



HAL
open science

Measuring Linguistic Complexity: Introducing a New Categorial Metric

Mehdi Mirzapour, Jean-Philippe Prost, Christian Retoré

► **To cite this version:**

Mehdi Mirzapour, Jean-Philippe Prost, Christian Retoré. Measuring Linguistic Complexity: Introducing a New Categorial Metric. LACompLing 2018 - Symposium on Logic and Algorithms in Computational Linguistics, Aug 2018, Stockholm, Sweden. pp.95-123, 10.1007/978-3-030-30077-7_5. hal-02146506

HAL Id: hal-02146506

<https://hal.science/hal-02146506>

Submitted on 4 Jun 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Measuring Linguistic Complexity: Introducing a New Categorical Metric

Mehdi Mirzapour, Jean-Philippe Prost and Christian Retoré

Abstract This paper provides a computable quantitative measure which accounts for the difficulty in human processing of sentences: why is a sentence harder to parse than another one? Why is some reading of a sentence easier than another one? We take for granted psycholinguistic results on human processing complexity like the ones by Gibson. We define a new metric which uses Categorical Proof Nets to correctly model Gibson's account in his Dependency Locality Theory. The proposed metric correctly predicts some performance phenomena such as structures with embedded pronouns, garden paths, unacceptable center embeddings, preference for lower attachment and passive paraphrases acceptability. Our proposal gets closer to the modern computational psycholinguistic theories, while it opens the door to include semantic complexity, because of the straightforward syntax-semantics interface in categorial grammars.

1 Introduction

Linguistics – especially generative grammar à la Chomsky, makes a distinction between *competence* and *performance* in the human processing of natural language [6]. The competence is, roughly speaking, our ideal ability without time and resource constraints to parse a sentence, i.e. to decide whether it is grammatical or not. Competence is described by a grammar, a lexicon, assuming an unlimited working memory and time. Performance is how we actually parse a sentence; whether we

Mehdi Mirzapour
LIRMM, Univ Montpellier, CNRS, Montpellier, France. e-mail: mehdi.mirzapour@gmail.com

Jean-Philippe Prost
LPL, CNRS & Univ Aix-Marseille, Aix-en-Provence, France. LIRMM, Univ Montpellier, CNRS, Montpellier, France. e-mail: jean-philippe.prost@univ-amu.fr

Christian Retoré
LIRMM, Univ Montpellier, CNRS, Montpellier, France. e-mail: christian.retore@lirmm.fr

succeed in achieving that and how much the sentence resists to our attempt to analyze it, in a limited amount of time for the conversation to work, with our present knowledge, and our memory limits. Computing the space and time algorithmic complexity is a false solution because no one knows the algorithm being used by human beings, whether it depends on the individual or on the kind of conversation. Even if it were so, nothing guarantees that space and time algorithmic complexity matches the degree of difficulty we experience when processing sentences. And finally there are other factors of complexity that are hard to take into account, like the size of the grammar and short time memory limitation [23].

This paper, following an earlier body of work [16, 25], tries to provide a formal and computable account of the results of psycholinguistic theories with regard to linguistic complexity. We focus on syntactic complexity of structures with embedded pronouns [11], garden pathing [1], unacceptability of center embedded relatives [13], preference for lower attachment [17] and passive paraphrases acceptability [25]¹.

Regarding the psycholinguistic aspects, we mainly follow the studies by Gibson of linguistic complexity of human parsing. Gibson first studied the notion of linguistic difficulty [12] through the maximum number of incomplete syntactic dependencies that the processor has to keep track of during the course of processing a sentence. We refer to this theory as Incomplete Dependence Theory (IDT), as coined by Gibson. IDT had some limitations for referent-sensitive linguistic phenomena, which justified the later introduction of the Syntactic Prediction Locality Theory [10]. A variant of this theory, namely Dependency Locality Theory (DLT), was introduced later [11] to overcome the limitations with respect to the new linguistic performance phenomena. In the original works, both IDT and DLT use properties of linguistic representations provided by Government-Binding Theory [7].

On the formal side, in order to compute the complexity of a sentence — in a way that matches Gibson’s results — we use Lambek Categorical Grammar [19] by means of proof nets construction [24, Chap 6]. Proof nets were originally introduced by Girard [14] to describe proofs in linear logic, up to some equivalences. Categorical proof nets are introduced by Roorda and Retoré [27, 26] to describe proofs in the Lambek calculus and related formalisms. They are to categorial grammar what parse trees are to phrase structure grammar. This kind of approach was initiated by Johnson [16], who defines a measure of the instantaneous complexity when moving from a word to the next one (in particular for center embedded relative clauses) in a way that matches Gibson’s and Thomas’ analysis [13]. To define the complexity of a sentence, Johnson considers the maximum complexity between the words in a given sentence. This approach was refined by Morrill [25], who re-interprets axiom links in categorial proof nets as incomplete (or unresolved) dependencies. We rename this technique as *IDT-based complexity profiling* since it clearly inherits many aspects of Gibson’s IDT, plus the new notion of profiling that exists in some psycholinguistic theories. This technique is quite successful at predicting linguistic performance phenomena such as garden paths, unacceptability of center embedding and preference for lower attachment. Nevertheless, there is some predictive limitation for referent-

¹ These linguistic phenomena are described in detail in sections (4) and (5).

sensitive phenomena such as structures with embedded pronouns. Our strategy to overcome this issue is to apply DLT instead of IDT on proof net constructions. This strategy leads us to introduce *DLT-based complexity profiling*. We will show how this reformulation can improve the predictive power of the existing models in favor of the referent-sensitive linguistic phenomena.

The purpose of developing our computational psycholinguistic model is not solely limited to measuring linguistic complexity. It is potentially applicable to some specific tasks in the domain of formal compositional semantics. For instance, ranking different possible readings of a given ambiguous utterance, or more generally translating natural language sentences into weighted logical formulas. The rest of the paper is organized as follows: Section 2 summarizes Gibson's ideas on modeling the linguistic complexity of human sentence comprehension, namely IDT and DLT. In section 3 we then define proof nets, and recall the success and limitation of IDT-based complexity profiling. In section 4 we define our DLT-inspired measure; we show how it fixes some problems in previous work and how it gives a correct account of those phenomena. In section 5 we provide more linguistic evidence to support our claim in favor of DLT-based complexity profiling over IDT-based approach. In section 6 we discuss limitations of our approach and possible future works for solving them. In the last section we conclude our paper and discuss possible future works.

2 Gibson's Theories on Linguistic Complexity

We provide a very quick review of Gibson's IDT and DLT in order to make the readers familiar with their underlying concepts. The question of how to automatically compute linguistic complexity based on both theories with categorial proof nets will be covered in the sections (3.2) and (4).

Incomplete dependency theory relies on counting incomplete dependencies during the incremental processing of a sentence, whenever a new input word must be connected to the current and pending linguistic structure. The main parameter in IDT is the number of incomplete dependencies when the new word integrates the existing structure. It explains the increasing complexity of examples (1a)-(1c) with nested relative clauses. In (1a), *the reporter* has 1 incomplete dependency; in (1b), *the senator* has 3 incomplete dependencies; in (1c) *John* has 5 incomplete dependencies at the point of processing. For room sake we only explain the most complex case, i.e. (1c), in which the incomplete dependencies at the moment of processing *John* are: (i) the NP *the reporter* depends on a verb that should follow it; (ii) the NP *the senator* depends on a different verb to follow; and (iii) the pronoun *who* (before *the senator*) depends on a verb to follow; (iv) the NP *John* depends on another verb to follow; and (v) the pronoun *who* (before *John*) depends on a verb to follow. These are five unsaturated or incomplete or unresolved dependencies. IDT in its original form suggests to calculate the maximum number of incomplete dependencies in a sentence. One can observe that the complexity is proportional to the number of incomplete dependencies.

Example 1.**1a.** The reporter disliked the editor.**1b.** The reporter [who the senator attacked] disliked the editor.**1c.** The reporter [who the senator [who John met] attacked] disliked the editor].**1d.** The reporter [who the senator [who I met] attacked] disliked the editor].

Dependency Locality Theory is a distance-based referent-sensitive linguistic complexity measure put forward by Gibson to supersede the predictive limitations of the incomplete dependency theory. DLT posits two different costs for integration and storage. In this paper, we have only focused on the integration cost. The linguistic complexity is interpreted as the locality-based cost of the integration of a new word to the dependent word in the current linguistic structure which is relied on the number of the intervened new discourse-referents. Through the count of these referents, we can predict the relative complexity, such as structures with embedded pronouns, illustrated in example (1d). The experiments [28] support the acceptability of (1d) over (1c). According to the discourse-based DLT structural integration cost hypothesis, since referents for the first-person pronoun *I* is already present in the current discourse, integrating across them consumes fewer cognitive resources than integrating across the new discourse referents before *John*. By means of just two aspects of DLT, namely the structural integration and the discourse processing cost, we can predict a number of linguistic phenomena, as we will see in details with some examples.

