Check for updates

METHOD ARTICLE

Developing neutral linguistic resources for the implementation of an automatic transformational analyzer

[version 1; peer review: 2 approved with reservations]

Max Silberztein[1], Cristina Mota[2], Anabela Barreiro[3]

[1]Universite de Franche-Comte SLHS, Besançon, Bourgogne-Franche-Comté, France
[2]Institute of Systems and Computer Engineering Research and Development in Lisbon, Lisbon, Portugal
[3]OpenLogos and Institute of Systems and Computer Engineering Research and Development in Lisbon, Lisbon, Portugal

v1 **First published:** 12 Jul 2024, **4**:144
https://doi.org/10.12688/openreseurope.17990.1
Latest published: 12 Jul 2024, **4**:144
https://doi.org/10.12688/openreseurope.17990.1

Abstract

Bakground

The linguistic pursuit of describing natural languages stands as a commendable scientific endeavor, regardless of immediate software application prospects. It transcends mere documentation of possible sentences to establish connections between sentences derived from transformations.

Methods

Amid the dominance of Large Language Models (LLMs) in research and technology, which offer intriguing advancements in text generation, the approaches presented in this article confront challenges like opacity, limited human intervention, and adaptation difficulties inherent in LLMs. The alternative or complementary approaches highlighted here focus on the theoretical and methodological challenges of describing linguistic transformations and are firmly rooted in the field of linguistics, the science of language. We propose two solutions to address the problem of language transformations: (i) the procedural approach, which involves representing each transformation with a transducer, and (ii) the declarative method, which entails capturing all potential transformations in a single neutral grammar.

Results

These approaches simplify the generation of complex sentences from elementary ones and vice versa.

Open Peer Review

Approval Status ? ?

	1	2
version 1 12 Jul 2024	? view	? view

1. **Shaina Raza**, University of Toronto, Toronto, Canada

2. **Kuldeep Vayadande**, Vishwakarma Institute of Technology, Pune, India

Any reports and responses or comments on the article can be found at the end of the article.

Conclusion

This work has benefited from research exchanges within the Multi3Generation COST Action (CA18231), and the resources produced can contribute to enhancing any language generation system.

Plain Language Summary
The Multi3Generation COST Action (CA18231) aims to propel Natural Language Generation research for societal advancement through collaborative and inclusive strategies, considering diverse methodologies and resources. Amid the dominance of Large Language Models (LLMs) in research and technological progress, which indeed yield intriguing development in text generation, we tackle challenges such as opacity, limited human intervention, and adaptation difficulties inherent in these LLMs. This article underscores alternative or complementary approaches that prioritize neutral linguistic resources, providing transparency, adaptability, and human manageability in language generation (humans in the loop). The alternative or complementary approaches outlined in this article direct their focus toward addressing the theoretical and methodological challenges associated with describing linguistic transformations. We propose two distinct solutions to tackle these challenges effectively: (i) firstly, the procedural approach, which entails representing each transformation using a transducer; and (ii) secondly, the declarative method, which involves capturing all potential transformations within a single neutral grammar. These approaches serve to streamline the generation of complex sentences from elementary ones and vice versa. This work has been enriched by collaborative research exchanges within the Multi3Generation COST Action, highlighting the importance of interdisciplinary cooperation in advancing our understanding and application of linguistic principles across various languages beyond English.

Keywords
Natural Language Processing, Automatic Text Generation, Automatic Transformational Analysis, Natural Language Generation, NooJ, eSPERTo, Paraphrase Generation, Multi3Generation COST Action

This article is included in the COST Actions gateway.

This article is included in the Advances in Natural Language Generation collection.

Corresponding author: Max Silberztein (max.silberztein@gmail.com)

Author roles: Silberztein M: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Mota C**: Methodology, Project Administration, Resources, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Barreiro A**: Funding Acquisition, Investigation, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Framework Programme for Research & Innovation as part of the COST Action [CA18231, Multi3Generation: Multitask, Multilingual, Multimodal Language Generation], as supported by the COST Association (European Cooperation in Science and Technology).

Copyright: © 2024 Silberztein M *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Silberztein M, Mota C and Barreiro A. **Developing neutral linguistic resources for the implementation of an automatic transformational analyzer [version 1; peer review: 2 approved with reservations]** Open Research Europe 2024, **4**:144 https://doi.org/10.12688/openreseurope.17990.1

First published: 12 Jul 2024, **4**:144 https://doi.org/10.12688/openreseurope.17990.1

1 Introduction

The Multi3Generation COST Action (CA18231) aims to enhance the importance of Natural Language Generation (NLG) research for societal benefit through collaborative efforts. It advocates inclusivity by embracing valuable resources and methodologies, irrespective of their origin, whether from Large Language Models (LLMs) or linguistically-oriented models. While LLMs have propelled text generation forward, they come with drawbacks that pose significant challenges when used without adequate knowledge and training and need scientific attention. The initial challenge lies in understanding the models due to their opacity, complicating the comprehension of generative processes. Human intervention is limited, resulting in bots prone to hallucination and misinformation. Adapting to various styles and domains also presents difficulties. Additionally, there is concern over the irresponsible use of undisclosed resources, amplifying worries about misuse, especially in domains reliant on trustworthy data processes. Potential risks may exert a more pronounced influence within the critical domains of education, health, law, and industry. In this article, we exemplify controllable linguistic-based approaches (to NLG and Natural Language Processing (NLP) in general) that prioritize the development of resources covering various language features, and characterized by a clearly defined design, distinguished traits, including transparency, adaptability, traceability, and human manageability. An important goal for linguists is to describe natural languages in a manner that enables automatic systems to differentiate between correct sentences, such as The lady ate these cookies from incorrect sequences of words, such as The ate lady cookie these. For a software application to function effectively, it requires access to several key components:

- the complete set of linguistic units present in the text, known as the language vocabulary: *the, lady*, *to eat*, *this*, *cookie*.
- the operations that establish connections between the word forms found in the text and the vocabulary, known as language morphology (*to eat* ⇄ *ate*, *this* ⇄ *these*, *cookie* ⇄ *cookies*).
- the structure of sentences in the language, referred to as language syntax, enables the system to analyze sentence structures, such as (The lady)Subject (ate)Verb (these cookies)Object.
- various distributional constraints among sentence elements, encompassing the semantics of the language, allowing the system to verify compatibility between sentence elements; for example, *to eat* as a verb requires a human subject and a edible (concrete) object.

However, knowing what constitutes a correct sentence is far from fully describing a language: one also needs to know how to express a negation (The lady did not eat these cookies), how to focus on an argument (It is these cookies that the lady ate), how to pronominalize an argument (She ate them), among others. The important question is: what transformations are applicable to any given sentence? In the realm of NLP application development, the objective is to create software capable of automatically generating any transformed sentence from a given elementary sentence, or conversely, breaking down any complex sentence by producing the elementary sentences containing its predicates, thereby analyzing its meaning.

The rest of the article is organized in the following way: Section 2 presents the procedural approach, which involves describing each transformation with a transducer. Section 3 outlines the declarative approach, where finite state automata describe transformations of the same predicate. Section 4 showcases eSPERTo, a paraphrasing system based on the procedural approach. Finally, Section 5 provides the main insights about both methods and discusses future directions.

Methods

2. The procedural approach

This application can be implemented using various computational devices, including Finite-State Transducers, Recursive Transducers, and Linear-Bounded Automata, such as the one represented by the graph in Figure 1.

This graph, which has the power of a Linear-Bounded Automaton, can be automatically applied to any text to perform a three-step computation:

1. it first matches sequences constituted by a pronoun *He* or *She*, or a singular determiner (<DET+s>) followed by a singular noun (<N+s>), followed by a verb conjugated in the preterit (<V+PT>), followed by a pronoun *them* or a plural determiner (<DET+p>) followed by a plural noun (<N+p>).
2. when a sequence has matched, its subject subsequence is stored as variable $Subject; its verb as the variable $Verb and its object subsequence <DET+p> <N+p> as the variable $Object.
3. the automaton then produces the output [$Subject *did not* $Verb_ $Object], in which it copies the value of the variable $Subject, followed by the sequence *did not*, followed by the lemma of the value of the variable $Verb, followed by the value of the variable $Object.

