an instance of social meaning in a given context of situation". It is important to apprehend that the interpretation is completed with the situational context in which semantic choices are made, closing the semiotic cycle: "the network that extends from the social system, as its upper limit, through the linguistic system on the one hand and the social context on the other, to the 'word', which is the text in its lexicogrammatical realization" (Halliday, 2006, p.4). The interpretation of language only in its logical (relational) aspect leads to decontextualized relationships, as is the case with BERT (Bidirectional Encoder Representations from Transformers, language representation tool for artificial intelligence, developed by Google) which sometimes makes relationships removed form a context, such as claiming that a bird has twice the probability of having four legs rather than two; or the case of GPT-3 (Generative Pre-trained Transformer 3, language tool that uses deep learning to produce humanlike text), which gives answers such as that grass blades have eyes or that a horse has four eyes Edwards (2021).

In computing, natural language has been related to different spoken languages like English,

French, German etc. One of the branches of computational science that works with this concept of language is called Natural Language Processing (NLP), considered a subfield of Artificial Intelligence that seeks to program computers to process and analyze large amounts of language data, using machine learning algorithms such as, for example, decision trees. This kind of research is mainly based on statistical models, which make probabilistic decisions, producing more reliable results when they are integrated with a larger system that takes advantage of multilingual textual corpora, but even so, there is a limitation in the success of these systems. Statistical training related to analysis techniques leads to results that modify the logical relationship between words; this corrupts the formal grammar. Researchers confuse (Bender et al., 2015) the distinction between meanings determined by the linguistic sign (sentence meaning) and meanings determined by a context of use (enunciation context). Linguistic semantics provides characterization of conventional content and the human activity of interpreting needs more knowledge such as sources and processes, in short, they need extralinguistic tasks.

The great majority of published research, even in NLP, deals with already established elements, replicating them to discover something different. This is not the case with this article, as we suggest a new conception of natural language to be considered as Natural Language Processing, NLP, by researchers and technicians. This is therefore a disruptive article in relation to the most linguistic theory applied to mathematical modeling. We propose that the representation of abstract meaning (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014). Many AI works give the spoken languages the name of natural language, seeking to be free from the inconsistency of the living language, and, at the same time, making the language static within the classification of concepts. It is worth remembering that, at the foundation of linguistic theory (Saussure, 1989), speech (parole) was despised to make way for writing (langue).

As it is intended, in this article, to teach the machine how the representation process takes place in human cognition, a broader concept of natural language should be sought so that it obtains reliable results in language analysis. It is established, from this point, the nomenclature of 'conventional language' for the languages spoken, such as English, Portuguese, French etc.; and 'natural language' for the entire cognitive process that involves human language, from the reception of stimuli to the mental representation that corresponds to an object of an external context.

Working with this broader concept of natural language, we believe that we can contribute to an improvement of what should be an abstract meaning representation (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014). We use an interdisciplinary theoretical approach of concepts related to the symbolization process in cognitive systems research, to better understand human-level cognition applying it to advance understanding, design and applications related to AI computer programs.

The theoretical foundations of abstract meaning representation (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014) deal with limitations in semantic representation because they are based on conventional language. They start from a restricted perspective of language and some failures may occur in their attempt to establish a universal quantifier. A paradox is established, for example, of semantic representation without reference to its own source (Banarescu et al., 2013 p.184). If we speak of 'semantics', we necessarily speak of context and source, because semantic knowledge is related to information processing taking the context into account (Squire & Wixted, 2011). This is one of the reasons why we suggest starting with the broad concept of natural language while validating quantifiers.

The fundamentals of abstract meaning representation (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014) tend to provide an abstraction by suggesting the existence of a common structure for whole sentences – paradoxically, in our opinion, regardless of the context in which they are inserted-, so that they can be used to formulate a comprehensive database. We promote a debate on this in the sense of integrating new ideas and concepts from human cognition, comparing the theoretical bases of abstract meaning representation with linguistic

ones, to locate the paradox and check if we are on the right path to form basic NLP resources that best fit the Portuguese language not yet investigated. The discussion of the difference between natural language and conventional language is important because it brings arguments to visualize what are the necessary elements for the formation of a simple dataset and also helps to elucidate papers that are supported in the language statistical generation.

