**SHORT TERM SCIENTIFIC MISSION (STSM) – SCIENTIFIC REPORT**

The STSM applicant submits this report for approval to the STSM coordinator
Action number: CA16105
STSM title: Exploring the Role of Paraphrases in Interpretability and Explainability on Language Learning Applications
STSM start and end date: 17/12/2018 to 21/12/2018
Grantee name: Anabela Barreiro

**PURPOSE OF THE STSM**

The main purpose of the proposed STSM was to explore the role of paraphrases in **interpretability and explainability on Language Learning Applications**. However, we saw fit to tighten our goal to paraphrase generation only.

The first task consisted of compiling a literature overview on the use of paraphrases for the distinct goals. One approach explored was to use **multiword units** and **idiomatic expressions** and transform them into several semantically equivalent ways of expression for the language learner, i.e., we generated several possibilities of writing the same expression to provide students with the option to choose among them, according to the purposes of each of their texts or simply according to their stylistic preferences.

Some texts require more formality, others, instead, require a certain level of informality. This formal/informal topic has been tackled briefly in a recent publication [1] in which the STSM grantee created a methodology for the automation process of paraphrasing and converting Portuguese constructions typical of informal or spoken language into a formal written language, a standard revision procedure in Portuguese.

The research work was performed with the support of NooJ linguistic environment through the application of dictionaries, morpho-syntactic grammars and generic transformational grammars. All transformations can be used to make a text clearer and more comprehensible. The linguistic resources produced were integrated into the eSPERTo paraphrasing tool, as illustrated in Figure 1.

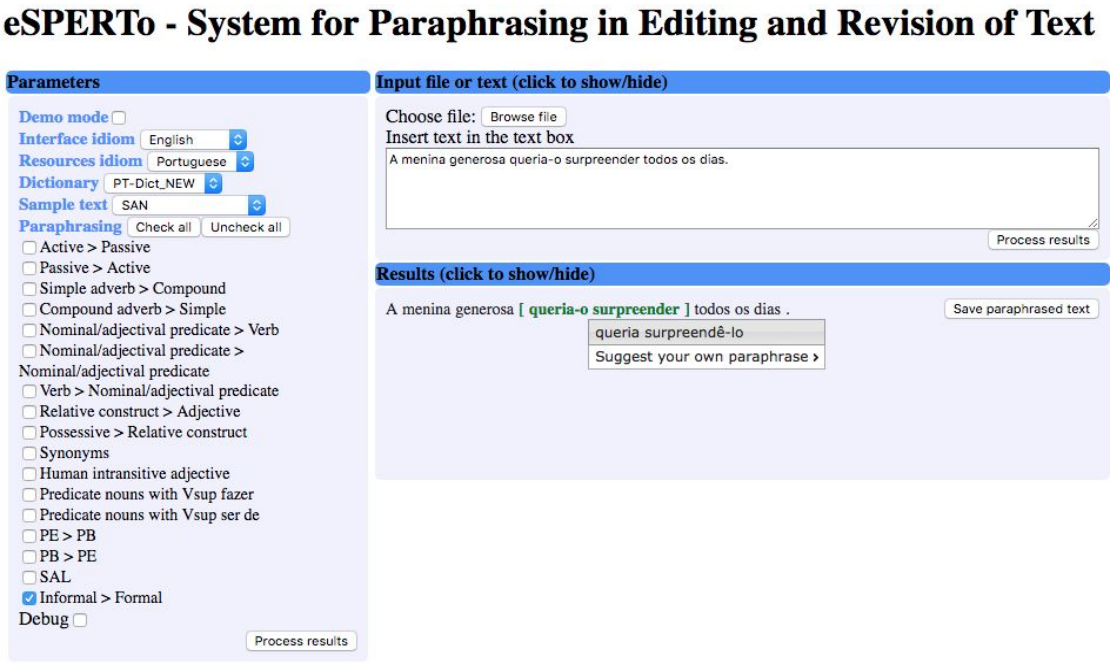


Figure 1: Conversion of an informal verbal compound with a clitic pronoun into a formal equivalent where the clitic appears after the main verb

The same methodology and the same type of grammars can be used to transform expressions or terms into other ways of expression for understandability purposes. There are already some resources available that can be used to paraphrase constructions with support verbs (for example, paraphrasing the verbal construction with the nominal construction and vice-versa, or alternation between the support verb and other stylistic or aspectual verbs), active and passive constructions, adverbial compounds, relatives and participles, among others.

In this STSM, we explored general texts that contained more common language and did not look into domain specific corpora (as first envisaged), but experimented adapting our existing resources into the generation of paraphrases for English, as illustrated in Figure 2. The paraphrastic knowledge acquired can be used for better understanding of a sentence or text, for writing a higher quality text or changing it stylistically. This research work targets the improvement of the eSPERTo paraphrasing tool into different languages, to promote linguistic independence, enabling students, especially those with more difficulties, to express themselves with a greater versatility. Efficient paraphrasing tools should have the capability to enable students to achieve quality sentences with a good understanding of them. Once paraphrastic correspondence is apprehended, then language learners can decide which one to use.