For instance, if the text contains the sentence *The lady ate these cookies*, the corresponding output sequence in the negative form will be *The lady did not eat these cookies*. If the text contains the sentence *She ate them*, the corresponding output sequence will be *She did not eat them*. The graph in Figure 1 is overly simplistic. A more complete grammar can be seen in Figure 2 (for negation) and in Figure 3 (for auxiliary verbs).

The graph NP contains a description of various noun phrases: nouns preceded by modifiers, such as *the large red pencil*, genitive constructions, such as *my cousin's friends*, and nouns with prepositional complements, such as *the demonstration for*

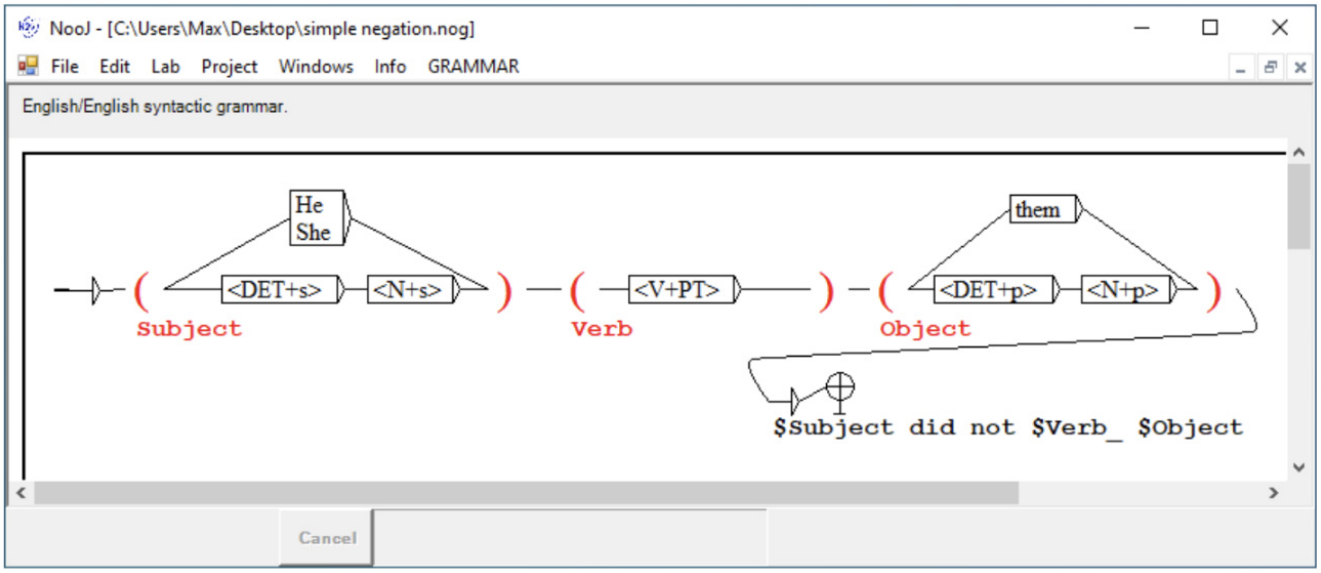

Figure 1. **Simple transformation.**

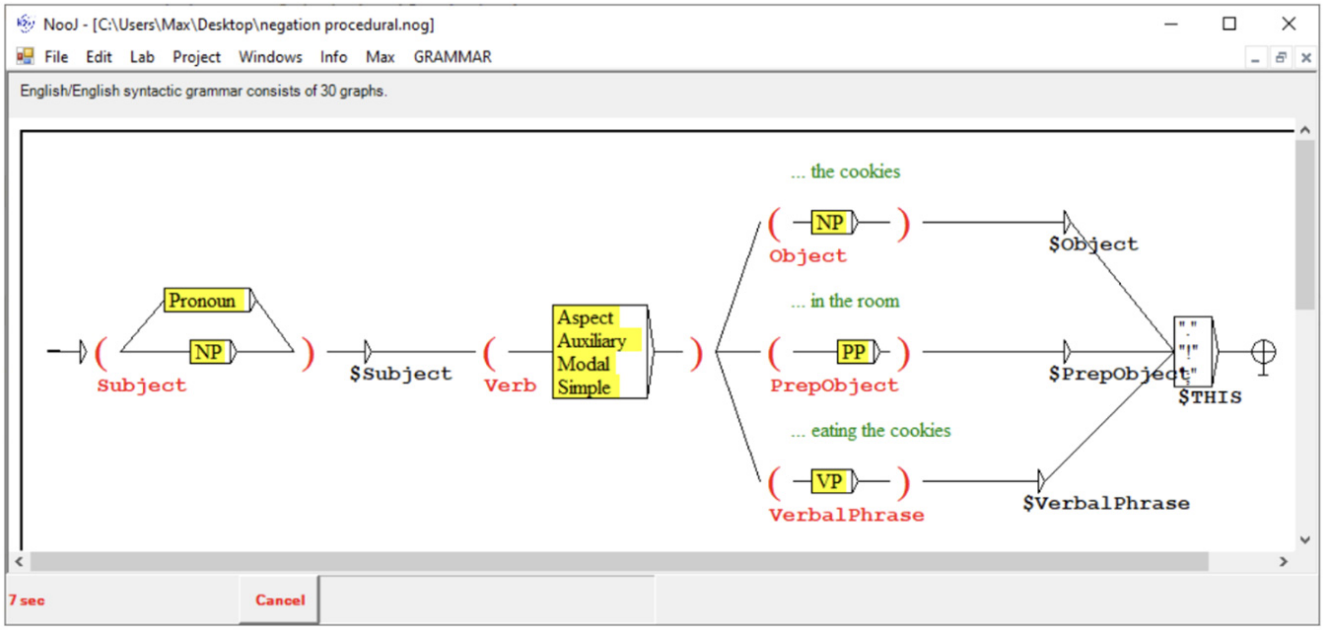

Figure 2. **Negation Transformation, main graph.**

the classroom. The noun phrase stored in $Subject is simply copied in the output of the graph, i.e., resulting negative sentence. Graph PP contains a description of prepositional phrases such as *for my friends*; graph VP contains a description of verbal phrases such as *to eat the cookies* and *eating the cookies*. The graph Aspect recognizes aspectual constructions such as *continues to eat* and *stopped eating*. The resulting negative structure produced by the transducer is *does not continue to eat* and *did not stop eating* respectively. The graph Auxiliary in Figure 3 recognizes verbal constructions with an auxiliary verb.

In Figure 3, the value of the special variable $THIS is the current word; <V+G> matches any verb in the Gerundive, <V+PP> matches any verb in the Past Participle and <V+INF> matches any verb in the Infinitive. The graph Neg recognizes the empty string and produces the negative adverbs *never|no longer|not*. The graph also recognizes