The content of this article follows this order: In Section 2 we tell how inputs are parsed and translated in the modeling of human reading comprehension according to the theory of Abstract Meaning Representation (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014) and we contrast it with a suggested model based on a dynamic language concept to be reproduced by AI in a way to solve real-world problems. Section 3 deals with more generic and structural annotation guidelines to facilitate the construction of dataset in Portuguese. For this, we reproduce some examples of AMR theory (Banarescu et al., 2013, pp.178–179), showing them simultaneously from the perspective of the linguistic theory and under the comprehensive concept of natural language. Some suggestions on how to design a dataset structure that is abstract are covered in Section 4. The AMR theory (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014) intends to share a universal structure but reducing semantics to categories makes the language static. We propose an abstract structure based on the dynamic nature of language (Verspoor, 2013; De Bot et al., 2007; Monte-Serrat & Cattani, 2021; 2021a) to arrive at a universal structure of meaning representation present in all spoken languages. In Section 5, we show that, while the AMR theory (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014) suggests extracting information from specific types of data, restricting itself to the labeling given previously and training machine learning to decontextualized patterns; information extraction can, on the other hand, follow recursive language patterns, conditioning the elements mutually to assume a generic purpose representation, which can occur from unrestricted text. We conclude, in Section 6, with the suggestion of applying labeling strategies with the following advantages: i) to concentrate the procedure on the structural basis of the construction of meaning, directing efforts to produce robust models suitable for unknown or error entries (misspellings or omitted words), but which have the same structure offering relative certainty of many different possible answers rather than only one. Thus, handwritten rules that take a long time and are prone to errors are avoided; ii) provide more accurate and reliable input data, instead of increasing the complexity of the rules, which makes the system unmanageable.

2 Theoretical approach on statistical Natural Language Understanding, NLU

Natural language understanding (NLU) analyzes and translates inputs according to natural language principles being, therefore, related to the modeling of human reading comprehension. According to abstract meaning representation theory (Banarescu et al., 2013 p.178; Damonte et al., 2016; Flanigan et al., 2014) syntactic treebanks are useful in language processing due to the combination of labeled phrases, but it is understood that the syntactic base is restricted due to the offer of only a few labels, leaving aside an immense list of possible labels given the complexity natural language under a broad concept. In this case, once working with representation, one should seek theoretical knowledge of linguistics to help bypass the problems arising from the construction of clauses, improving, this way, the accuracy of the meaning formation. Knowledge, on topics such as relationship and value (Monte-Serrat, 2021), clarifies why abstract meaning representation theory (Banarescu et al., 2013 p.178; Damonte et al., 2016; Flanigan et al., 2014) agree that better results come from the analysis of whole sentences instead of separate identification (such as those that evaluate base noun, prepositional sentence attachment, dependencies on verbal arguments etc.). It is understood that these minor tasks should be investigated as a by-product of analyzing entire sentences, since only the latter provide the context that avoids ambiguity.

Separate notes for named entities interfere with the assessment associated with them (Banarescu et al., 2013, p.178; Damonte et al., 2016; Flanigan et al., 2014). Thus, it is expected that the multiplicity of characteristics of these notes will contribute to the formation of the logical meaning of the sentence. On the other hand, a comprehensive conception of language that includes the cognitive capacity of mental representation, helps to explain how meanings are formed. This strategy replaces the multiplicity of annotation characteristics, giving the analyst better means to search for a simple and readable sentence structure from the dataset to be combined with the sentence logical sequence. This is a procedure of reproduction, in AI, of cognitive architecture to solve real-world problems. We believe that this solution path offers insights for new works on understanding conventional language that sheds light on new models of machine learning.

Some strategic suggestions based on an interdisciplinary approach are provided due to the complex structure of natural language. A theoretical approach is adopted in which sentences are paired with their logical meanings, producing a better result for natural language generation (NLG). At the same time, arguments are offered, based on linguistic theory, justifying why the lack of contextualization of the separate notes also requires a logical semantic input appropriate to the context.

3 Expected Outcome: Providing more generic and structural annotation guideline to facilitate dataset construction in Portuguese

The cognitive system - understood as a system that incorporates contextual reality into linguistic elements (Monte-Serrat & Cattani, 2021) offers the opportunity to deepen this research, making it both theoretical and empirically informed, for contributions to the development of the theory computational modeling. For considerations on how to rethink an 'abstract' (decontextualized) meaning representation, some of its principles are reproduced (Banarescu et al. 2013, p.178–179; Damonte et al., 2016; Flanigan et al., 2014), to be showed concomitantly from the perspective of majoritarian linguistic theory and under the comprehensive concept of natural language. The focus is to direct these principles to a more general and structural annotation guideline to facilitate their use in the Portuguese language and suggest indicators for assessment tools and software:

a) Abstract meanings are labeled and rooted graphics to make it easier for people to read and to be machine readable (Banarescu et al., 2013, pp.178–179).

This feature would be justified to avoid setting up meaningless sentences. It is important to remember Chomsky's (Chomsky N.1957) lesson that grammaticality does not correspond to making sense, that is, grammaticality alone is not sufficient to construct meaning. The meaning also stems from an underlying sentence structure, which is common to all languages spoken (Pinker, 2000; Devlin, 2000, p.173; Chomsky, 2001; Jespersen, 1992; Caramazza & Shapiro, 2004; Halliday, 2006), as will be explained later on.

b) 'AMR [abstract meaning representation] seeks to abstract from syntactic idiosyncrasies. A sentence with the same basic meaning can be written in many ways and belong to the same AMR' (Banarescu et al., 2013 p.178).