3 Complexity Profiling in Categorical Grammars

3.1 Proof-nets as parse structures

Our exposition of the henceforth classical material on proof nets for categorial grammar follows [24] — the main original papers on this topic are [19, 14, 26, 27]. Categorial grammars are defined from a set C of grammatical categories, defined from base categories (for instance $\mathcal{B} = \{np, n, S\}$) including a special symbol S (for sentence) and operators, for instance :

$$C ::= \mathcal{B} \mid C \backslash C \mid C / C$$

The symbols \backslash and $/$ can be viewed as logical connectives, namely implication(s) of a logic, namely intuitionistic non-commutative multiplicative linear logic better known as the Lambek calculus. Such formulas can be viewed as formulas of linear logic, with conjunction \otimes disjunction \wp and negation $(_)^\perp$ because implications can be defined from negation and disjunction:

$$\begin{aligned} \text{Definition of } \backslash \text{ and } /: \quad A \backslash B &\equiv A^\perp \wp B & B / A &\equiv B \wp A^\perp \\ \text{De Morgan equivalences } (A^\perp)^\perp &\equiv A & (A \wp B)^\perp &\equiv B^\perp \otimes A^\perp \\ (A \otimes B)^\perp &\equiv B^\perp \wp A^\perp \end{aligned}$$

Some formulas have a polarity. Formulas are said to be positive (output) \circ or negative (input) \bullet as follows:²

	$a : \circ,$	
\otimes	\bullet	\circ
\bullet	<i>undefined</i>	\bullet
\circ	\bullet	<i>undefined</i>

	$a^\perp : \bullet$	
\wp	\bullet	\circ
\bullet	<i>undefined</i>	\circ
\circ	\circ	<i>undefined</i>

So $a \wp b$ has no polarity; $a^\perp \wp b$ is positive which is equal to $a \setminus b$; while $b^\perp \otimes a$ is negative and it is equal to the negation of $a \setminus b$. Categories are, roughly speaking, analogous to non-terminal categories in phrase structure grammars. But observe that they are endowed with an internal structure, i.e. $(np \setminus S) / np$ is a compound category and the rules make use of this internal structure, connectors \setminus and $/$ and subcategories n , np and S . The rules (of the logic) do not depend on the language generated (or analyzed) by the grammar. They are the same for every language, and the lexicon makes the difference. The lexicon maps every word to a finite set of categories. A parse structure in a categorial grammar defined by a lexicon \mathcal{L} for a sequence of words w_1, \dots, w_n simply is a proof of $c_1, \dots, c_n \vdash S$ with $c_i \in \mathcal{L}(w_i)$ in some variant of the Lambek calculus. The rules for the basic (associative) Lambek calculus are:

$$\frac{}{A \vdash A}$$

$$\frac{H_1, H_2, \dots, H_{n-1}, H_n \vdash C}{H_1, H_2, \dots, H_{n-1} \vdash C / H_n}$$

$$\frac{H_1, H_2, \dots, H_{n-1}, H_n \vdash C}{H_2, \dots, H_{n-1}, H_n \vdash H_1 \setminus C}$$

$$\frac{H_1, \dots, H_i, \dots, H_n \vdash C \quad G_1, \dots, G_n \vdash A}{H_1, \dots, H_{i-1}, G_1, \dots, G_n, A \setminus H_i, H_{i+1}, \dots, H_n \vdash C}$$

$$\frac{H_1, \dots, H_{i-1}, H_i, H_{i+1}, \dots, H_n \vdash C \quad G_1, \dots, G_n \vdash A}{H_1, \dots, H_{i-1}, H_i / A, G_1, \dots, G_n, H_{i+1}, \dots, H_n \vdash C}$$

Since the Lambek sequent calculus enjoys the cut-elimination property, whereby a sequent is provable if and only if it is provable without the cut rule, we do not mention the cut rule. Categorical grammars are known for providing a transparent and computable interface between syntax and semantics. The reason is that the categorial parse structure is a *proof* in some variant of the Lambek calculus, and that this *proof* gives a way to combine semantic lambda terms from the lexicon into a lambda term which encodes a formula expressing the meaning of the sentence.

² Here we are stricter than in other articles, i.e. we neither allow \otimes of positive formulas nor \wp of negative formulas, because we only use the \setminus and $/$ symbols in categories (and not \otimes): only combining heterogeneous polarities guarantees that a positive formula is a category, and that a negative formula is the negation of a category.

We cannot provide more details herein, but the reader is referred to [24, Chapter 3]. For instance, the categorial analysis of *Every barber shaves himself* with the proper semantic lambda terms for each word in the sentence yields the logical form $\forall x. \text{barber}(x) \Rightarrow \text{shave}(x, x)$.

It has been known for many years that categorial parse structures, i.e. proof in some substructural logic, are better described as proof nets [27, 26, 22, 24]. Since categorial grammars follow the parsing-as-deduction paradigm, an analysis of a c phrase w_1, \dots, w_n is a proof of c under the hypotheses c_1, \dots, c_n , where c_i is a possible category for the word w_i ; and proofs in those systems are better denoted by graphs called proof nets. The reason is that different proofs in the Lambek calculus may represent the same syntactic structure (constituents and dependencies), but these essentially similar sequent calculus proofs correspond to a unique proof net. A proof net is a graph, whose nodes are formulas, and it consists of two parts:

subformula trees of the conclusions, in the right order, whose leaves are the base categories and branching are two connectives \wp and \otimes — as we saw earlier formulas with \backslash and $/$ can be expressed from base categories and their negations with \wp and \otimes — for nodes that are not leaves the label can be limited to the main connective of the subformula instead of the whole formula, without loss of information;

axioms that are a set of edges, pairwise disjoint, and connecting a leaf z to a leaf z^\perp in such a way that every leaf is incident to some axiom link.

However not all such graphs are proof nets, only those satisfying:³

Acyclicity every cycle contains the two edges of the same \wp branching.

Connectedness There is a path not involving the two edges of the same \wp branching between any two vertices.

Intuitionism Every conclusion can be assigned some polarity.

Non commutativity The axioms do not cross (are well bracketed).

The advantage of proof-nets over sequent calculus is that they avoid the phenomenon known as spurious ambiguities— that is, when different parse structures correspond to the same syntactic structure (same constituent and dependencies). Indeed proofs (i.e. parse structures) with unessential differences are mapped to the same proof net. A (normal) deduction of $c_1, \dots, c_n \vdash c$ (i.e. a syntactic analysis of a sequence of words as a constituent of category (c)) maps to a (normal) proof net with conclusions $(c_n)^\perp, \dots, (c_1)^\perp, c$ [27, 24]. Conversely, every normal proof net corresponds to at least one normal sequent calculus proof [26, 24].

³ This list is redundant: for instance intuitionism plus acyclicity implies connectedness.

3.2 Incomplete Dependency-Based Complexity Profiling and its Limitation

In this subsection we recall the IDT-based measure of linguistic complexity by Morrill [25], which itself improves over a first attempt by Johnson [16]. Both measures are based on the categorial proof nets. The general idea is simple: to re-interpret the axiom links as dependencies and to calculate the incomplete dependencies during the incremental processing by counting the incomplete axiom links for each word in a given sentence. This is almost the same as Gibson’s idea in his IDT, except the fact that he uses some principles of Chomsky Government-Binding theory [7] instead of the categorial proof nets. The notion of counting incomplete dependencies for each node, called complexity profiling, is more effective in terms of prediction than approaches that only measure the maximum number of incomplete dependencies or the maximum cuts [16]. We can rewrite IDT-based complexity profiling [25] with the following definitions:

Definition 1 Let π be a syntactic analysis of w_1, \dots, w_n with categories C_1, \dots, C_n — that is a categorial proof net with conclusions $(C_n)^\perp, \dots, (C_1)^\perp, S$. Let C_{i_0} be one of the C_i ($i \in [1, n]$). The incomplete dependency number of C_{i_0} in π , written as $ID_\pi(C_{i_0})$, is the count of axioms $c - c'$ in π such that $c \in (C_{i_0-m} \cup S)$ ($m \geq 0$) and $c' \in C_{i_0+n+1}$ ($n \geq 0$).

Definition 2 Let π be a syntactic analysis of w_1, \dots, w_n with categories C_1, \dots, C_n — that is a categorial proof net with conclusions $(C_n)^\perp, \dots, (C_1)^\perp, S$. We define the IDT-based linguistic complexity of π , written $f_{idt}(\pi)$ by $(1 + \sum_{i=1}^n ID_\pi(C_i))^{-1}$.