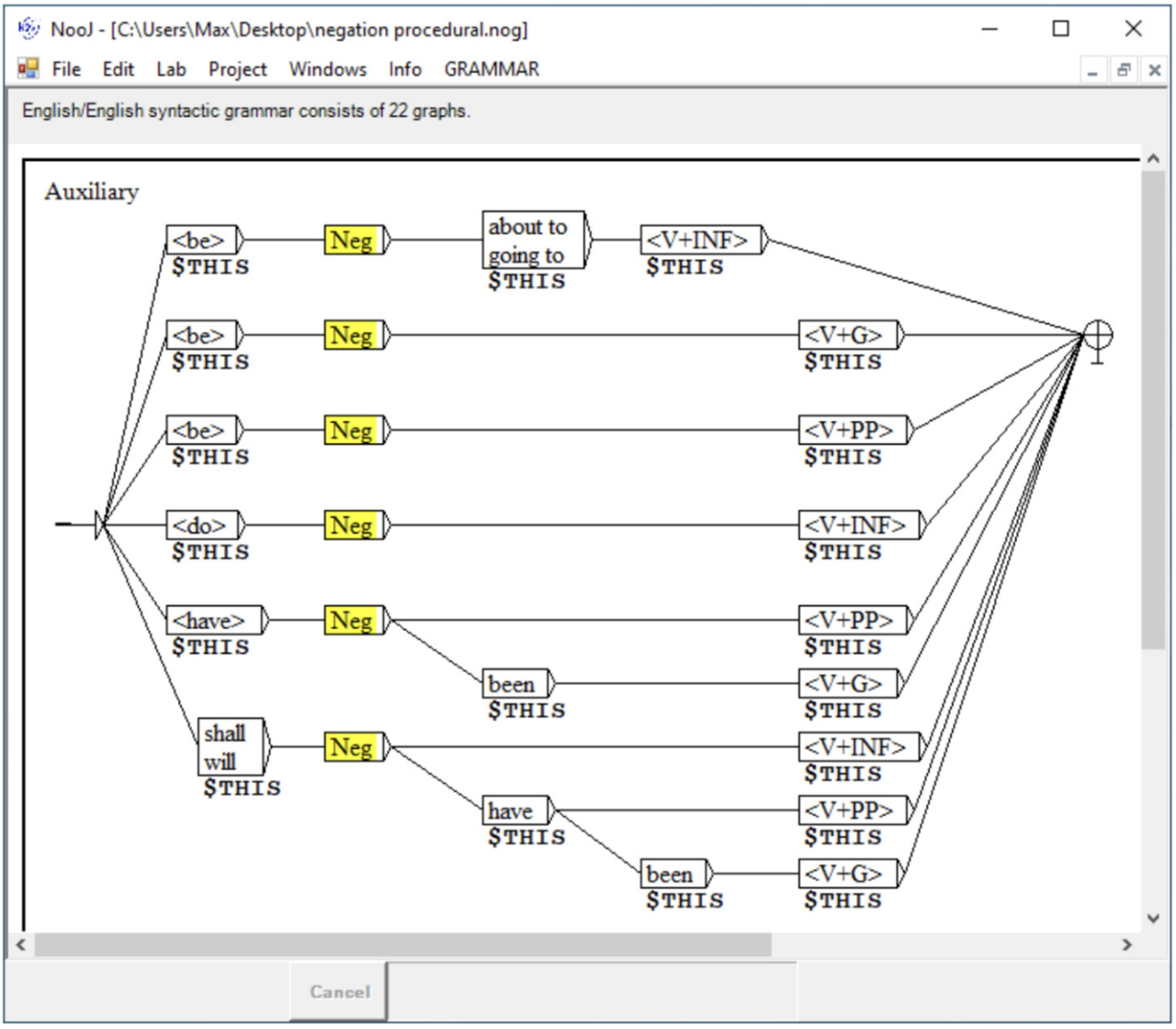

Figure 3. Graph Auxiliary.

sequences such as *has eaten* and *is going to eat*. The resulting negative structure produced by the transducer is *has not eaten*, *is not going to eat*, and *will not eat* respectively. The graph Modal recognizes verbal constructions with a modal verb *may*, *might*, *must*, *should*, and *would*, and produces the corresponding negative sentences. The graph Simple recognizes simple verb conjugated forms such as *eat* and *ate* and produces the corresponding negative sequences. This graph contains a special treatment for the verbs *be*, *have*, and *need* to produce the correct negative sequences *we are not* (instead of **we do not be*), *we have not*, and its variant *we do not have*, and *we need not* with its variant *we do not need*. Figure 4 displays the sequences that have been automatically rewritten in the negative form by applying the grammar to the text "The Portrait of a Lady" (Henry James, 1881). Two phenomena should be described in this grammar to make it complete:

- Negations using the determiner, e.g., *He eats meat* → *He eats no meat*.
- Negations do not involve replacing positive aspectual verbs (*continues to smoke*) with a negative aspectual verb (*stopped smoking*). The present grammar produces only the sentence *does* (*never*|*no longer*|*not*) *continue to smoke*.

The grammar produces only the sentence *He does* (*never*|*no longer*|*not*) *eat meat*. The grammar does not process verbal

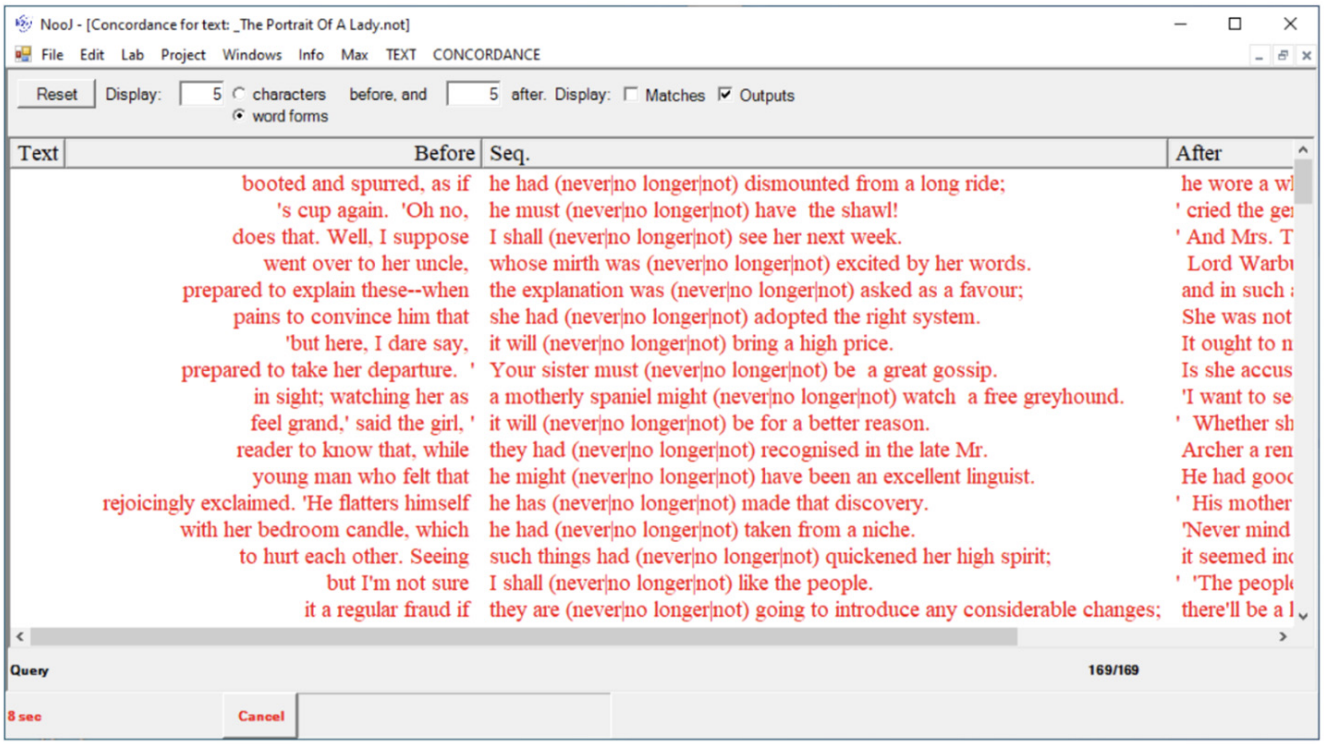

Figure 4. Applying the Negation Transformation.

groups that contain an adverb, such as *She is still smoking*, to not produce questionable sentences, such as *She is not still smoking*.