Syntax has been losing its importance for computer programming due to the complexity imposed by the large number of rules and the difficult handling. Because of this, contradictory rules require additional formulas to resolve the conflicts. In order not to need to rely on syntax, a sentence structure is sought that helps in the construction of meaning (for example, of affirmation, opinion, prohibition etc.), so that the order of words in the sentence does not interfere with the final meaning. In the following example (Devlin, 2000, p.175) the order of the sentence components does not change the declarative structure that gives it meaning and exemplifies:

The major killed the maid in the library with a dagger.
The major killed the maid with a dagger in the library.
The maid was killed by the major in the library with a dagger.
The maid was killed by the major with a dagger in the library.

It is important to say that the abstraction of idiosyncrasies leads to an abstract form of the official language, making it a little artificial. This language sanitization removes variations to facilitate decoding. Language is commonly undergoing variations in its conventional forms. Dataset will never cover all these innovations. For this reason, this research seeks more general rules capable of circumventing these differences.

c) The derivation of string meanings (Banarescu et al., 2013, p.178) is done independently, that does not follow a specific sequence to apply rules or alignments. The authors state that this facilitates the use of dataset with varying use of strings for meanings. This rule stems from the use of the underlying structure of the sentence, which gives it meaning regardless of the order of the words, as explained in the previous item. It should be emphasized that this independence is relative, since the sentences are naturally formed by mental processes of language production that are closely related to logical reasoning (Monte-Serrat & Cattani, 2021; Monte-Serrat & Cattani, 2021a). Word order is a superficial feature of language; the important thing is to see the sentence, not the word, as the basic unit in the construction of these phrases (Devlin, 2000, p.176; Halliday, 2006).

These excerpts from the abstract meaning representation principles followed by our comments show some challenges to be faced in any machine learning, ML, and data analysis project. This paper focuses on searching for some particularities of the broad sense of natural language that meets the needs of conventional language analysis that circumvent the problems pointed out and that work in a more intuitive way, reducing the probability of ambiguity.

We understand that language processing will be successful when the technology is applied according to the structure of human language (under the broad concept of natural language adopted in this article), because this structure somehow addresses the issue of value and context (Monte-Serrat & Cattani, 2021; Monte-Serrat, 2021), decreasing the appearance of ambiguity. Little is known about the problems of choosing texts in the early stages of analysis. It is expected that the suggestions brought in this article operate on a series of results in such a way that they reach the state-of-the-art in order that the system allows the communication of the expected information similar to the way of human beings thinking.

4 Method to overcome the challenge: Propose a 'real' abstract dataset structure

Writing abstract meaning representation as tree-shaped graphs (representing the hierarchical nature of a structure) is not new in resource structures (Pinker, 2000; Devlin, 2000). Abstract meaning theory (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014) 'intends' to share a universal structure. We emphasize 'intend' because that theory reduces semantics to categories, which means making the language static.

Natural language has a dynamic nature because it presents a set of variables that interact with the context (Monte-Serrat & Cattani, 2021; 2021a; Monte-Serrat, 2021), have temporal dependence (Monte-Serrat & Cattani, 2021; Verspoor, 2013; De Bot et al., 2007), depend on the initial conditions, are interconnected with the subsystems, vary between individuals (Verspoor, 2013; De Bot et al., 2007). Therefore, we propose that the tree-shaped graphs share a universal structure whose existence is recognized by some authors (Jespersen, 1992; Devlin, 2000, p.173; Pinker, 2000; Chomsky, 2001; Caramazza & Shapiro, 2004), that is, a structure that is present in all languages: The axiomatic-logical structure (Monte-Serrat & Cattani, 2021). Instead of categorizing semantic groups, we seek a solution in the way the linguistic process attributes value to words (Monte-Serrat, 2021).

As suggestion we present an analysis (Devlin, 2000, pp.179–180) of basic structures such as nouns, verbs, modifiers (adjectives, adverbs), pronouns and prepositions that, generally (in addition to differences in word order and some other details), offers the formation of minor sentences constituting longer sentences in essentially the same way. Putting these elements in the tree structure we have: head, specifier and complement (Devlin, 2000, p.322), a structure that comes from a single combinatory rule and is repeated indefinitely (Figure 1).

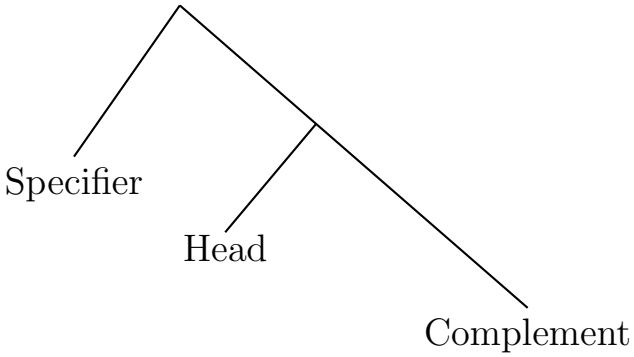

Figure 1: (Devlin's tree structure. 2000, p. 322)

The figure 1 shows (Devlin, 2000, p.322) that the head is the only element of the sentence that must be present; the other two (specifier and complement) are optional. The head should have no more than a single word. The head of a noun phrase is a name, of a verbal phrase is a verb, and of an adjective phrase is an adjective. The head has considerable control over the sentence. This is an important linguistic principle: That words acquire meaning because they are related to other words in a sentence (Saussure, 1916; Monte-Serrat & Cattani, 2021). The tree structure adequately represents the structuralist view of language proposed by Saussure (Saussure, 1916), which encompasses underlying abstract rules to generate observable linguistic structures.