Definition 3 Given two syntactic analyses π_i and π_j , not necessarily of the same words and categories, we say that π_i is IDT-preferred to π_j whenever $f_{idt}(\pi_i) > f_{idt}(\pi_j)$.

Example 2.

2a. The reporter who sent the photographer to the editor hoped for a good story.

2b. The reporter who the photographer sent to the editor hoped for a good story.

Figures (1) and (2) show the two relevant proof nets for examples (2a) with subject-extracted relative clause and (2b) with object-extracted relative clause (examples from [11]). The relevant complexity profiles for (2a) and (2b) are illustrated in the figure (3). As shown in the figures, the total sum of the complexity for (2b) is greater than (2a). Therefore it can predict correctly that (2a) is preferred over (2b). That preference is confirmed by reading time experiments [9].⁴

Obviously, IDT-based account does not use DLT as its underlying theory. Not surprisingly, the linguistic phenomena that can only be supported by DLT would not be supported by IDT-based complexity profiling. Figure (4) shows this failure. We can verify this by applying the definitions on the relevant proof nets as it is illustrated

⁴ The same procedure can show the increasing complexity of the examples (1a)-(1c) by drawing the relevant proof-nets. This practice is avoided in this paper due the space limitation and its simplicity comparing to the running examples.

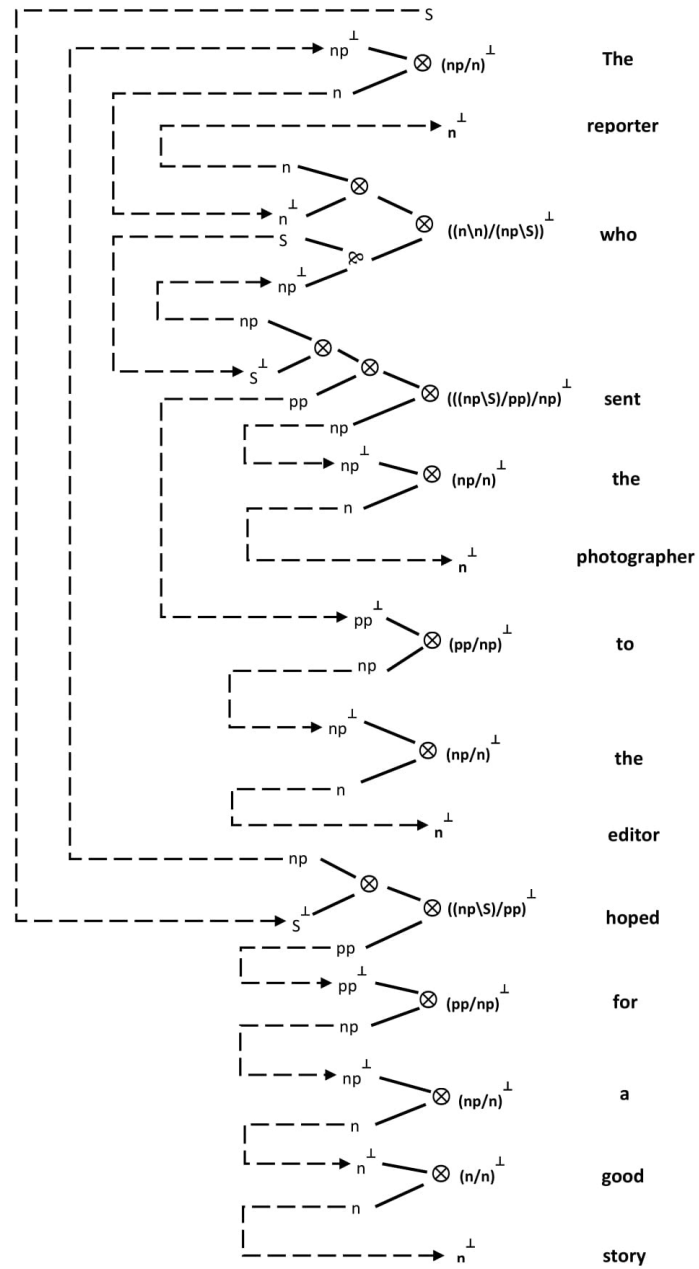


Fig. 1: Proof net analyses for (2a) subject-extracted relative clause.

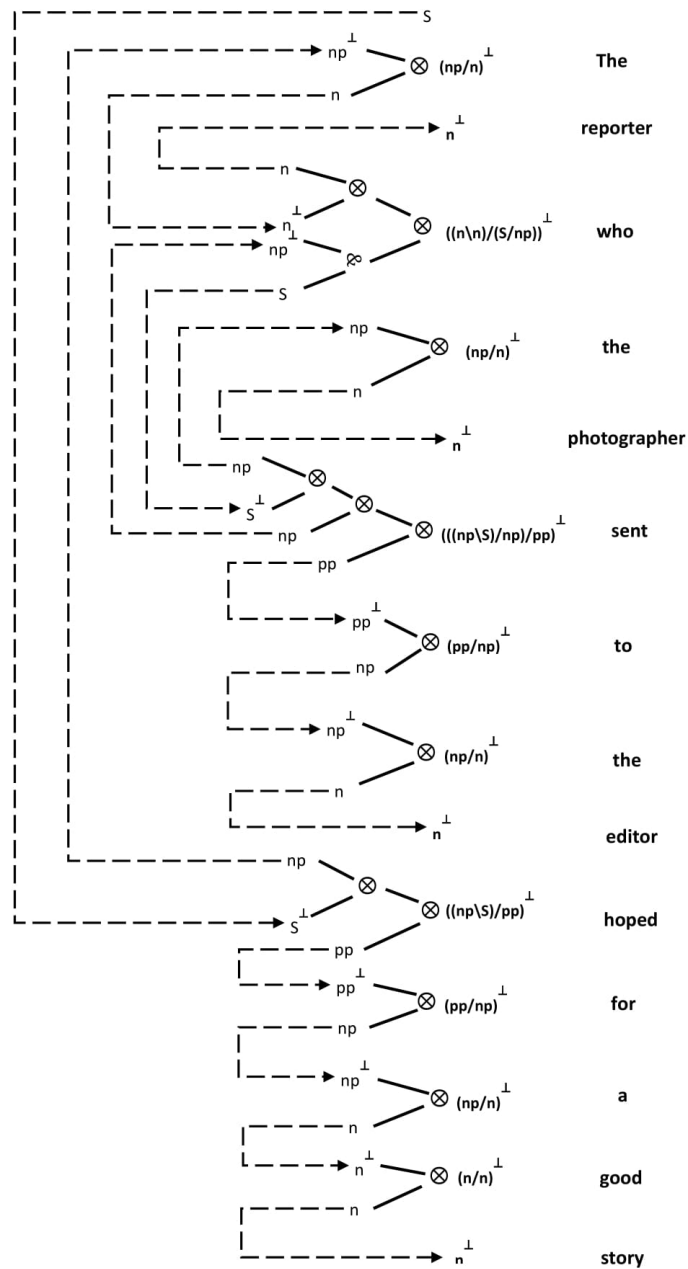


Fig. 2: Proof net analyses for (2b) object-extracted relative clause.

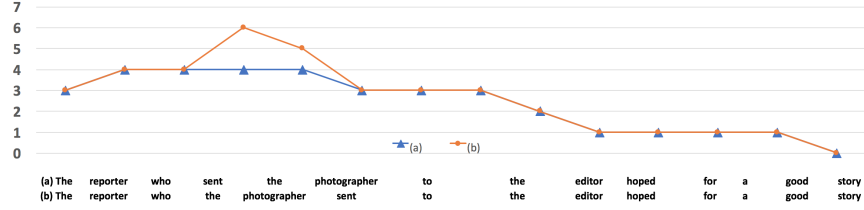


Fig. 3: IDT-based Complexity Profiles for (2a) and (2b).

in the figures (5) and (6). As one may notice, the corresponding proof nets for the examples (1c) and (1d) are almost the same. Consequently, IDT-based complexity profiling cannot discriminate both examples, i.e. it generates the same number for both sentences in contrast to the experiments [28] as it is shown in the figure (4). This shows the importance of introducing DLT-based complexity profiling for proof nets in order to make more predictive coverage—as we will do so.

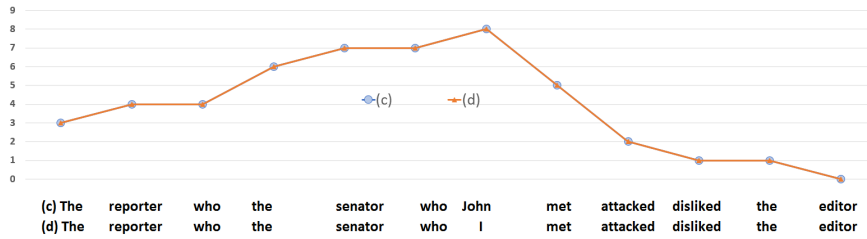


Fig. 4: IDT-based Complexity Profiles for (1c) and (1d).

