Using an Ontology-based Approach to Build Open Assisting Tools in Foreign Language Writing

Hugo Sanjurjo-González [1]*, Héctor Alaiz-Moretón [2], Noelia Ramón [3], Belén Labrador [3], Isaías García [2]

[1] *Department of English, Linguistics and History, University of Huddersfield*, UK
[2] *Department of Electrical, Systems and Automatics Engineering, University of León*, SPAIN
[3] *Department of Modern Languages, University of León*, SPAIN

***Corresponding Author:** H.SanjurjoGonzalez@hud.ac.uk

ABSTRACT

In today's globalised world where there is a growing need for international communication, non-native speakers (NNS) from a wide range of professional fields are increasingly called upon to write specialised texts in English. More often than not, however, the linguistic competence required to do so is well beyond that of the majority of NNS. While software applications can serve to assist NNS in their English writing tasks, most of the applications available are designed for users of English for general purposes as opposed to English for professional purposes. Therefore, these applications lack the specific vocabulary, style guidelines and common structures required in more specialised documents. Necessary modifications to meet the needs of English for professional purposes tend to be viewed as representing an overly complex and expensive task. To overcome these challenges, we present a software called O-WEAA (Ontology-Writing English Assistant Architecture) which makes use of an ontology that represents the knowledge which, according to our formalisation, is required to write most types of specialised professional documents in the English language. Our formalisation of the required knowledge is based on an exhaustive linguistic analysis of several written genres. The proposed software is composed of two parts: i) a web application named Acquisition Interface Module, which allows experts to populate the ontology with new data and ii) a user-friendly, general web interface named Writing Assistant Interface Module which guides the user throughout the writing process of the English document in the specific domain described in the ontology.

Keywords: knowledge acquisition, knowledge representation, corpus linguistics, ontology classification and retrieval, natural language processing

INTRODUCTION

Because of globalisation, more and more people from different parts of the world are working together on projects related to their industry, economy or academic field. Most of these international relationships make use of English as a *lingua franca*, even in written contexts. This means that often NNS (non-native speakers) are required to write specialised documents in English and in a wide range of different areas. Some examples of such technical documents include medical abstracts, electronic product descriptions, and technical brochures. Writing such texts requires adherence to a strict set of conventions and rules, different in each domain. Achieving the competence and the proficiency required to do so is necessarily a long, complex and time-consuming process.

From our point of view, NNS have to face two main problems: first, inadequate knowledge of the English language to write acceptable texts, and second, lack of awareness of the particular conventions related to some specific domain.

Several studies have focused on the use of computer-based resources to improve the writing skills of NNS and deal with the aforementioned issues, e.g., Sullivan and Pratt, (1996), Wang et al. (2016) and Ducate and Lomicka (2008). Ideally, software must not merely assist in the writing process by offering certain guidelines concerning the genre-specific features of a particular field, but facilitate the writing task by reducing the users' amount of typing. We refer to these types of applications as software for professional writing.

In order to be useful in most circumstances, software for professional writing offers default options that allow users to write documents in English related to any topic but without providing very specific information about it. For instance, one user may wonder if technical documents have an introductory paragraph, another user may find it interesting that software includes some type of bilingual dictionary related to the specific jargon used in a particular technical document.

Another recurring problem is that any modification of the software may be a complex, and therefore expensive task; it may require a programmer to make changes in the source code or in the database schema. Modifications to include new functionalities or merely adding new entries or fields in the internal dictionary of any software for professional writing involves a technical or/and economic effort. For these reasons, it is relatively understandable that most software for professional writing tends to be a general-purpose tool that merely assists users in writing texts in English without a specific context, as it is cheap to build and may fulfil its purpose to some extent at least.

To overcome these problems, in the present study we follow an ontology-based approach. The ontology designed contains the representation of required knowledge that according to our study is suitable for writing technical documents in the English language from a Spanish native speaker's perspective. On this basis, linguistic specialists can create new writing assistants. To do that, they just have to carry out an analysis of the desired genre in order to gather the necessary linguistic information to populate (fill up) the ontology.

In order to maintain and reuse the basic structure of our ontology for each additional professional field, we present a software called O-WEAA (Ontology-Writing English Assistant Architecture). This software has two web modules:

i) *Acquisition Interface Module* allows linguistic experts to populate the ontology with new data related to any technical genre in the English language that they wish to include. These new data are added through the creation of new instances or individuals in the ontology. No technical skills are required.

ii) *Writing Assistant Interface Module* is a user-friendly interface that guides the user throughout the writing process of an English document in the specific technical English domain that is described in the ontology. As words, structures and common sentences are provided, the amount of typing is substantially reduced.

The remainder of this paper is structured as follows: Section II introduces the state of the art of software for professional writing. Section III describes the domain model of technical documents. Section IV explains components of the proposed ontology. Section V shows the implementation of the software. Finally, Section VI outlines conclusions and future directions.