There is (Chomsky, 1957) another path, studying language from the prescription of correction norms, arguing that a set of phrases can be generated from a given set of formulas. Chomsky's works have a short validity, as they do not face the difficulties caused by the dynamic system of natural language.

To work on the symbolic nature of language in Artificial Intelligence, AI, it is necessary to consider elements arising from natural language under the comprehensive concept of a dynamic system (Verspoor, 2013; Monte-Serrat & Cattani, 2021; 2021a; De Bot et al., 2007). The representation of language (Chomsky, 2006, p.107) is established by leaving aside the linguistic representation of intonations, which have the semantic function of interfering in the construction of meaning.

Computer languages based on grammars without context (Wolfram, 2002, p.1103) lead to occasional deviations from the model, which makes interpretation difficult:

The idea of describing languages by grammars dates back to antiquity [. . .]. And starting in the 1800s extensive studies were made of the comparative grammars of different languages. But the notion that grammars could be thought of like programs for generating languages did not emerge with clarity until the work of Noam Chomsky beginning in 1956. And following this, there were for a while many efforts to formulate precise models for human languages, and to relate these to properties of the brain. But by the 1980s it became clear - notably through the failure of attempts to automate natural language understanding and translation - that language cannot in most cases (with the possible exception of grammar-checking software) meaningfully be isolated from other aspects of human thinking.

The mental representations corresponding to real-world experiences are structured in the brain in the form of a chain of ideas (Monte-Serrat & Cattani, 2021), not in separate words.

This is an important principle of natural language conceived as a dynamic system that interacts with the context in which it is used (Voloshinov, 1986; Verspoor, 2013; De Bot et al., 2007; Pecheux, 1975; Pecheux, 1988; Lacan, 1949; Monte-Serrat & Cattani, 2021). The receiver of the message decodes it by reconstructing the mental representation as it receives information, so that the 'whole' of the mental representation is an addition to the 'pieces' of information. Mental representation is part of a linguistic process by which an idea is formed in the human mind to correspond to the real world (Pitt, 2020). For these reasons, the suggestions given in this article have more support in semantics than in syntax, giving preference to the tree structure of the phrases, instead of focusing on the branches without considering the whole structure.

Just to illustrate, the logical structure of language (Chomsky, 1961, p.10) describes grammar containing left-recursive symbols, which generate P-markers that branch indefinitely to the left; right-recursive symbols, which branch to the right; or self-embedding symbols, that generate configuration (iii) containing nested dependencies of arbitrary depth in the resulting terminal strings (Figure 2):

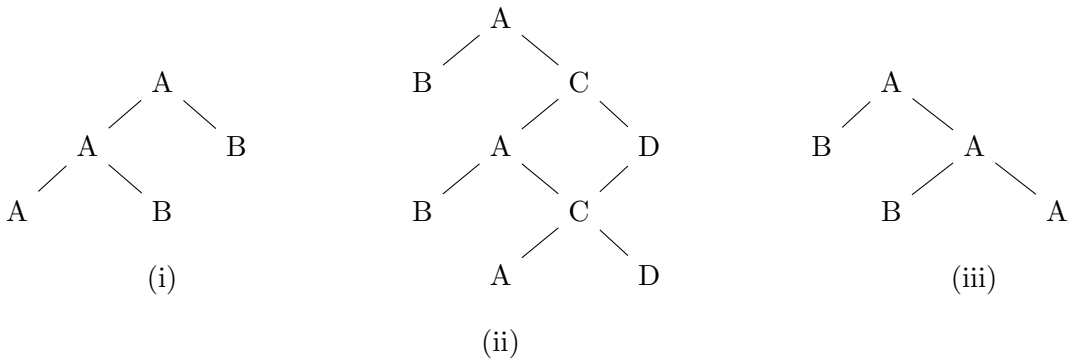

Figure 2: Chomsky's (1961, p. 10) logical structure of language.

The natural language, under the comprehensive concept, goes beyond Chomsky's structure, housing semantics and logic. The concept of conventional language (Chomsky, 1961) is more restricted and deals only with logical rules of sentence construction, standing out from semantics (Monte-Serrat, 2021; Monte-Serrat & Belgacem, 2017, p.21; Hamp Eric et al., 2022). Grammar has the task of determining what must be true through logical reasoning, since logic has the role of determining the laws according to which the way how one thought will affect another thought (Derrida, 1967; Monte-Serrat, 2017).

On the one hand, conventional language gives more importance to the logical aspect of language, as it is based on grammatical rules of sentence formation (Derrida, 1967; Monte-Serrat, 2017). On the other hand, if the starting point of a machine learning operator is natural language as a dynamic system, semantics must be considered, that is, the relationship between what is said and the context in which it is said must be included in language analysis (Monte-Serrat, 2021; Monte-Serrat & Cattani, 2021). The contextualization gives greater security to the study of language in the search for meaning formation (interpretation). Logic takes the sentence out of context; semantics, in turn, takes this context into account in the meaning construction, reducing the occurrence of ambiguity. The search for disambiguation in Artificial Intelligence, AI, must be attentive to the right way to design the machine's core (algorithm) that best meets this need.