4 A New Proposal: Distance Locality-Based Complexity Profiling

As we discussed, IDT-based complexity profiling is a distance-based measure. However, it does not have a referent-sensitive criterion and due to this fact, it cannot support some of the linguistic phenomena such as structures with embedded pronouns. One plausible strategy to overcome this lack is introducing DLT-based complexity profiling. This will allow us to have a referent-sensitive measure. In this section, we provide the precise definitions of our DLT-based proposal on the basis of the categorial proof nets. Here they are:

Definition 4 A word w is said to be a discourse referent whenever it is a *proper noun*, *common noun* or *verb*.

Definition 5 Let π be a syntactic analysis of w_1, \dots, w_n with categories C_1, \dots, C_n — that is a categorial proof net with conclusions $(C_n)^\perp, \dots, (C_1)^\perp, S$. Let $c - c'$ be an axiom in π such that $c \in C_i$ and $c' \in C_j$ ($i, j \in [1, n]$). We define the **length** of axiom $c - c'$ as the integer $i + 1 - j$.

Definition 6 Let π be a syntactic analysis of w_1, \dots, w_n with categories C_1, \dots, C_n — that is, a categorial proof net with conclusions $(C_n)^\perp, \dots, (C_1)^\perp, S$. Let C_{i_0} be one of the C_i , and let consider axioms $c - c'$ with c in C_{i_0} and c' in some C_{i_0-k} . Let us consider the largest k for which such an axiom exists — this is the longest axiom starting from C_{i_0} with the previous definition. The dependency locality number of C_{i_0} in π , written $DL_\pi(C_{i_0})$ is the number of discourse referent words between $w_{i_0} : C_{i_0}$ and $w_{i_0-k} : C_{i_0-k}$. The boundary words, i.e. $w_{i_0} : C_{i_0}$ and $w_{i_0-k} : C_{i_0-k}$ should also be counted. Alternatively, it may be viewed as $k + 1$ minus the number of non-discourse references among those $k + 1$ words.

Definition 7 Let π be a syntactic analysis of w_1, \dots, w_n with categories C_1, \dots, C_n — that is, a categorial proof net with conclusions $(C_n)^\perp, \dots, (C_1)^\perp, S$. We define the DLT-based linguistic complexity of π , written $f_{dl}(\pi)$ as $(1 + \sum_{i=1}^n DL_\pi(C_i))^{-1}$.

Definition 8 Given two syntactic analyses π_i and π_j , not necessarily of the same words and categories, we say that π_i is DLT-preferred to π_j whenever $f_{dl}(\pi_i) > f_{dl}(\pi_j)$.

Examples: We apply our new metric on examples (1c) and (1d). Figures (5) and (6) show the relevant proof nets for (1c) and (1d). The proof nets for both examples are the same except a difference in one of the lexicons in each example, i.e. *John* and *I*.⁵ Figure (7) shows the accumulative chart-based representation of our measure for each example. The axis Y shows the accumulative sum of the dependency locality function applied to each category on axis X. The quick analysis of the profiles shows the total complexity numbers 14 and 11 for (1c) and (1d) respectively. This correctly predicts the preference of example (1d) over (1c), which was not possible in the IDT-based approaches.

The measure for dependency locality number is quite straightforward. As an example, we calculate the dependency locality number for the word *attacked* in figure (6) for (1d). We can find the longest axiom link starting from *attacked* and ended to its uppermost category, namely *who*. This axiom link is highlighted in figure (6) for the reader's convenience. Now, we can count the number of discourse referents intervened in the axiom link, which is actually three: *attacked*, *met* and *senator*.

5 More Linguistic Support

We can evaluate our proposal for measuring the linguistic complexity against other linguistic phenomena. We basically illustrate the examples based on a survey provided in [25]. As we will see, the new metric supports both referent-sensitive and

⁵ Following Lambek [19], we have assigned the category $S/(np \setminus S)$ to relative pronoun *I*. Note that even assigning np , which is not a type-shifted category, would not change our numeric analysis at all.

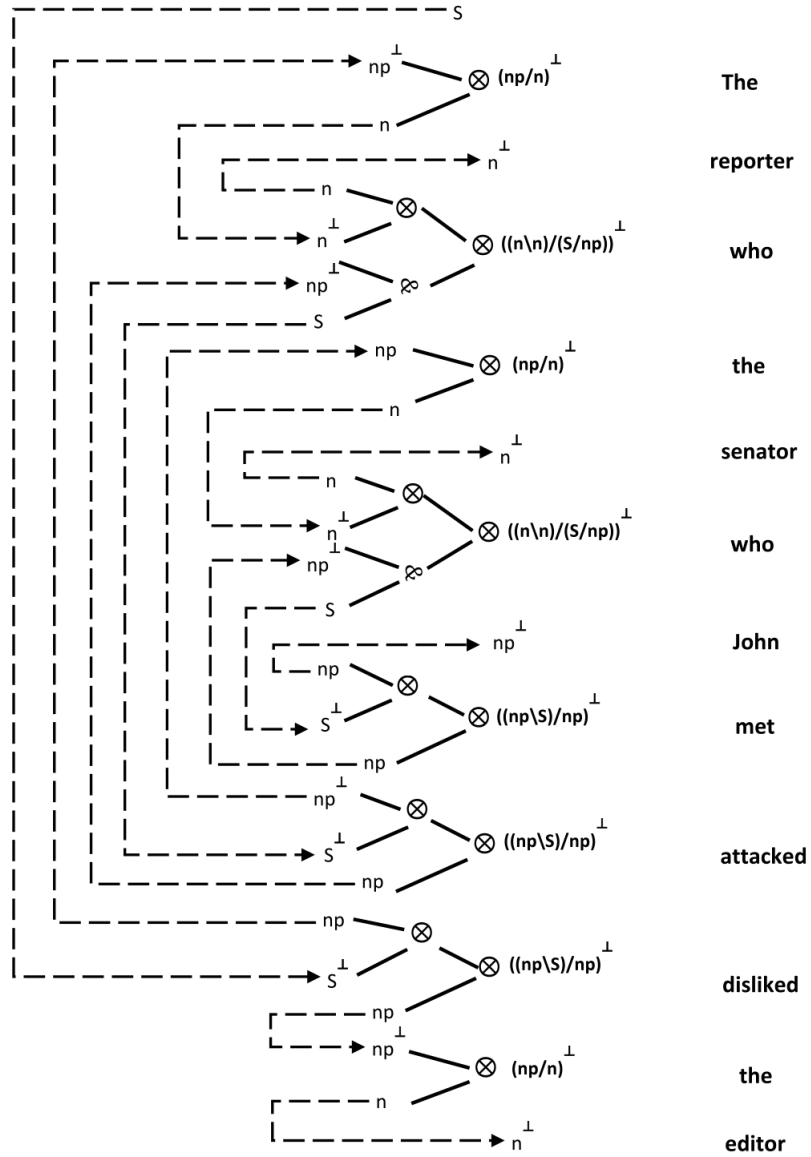


Fig. 5: Proof net analysis for example (1c).