OVERVIEW OF SOFTWARE FOR PROFESSIONAL WRITING

From our point of view, one of the main aims of software for professional writing is to facilitate and speed up the writing process. Currently available applications to produce texts in a particular professional field in a foreign language differ from each other in the degree of assistance offered. This degree of assistance depends on the competence of the target user in the foreign language, and on the type of text in question. Although there are several programs that aim to assist users in writing texts in English, they are of little use when the context is very precise, for instance, when it is related to some professional topic or a specific purpose.

Machine Translation Tools

Machine translation (MT) tools are one of the most widely used types of applications. Although it might seem that most of the users of machine translation tools are people that do not understand the target language, some of these users are, in fact, translators, even professionals, who hope to benefit from a machine-generated draft (García-Peñalvo et al., 2014; García and Pena, 2011). These tools are relatively easy to use, because they have a user-friendly interface that makes them visually appealing. MT tools offer good results when the linguistic context is general, and sentences are simple, short and unambiguous. However, when the communication is highly specialised, for instance in the case of a particular genre that employs very specific terminology, this type of software does not recognise the most frequent words in that field. The program employs the most common meaning and cannot deal with polysemy properly. As a consequence, the resulting translations are not always of high-quality. MT tools also fail to provide a template of the prototypical textual structure in any specific field.

Internally, it is very difficult to know exactly the behaviour of commercial MT tools. They work as a black box and they do not provide scientific specifications describing their functionality. They tend to be based on billions of words, from both monolingual texts in the target language, and aligned texts consisting of examples of human translations between the two languages involved. Statistical MT techniques (e.g., word-based models, phrase-based models or tree-based models (Koehn, 2009)) are then applied to build a proper translation model. Some well-known examples of these translation tools are Google Translate and DeepL.

Dictionaries

Dictionaries serve to assist users to find a correspondence between two words in different languages. Classical dictionaries offer equivalents and examples of use. Although there are some specific dictionaries that provide specialised terminology, none of them gives instructions about the structure of specific text types. The writing process with dictionaries is tedious and time-consuming. Some examples of general-purpose dictionaries are Collins and WordReference. On the other hand, EuroTermBank and Black's Law Dictionary are dictionaries of language for specific purposes.

Word Processing Systems

Word processing systems are not specifically designed for any translation process. However, they usually work as grammar and syntax checkers that allow users to write in any language and make corrections or suggestions on the fly or on request. Most of these tools employ text quality metrics based on collocations from corpus data (Wanner et al., 2013) or similar, in a general topic area, e.g., Microsoft Word, Grammarly Writer's Workbench and SpellCheckPlus. On the other hand, there are some advanced tools, like StyleWriter4, which take into account contextual information such as type of document, e.g. financial or scientific, as well as patterns defined by the user. Nonetheless, the users must write all the text by themselves. The software only works as a kind of proofreading device and does not provide information about the rhetorical structure of that particular text type. Like MT tools, most word processors are built for commercial purposes, and no descriptions of their internal structure are publicly available.

Writing Assistance Systems

Writing assistance systems are the more useful software when users have to write a specialised document in a very specific field and do not have a high command of the target language and/or do not know the specific guidelines in languages other than their own. Writing assistance systems are usually built as a result of collaboration between an interdisciplinary group of linguists and computer science engineers. The former decide the features that the application has to include, and the latter decide on the software necessary to meet these requirements. After a long process involving changes and adaptations to the initial design, a prototype is launched. These tools feature the highest degree of automation, together with relevant contextual information and they include rhetorical information on the structure of a specific text type. Consequently, these applications might not require a high level of competence in the target language. Writing assistance systems in a foreign language are directed towards semiautomatic text generators, e.g., Aluisio and Oliveira (1996), Oliveira et al. (2001), Aluísio et al. (2011) and Chang and Chang (2004). One of the best-known programs to assist users in writing with a specific goal is SWAN (Kinnunen et al. 2012). This tool helps scholars to identify and correct potential writing problems in their scientific papers, and is intended to aid writers with the content of a document, not only with grammar or spelling. SWAN is powered by Stanford CoreNLP (Manning et al., 2014), a suit of core NLP (Natural Language Processing) tools. The main drawback of SWAN is that it is aimed exclusively at scientific writing in English; therefore, it is not very flexible in the development of new applications. It is necessary to start a new design process to achieve that. In Narita et al. (2015) a tool designed to organise the writing structure of specialised texts is described. This tool is based on rhetorical templates extracted from a parallel corpus. Lastly, Omar et al. (2009) have developed a tool for Malaysian speakers to check their English grammar.

As can be seen in the previous classification, none of the tools provides an environment that allows users to write their own genre-specific English text with a proper structure and vocabulary. All of the categories analysed present several problems regarding the degree of automatism, the necessary knowledge about the topic, and the competence in the foreign language. Software belonging to general dictionaries and MT tools are not really assistants, but mere browsers, and require a high command of the language and of the specialised terminology because they are not designed for any specific purposes. Word processing software improves the level of automation in relation to MT tools and dictionaries, but they remain insufficient and also require a higher degree of competence in the foreign language. Finally, writing assistance systems also have an excellent degree of automation, require less knowledge of the foreign language, and may include genre-specific terminology and textual conventions. However, these systems have usually been developed to assist users in writing only one particular type of technical document and hence this software is only valid for the specific communicative goal it has been

built for, e.g., scientific papers, meeting minutes, etc. No changes can be performed in order to adapt the existing software to a different writing task, and consequently these applications have a short software life-cycle. For example, if somebody wants to add a new feature to improve the application, a change of source code would be needed.

