EcoLexiCAT: a Terminology-enhanced Translation Tool for Texts on the Environment

Pilar León-Araúz, Arianne Reimerink, Pamela Faber

Department of Translation and Interpreting, University of Granada, C/ Buensuceso, 11, 18002 Granada, Spain

E-mail: pleon@ugr.es, arianne@ugr.es, pfaber@ugr.es

Abstract

Although machine translation and computer assisted translation (CAT) are now a reality in the workflow of professional translators, terminology management is still considered complex and time-consuming and is often not seamlessly integrated into the translation process. Most terminographic resources are not designed to take into account the real search behavior of end users such as translators (Tudhope et al., 2006), and in many cases CAT tools do not provide terminological modules that go beyond a simple glossary with interlinguistic equivalents. Furthermore, corpus consultation is rarely possible in most CAT tools, despite the fact that the phraseological information extracted from a corpus is of great help for translators. To address these issues, we created a web-based tool for the terminology-enhanced translation of specialized environmental texts for the language combination English-Spanish-English. EcoLexiCAT uses the open source version of the web-based CAT tool MateCat and enriches a source text with information from: (i) EcoLexicon, a multimodal and multilingual terminological knowledge base on the environment (Faber et al., 2014; Faber et al., 2016); (ii) BabelNet, an automatically constructed multilingual encyclopedic dictionary and semantic network (Navigli & Ponzetto, 2012); (iii) and Sketch Engine, the well-known corpus query system (Kilgarriff et al., 2004).

Keywords: computer assisted translation; terminology management; specialized translation

1. Introduction

In today's world, machine translation (MT) and computer-assisted translation (CAT) are a consolidated part of the professional translation workflow. Nevertheless, terminology management is still considered complex and time-consuming and is often not seamlessly integrated into the translation process. Furthermore, most terminological tools do not take into account the real search behavior of end users such as translators (Tudhope et al., 2006; Durán Muñoz, 2012: 78) and most terminological modules in CAT tools do not go beyond a simple list of equivalences. Apart from that, access to corpora is generally not provided in most CAT tools, despite the valuable phraseological information that a corpus can provide. An exception to this is the recently added Sketch Engine plug-in (available from the SDL AppStore) in SDL Trados Studio but, generally speaking, loss of translation quality and precious time are the inevitable consequences.

An excellent example of how to improve on the current situation is the initiative carried out by the TaaS project[1]. TaaS (Terminology as a Service) is a European project developed by a group of institutions and companies in the translation technology field who conceive 21st century terminology in a user-friendly, collaborative, cloud-based environment (Gornostay, 2014). Their aim is to create a platform for instant access to the most up-to-date terms and for user participation in the acquisition, sharing and reuse of multilingual terminological data. TaaS targets all types of language professionals, but specifically focuses on translators as end users, as it provides the following terminology services: (1) automatic extraction of term candidates; (2) automatic recognition of translation equivalents in different public and industry terminology databases; (3) automatic acquisition of translation equivalents for terms not found in term banks from parallel/comparable web data using the state-of-the-art terminology extraction and alignment methods; (4) facilities for terminology sharing and reusing within CAT tools; and (5) improvement of statistical machine translation systems through terminological data integration.

As an improvement, we developed EcoLexiCAT, a terminology-enhanced CAT tool that provides easy access to domain-specific terminological knowledge in context. This application integrates different features of the professional translation workflow in a stand-alone interface where a source text is interactively enriched with terminological information (i.e. definitions, translations, images, compound terms, corpus access, etc.) from different external resources: (i) EcoLexicon, a multimodal and multilingual terminological knowledge base on the environment (Faber et al., 2014; Faber et al., 2016); (ii) BabelNet, an automatically constructed multilingual encyclopedic dictionary and semantic network (Navigli & Ponzetto, 2012); (iii) and Sketch Engine, the well-known corpus query system (Kilgarriff et al., 2004).

The remainder of this paper is organized as follows. Section 2 explains terminology management from the perspective of the needs and expectations of professional translators. Section 3 concisely describes the web-based open source CAT Tool MateCat on which EcoLexiCAT is based as well as the external resources used for terminology enhancement. Section 4 provides a detailed explanation of EcoLexiCAT and its different modules. Finally, Section 5 presents the conclusions that can be derived from this study and outlines ideas for future research.

2. Translators' needs and expectations for terminology management

Any lexicographic or terminographic tool should take into account the needs of end users, in its structure and content as well as the way that the information is represented so that users can search and interact with the tool (Tarp, 2013). When

[1] http://taas-project.eu/

translators query a resource and do not find the information needed, they lose time and their productivity decreases (*search costs*; Nielsen, 2008). Similarly, when translators obtain too much data (*infoxication*; Cornellà, 1999), which lengthens the knowledge construction time, their *comprehension costs* (Nielsen, 2008) increase. In addition, translators do research in all phases of the translation process. This occurs during the pre-translation phase in order to understand the original text and its terminology. Research is also performed when the original message is encoded in the target text, with a view to fulfilling pragmatic requirements and searching for equivalents. Finally, in the revision phase, translators must check terminology and generally ensure the quality of their translation (Durán Muñoz, 2012: 80). Accordingly, one of the major challenges of lexicographic and terminographic resources for translators is to find the right balance between search costs and comprehension costs.