For illustrations of transformational grammars in French, refer to 1. This automatic generation system has been successfully used in several applications, including automatic paraphrasing of Portuguese[2,3], Arabic-French machine translation[4], and generation of a semantic representation from sentences written in Italian[5]. However, 1 has shown several of its limitations if the goal is to implement an automatic transformational system. The description of transformations in this manner is procedural, requiring the sequential application of multiple transducers from an initial sentence to recognize or generate a complex sentence. However, this approach does not scale up, which is why procedural approaches to Computational Linguistics were abandoned in the 1970s. 6 addresses the problem of describing the behavior of chains of transformations. Sometimes, altering the order of applying a transducer within a chain of transformations can dramatically change the resulting sentence, while in other cases, it may not have a significant impact.

For example, from the original sentence *He eats the cookie*, the three chains of transformations [Pronoun0] [Focus0] [Preterit], [Focus0] [Pronoun0] [Preterit], and [Focus0] [Preterit] [Pronoun0] produce the same sentence: *It is he who ate the cookie*. However, the two chains of transformations [Negation] [Focus0] [Preterit], and [Focus0] [Negation] [Preterit] produce two different results, respectively: *It is not he who ate the cookie*, and *It is he who did not eat the cookie*. Describing classes of equivalence for chains of transformations is an unsolved problem. The sheer quantity of transducers that would need development is astronomical. Our estimate suggests that there are approximately 1,000 productive transformations in English alone. These transformations include actions such as pronominalizing an argument to alter sentences (*The man ate the cookie He ate the cookie*), by changing the tense (*She will eat the cookie*), the aspect (*He goes on eating the cookie*), the modality (*She should eat the cookie*), the focus (*It is the cookie that he ate*), by nominalizing the verb (*She is a cookie eater*), by adding intensifiers (*He is a formidable cookie eater*), constructing passive sentences (*The cookie has been eaten*), questions (*Who ate the cookie?*), and so on. However, the transducers shown in Figure 1 and Figure 2 cannot be applied to a negative sentence to produce the underlying elementary sentence, i.e., its semantic analysis. In a broader context, since these transformational transducers are not symmetric, one must construct 2 × 1,000 = 2,000 transducers to comprehensively describe the transformational system of the English language.

2 Declarative approach

Transducers that recognize some sentences (as their input) to produce transformed sentences (as their output) are fundamentally asymmetric, which is against the linguistic goal of

describing transformations as symmetric pairs of sentences. We need a device that represents and processes both sentences of a transformation in the same way. Consider Figure 5. This transducer recognizes both sentences *The boy eats meat* and *The boy does not eat meat*. The transformational analysis of a sentence is therefore performed in two steps:

- The parser applies the transducer to the sentence, e.g., *The boy does not eat meat*. The transducer then produces the output [Negation] and sets variable $Subject to *The boy*, $Verb to *eat*, and $Object to *meat*.

- The generator then produces the language recognized by the transducer by exploring all its paths according to the values of the variables set at the first step. Because the variables were set during the first step of computation, their value can be computed during the generation step. For example, as variable $Verb stores the lexeme *eat* the symbol <V+PR> for the same variable/lexeme will be computed as *eats*. Any inflectional or derivation described in the system can then be applied to a given lexeme. For example, the derivation *eat eater* will be automatically produced by instantiating the symbol <N+Hum+s> (human noun, singular) for variable $Verb, whereas the combination of derivation and inflection *eat eateries* would be produced from the same variable $Verb by the symbol <N+Coll+p> (collective noun, plural).

Among all potentially generated sentences, one can produce the sentence that produces the [Elementary] output, i.e., *The boy eats meat*. The transformational analysis of the sentence is then presented as: [Negation]: *The boy does not eat meat* *The boy eats meat*. In other words, the system analyzes the sentence *The boy does not eat meat* as a negation of the sentence *The boy eats meat*. Figure 5 represents a very simplistic transformational grammar. Figure 6 presents a more complete transformational grammar that describes over 1,2 million declarative sentences transformed from the elementary sentence *Joe loves Lea*, including *Joe has never stopped loving Lea*, *Joe fell madly in love with her*, etc.,[7].

We are in the process of developing similar transformational grammars for a dozen languages. For instance, sentences produced by the grammar for the French sentence *Luc aime Léa* (Luc loves Lea) are shown in Figure 7.

Many of the sentences produced by transformations, even though syntactically correct, are awkwardly phrased, e.g., "*Ce n'est pas elle qui n'aura pas été l'amour fou de Luc*" [It is not she who will not have been Luc's passionate love]. They could be filtered out by some stylistic constraints, such as restraining the use of [Passive], avoiding performing more than one [Negation] in a sentence, and strategies alike. The declarative approach can also be used to automatically generate text in English from a semantic predicate represented by an RDF Turtle triplet[8]. The principle is to replace the description of the elementary sentence with a semantic predicate written in the RDF format: the system can then parse and generate both the predicate in its RDF form and the sentences that represent all the ways to state it in English. The same process can both generate English sentences from the RDF semantic representation and generate the RDF representation from any English sentence recognized by the grammar, i.e., compute its meaning.

3 A Case Study with eSPERTo Paraphrasing System

In this Section, we showcase how NooJ was used to create resources for the eSPERTo paraphrase generation system. The generation system allows, among other transformations, the inversion of subject-object arguments in a sentence with a reciprocal verb, such as the Portuguese verb *parecer-se com* (*look alike*, *resemble*), as illustrated in Figure 8.

4.1 The Interface

When presented with a text, eSPERTo identifies word sequences that can be replaced with equivalent alternatives, known

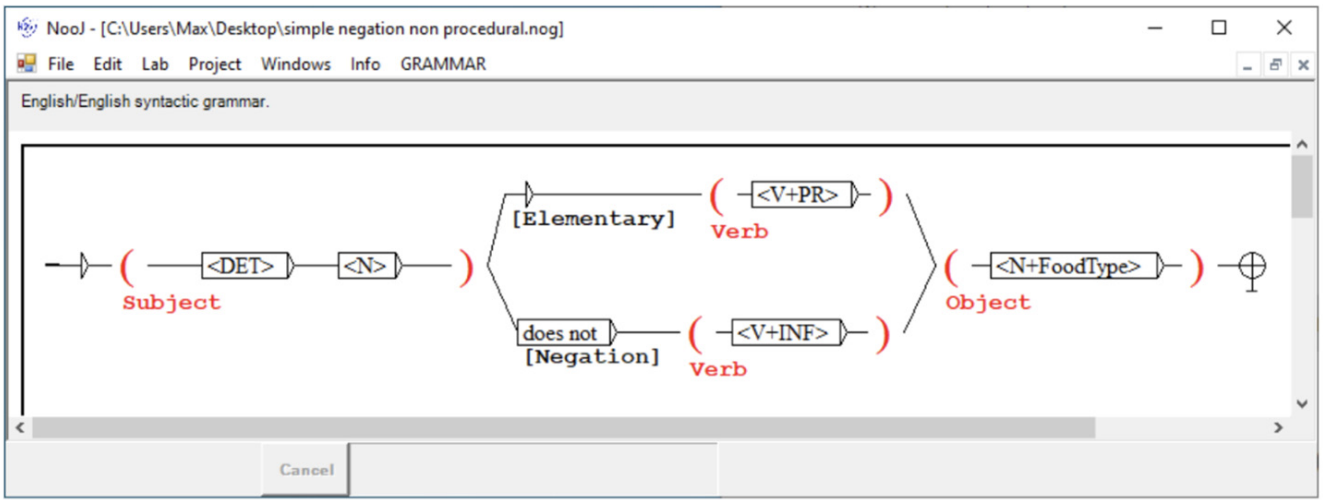

Figure 5. Describing Transformations in a Declarative way.