O-WEAA allows us to define main features of any type of technical documents by means of the Acquisition Interface Module. After that, this knowledge is employed as a database in a writing assistance system software called Writing Assistant Interface Module. By using this knowledge, the module is able to:

i) Provide relevant sections of the document, also known as rhetorical structure.
ii) Offer prototypical examples, that we named model lines.
iii) Give access to a semi-specialised lexicon of the document, also named glossary.

As we will show in detail in this document, our software overcomes the main disadvantages of most writing assistance systems, namely a short life-cycle and the fact that they are generally restricted to only one text type. It also reduces the required amount of typing.

The software presented here has been tested and several writing assistant versions have been developed in different domains such as scientific abstracts or technical brochures.[1]

DOMAIN KNOWLEDGE OF TECHNICAL DOCUMENTS

In our proposal, this process has a strong linguistic influence and is based on well-known approaches and theories accepted by the majority of the specialists in corpus linguistic research. For an in-depth, comprehensive analysis of the linguistic methods employed, see Labrador et al. (2014), where a purely linguistic analysis in the subgenre of online advertisements is described. Therefore, this knowledge extraction provides a significant scientific added-value.

Domain modelling process consists of three main steps:

i) The identification of professional fields where NNS are required to write in English.
ii) The compilation of a specialised comparable corpus in that particular professional field in the native language of the speaker, in this case Spanish, and in English.
iii) The implementation of a corpus-based analysis that includes rhetorical analysis.

Due to this strong linguistic influence, these steps must be carried out by linguistic experts with extensive experience of and familiarity with this type of analysis. Experts extract linguistic knowledge based on quantitative and qualitative statistics from annotated corpora. Three documents must be collected:

i) The prototypical rhetorical structure of each text type based on the most frequent sections found in the corpus. These sections are also known as moves and steps. For example, the rhetorical structure obtained for technical brochures, which was used to tag the corresponding corpus can be found in Ramón and Labrador (2015).

ii) The list of model lines based on the most common phraseological constructions in each move and step. Users tend to employ a narrow range of constructions which seem to have become conventionalised for the expression of a particular function. These model lines are composed of plain text together with several gaps.

Three types of gaps are considered: 1) obligatory gaps, where a lexical unit in English has to be provided to replace the Spanish indication, 2) optional gaps, where one option in English is already available, but not required obligatorily, and 3) selectable gaps, which indicate that one option in English out of two given alternatives must necessarily be chosen to complete the model line. **Table 1** shows an example of each type of gap.

Table 1. Types of gap to be filled by the user of the writing aid application

Type of gap	Meaning	Example
Obligatory	It has to be filled with a specific word in English	The use of (ELEMENTO) is recommended.
Optional	Word or sentence is not obligatory	The fan must be clear to work [CORRECTLY]
Selectable	One choice has to be selected	The fan {SHOULD/MUST} be disconnected

iii) The list of the most frequent technical and semi-technical lexical items extracted from the corpus, i.e., the glossary. In order to create a relevant glossary, the linguists select a list of specialised terms in Spanish that occur frequently in the corpus and whose equivalents in English may be required in the writing procedure.

THE PROPOSED ONTOLOGY

Based on this information, an ontology of eight main concepts or classes (Sanjurjo et al. 2019) was developed using Protégé software (Musen, 2015):

1. *Context:* That is the context related to a specific domain of writing (KnowledgeDomain).

---

[1] Visit http://actres.unileon.es/ for more information

2. *KnowledgeDomain:* The specific domain of writing, usually related to a professional field. It is characterised by a tag naming the specific domain, for example, "Technical brochures".

3. *RhetoricalStructure:* A class reflecting rhetorical information on the structure of one specific text type. It has a number of Sections similar to building blocks.

4. *Section:* Formed by a number of elements of the type ModelLine, each section also has a title consisting of a string.

5. *ModelLine:* A common phraseological construction that includes some elements of the type FillingGap.

6. *FillingGap:* Represents placeholders in the phraseological constructions (ModelLine), that is, places where a given set of terms can appear when writing common *expressions*. As described previously, there are three types of FillingGap:

   i) *ObligatoryFillingGap:* this has to be filled with a specific word in English - which is modelled in the ontology.

   ii) *SelectableFillingGap:* one choice has to be selected from a list of terms.

   iii) *OptionalFillingGap:* a word or sentence which is not obligatory.

7. *Term:* Found in a glossary, it is composed of a text in Spanish, a corresponding text in English and a usage example text for the term. Terms are associated to instances of the class Glossary and are related to ObligatoryFillingGap as the words that will be shown to the user in order to fill the gap in the given ModelLine.

8. *Glossary:* A grouping of different terms deemed by the expert to occur frequently in the corpus for the domain being modelled.