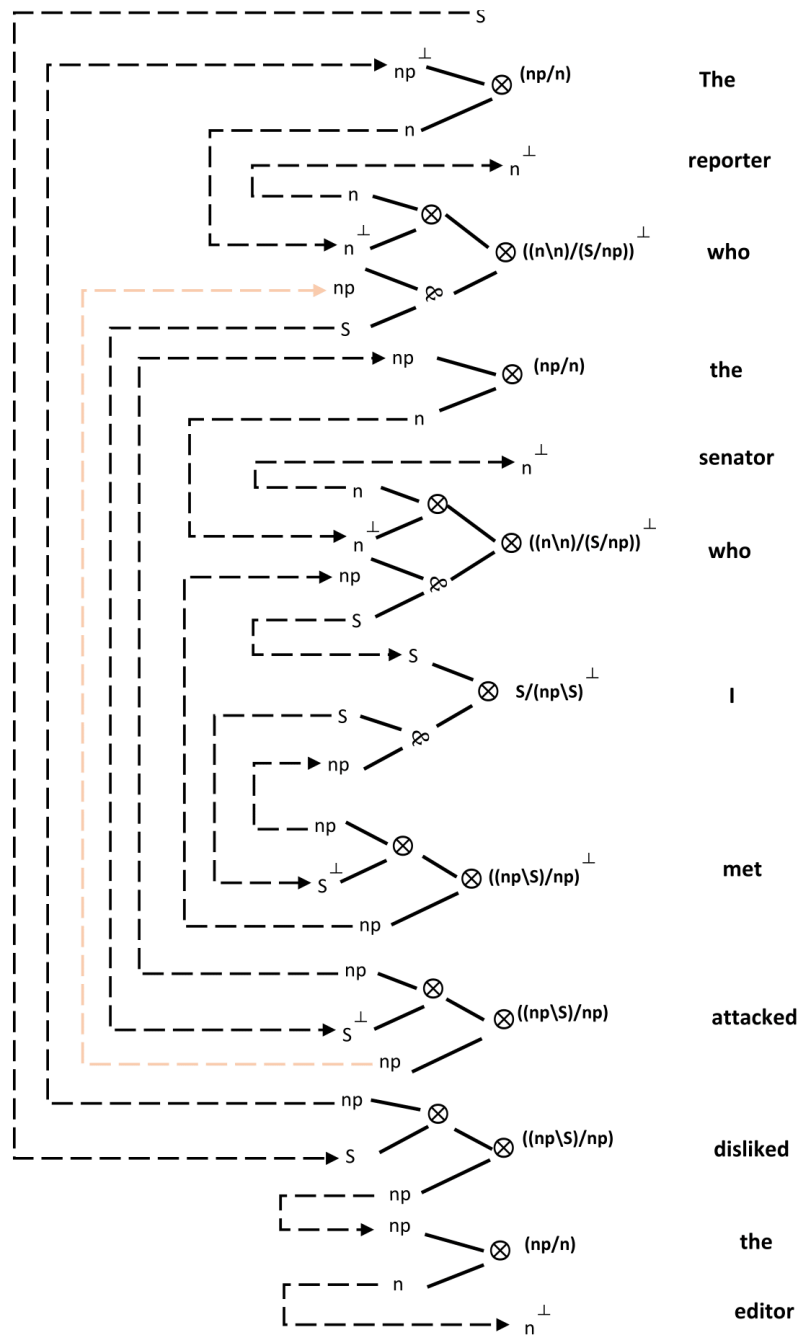


Fig. 6: Proof net analysis for example (1d).

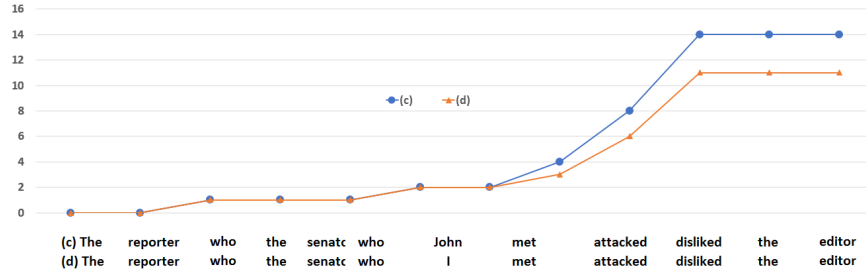


Fig. 7: DLT-based Complexity Profiles for (1c) and (1d).

some of the non-referent-sensitive phenomena⁶. We start with the Garden pathing [1], which can be illustrated by the following examples:

Example 3.

3a. The horse raced past the barn.

3b. ?The horse raced past the barn fell.

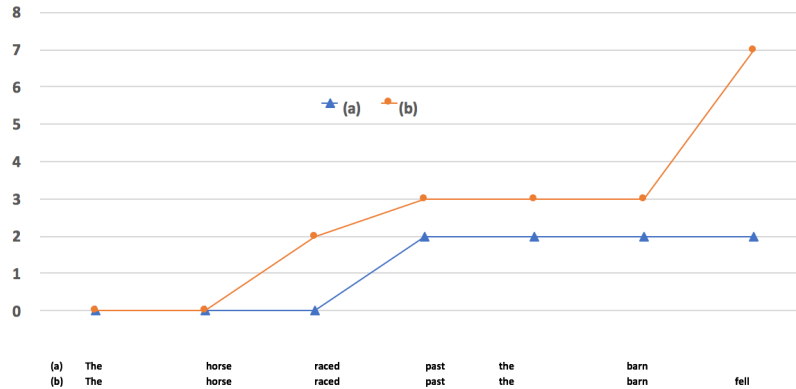


Fig. 8: Accumulative DLT-based Complexity Profiles for (3a) and (3b).

Although (3b) is grammatical, it is perceived ungrammatical due to strong tendency to interpret the initial segments as in (3a). The sentence (3b) leads the reader toward a seemingly familiar meaning, namely (3a), that is actually not the one intended. This often requires re-reading so that the meaning may be fully grasped after careful parsing. The difficulty in correctly parsing the sentence (3b) results from the fact that *race* can be interpreted transitively or intransitively. The figure (9) illustrates the proof net analyses for (3a) and (3b). Figure (8) shows the complexity profiles in which the total complexity numbers are 2 and 7 for (3a) and (3b), respectively. This correctly predicts the high level of complexity in (3b) and the preferred reading of (3a) compared to (3b) as it happens in the real sentence comprehension.

⁶ It is worth mentioning that DLT-based complexity profiling can not support two linguistic phenomena: Multiple Sentences and Heavy Noun-Phrase Shift. For more details on the problem related to the Multiple-Quantifier Sentences and a possible solution consider [21, Chap 5,7].

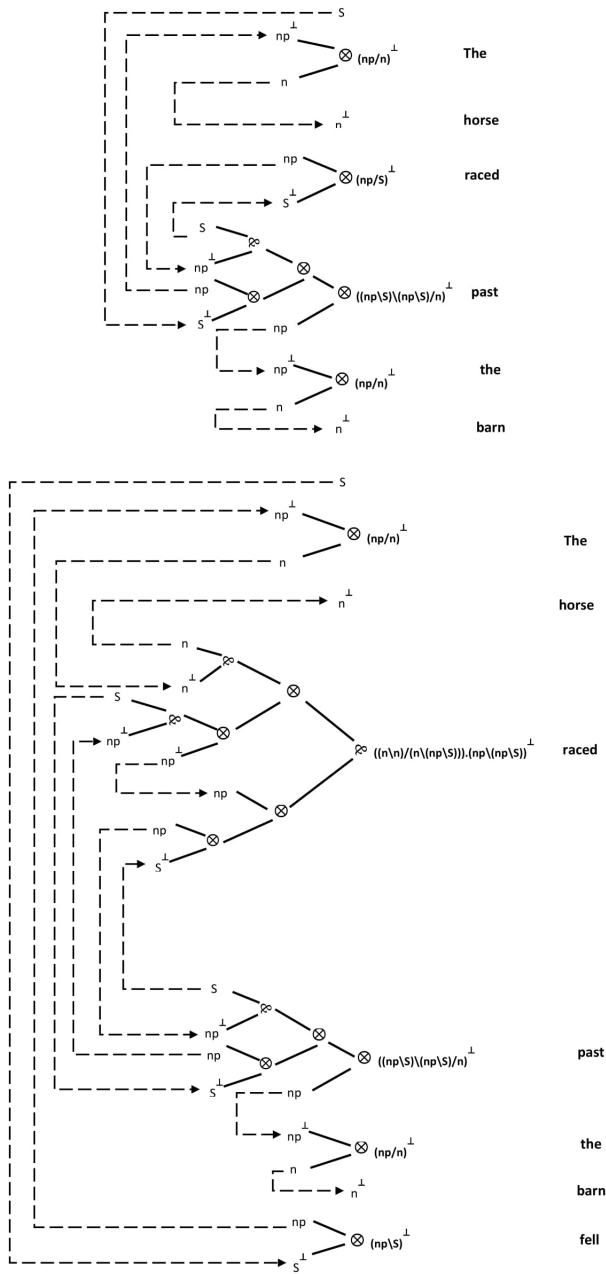


Fig. 9: Proof net analyses for (3a) located in top (first attempt reading) and (3b) in bottom (full garden path sentence).

Example 4.

4a. ?The cheese that the rat that the cat saw ate stank.

4b. The dog that chased the cat that saw the rat barked.

The unacceptability of center embedding phenomena is illustrated by the examples (4a) and (4b). The example (4a) shows object relativization, which exhibits deterioration in acceptability, while the sentence (4b) exhibits little variation in acceptability since it carries subject relativization [8, Chap. 1]. This linguistic phenomenon can be captured in our model. The figures (12) and (13) illustrate the proof net analyses for (4a) and (4b) respectively. Figure (10) shows the complexity profiles in which the total complexity numbers are 14 and 10 for (4a) and (4b) respectively.