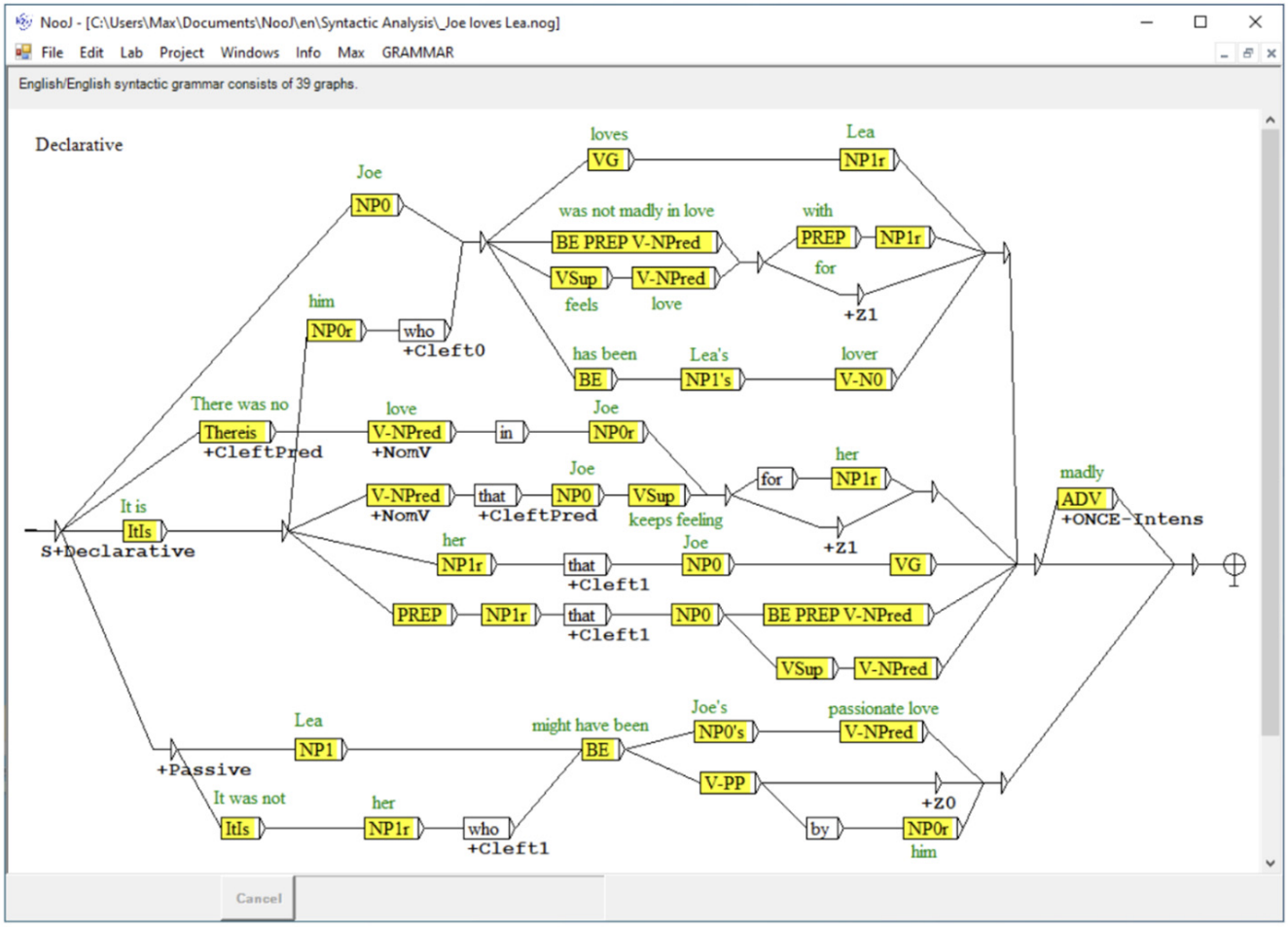

Figure 6. A Transformational description of direct transitive sentences.

as paraphrases. In Figure 5, these identified sequences are enclosed within red square brackets. Upon hovering the cursor over them, eSPERTo highlights the sequences in green and presents a list of suggested paraphrases that can replace that sequence. In our example of the reciprocal verb in Portuguese, *parecer-se com* in the sentence *a mulher é parecida com o irmão* (the woman looks like the/her brother), eSPERTo suggests the inversion *o irmão é parecido com a mulher* (*the brother looks like the woman*) and the coordination of the two arguments *a mulher e o irmão* (*the woman and the brother*) *são parecidos* (*the woman and the brother look alike)* with the correct change of the verb to the plural form (*é* (*is*) transforms into *são* (*are*). Then, the different constructions of the sentence can also be paraphrased by replacing words with synonyms and phrases with similar phrases.

4.2 The Generative Process

To elucidate the features of a generative system like eSPERTo alongside ChatGPT, the aim is to outline the differences between their generative processes and the resulting outputs of each system. In ChatGPT (version 3.5), the user can see the generated paraphrases illustrated in Example 1 for the Portuguese sentence illustrated in Figure 4.