Figure 1 shows the OWL class hierarchy in the Protégé software interface.
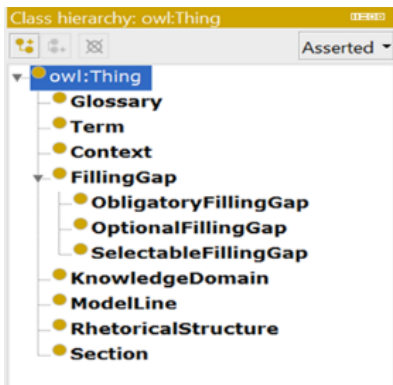

Figure 1. A screenshot of the OWL class hierarchy in Protégé

IMPLEMENTATION

The proposed ontology is not yet useful for our purpose, i.e., that users without programming skills should be able to create their own writing assistant software. To this end, we developed O-WEEA. As mentioned before, O-WEEA consists of two modules: Acquisition Interface Module and Writing Assistant Interface Module. **Figure 2** shows the architecture of the software.
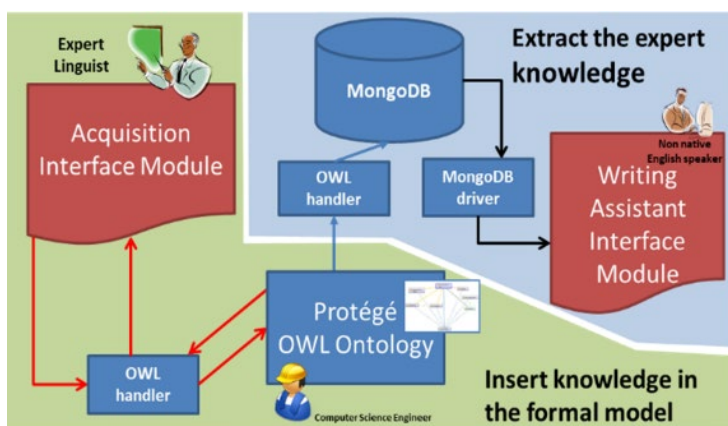

**Figure 2.** O-WEAA: Ontology-Writing English Assistant Architecture general schema

Acquisition Interface Module

In order to be useful, the ontology must be populated with particular data of a specific domain, for example, with the terms and expressions used for writing technical brochures. To achieve that, the Acquisition Interface Module allows users to populate and/or modify the expert knowledge contained in the ontology without needing technical skills in computer science.

The module works in two different directions as shown in **Figure 2** (on the green background):

i) It exposes the structure of concepts in the ontology to the linguistic expert in a transparent way, showing text fields, drop-down lists, clickable tags, etc., so that the expert can populate the conceptual model with the specific terms and expressions.

ii) It inserts these terms and expressions as instances in the ontology.

A screenshot of the Acquisition Interface Module can be seen in **Figure 3**. In this case, the linguistic expert adds a new instance of the concept ModelLine to the ontology.
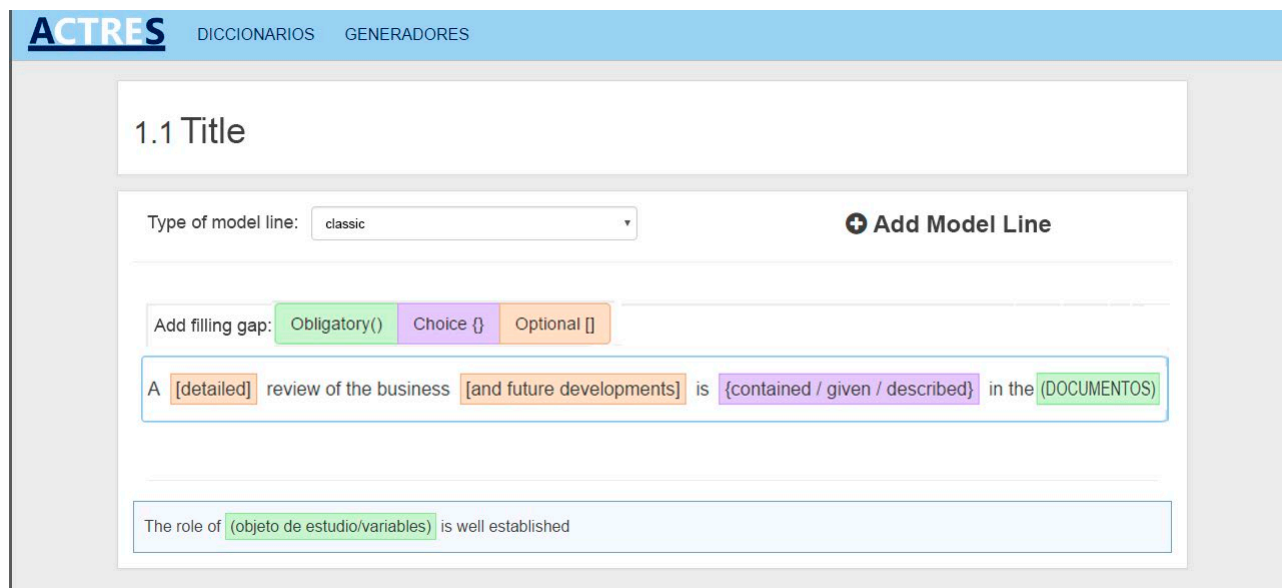
Figure 3. Acquisition Interface Module

Writing Assistant Interface Module

Previously inserted knowledge is used to assist non-native English users in writing documents in English. This is achieved by a web application named Writing Assistant Interface Module. This module exposes the terms and expressions that were stored as instances in the ontology by the linguists and inserted into a database.