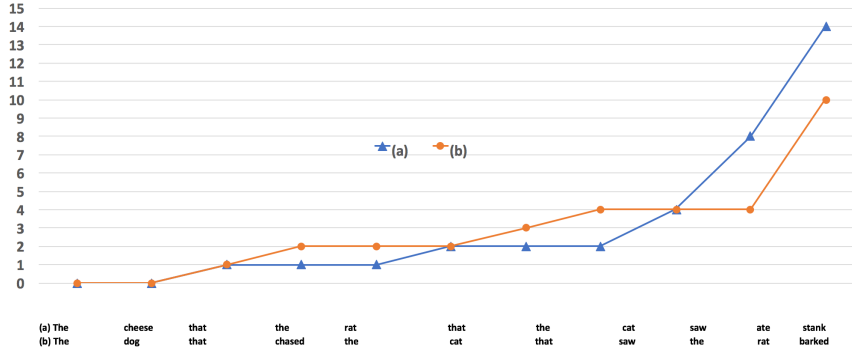


Fig. 10: Accumulative DLT-based Complexity Profiles for 4a and 4b

Example 5. Joe said that Marthus believes that Ingrid fell today.

Example 6. The book that shocked Mary’s title.

Examples (5) and (6) carry different readings due to the syntactical ambiguity that exists in each of them. As for Example (5), the preference goes to the lowest

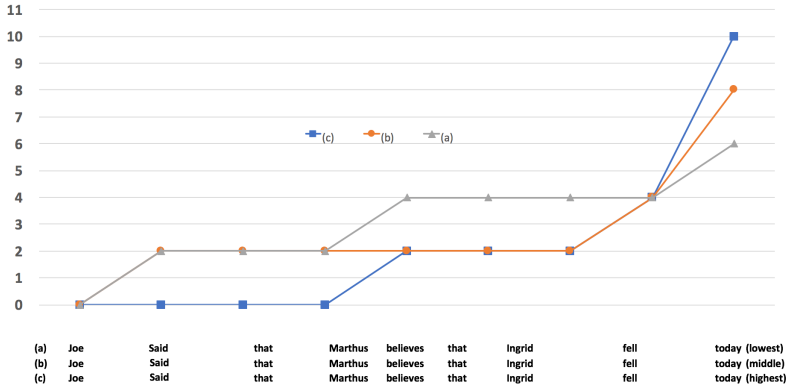


Fig. 11: Accumulative DLT-based Complexity Profiles for Three Readings of 5.

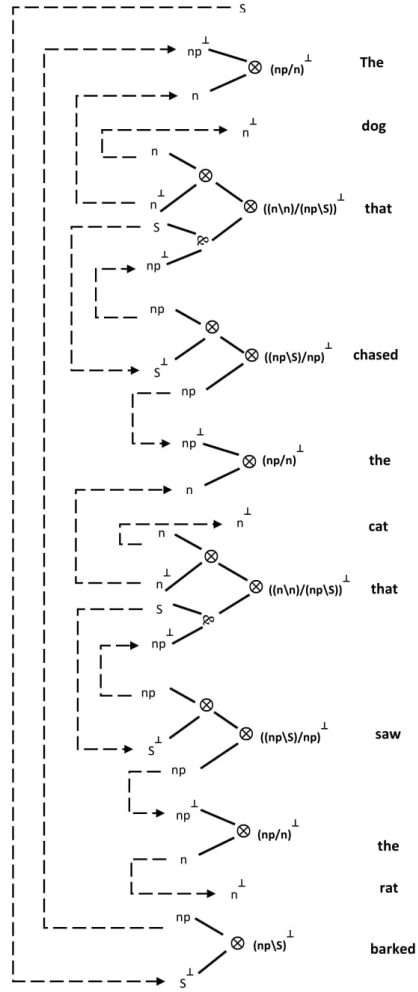


Fig. 13: Proof net analyses for 4b (subject relativization).

attachment [17]. The three profiles in figures (14), (15) and (16) show the highest, middle and lowest attachment of the adverb *today* to the verbs *said*, *believes* and *fell*, respectively. As it is illustrated in Figure (11), our model can predict the preference of (5a) over (5b), and (5b) over (5c). In other words, the lower the attachment of the adverb, the higher the preference [17].

As for Example (6), two proof nets in Figure (17) show the lowest and highest attachment of the phrase *'s title* to the sentence *the book that shocked Mary*. As illustrated in Figure (18), our model can predict the preference of the right interpretation in Figure (17) over the left one, despite the fact that the right one is non-sensical [25].

Example 7.

7a. Ingrid was astonished that Jack was surprised that two plus two equals four.

7b. ?That that two plus two equals four surprised Jack astonished Ingrid.

The last linguistic performance phenomena that we want to discuss is Passive Paraphrases Acceptability [25], exemplified in (7a) and (7b). Notice that the DLT-based complexity profile of (7a) is lower, even though the number of sentences and axiom links are more than (7b). The real preference is on the syntactic forms, namely, the one which is less complex. In our case the example (7a) is preferred to (7b) since it is simpler. The relevant proof nets and the accumulated complexity profiles are illustrated in the figures (19)-(20) and (21), respectively.

6 Limitation

As discussed in this paper, our DLT-based complexity profiling is quite successful at predicting the referent-sensitive phenomenon, i.e., *structures with embedded pronouns*, that cannot be treated by the IDT-based approach. Moreover, it can also predict almost all the linguistic phenomena that the IDT-based approach can predict. To be more specific, the covered phenomena are: *garden pathing*, *unacceptability of center embedding*, *preference for lower attachment*, and *passive paraphrases acceptability*. The only exception— in terms of prediction— is *Heavy Noun Phrase Shift*, a phenomenon that our proposed approach cannot predict. This limitation is reported in [21, Chapter 7] and consequently, some treatments are considered in the aggregation phase when different metrics are collected, compared and preferred. Nevertheless, we admit this problem as one of the limitations of the DLT-based complexity approach, which can not supersede the IDT-based complexity profiling on this point.

The second limitation is the problem of ranking valid semantic meanings of a given multiple-quantifier sentence, which cannot be supported by our proposal. A study [21, Chapter 3] shows that the same problem exist in the IDT-based approach when dealing with some type of the expressions such as sentence-modifier adverbials and nested sentences. Our experiment shows that the same limitations hold in our DLT-based approach. Thus, both IDT-based and DLT-based complexity profiling cannot correctly predict ranking the quantifier scoping problem. Hopefully, this can

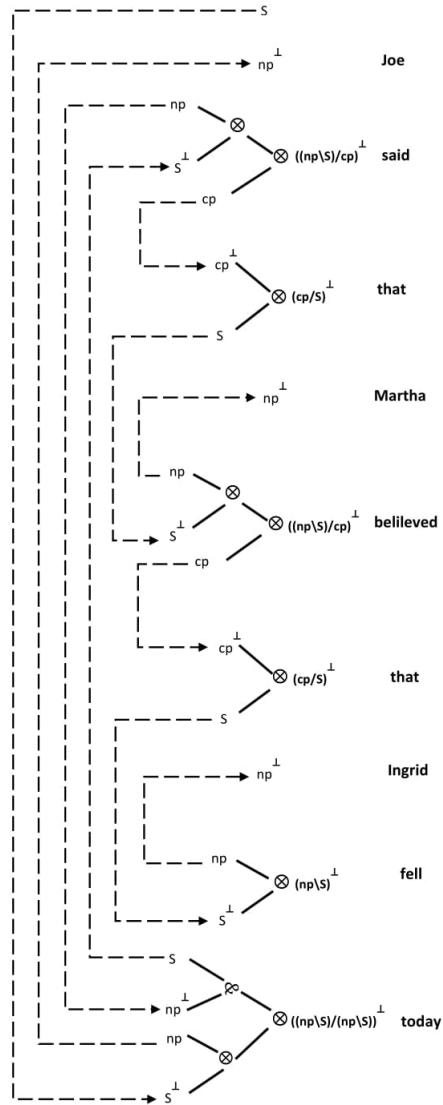


Fig. 14: Proof net analyses for 5 with highest adverbial attachments.

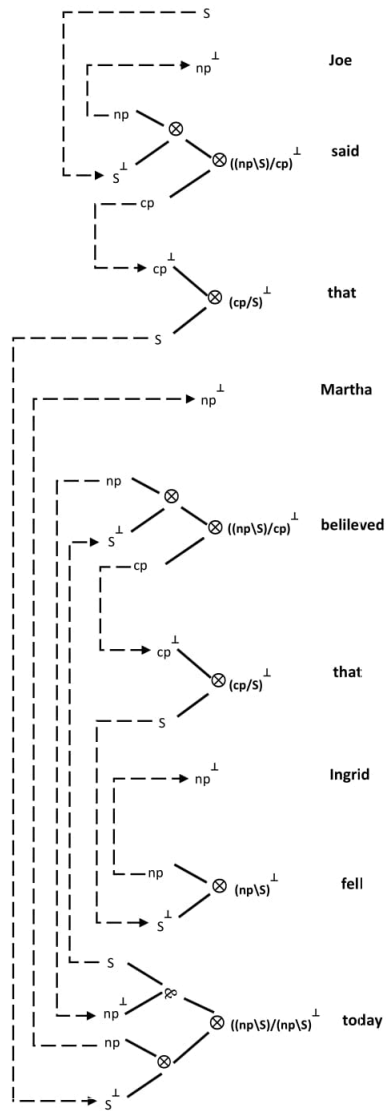


Fig. 15: Proof net analyses for 5 with middle adverbial attachments.