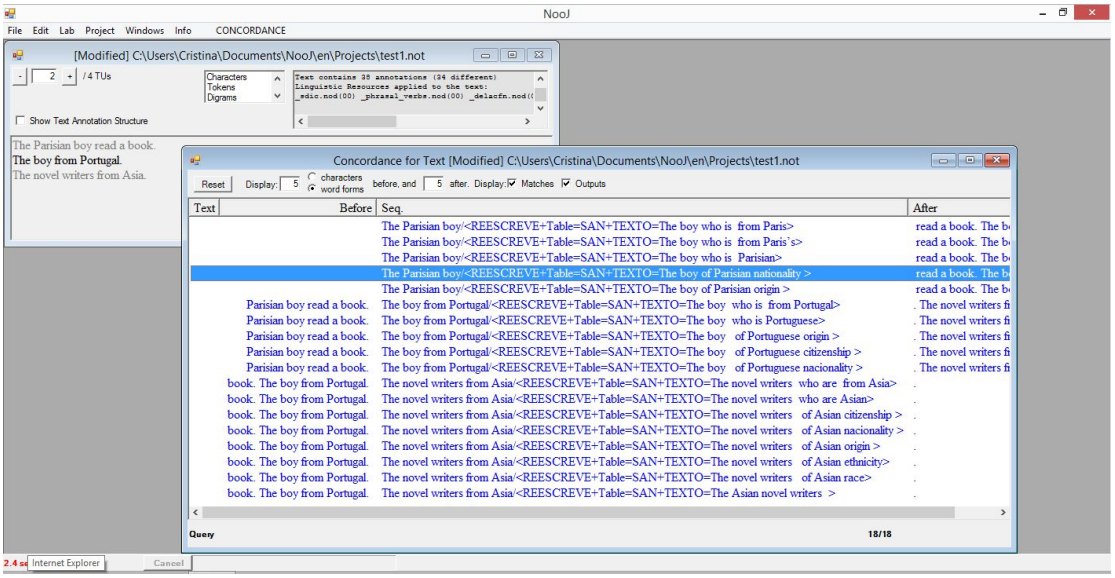


Figure 1: Conversion of an informal verbal compound with a clitic pronoun into a formal equivalent where the clitic appears after the main verb

The STSM also aimed to (i) extend and improve paraphrase-oriented resources and achieve enhanced collective quality sets of resources for advancing the state of the art in paraphrasing through a combination of techniques; and (ii) enhance the potential to address more complex challenges, testing systems with larger volumes of data studied and validated by linguists; (iii) apply machine learning to identify suitable pairs of paraphrastic units.

The outcome of the STSM led to some research experiments, such as exploring natural language processing techniques to apply linguistic rules to language learning, focusing on paraphrase methods that use structured definition graphs, use a navigation algorithm based on distributional semantic models to find a path in the graph which links text and entailment hypothesis. This provides a greater differentiation with regards to related works of lexical text simplification.

**DESCRIPTION OF WORK CARRIED OUT DURING THE STSM**

The STSM at University of Manchester provided the opportunity to regular meetings with the host, Professor André Freitas and his PhD student Mokanarangan Thayaparan, who have experience in the area of expertise, namely in textual entailment and semantic relations. During the STSM that took place between 17/12/2018 and 21/12/2018 (5 days), the following work was carried out:

- Brainstorming on the role of paraphrases in language learning;

- Overview of paraphrasing functions and their role in Natural Language Processing (NLP) and NLP Applications, namely language learning, language writing and understanding;
- Revision of state of the art literature in new techniques on text entailment and semantic relations, which represent the core of a paraphrasing system and revision of the development of paraphrasing systems taking into consideration language learning applications;
- Initiated a summary of literature review on the STSM related topics;
- Brainstorming on how to use hand-crafted paraphrases to improve language learning techniques;
- Development of script for collecting multilingual dictionaries from the OpenLogos system for several languages, with priority for English and Portuguese;
- Discussed of how OpenLogos Semantico-Syntactic Abstraction Language (SAL) can be used in the creation of rules/local grammars to generate paraphrases and how it can be combined with the techniques of knowledge extraction used by the host institution researchers;
- Discussed the possibility of re-using the OpenLogos SAL Tutorial and lexical resources, making them a public standalone resource to be used by the research community;
- Discussed the type of paraphrastic resources that can be built in a collaborative way, namely an automated way creating transformational grammars that use SAL as semantic constraints;
- Discussion on how to improve the eSPERTo paraphrasing system creating a module suitable for language learning, area of research that is very relevant for enetCollect's WG3 group as it addresses the topic of user-design strategies for a competitive solution.
- Compilation of an extended abstract and draft of a scientific article, co-authoring the host and participating colleagues. A discussion has also taken place about how/where the collaborative research work can be published -- aiming ACL 2019 or CICLing 2019 conference research paper and an oral presentation at the 3rd Annual MC Meeting of the enetCollect COST Action, which will take place on March 14—15 in Lisbon, Portugal with the grantee holder being its local organizer;
- Outlined the set of Machine Learning frameworks to support the follow-up experimentation, focusing on neuro-symbolic models and program synthesis methods. The main target of the work is the induction of transformational grammar rules for paraphrases in the presence of sparse and unbalanced data (few-shot learning).
- Definition of an initial target gold-standard to use in the Machine Learning experiments.
- The development and evaluation of the machine learning methods for learning transformational grammars from a low number of paraphrases will be the main target of the follow-up collaboration.
- Future collaboration will include developing the topic of NLP tools including paraphrasing tools to LL applications.