O-WEAA's Writing Assistant Interface Module offers the different moves/steps (sections and subsections) which correspond to the rhetorical structure of the specialised text. First, the user chooses one of them by clicking on the list on the left-hand side of the screen (**Figure 4**) or moving through the arrows of the interface.
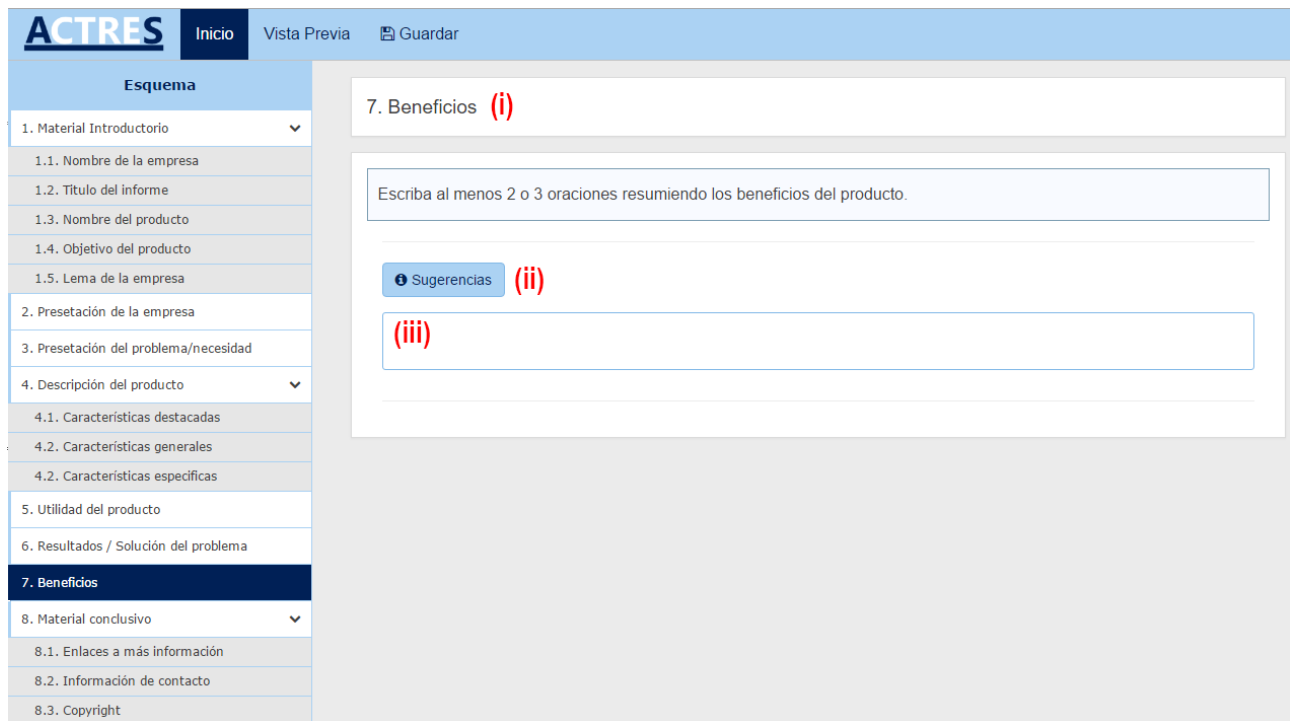
Figure 4. Drafting area of O-WEAA's Writing Assistant Interface Module

Each move/step has interactive areas that the user must complete. Most of them are simple drafting fields, but there can also be multimedia sections, e.g., to upload images. Every drafting field has the following features, as shown in **Figure 4**: i) section identification; ii) a selector that shows the model lines available; and iii) a drafting area where the user completes their selected model line assisted by the semi-specialised lexicon or glossary.