The constituent logical structure of grammar is represented by the tree format in the following diagram of the phrase "The man will hit the ball", explaining the way the sentence is generated (Figure 3):

Although there is a use of (Chomsky, 1961) a recursive figure similar to a tree, Chomsky dwells on the logical characteristic of language. The term 'treebank' was constructed for the

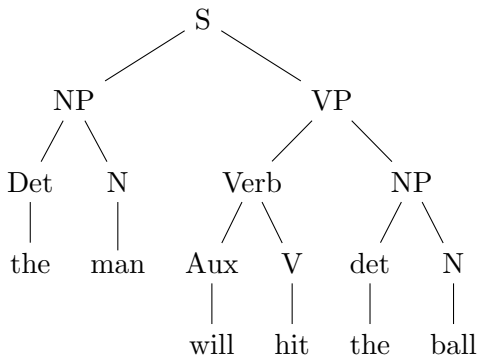

Figure 3: Structural description of a phrase-structure grammar. Adapted from (Britannica, 2018).

syntactic and semantic analysis of a corpus represented by the shape of a tree (Leech, 1981). Syntactic parsing has its own value for grammatical and statistical verification, but when it comes to the construction of meaning, semantic representation is essential (Monte-Serrat, 2021). The way in which semantic representation is worked, in the interaction between grammatical surface and use of background knowledge, is fundamental to the objective of reaching the state of art.

A semantic effect that makes a sentence meaningful, for example, comes from connecting elements between words or between short sentences. They are the so-called connectives, considered essential in the sentence / text hierarchy, as they are capable of leading to more generic structures of meaning. Attention to connectors facilitates abstraction, which allows one to deal with multifunctional processors. The connection elements are responsible for assigning a value (meaning) between one layer and another in the hierarchy, directing the formation of the ultimate meaning. Attention to the hierarchy of meaning-building helps to avoid what happened in the failed AI version of the Harry Potter book, as noted in the following excerpt (McCall, 2021):

'What about Ron magic?' Offered Ron. To Harry, Ron was a loud, slow, and soft bird. Harry did not like to think about birds. 'Death Eaters are on top of the castle!' Ron bleated, quivering. Ron was going to be spiders. He just was. He wasn't proud of that,but it was going to be hard to not have spiders all over his body after all is said and done (McCall, 2021; Chapter thirteen, Harry Potter).

That book McCall (2021) lacks the effect of the tree shape that deals with syntax and semantics concurrently, which would condition the meaning to a truth provided by the context. The emphasis on connectives between the hierarchical layers of sentences or texts signals whether this or that information is interesting or controversial. The incorporation of meaning throughout the text / phrase results from a mapping of the values carried by the connectors. The accommodation of this mapping of values is possible with the use of neural models (Zhang, et al., 2018; Vinyals et al., 2016) that adjust to the tree structure through which a previous condition interferes with the interpretation of the next layer.

The neural structure is presented in layers (3D) (see next topic), differentiating, therefore, from the tree structure in two dimensions (2D) much explored in the application of conventional language for Artificial Intelligence. In the next topic, we demonstrate the incorporation of meaning as the assumptions introduce information linking them to the referents previously provided in the sentence / text.

5 How a phrase assumes general purpose representation from unrestricted text

Abstract meaning theory (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014) suggests extracting information from specific types of data to fill databases. These databases are limited to answers intended for specific questions, restricting themselves to the contextual situations of the words examined in relation to the words next to each other to determine the specific relationships among these entities (in view of the labelling previously given). This extraction of relationships, based on specific patterns that connect entities and intermediate words, trains machine learning to decontextualized patterns, which leads to ambiguous phrases.

This limitation is due to the use of conventional language (static language) as the basis of the analysis (since tools based on the logical rules of the spoken languages are sought, disconnecting them from the context). It is suggested in this article that the basis be natural language as a dynamic system (Verspoor, 2013; Monte-Serrat & Cattani, 2021; De Bot et al., 2007). In this case, instead of looking for tools based on the logical rules of language, an appropriate tool is suggested for the dynamic process of language, which can unite its semantic aspects (dynamic / axiological / added to the contextual reality) to the logical aspects, helping thus the verification of which upper hierarchical structure the lower structures originate from; this means considering the context.

The dynamic aspect of natural language through recursion (Monte-Serrat & Cattani, 2021) integrates the input stimulus into the linguistic system until the construction of information. The recursive patterns of language act in the formation of meanings, conditioning the elements mutually (Monte-Serrat & Cattani, 2021a). As it is in the language structure, recursion should also be present in the parsing tools.

5.1 How to build treebanks

From a computational perspective, treebanks are built from evidence of the frequency of conventional language rules, such as markers and analyzers of parts of speech. Computer systems use this data to form the treebank pattern.