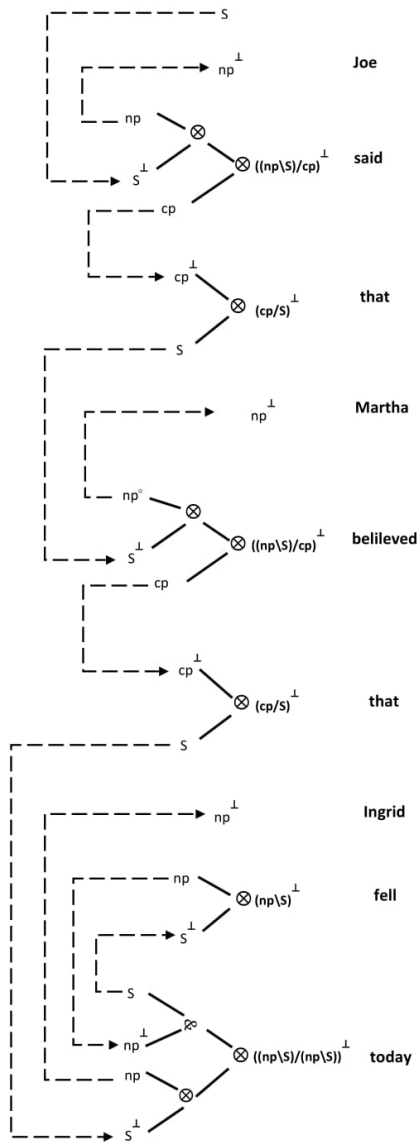


Fig. 16: Proof net analyses for 5 with lowest adverbial attachments.

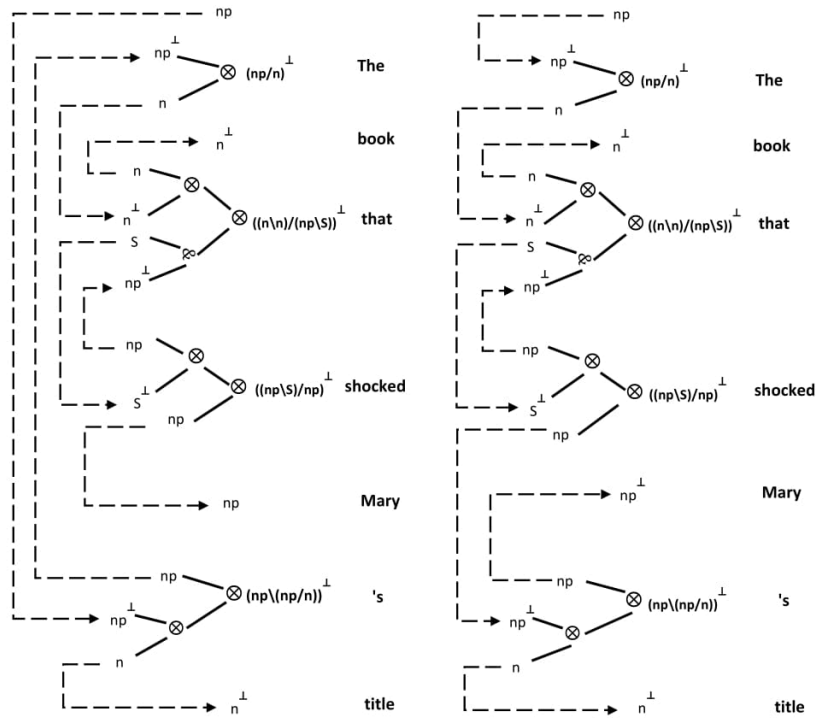


Fig. 17: Proof net analyses for 6 with sensical (left) and nonsensical (right) interpretations.

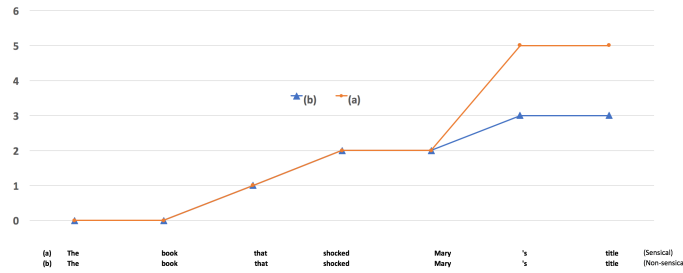


Fig. 18: Accumulative DLT-based complexity profiles for two readings of 6

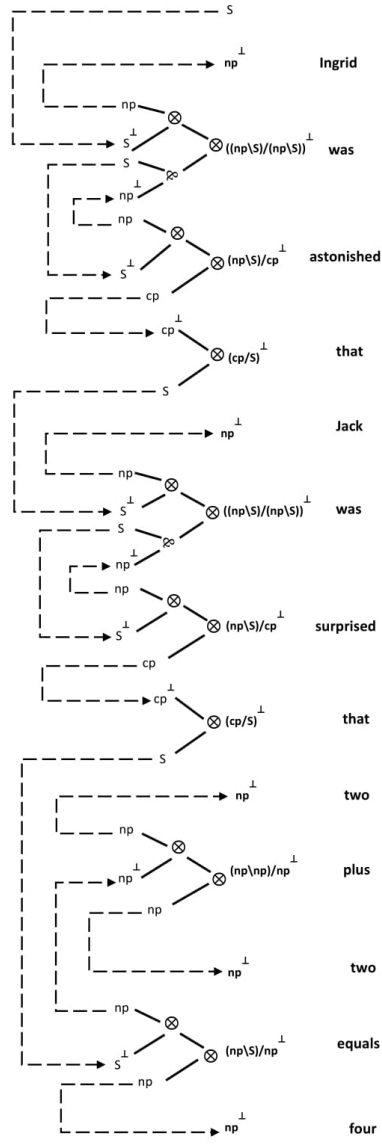


Fig. 19: Proof net analyses for 7a.

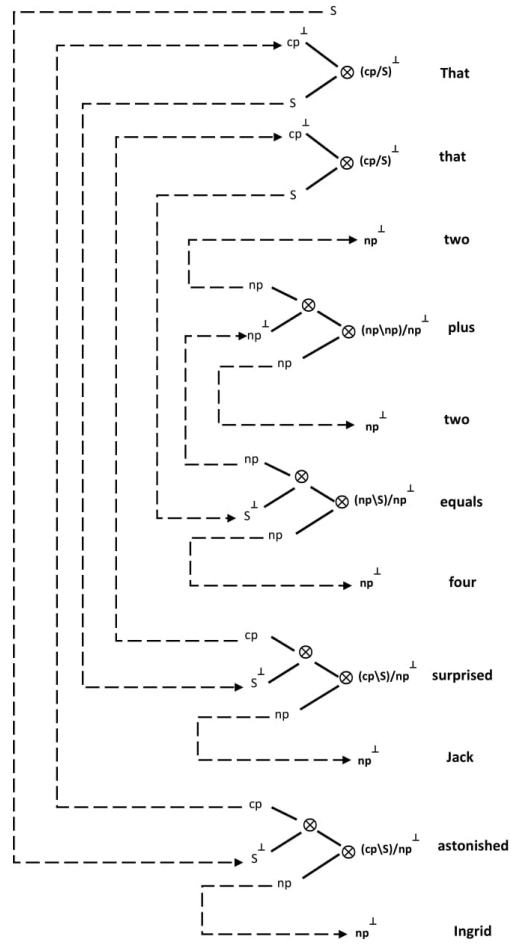


Fig. 20: Proof net analyses for 7b.

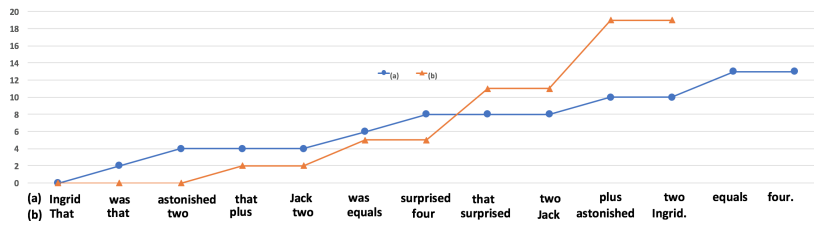


Fig. 21: Accumulative DLT-based complexity profiles for 7a and 7b

be treated with the hybrid models (discussed in [21, Chapter 3]) in which Hilbert's epsilon and tau [15, 5] are exploited in order to neutralize the quantifier effect, and after that, the complexity is measured on the basis of the penalty of the quantifiers re-ordering.