When the user clicks on "Sugerencias", a pop-up menu appears with the list of model lines associated with that particular move/step and their corresponding examples (**Figure 5**).
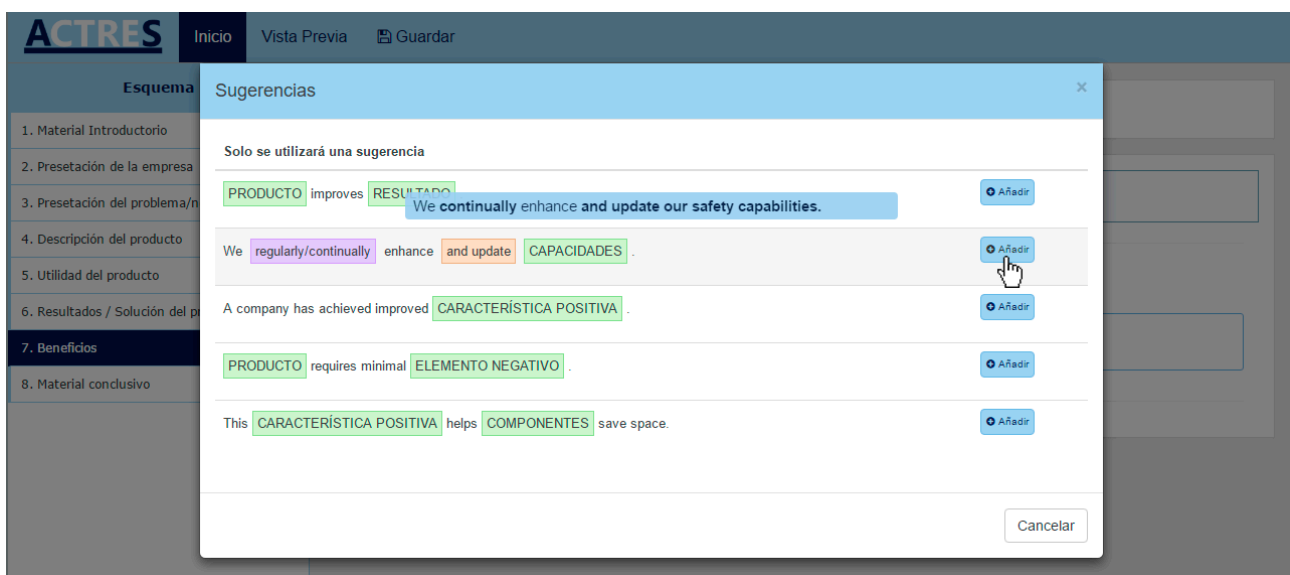

Figure 5. Suggestions of O-WEAA's Writing Assistant Interface Module

When a specific model line is selected by clicking on it, the model line appears in the drafting area below to be completed and edited by the user as required, as shown in **Figure 6**.
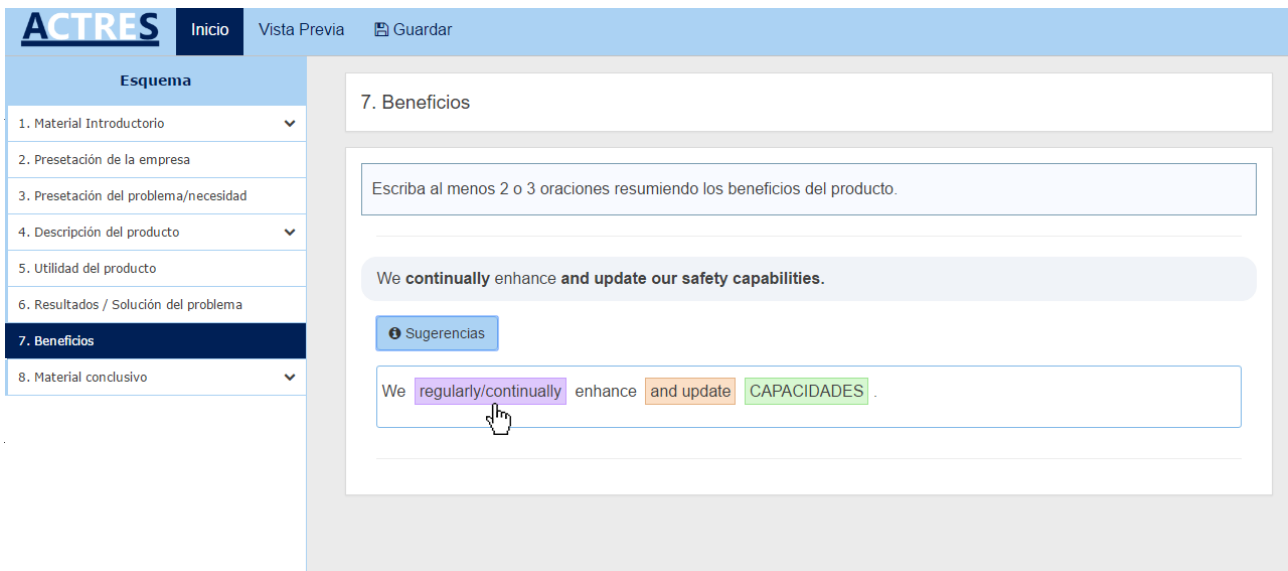
Figure 6. Completing and editing the model line selected with O-WEAA's Writing Assistant Interface Module

As was explained in the Section III, three types of gaps are considered: i) obligatory, in green ii) optional, in orange, and iii) selectable, in purple. User must edit the model line by clicking on the gaps.