The fact that ambiguity persists leads us to suggest a new criterion for the construction of a semantic treebank: Using the model of neural networks in such a way that meaning and context are linked before moving on to the next layer of meaning. We suggest that treebanks be built with inspiration in the structure of argumentative discourses in which there is an opinion based on statements present in other layers of meanings, using neural network techniques, as they are recursive in nature, corresponding to the dynamic characteristic of language.

Theoretical bases of the abstract representation of meaning have been rethought to suggest that parts of the text be selected along with its context. This strategy removes or inhibits the meaning misunderstanding by following a line of reasoning that observes the structural dependence (context in which the meaning was structured). Priority is given to the evidence of the relationship between the hierarchical levels of the sentence or the text, which helps to visualize the relationship of the tree 'branches' in the formation of the decision content. In this way, one can evaluate the impact of semantic phenomena on the grammatical choice.

The accuracy of an argument leads to the formation of logical reasoning appropriately on a topic whose validity can be verified through linguistic representations. The arguments will be correct (believable) if they are in the correct logical form and will lead to a logical truth. This treebank construction technique suggestion attempts to capture the meaning of linking the database branches (layers of the discourse) in a semantic approach with a focus on logical structure. We seek to understand how meaning is constructed through propositions and logical connectives in an argument. At some point in the argumentative text or sentence a proposition or logical connective is revealed that triggers, within the tree structure (layers), an inference operation that will result in a specific meaning given by the context.

The truth conditions of the argument are found in the logic that determined the construction of meaning (trunk in relation to the branch of the tree under analysis), that is, in the proposition that gave rise to this meaning. The constructed meaning is a logical consequence (implication), it is a consequence of the relationship between propositions that can be true or false (in the latter case we will obtain a fallacy). Linguistically speaking, connectives play an important role in linking the conclusion to its premises in a sentence, in a paragraph and in a text. At this point, our study provides improvement to the limitations of the abstract meaning representation theory (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014), since the latter does not refer to the string-to-meaning derivation, on the contrary, it compares semantic representations directly, without reference to the source , remaining 'agnostic about the relation between strings and their meanings, considering this a topic of open research' (Banarescu et al., 2013, p.184). In short, if the meaning is attached to the syntactic properties in a disconnected way from the rest of the tree (context/layers), there will be a deviation from its specificity and most likely will lead to ambiguity.

We believe we are on the right path, because (Goodfellow et al., 2016, p.721), in computational terms, the probability that a model attributes to the data does not seem to measure any attributes of the model that we really care about. Authors (Goodfellow et al., 2016, p.721) arrived at the conclusion that many of the output values are trivial to predict, suggesting the need to develop other ways of evaluating generative models.

If one works with the recursive structure in syntactic parsing, combined with the neural model, they reduce the possibility of having more than one interpretation. In the argumentative example - If it is admitted that human life is man's most precious asset, the death penalty cannot be accepted (Fiorin & Savioli, 2006), the meaning of not accepting the death penalty is in the condition (other ˜ layer) of accepting that life is man's most precious asset; this sense is found at another hierarchical level. A meaning link (Zhang et al., 2018) be made at one level before continuing to interpret the next level. When this link does not occur, ambiguity arises, as in the following examples: 'The old man the boat' (where man designates verb to maneuver); 'The complex houses married and single soldiers and their families' (where the word 'houses' designates shelter). The treebank combined with the layered model of neural networks allows us to find the necessary logical consequence in a sentence or text. In this case, we must look for the common trunk from which the branches started.

In short, while the abstract meaning representation theory compares semantic representations directly without reference to the source (Banarescu et al., 2013, p.183; Damonte et al., 2016; Flanigan et al., 2014), our suggestions fall on the head-complement relationship or on the use of the connective, under the strategy of using a structure minimal neural value of the treebank, decreasing the task of finding a relation between the phrases / words in context.

This procedure by means of a minimal treebank structure investigates the construction of meaning due to the relationship of an expression with its antecedent that may be in another part of the sentence or discourse.

5.2 2D tree shapes

Our suggestion, therefore, is to work with 3D tree shapes inspired by generic neural network techniques that can be successfully applied to the processing of the Portuguese language or any other conventional language. As this article deals with the generic bases of meaning formation in a text or sentence, we focus on an approach consistent with the content developed here. For an excellent performance of Artificial Intelligence, we suggest as a strategy the processing of sequential data through language or translation models based on n-grams.

The language or translation model based on n-grams (Goodfellow et al., 2016, p.463) has the set of symbol strings partitioned according to a tree structure so that the later sharing corresponds to the previous one. We can adapt n-grams (Goodfellow et al., 2016, pp.565-568) to the abstract meaning representation, which can be rethought as a model in which the direction of

the arrow indicates the probability of distribution of meaning, defined in context terms (Figure 4).

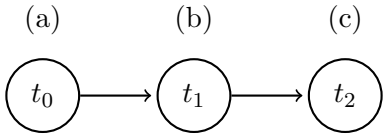

Figure 4: Explaining the probability of meaning formation (Goodfellow, 2016, pp. 565-568).