The third drawback of the DLT-based motivated approaches is that they are not applicable cross-linguistically for human parsing processes. One study [35] has shown the failure of DLT-based approaches in modeling some human performances in the Hindi language. DLT uses the distance hypothesis: the increasing distance between dependents and heads results in increased processing difficulty. This hypothesis is not a universal linguistic feature as it is claimed [35]. This cross-linguistic problem in DLT-based approaches has motivated the activation-based models [33], which are formalized in computational form in the Adaptive Control of Thought-Rational (ACT-R) architecture [29]. This theory construes parsing as an efficient series of the guided memory retrievals, as opposed to a proof net construction in a way we have practiced in this chapter. Nevertheless, activation-based approaches cannot properly treat the structures with embedded pronouns [33]. Moreover, it has no systematic approach for automatically deriving meaning representations in logical formulas as one is generally capable to do in the Categorical Grammar frameworks and specifically in our proposed DLT-based complexity profiling.

7 Conclusion and Possible Extensions

In this paper we explored how our DLT-based complexity profiling on proof nets can give a proper account of the complexity of a wide range of linguistic phenomena. We have also shown that the IDT-based method could not support phenomena of referent-sensitive linguistic performance. It was one of the main reasons for introducing the DLT-based complexity profiling technique within the framework of Lambek calculus. There are some extensions for our study and research:

- As we mentioned it is possible to bridge our model with other study [4] to overcome the problem of ranking quantifier scoping. As we discussed, both IDT-based and DLT-based complexity profiling fail to predict quantifier scoping. This bridging strategy can help us to solve this issue by exploiting Hilbert's epsilon and tau operators [15, 5] for neutralizing the quantifier effect and making possible the complexity measure by a penalty cost of the quantifiers re-ordering.
- Another important direction is to take into account not only the axioms of the proof-nets but also the logical structure, i.e., par-links, tensor-links and the correctness criterion. This is important indeed, because this structure is needed to compute the logical form (semantics) from the syntactic structure given by proof nets. For instance, nesting Lambek slashes (that are linear implications, and therefore par-links in the proof net) corresponds to higher order semantic constructions (e.g. predicates of predicates) and consequently this nesting of par-links increases the complexity of the syntactic and semantic human processing.

- Our method can be combined with studies in other directions, e.g. the task of sentence correction/completion in Lambek Calculus [20, 21]. The other task is measuring semantic gradience in natural language. Some line of research suggests this feature within lexical/compositional frameworks by creating and enrichment of the wide-coverage weighted lexical resources from crowd-sourced data [18].
- We have solely relied on Gibson's theories in this paper. However, we can enrich our grounds with some of the existing tracks. Our work can be extended by adapting the best-fitting model introduced in [30], which has three constructs of total cognitive effort, cognitive strain, and time. Also, our future model can potentially use the outcomes of an experiment [31] that shows how different linguistic structures may cause measurable changes in cognitive workload. There are a couple of works in linguistic complexity surveyed in [34] and [32], which can be potentially insightful for our future study.

Acknowledgement We would like to show our gratitude to Philippe Blache for his insightful discussion at our lab and also for inspiration that we got from his papers [2, 3]. We would like to thank our colleague Richard Moot as well for his numerous valuable comments on this work.

References

1. Bever, T.G.: The cognitive basis for linguistic structures. *Cognition and the development of language* (1970)
2. Blache, P.: A computational model for linguistic complexity. In: *Proceedings of the first International Conference on Linguistics, Biology and Computer Science* (2011)
3. Blache, P.: Evaluating language complexity in context: New parameters for a constraint-based model. In: *CSLP-11, Workshop on Constraint Solving and Language Processing* (2011)
4. Catta, D., Mirzapour, M.: Quantifier scoping and semantic preferences. In: *Proceedings of the Computing Natural Language Inference Workshop* (2017)
5. Chatzikyriakidis, S., Pasquali, F., Retoré, C.: From logical and linguistic generics to hilbert's tau and epsilon quantifiers. *IfCoLog Journal of Logics and their Applications* **4**(2), 231–255 (2017)
6. Chomsky, N.: *Aspects of the Theory of Syntax*. MIT Press, Cambridge (1965)
7. Chomsky, N.: Some concepts and consequences of the theory of government and binding, vol. 6. MIT press (1982)
8. Chomsky, N.: *Aspects of the Theory of Syntax*, vol. 11. MIT press (2014)
9. Gibson, E., Ko, K.: An integration-based theory of computational resources in sentence comprehension. In: *Fourth Architectures and Mechanisms in Language Processing Conference*, University of Freiburg, Germany (1998)
10. Gibson, E.: Linguistic complexity: Locality of syntactic dependencies. *Cognition* **68**(1), 1–76 (1998)
11. Gibson, E.: The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain* pp. 95–126 (2000)
12. Gibson, E.A.F.: A computational theory of human linguistic processing: Memory limitations and processing breakdown. Ph.D. thesis, Carnegie Mellon University Pittsburgh, PA (1991)
13. Gibson, E., Thomas, J.: The processing complexity of english center-embedded and self-embedded structures. In: of Massachusetts, U. (ed.) *The proceedings of the North-Eastern Linguistic Society* 1996 (1996)
14. Girard, J.Y.: Linear logic. *Theoretical Computer Science* **50**, 1–102 (1987)

15. Hilbert, D.: Die logischen Grundlagen der Mathematik. *Mathematische Annalen* **88**(1), 151–165 (1922)
16. Johnson, M.E.: Proof nets and the complexity of processing center-embedded constructions. In: Retoré, C. (ed.) Special Issue on Recent Advances in Logical and Algebraic Approaches to Grammar. *Journal of Logic Language and Information*, vol. 7(4), pp. 433–447. Kluwer (1998)
17. Kimball, J.: Seven principles of surface structure parsing in natural language. *Cognition* **2**(1), 15–47 (1973)
18. Lafourcade, M., Mery, B., Mirzapour, M., Moot, R., Retoré, C.: Collecting weighted coercions from crowd-sourced lexical data for compositional semantic analysis. In: *New Frontiers in Artificial Intelligence, JSAI-ISAI 2017. Lecture Notes in Computer Science*. vol. 10838, pp. 214–230. Springer (2018)
19. Lambek, J.: The mathematics of sentence structure. *The American Mathematical Monthly* **65**(3), 154–170 (1958)
20. Mirzapour, M.: Finding missing categories in incomplete utterances. In: *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*. p. 149
21. Mirzapour, M.: Modeling Preferences for Ambiguous Utterance Interpretations. (Modélisation de préférences pour l'interprétation d'énoncés ambigus). Ph.D. thesis, University of Montpellier, France (2018), <https://tel.archives-ouvertes.fr/tel-01908642>
22. Moot, R.: Proof nets for linguistic analysis. Ph.D. thesis, UIL-OTS, Universiteit Utrecht (2002)
23. Moot, R., Retoré, C.: Natural language semantic and computability. *Journal of Logic Language and Information* (to appear — preliminary version: <https://arxiv.org/abs/1605.04122>)
24. Moot, R., Retoré, C.: The logic of categorial grammars: a deductive account of natural language syntax and semantics, LNCS, vol. 6850. Springer (2012), <http://www.springer.com/computer/theoretical+computer+science/book/978-3-642-31554-1>
25. Morrill, G.: Incremental processing and acceptability. *Computational Linguistics* **26**(3), 319–338 (2000)
26. Retoré, C.: Calcul de Lambek et logique linéaire. *Traitement Automatique des Langues* **37**(2), 39–70 (1996)
27. Roorda, D.: Proof nets for Lambek calculus. *Logic and Computation* **2**(2), 211–233 (1992)
28. Warren, T., Gibson, E.: The effects of discourse status on intuitive complexity: Implications for quantifying distance in a locality-based theory of linguistic complexity. In: Poster presented at the Twelfth CUNY Sentence Processing Conference, New York (1999)
29. John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological review*, 111(4):1036, 2004.
30. Elizabeth Cooper-Martin. Measures of cognitive effort. *Marketing Letters*, 5(1):43–56, 1994.
31. Nikolaos Engonopoulos, Asad Sayeed, and Vera Demberg. Language and cognitive load in a dual task environment. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013.
32. Claire Kramsch. Why is everyone so excited about complexity theory in applied linguistics. *Mélanges Crapel*, 33:10–24, 2012.
33. Richard L Lewis and Shrawan Vasishth. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29(3):375–419, 2005.
34. Charles A. Perfetti. Reviewed work: Linguistic complexity and text comprehension: Readability issues reconsidered by Alice Davison, Georgia M. Green. *Language*, 65(3):643–646, Sep., 1989.
35. Shrawan Vasishth et al. Quantifying processing difficulty in human language processing. In *Rama Kant Agnihotri and Tista Bagchi*, 2005.