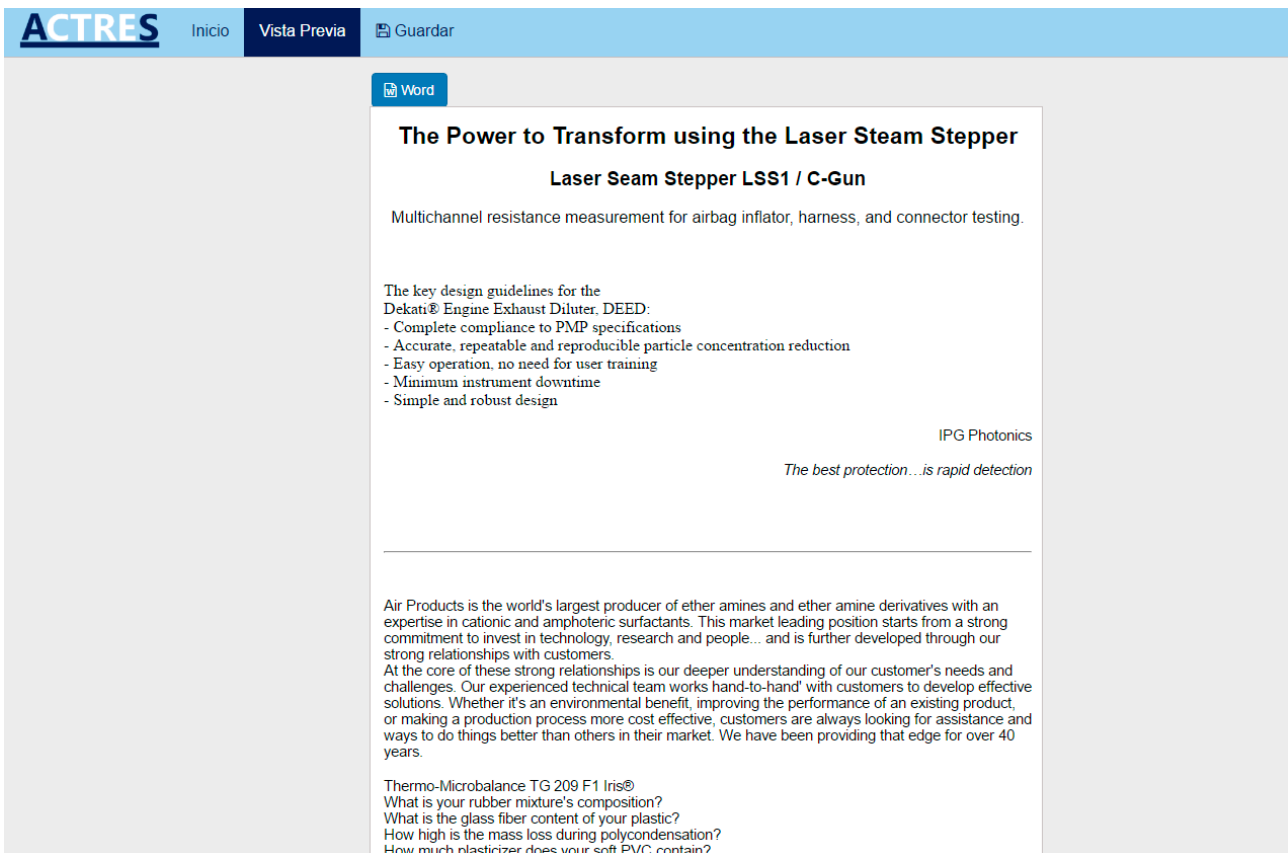An example of a final document is shown in **Figure 7**.

Figure 7. Final document produced with the O-WEAA's Writing Assistant Interface Module

A demo can be found at https://actres.unileon.es/demos/generadores2.

CONCLUSIONS AND FUTURE DIRECTIONS

Writing in a foreign language is a particularly difficult skill to acquire, especially in professional contexts involving language for specific purposes. Developing software tools to help users to write these types of texts is one of the possible solutions. However, most of the analysed tools have proved to be useful only in a general context, or in the case of some writing assistance systems, only in one type of genre. Any modification of software implies a technical and/or economic effort, as they often work like a black box, typical in commercial software environments.

In order to overcome these inconveniences, this paper has presented an innovative tool called O-WEAA. O-WEAA has an ontology-based approach, where the ontology has been designed using a linguistic method relying on the premise that any specific domain may be characterised by a set of rhetorical, phraseological and terminological conventions (Labrador et al., 2014). As a consequence, our ontology is defined by the most common sections included in the text, a set of prototypical sentences and a Spanish-English semi-specialised glossary. The ontology designed can be reused for developing new software for professional writing in as many fields as required. All that is required is that linguists extract linguistic knowledge following the steps described in this paper. After that, any user without technical programing skills can populate and update the ontology by means of the Acquisition Interface Module developed. Furthermore, non-native English users can write their own English text using stored knowledge of ontology via Writing Assistant Interface Module.

The main contribution of this application is the use of linguistic analyses that generate knowledge instances that are later formalised into an ontology. This ontology can be populated with knowledge related to any specific genre o subgenre by any person, even if that person does not have programming skills. After that, the populated ontology can be used in our user-friendly writing assistant software module.