The arrow drawn from 'a' to 'b' means that we define the probability of meaning construction on 'b' by means of a conditional distribution, with 'a' being one of the conditioning variables that builds meaning. In other words, the distribution of meaning on 'b' depends on the meaning value of 'a'. This will certainly reduce the number of parameters used, diminishing the occurrence of ambiguity, determining the starting point of the meaning representation, and defining which variables are allowed as arguments.

In line with the suggestions in this article we can see Artificial Intelligence strategies (Kaliyar R; Goswami, 2020) for detecting fake news: A deep convolutional neural network called FNDNet was designed according to principles that we defend here by making use of the structure of the various hidden layers built in the deep neural network, achieving cutting-edge results with 98.36% accuracy in the tested data.

6 Conclusion

This article deals with meaning representation in natural language processing, NLP. Research on representation of meaning is important because machines still do not deal effectively with disambiguation as humans do. A universal structure of meaning representation (present in all languages) is proposed, based on the dynamic nature of language acting in the formation of meanings. The dynamic aspect of natural language is approached under the suggestion of a tool that joins semantic aspects to logical aspects, to consider the context through the verification of which superior hierarchical structure gives rise to inferior structures.

The authors seek, in Linguistic Theory, inspiration to improve the precision of meaning formation and to shed light on new models of machine learning. For this, the authors provide generic annotation guidelines that facilitate the construction of databases in which semantics is considered an essential part of the meaning construction. Frequency evidence for pattern formation is not sufficient, as ambiguity persists. Semantics are built into the database when using the neural network model, so that meaning, and context remain linked before moving on to the next layer of meaning. This three-dimensional neural model, as opposed to the logical one (two-dimensional), gives consistency to the representation of meaning.

It is intended that the treebank to be used in the process of interpreting sentences and phrases for the meaning construction is considered as close as possible to the natural language conceived as a dynamic system. In this way, the theory of abstract meaning representation would suit a necessary and simple strategy to perform interpretation in a more contextualized and unambiguous way. To exemplify this strategy, this article considers semantic treebanks based on the repetition of a minimal structure that contains the main element (head) modified by a specifier and a complement, as well as neural models. This model suggestion is more effective in analysis because of their context involvement in forming meaning, rather than analyzing language in separate tasks.

The theoretical contribution of linguistics makes it easier to understand how semantic meaning is constructed. While the abstract meaning representation theory (Banarescu et al., 2013; Damonte et al., 2016; Flanigan et al., 2014) proposes sets of concepts and relationships, we work with the minimum sentence structure and neural model that allow the analysis of word and sen-

tence relationship in a contextualized way. This makes the abstract representation of meaning more consistent, meeting the working structure of the linguistic process (Monte-Serrat, 2021; Monte-Serrat & Cattani, 2021; 2021a). The abstract meaning representation theory recognizes that sometimes it does not have an adequate relationship or standardizes the treatment for titles, appositives, and other constructions (Banarescu et al., 2013, p.182). This procedure sanitizes language, pushing it away from the natural language, and leads to building ambiguous meaning. The abstract meaning representation theory also lacks the representation of a semantic target for articles, tense, and number; nor does it distinguish between a hypothesis of real events. These are some topics that can be circumvented with the use of the minimum structure that we propose in this study, analyzing the construction of meaning in its own context. Based on these theoretical references, we can develop treebanks not only in Brazilian Portuguese, but also in other languages, as we focus on the structure of natural language understood as a dynamic system.

7 Acknowledgement

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the Sao Paulo Research Foundation (FAPESP grant 2019/07665-4) and by the IBM Corporation.

References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ...& Schneider, N. (2013, August). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 178–186).

Bender, E. M., Flickinger, D., Oepen, S., Packard, W., & Copestake, A. (2015, April). Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th international conference on Computational Semantics* (pp. 239-249).

Caramazza, A., Shapiro, K. (2004). Language categories in the brain: Evidence from aphasia. *Structures and Beyond: The cartography of syntactic structures*, 3, 15–38.

Chomsky, N. (2009). Syntactic structures. In: *Syntactic Structures*. De Gruyter Mouton.

Chomsky, N. (1961). On the notion" rule of grammar" (pp. 155-210). American Mathematical Society.

Chomsky, N. (2006). *Language and mind*. Cambridge University Press.

Copstead-Kirkhorn, L.E.C., Banasik, J.L. (2012). *Pathophysiology-E-Book*. Elsevier Health Sciences.

Damonte, M., Cohen, S.B., & Satta, G. (2016). An incremental parser for abstract meaning representation. arXiv preprint arXiv:1608.06111.

De Bot, K., Lowie, W., &Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and cognition, 10*(1), 7-21.
De Saussure, F. (1989). *Cours de Linguistique Générale* (Vol. 1). Otto Harrassowitz Verlag.

Derrida, J. (1967). La structure, le signe et le jeu dans le discours des sciences humaines. *L'ecriture et la difference*, 409-428.

Devlin, K.J. (2000). *The math gene: How mathematical thinking evolved and why numbers are like gossip* (Vol. 329). New York: Basic Books.

Edwards, C. (2021). The best of NLP. *Communications of the ACM, 64*(4), 9-11.