(1) $_{PT}$ *- a mulher é parecida com o irmão*

$_{PARA}$ - ?a irmã se assemelha ao irmão

$_{PARA}$ - a mulher possui semelhanças com o irmão

$_{PARA}$ - a mulher tem semelhanças com o irmão

$_{PARA}$ - a mulher apresenta semelhanças com o irmão

$_{PARA}$ - a mulher é similar ao irmão

$_{PARA}$ - ?a irmã tem traços semelhantes ao irmão

$_{PARA}$ - a mulher e o irmão possuem semelhanças

$_{PARA}$ - a mulher possui características que se assemelham ao irmão

$_{PARA}$ - a mulher e o irmão compartilham semelhanças

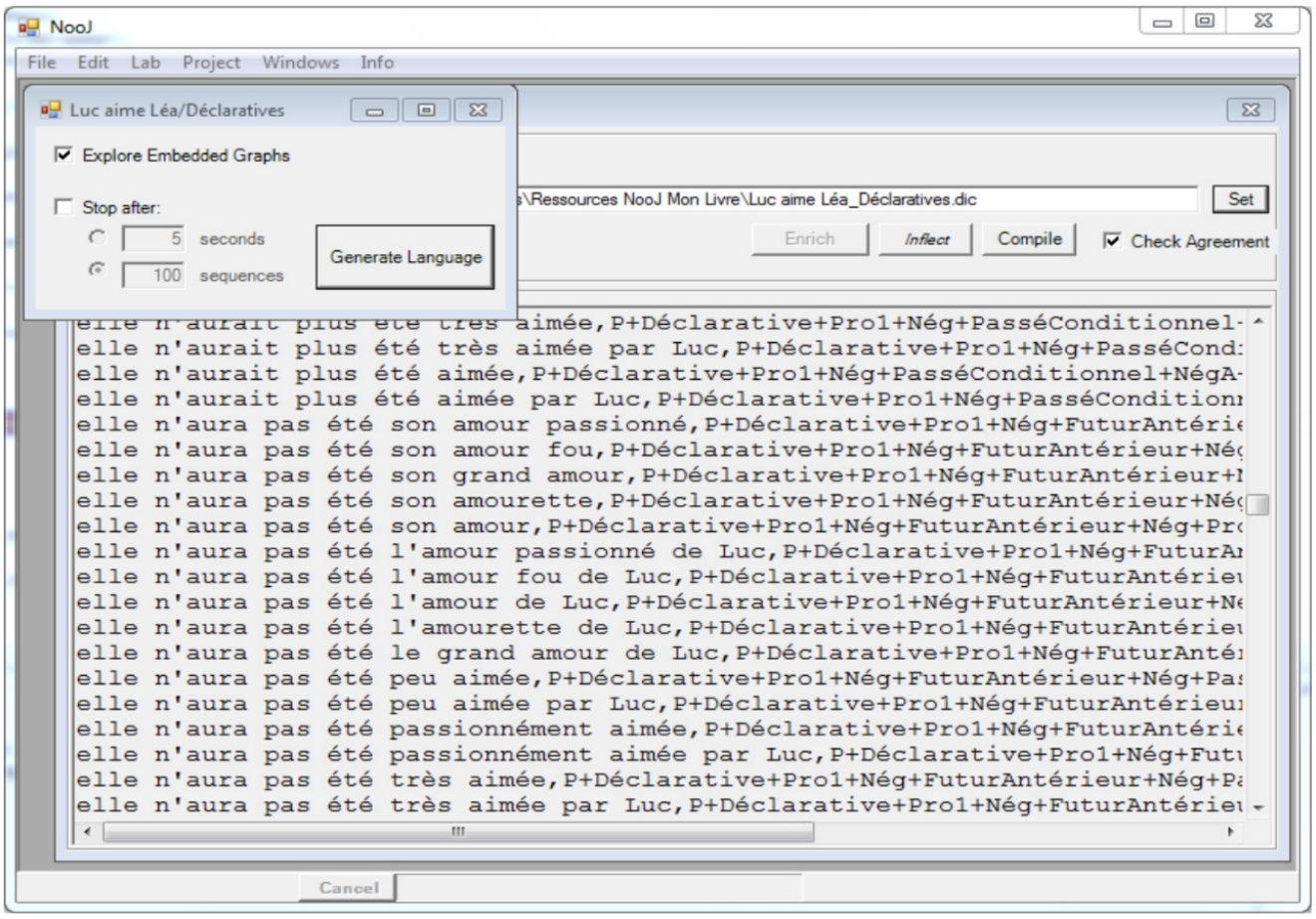

Figure 7. **Producing transformations automatically.**

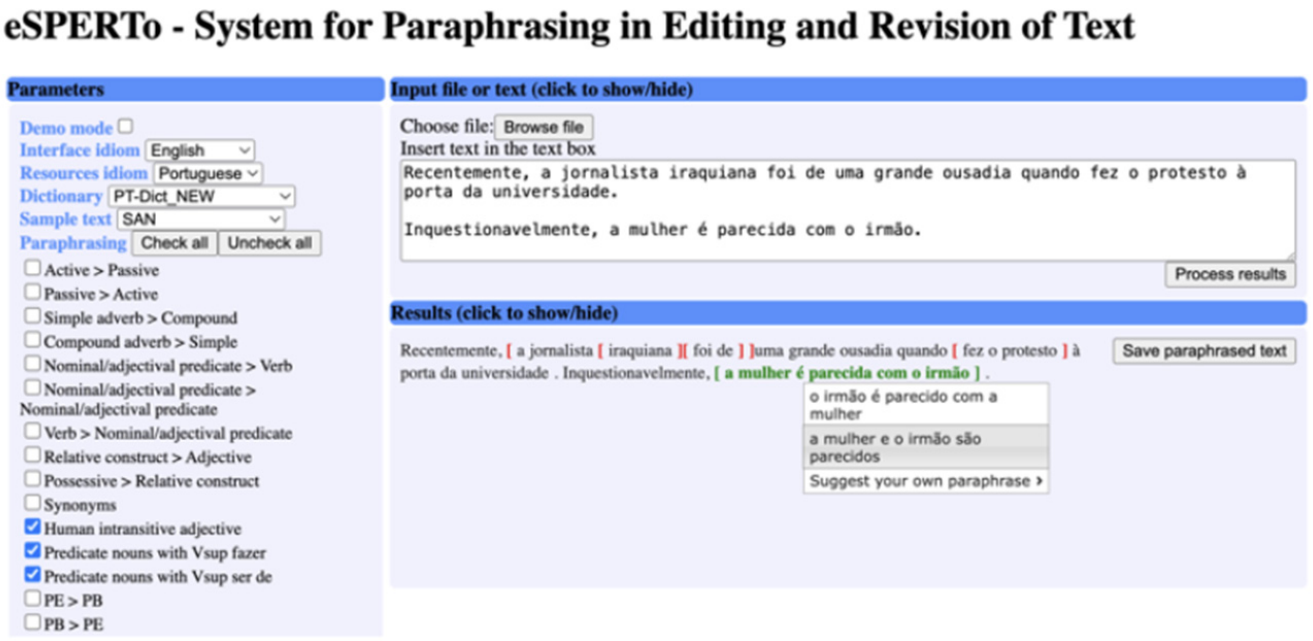

Figure 8. **eSPERTo paraphrasing system.**

PARA - a mulher tem uma semelhança com o irmão

PARA - a mulher tem características parecidas com as do irmão

PARA - a mulher tem semelhanças com seu irmão

PARA - a mulher se assemelha ao irmão

PARA - a mulher possui traços semelhantes ao irmão

PARA - a mulher apresenta semelhanças com seu irmão

PARA - a mulher possui características que são parecidas com as do irmão

PARA - a mulher e o irmão compartilham semelhanças

PARA - a mulher e o irmão têm semelhanças

PARA - a mulher tem características que se assemelham ao irmão

PARA - a mulher possui semelhanças com o irmão

PARA - a mulher apresenta traços similares ao irmão

PARA - a mulher e o irmão possuem características semelhantes

PARA - a mulher possui traços que se assemelham ao irmão

The paraphrases generated from the original sentence showcase paraphrastic units that have been altered across different segments of the sentence. In eSPERTo, the linguistic processes involved in transforming the original sentence are transparent, with clearly delineated boundaries indicating where transformations can occur. This shift signifies visible paradigmatic changes. The methodology maintains consistency and has been validated by humans. Additionally, it's worth noting that all the paraphrases produced by ChatGPT can also be generated by eSPERTo.

4.3 The resources

A significant part of the resources used to generate paraphrases is based on the lexicon-grammar theory[1]. More specifically, the resources include the lexicon-grammars for (i) human intransitive adjectives[9], (ii) nominal predicates supported by the verb *fazer* (*to do/to make*)[3], and (iii) nominal predicates supported by the verb *ser de* (*to be of*)[10]. Figure 9 shows an excerpt of the lexicon-grammar tables for patronymic adjectives highlighting two properties that are different for two different adjectives: *açoriano* (*Azorean*) which refers to people from the Azores, a Portuguese island, and *iraquiano* (*Iraqi*) which refers to people from the country Iraq. When a predicate (in this case, human intransitive adjective) has a certain property, it is marked '+', otherwise, it is marked '-' under that property. For instance, according to this lexicon-grammar, the copulative verb *ser* (*to be*) in *ser açoriano* cannot be replaced with the auxiliary verb *tornar-se* (*become*) nor the adjective be rewritten with a construction with the classifier noun *nacionalidade* (e.g., **ter nacionalidade açoriana* - *to have Azorean nationality) whereas both constructions are possible with the adjective *Iraqi*: *tornar-se iraquiano* (*become Iraqi*) and *ter nacionalidade iraquiana* (*to have Iraqi nationality*).

The lexicon-grammars were semi-automatically converted to NooJ dictionaries by transforming each property marked with a '+' into a dictionary entry attribute +property_name as highlighted in bold in Figure 10. This figure shows the converted NooJ entries of the adjectives highlighted in Figure **??**6. As can be seen, property [+Vcoptornarse] and [+NclassPnacionalidade] highlighted in orange were added to *iraquiano* dictionary entry but not to *açoriano* as this predicate does not have those properties marked '+' in the lexicon-grammar.