Future research should examine the use of O-WEAA in language-pairs other than English-Spanish. This would require some linguistic considerations to be addressed, for instance, extreme differences between grammatical structures of the languages. The possibility of automate the extraction of model lines, that is, more frequent/useful sentences, as well as automatic recognition of gap filling warrants further investigation. Finally, improving the acquisition model through the implementation of queries, functionalities or statistics may be an excellent way to obtain the most accurate knowledge structures.

ACKNOWLEDGEMENT

REFERENCES

Aluisio, S. M. and Oliveira Jr, O. N. (1996). A detailed schematic structure of research paper introduction: An application in support-writing tools. *Procesamiento del Lenguaje. Natural*, *19*, 141-147.

Aluísio, S. M., Barcelos, I., Sampaio, J. and Oliveira Jr., O. N. (2001). How to learn the many unwritten 'rules of the game' of the academic discourse: a hybrid approach based on critiques and cases to support scientific writing. *Proceedings IEEE International Conference on Advanced Learning Technologies*, 257-260. https://doi.org/10.1109/ICALT.2001.943916

Chang, J. and Chang, Y. (2004). Computer assisted language learning based on corpora and natural language processing: the experience of project CANDLE. *An Interactive Workshop on Language e-Learning*, 15-23.

Ducate L. C. and Lomicka, L. L. (2008). Adventures in the blogosphere: from blog readers to blog writers. *Computer Assisted Language Learning*, *21*(1), 9-28. https://doi.org/10.1080/09588220701865474

Garcia, I. and Pena, M. I. (2011). Machine translation-assisted language learning: writing for beginners. *Computer Assisted Language Learning*, *24*(5), 471-487. https://doi.org/10.1080/09588221.2011.582687

García-Peñalvo, F. J., Johnson, M., Alves, G. R., Minović, M. and Conde-González, M. A. (2014). Informal learning recognition through a cloud ecosystem. *Future Generation Computer Systems*, *32*(1), 282-294. https://doi.org/10.1016/j.future.2013.08.004

Kinnunen, T., Leisma, H., Machunik, M., Kakkonen, T. and Lebrun, J. (2012). SWAN-scientific writing AssistaNt: a tool for helping scholars to write reader-friendly manuscripts. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 20-24. https://doi.org/10.13140/2.1.3284.4808

Koehn. P. (2009). *Statistical Machine Translation.* Cambridge University Press: Cambridge. https://doi.org/10.1017/CBO9780511815829

Labrador, B., Ramón, N., Alaiz-Moretón, H. and Sanjurjo-González, H. (2014). Rhetorical structure and persuasive language in the subgenre of online advertisements. *English for Specific Purposes*, *34*, 38–47. https://doi.org/10.1016/j.esp.2013.10.002

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55-60. https://doi.org/10.3115/v1/P14-5010

Musen, M. A. (2015). The Protégé project: A look back and a look forward. *AI Matters*, *1*(4), 4-12. https://doi.org/10.1145/2757001.2757003

Narita, M., Kurokawa, K. and Utsuro, T. (2002). A Web-based English Abstract Writing Tool Using a Tagged EJ Parallel Corpus. *Proceedings of 3rd International Conference on Language Resources and Evaluation*, 2115-2119.

Oliveira Jr, O. N., Zucolotto, V. and Aluísio, S. M. (2006). *Developing strategies to produce better scientific papers: a Recipe for non-native users of English.* ArXiv preprint cs/0611013

Omar, N., Razali, N. A. M. and Darus, S. (2009). Automated grammar checking of tenses for ESL writing. *International Conference on Rough Sets and Knowledge Technology*, 475-482. https://doi.org/10.1007/978-3-642-02962-2_60

Ramón, N. and Labrador, B. (2015). The rhetorical structure of technical brochures: A proposal for technical writing. *Procedia-Social and Behavioral Sciences*, *173*, 241-245. https://doi.org/10.1016/j.sbspro.2015.02.059

Sanjurjo-González, H., Alaiz-Moretón, H. Ramón, N., Labrador, B. and García, I. (2019). Developing an open writing aid for non-native speakers of English: an ontology-based approach. *14th Iberian Conference on Information Systems and Technologies (CISTI),* Coimbra. https://doi.org/10.23919/CISTI.2019.8760895

Sullivan N. and Pratt, E. (1996). A comparative study of two ESL writing environments: A computer-assisted classroom and a traditional oral classroom. *System*, *24*(4), 491-501. https://doi.org/10.1016/S0346-251X(96)00044-9

Wang, W., Chen, Y., Zhang, Q. and Jiang, T. (2016). A Software-Defined Wireless Networking Enabled Spectrum Management Architecture. *IEEE Communications Magazine*, *54*(1), 33-39. https://doi.org/10.1109/MCOM.2016.7378423

Wanner, L., Verlinde. S. and Ramos, M. A. (2013). Writing assistants and automatic lexical error correction: word combinatorics, in I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (eds.), *Electronic lexicography in the 21st century: thinking outside the paper: Proceedings of the eLex 2013 conference.* Tallinn: Institute for Applied Slovene Studies/Eesti Keele Instituut, 472–487.