Fiorin, J.L., Savioli, F.P. (2006). Licoes de texto: leitura e redacao.

Flanigan, J., Thomson, S., Carbonell, J.G., Dyer, C., Smith, N.A. (2014, June). A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 1426-1436).

Friston, K.J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Human Brain Mapping, 2*(1-2), 56-78.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT press.

Halliday, M. A. K. (2061). Categories of the theory of grammar. In: *Readings in Modern Linguistics* (pp. 157–208). De Gruyter Mouton.

Halliday, M.A. (1985). Systemic background. *Systemic perspectives on discourse*, 1, 1–15.

Halliday, M.A.K. (1995). A recent view of 'missteps' in linguistic theory. *Functions of language, 2*(2), 249–267.

Halliday, M.A.K. (2006). *Linguistic studies of text and discourse* (Vol. 2). A & C Black.

Halliday, M.A.K., Webster, J.J. (2003). *On Language and Linguistics*. Volume 3. A & C Black.

Hamp Eric, P., Pavle, I., John, L. (2022). *Linguistics*. Encyclopedia Britannica.

Jespersen, O. (2013). *Language: Its nature, development, and origin*. Routledge.

Kaliyar, R.K., Goswami, A., Narang, P., Sinha, S. (2020). FNDNet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research,* 61, 32–44.

Lacan, J. (1966). Le Stade du miroir comme formateur de la fonction du Je: telle qu'elle nous est révélée dans l'expérience psychanalytique. Ècrits.

Leech, G. (1976). *Semantics*. Penguin Books Ltd.

Maia, B., Santos, D. (2018). Language, emotion, and the emotions: The multidisciplinary and linguistic background. *Language and Linguistics compass, 12*(6), e12280.

McCall, R. (2017). AI attempts to write Harry Potter and it goes hilariously wrong. *Open Library of Humanities, 6*(2), 1–23.

Monte-Serrat, D. (2017). Neurolinguistics, Language and Time: investigating the verbal art in its amplitude. International *Journal of Perceptions in Public Health, 1*(3), 162–171.

Monte-Serrat, D. (2021). Operating language value structures in the intelligent systems. *Advanced Mathematical Models & Applications, 6*(1), 31–44.

Monte-Serrat, D.M., Belgacem, F.B.M. (2017). Subject and time Movement in the Virtual reality. *International Journal of Research and Methodology in Social Science, 3*(3), 19–26.

Monte-Serrat, D.M., Cattani, C. (2021). *The Natural Language for Artificial Intelligence*. Academic Press.

Monte-Serrat, D.M., Cattani, C. (2021a). Interpretability in neural networks towards universal consistency. *International Journal of Cognitive Computing in Engineering*, 2, 30–39.

Pecheux, M. (1975). *Les Verites de La Palice: linguistique sémantique, philosophie*. FeniXX.

Pecheux, M. (2013). *Discourse: structure or event?. In Lacan, discourse, event: New psychoanalytic approaches to textual ˆ indeterminacy* (pp. 91–112). Routledge.

Pinker, S. (2003). *The Language Instinct: How the Mind Creates Language*. Penguin UK.

Pitt, D. (2022). Mental representation. In: *The Stanford Encyclopedia of Philosophy. Edward N. Zalta and Uri Nodelman* (editors). 2022.

Rosner, M., Johnson, R. (Eds.). (1992). *Computational Linguistics and Formal Semantics*. Cambridge University Press.

Sporns, O. (2013). Network attributes for segregation and integration in the human brain. *Current Opinion in Neurobiology*, *23*(2), 162-171.

Squire, L.R., Wixted, J.T. (2011). The cognitive neuroscience of human memory since HM. *Annual Review of Neuroscience*, 34, 259.

Terlow, E.M.C. (2020). The physiology of the brain and determining insensibility and unconsciousness. The slaughter of farmed animals: Practical ways of enhancing animal welfare, 202–228.

Tononi, G., Sporns, O., Edelman, G.M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences, 91*(11), 5033–5037.

Verspoor, M. (2013). Dynamic systems theory as a comprehensive theory of second language development. *Contemporary Approaches to Second Language Acquisition*, 9, 199.

Vinyals, O., Bengio, S., & Kudlur, M. (2015). Order matters: Sequence to sequence for sets. arXiv preprint arXiv:1511.06391.

Voloshinov, V.N., Bakhtin, M.M. (1986). *Marxism and the Philosophy of Language*. Harvard University Press.

Wallon, G.H. (1949). *Les notions morales chez l'enfant*. Presses universitaires de France.

Webster, J. (2002). *Linguistic studies of text and discourse*. Continuum.

Wolfram, S., Gad-el-Hak, M. (2003). A new kind of science. *Appl. Mech. Rev., 56*(2), B18-B19.

Zalta, E.N. (2017). *Metaphysics Research Lab*. Center for the Study of Language and Information, Stanford University, Stanford, CA.

Zhang, R., Santos, C.N.D., Yasunaga, M., Xiang, B., Radev, D. (2018). Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. arXiv preprint arXiv:1805.04893.