In addition to converting the lexicon-grammar into NooJ dictionary entries, a few transformational properties described in the lexicon-grammars were represented in NooJ syntactic grammars following a procedural approach. These grammars generate paraphrases. In Figure 11, the grammar identifies constructions with the structure [N0 ser de DET A Npred com N1] (N0 to be of DET A Npred with N1) such as *O Pedro é de uma grande parecença com a Maria* (*John is of a great resemblance with Mary*) and generates the equivalent constructions with the arguments reversed ([N1 ser de DET A

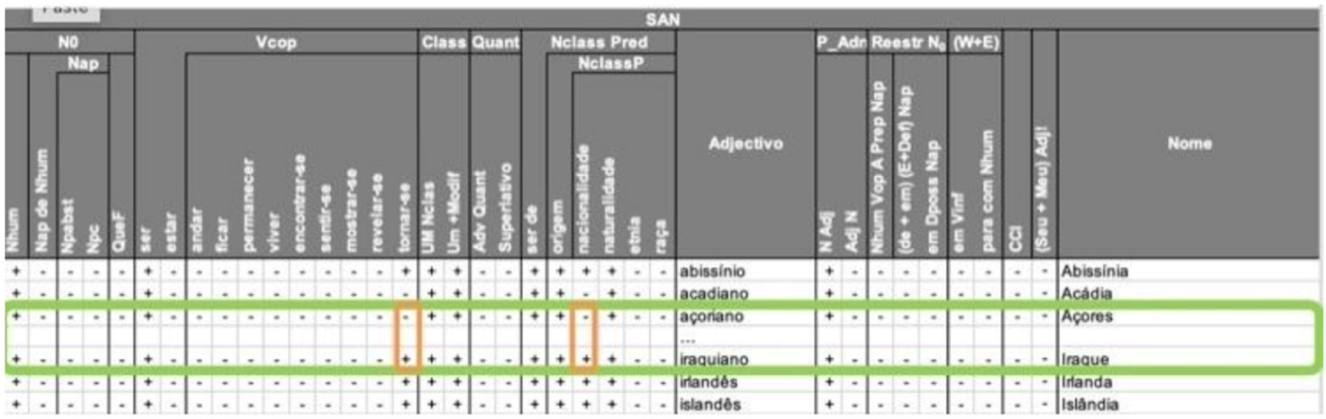

Figure 9. Excerpt of the lexicon-grammar of patronymic adjectives.

```
açoriano,A+FLX=ALTO
+Table=SAN+Nhum+Vcopser+UMNclas+UmModi
f+NclassPserde+NclassPorigem+NclassPnatural
idade+NAdj+DRV=nometop:Açores
…
iraquiano,A+FLX=ALTO+AN+lang+EN=Iraqi
+Table=SAN+Nhum+Vcopser+Vcoptornarse+UM
Nclas+UmModif+NclassPserde+NclassPorigem+
NclassPnacionalidade+NclassPnaturalidade+NA
dj+DRV=nometop:Iraque
```

Figure 10. Converted NooJ entries from adjectives described in the lexicon-grammar highlighted in Figure 9.

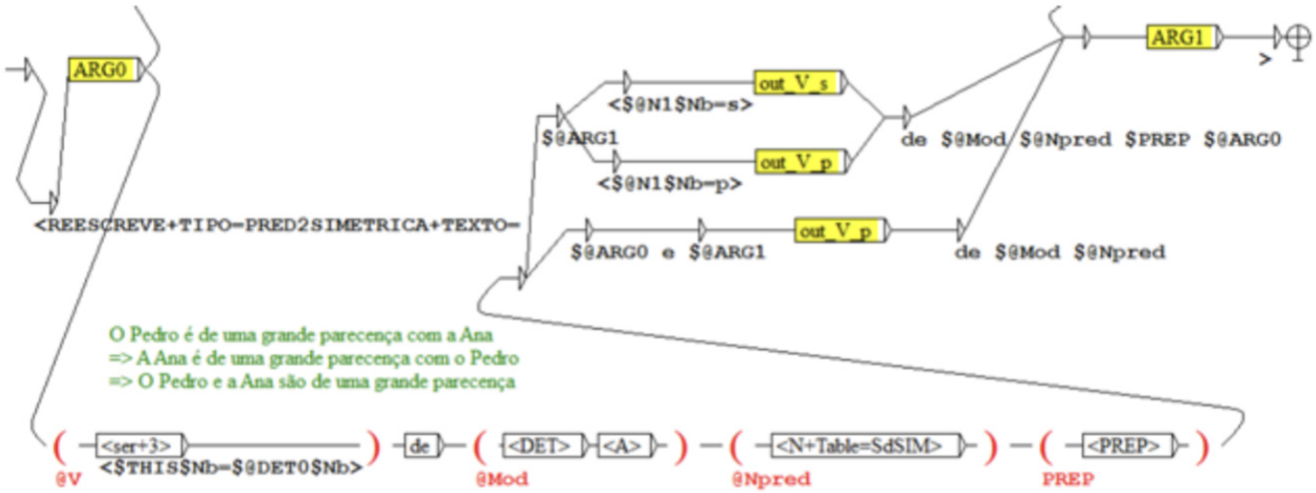

Figure 11. Excerpt from the grammar that generates paraphrases for the predicates *ser de* (to be of) that are symmetric: ARG0 and ARG1 are subgraphs that identify simple human noun phrases that are assigned to the global variables @ARG0 and @ ARG1, respectively; the head of the noun phrases are also assigned respectively to global variables @N0 and @N1. The subgraph :out_V_s creates the corresponding paraphrases when @N1 is singular, and :out_V_p creates the paraphrases when @N1 is plural and for the coordinated arguments so that the subjects @ARG1 or @ARG0 e @ARG1 (@ARG0 and @ARG1)agree in number with the support verb (@V).

Npred com N0] – [N1 to be of DET A Npred with N0]) and coordinated ([N0 e N1 ser de DET A Npred] – [N0 and N1 to be of DET A Npred]). In the case we are illustrating, it generates the two paraphrases: *A Maria é de uma grande parecença com o João* (*Mary is of a great resemblance with John*) and *O João e a Maria são de uma grande parecença* [*um com o outro*] (John and Mary are of a great resemblance [with one another]). Moreover, it was also necessary to create a grammar to identify the coordinated arguments and then generate the constructions ([N0 ser de DET A Npred com N1] – [N0 to be of DET A Npred with N1]) and ([N1 ser de DET A Npred com N0] – [N1 to be of DET A Npred with N0]).

However, as pointed out in 2, which includes a detailed description of how the lexicon-grammar information was integrated into NooJ, eSPERTo will benefit greatly by using the declarative approach as for any set of *n* sentences that are paraphrases of each other one needs to develop only one graph stating that the *n* paraphrases are equivalent whereas with the procedural approach, it is necessary to create *n* graphs with (*n -1*) paths in a total of *n x* (*n -1*) paths, because each graph identifies one sentence only in the set of paraphrases and generates all other *n-1* paraphrases in the set.

In addition to the lexicon-grammar resources already in existence, INESC-ID's scholars introduced a new concept and coined it "paraphrasary"[11,12], which encompasses advanced expansions of dictionaries incorporating resources spanning multiword and phrasal levels. The Multi3Generation COST Action has propelled research in monolingual and multilingual applications, particularly through the development of paraphrasaries. These resources aim to address the gap in linguistically sophisticated tools. Experimental tests, aided

by alignment tools, have informed the creation of guidelines for aligning paraphrases and translations to improve semantic equivalence. Expanding the volume of paraphrastic units and extending alignment processes to more creative text types are critical endeavors. Collaborative efforts in building paraphrasaries stand as a significant advancement in linguistic research, promising enhancements in NLP applications.

4 Conclusion

The linguistic objective of describing natural languages is a valuable scientific pursuit, regardless of any immediate or short-term applications. Accessing an exhaustive and precise description of a language benefits all NLP software applications. Simply enumerating sentences within a language is insufficient; it is crucial to establish connections between sentences derived through transformations, such as active and passive versions or affirmative and negative forms, among others. This article presents two methods for formalizing these connections: the procedural approach and the declarative approach. The procedural method uses simple transducers to generate transformed sentences from elementary ones, taking an elementary sentence as input and producing the corresponding transformed sentence as output. The procedural approach, exemplified by eSPERTo, has demonstrated success in automatically producing paraphrases (using monolingual paraphrasaries) and translations (when associated with bilingual dictionaries). The declarative approach employs a single "neutral" grammar to represent both elementary and transformed sentences. Future iterations of eSPERTo could explore adopting the declarative approach to represent equivalent transformations within a single grammar. "Neutral" grammars can function in both "parsing" and "generating" modes using a single computational device to process (parse or generate) texts. Their advantage lies in enabling linguists to describe transformations in a purely declarative manner. A comprehensive transformational grammar of a language would be the linguistic equivalent of the human genome project. The NooJ community has made strides in developing transformational grammars for numerous languages over several years. These grammars can further improve the generative capabilities of paraphrase generation in eSPERTo and other NLP applications, regardless of their model type.

Data and software availability

The procedural approach is a methodology implemented in NooJ, a free and open-source software that can perform the functions described in the article (https://nooj.univ-fcomte.fr). [illegible] has a large international multilingual community that uses and shares results at the annual NooJ internal conference. To ensure the reproducibility of our research, we provide detailed descriptions of the materials and methods used, allowing others to replicate our work. All data supporting the findings of this study are available within the article. [illegible] relevant materials are already accessible through the website and other channels. [illegible] tutorial and resources on the NooJ website, known for creating high-quality linguistic resources for many languages. There is no need for an additional description of extended data or novel analysis code. [illegible] Tutorial is included. Sufficient and relevant references to the most pertinent publications for this article are provided [illegible] and other visual data were created using software developed by the authors. Any additional inquiries regarding data access can be directed to the corresponding author. [illegible] has a large international multilingual community that uses and shares results at the annual NooJ internal conference. To ensure the reproducibility of our research, we provide detailed descriptions of the materials and methods used, allowing others to replicate our work. All data supporting the findings of this study are available within the article. [illegible] relevant materials are already accessible through the website and other channels. [illegible] tutorial and resources on the NooJ website, known for creating high-quality linguistic resources for many languages. There is no need for an additional description of extended data or novel analysis code. [illegible] Tutorial is included. Sufficient and relevant references to the most pertinent publications for this article are provided [illegible] and other visual data were created using software developed by the authors. Any additional inquiries regarding data access can be directed to the corresponding author.

Nooj Tutorials and resources can be accessed as by the following link:

https://nooj.univ-fcomte.fr/tutorials.html

eSPERTo paraphrase generation system -

eSPERTo code and resources can be accessed freely for research purposes by requesting access to the principal investigator of the eSPERTo project, Anabela Barreiro (anabela.barreiro@inesc-id.pt)

Author contributions

Max Silberztein, Multi3Generation's Grant Coordinator, wrote most of Section 1, and Section 2 and Section 3. Cristina Mota, Multi3Generation member, wrote part of Section 4.

Anabela Barreiro, Multi3Generation's chair, coordinated the article, organised the structure of the manuscript, wrote the Abstract, Plain Language Summary, Introduction, part of Section 4, Subsection 4.2, part of Subsection 4.3, Conclusion, and revised the entire article.

Acknowledgment

This article is based upon work from COST Action Multi3Generation (CA18231) supported by COST (European Cooperation in Science and Technology).

References

1. Gross M: **Les bases empiriques de la notion de prédicat sémantique.** *Langages.* Included in a thematic issue: Formes syntaxiques et prédicats sémantiques. 1981; **15**(63): 7–52.
Reference Source
2. Barreiro A, Mota C, Baptista J, *et al.*: **Linguistic resources for paraphrase generation in portuguese: a lexicon-grammar approach.** *Lang Resour Eval.* 2022; **56**(1): 1–35.
Publisher Full Text
3. Chacoto L: **O verbo fazer em construções nominais predicativas.** PhD thesis, Universidade do Algarve, 2005.
Reference Source
4. Essid M, Fehri H: **A semantico-syntactic disambiguation system of Arabic movement and speech verbs and their automatic translation to French using NooJ**. *12th International Conference, NooJ 2018, Palermo, Italy, June 20– 22, 2018*. Revised Selected Papers. 2019; **987**: 167–179.
Publisher Full Text
5. Vietri S: **The annotation of the predicate-argument structure of transfer nouns**. Cambridge Scholars Publishing, 2013; 88–101.
Reference Source
6. Silberztein M: **Formalizing natural languages: the NooJ approach**. 2016.
Reference Source
7. Silberztein M: ***Joe Loves Lea*: transformational analysis of direct transitive sentences.** *Automatic Processing of Natural-Language Electronic Texts with NooJ.* 2016; **607**: 55–65.
Publisher Full Text
8. Silberztein M: **From FOAF to english: linguistic contribution to web semantics**. 2017; 1–9.
Reference Source
9. Carvalho P: **Análise e representacão de construcões adjectivais para processamento automático de texto. Adjectivos intransitivos humanos**. 2007.
Reference Source
10. Baptista J: **Sintaxe dos nomes predicativos com verbo-suporte SER DE**. Fundação para a Ciência e a Tecnologia/Fundação Calouste Gulbenkian, 2005.
Reference Source
11. Barreiro A, Rebelo-Arnold I, Mota C: **Parafrasário: a variety-based paraphrasary for portuguese**. In: *Proceedings of the 17th NooJ International Conference 2023 – Book of Abstracts.* Zadar, Croacia, University of Zadar, Department of Classical Philology and Department of Information Sciences, 2023; 44–45.
Reference Source
12. Barreiro A, Mota C: **A multilingual paraphrasary of multiwords**. In: *Proceedings of the 1st International Workshop on Multilingual, Multimodal and Multitask Language Generation (Multi3Generation).* Tampere, Finland, The Association for Computational Linguistics. 2023; 47–56.
Reference Source

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 13 August 2024

https://doi.org/10.21956/openreseurope.19444.r42723

© **2024 Vayadande K.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Kuldeep Vayadande
Vishwakarma Institute of Technology, Pune, Maharastra, India

The paper provides a conceptual overview of the methods used, including the use of finite-state transducers (FSTs) and finite state automata (FSA) for linguistic transformations. However, while the theoretical framework is well-explained, the paper does not provide sufficient technical details, such as specific algorithms, parameters, or step-by-step instructions, that would be necessary for full replication by others. To enable replication, the paper would need to include more detailed explanations of the procedures, including code snippets, data sets, or examples of how the transducers and automata are implemented and applied to specific linguistic tasks.

The paper does not include comprehensive data sets or detailed examples that would allow full reproducibility. For full reproducibility, the paper would need to include access to the source data, such as the specific linguistic datasets used in the experiments, as well as the exact configurations of the FSTs and FSAs used in the transformations.

The conclusions about the method and its performance are supported by the findings, but there are some limitations. While the paper provides a clear demonstration of the method's effectiveness through examples and experiments, the lack of comprehensive source data and detailed performance metrics means that the conclusions are not fully substantiated. More detailed experimental results, comparisons with other methods, and additional data would be necessary to fully support the conclusions.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial Intelligence, Machine Learning, Deep Learning, LLM, NN, Security, Cloud Computing, Web Serices, Software Engineering

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 27 July 2024

https://doi.org/10.21956/openreseurope.19444.r42161

© 2024 Raza S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Shaina Raza
University of Toronto, Toronto, Ontario, Canada

The paper presents interesting and useful research, but there are a few concerns that need to be addressed:

1) The authors must clearly define their scope and contribution in the introduction. A key question that remains unanswered is how their approach improves upon the state of the art. This is a critical aspect that appears to be missing from the paper and needs to be explicitly addressed.

2) The responsible use of language generation is another area that requires attention. The authors should consider incorporating key takeaways from the following source:

- Raza S et al. (2024 [Ref - 1])

This will provide valuable insights into the ethical considerations and best practices in language generation.

3) The paper seems to focus more on compiler level studies. If this is indeed the main contribution, it should be explicitly stated.

4) Furthermore, with the advent of LLMs, it is important to discuss whether bulk analysis is feasible. If the LLM approach is prone to bias, this should be highlighted in the contributions section that authors works have another line of direction.

5) Additionally, the scalability of the current approach should be examined. Have the authors found any evidence or made any observations about the scalability of their method?

6) The term "linguistics" in the paper needs further clarification. It is essential to explain what is meant by linguistic analysis in this context. Is it a more granular level of analysis? The authors need to shed light on this aspect to ensure readers understand the depth and specificity of their linguistic study.

7) Is the approach applicable to other languages?

References

1. Raza S, Ghuge S, Ding C, Dolatabadi E, et al.: FAIR Enough: Develop and Assess a FAIR-Compliant Dataset for Large Language Model Training?. *Data Intelligence*. 2024; **6** (2): 559-585 Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: NLP and responsible AI

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.