Durán Muñoz (2010, 2012) affirms that translators prefer to solve their terminological problems by consulting ready-made resources. According to her study, the most frequent resources used are (in this order): bilingual specialised dictionaries or glossaries, searches in search engines, terminological databases, monolingual specialised dictionaries, and Wikipedia (Durán Muñoz, 2012: 81). She mentions that translators do not trust the quality of multilingual resources and that searches in parallel corpora are not high on the list of preferences. However, when asked to classify the most frequent ISO fields (ISO 12620:1999) in the microstructure of terminological resources, translators considered the following to be most essential: clear and concrete definitions, equivalents, derivatives and compounds, domain specification, examples, phraseological information, definition in both languages for bilingual resources, and abbreviations and acronyms (Durán Muñoz, 2012: 82). Finally, when asked for their opinion, translators said that terminological resources should be able to do the following: (i) permit exportability and/or importability in different formats; (ii) include more pragmatic information about usage and tricky translations (old usage, false friends, specific usage in a domain or region, etc.); (iii) offer links to other resources to improve or increase the results; (iv) improve search options; and (v) provide examples taken from real texts (idem). Quite surprisingly, although the translators in this study did not show much interest in having access to corpora, they did highlight the need for more phraseological information, pragmatic information and examples taken from real texts. Even though this information can be extracted from corpora, translators were probably reticent to use them because it can take a long time if the right query methods are not provided.

Translation-oriented terminology management, or terminology-enhanced translation, should take into account all of the above. As shown in Section 4, EcoLexiCAT is a tool that includes the essential fields mentioned, links to other resources and improved search options for corpus analysis that provide the necessary pragmatic information and real text examples. All of this is available in a single-platform web-based CAT environment that has the capabilities of importing and exporting different file types and formats.

3. EcoLexiCAT sources

3.1 MateCat

MateCat, acronym of Machine Translation Enhanced Computer Assisted Translation, was originally a three-year research project led by a consortium composed of the international research center FBK (Trento, Italy), Translated SRL, the Université du Maine and the University of Edinburgh. The objective was to improve the integration of MT and human translation (Federico et al., 2014: 129). Within the project a computer-aided translation tool was developed, The MateCat Tool. This application is not only an industrial tool but also an open source platform[2]. It offers all the features of a modern CAT tool, such as a text editor that divides the text to be translated in source and target segments and saves them along with their translation in a translation memory (TM).

MateCat runs as a web server and communicates with other services through open APIs. It allows communication with pre-existing TMs, terminological databases, concordance searches within the TMs and MT engines, from which the MT provider MyMemory (a combination of Google Translate and Microsoft Translator) is freely available. The tool has been tested in professional settings and adapted for research in MT (e.g. Bertoldi et al., 2013 *apud* Federico et al., 2014: 131) and for educational purposes. The fact that it has an open-source version as well as a high level of flexibility made it a suitable option for the development of EcoLexiCAT. In addition, the features and operation of MateCat are basically the same as those found in most CAT tools used nowadays. Therefore, professional translators will not need to invest much time in learning how to use the tool and will benefit from the interoperability of CAT-related formats (TBX for glossaries, XLIFF for bilingual files, TMX for TMs, etc.). This enables them to use the resources generated during the translation process in other similar tools and reuse pre-existing resources (i.e. glossaries, bilingual files and TMs) in EcoLexiCAT.

3.2 EcoLexicon

EcoLexicon[3] is a multilingual and multimodal terminological knowledge base on environmental science (Faber et al., 2014; 2016). It is the practical application of Frame-based Terminology (Faber et al., 2011; Faber, 2012, 2015), a theory of specialized knowledge representation that uses certain aspects of Frame Semantics (Fillmore, 1982; Fillmore & Atkins, 1992) to structure specialized domains and create non-language-specific representations. Frame-based Terminology focuses on

[2] https://www.matecat.com/open-source/

[3] ecolexicon.ugr.es

conceptual organization, the multidimensional nature of specialized knowledge units, and the extraction of semantic and syntactic information through the use of multilingual corpora.

EcoLexicon is an internally coherent information system, which is organized according to conceptual and linguistic premises at the macro- as well as the micro-structural level. It currently has 3,601 concepts and 20,211 terms in Spanish, English, German, French, Modern Greek, and Russian. This terminological resource was conceived for language and domain experts as well as for the general public. It targets users such as translators, technical writers, and environmental experts who need to understand specialized environmental concepts with a view to writing and/or translating specialized and semi-specialized texts.

End users interact with EcoLexicon through a visual interface with different modules that provide conceptual, linguistic, and graphical information. Instead of viewing all information simultaneously, they can browse through the windows and select the data that is most relevant for their needs. Figure 1 shows the entry in EcoLexicon for the word FAN. When users open the application, three zones appear. The top horizontal bar gives users access to the term/concept search engine. The vertical bar on the left of the screen provides information regarding the search concept, namely its definition, term designations, associated resources, general conceptual role, and phraseology.

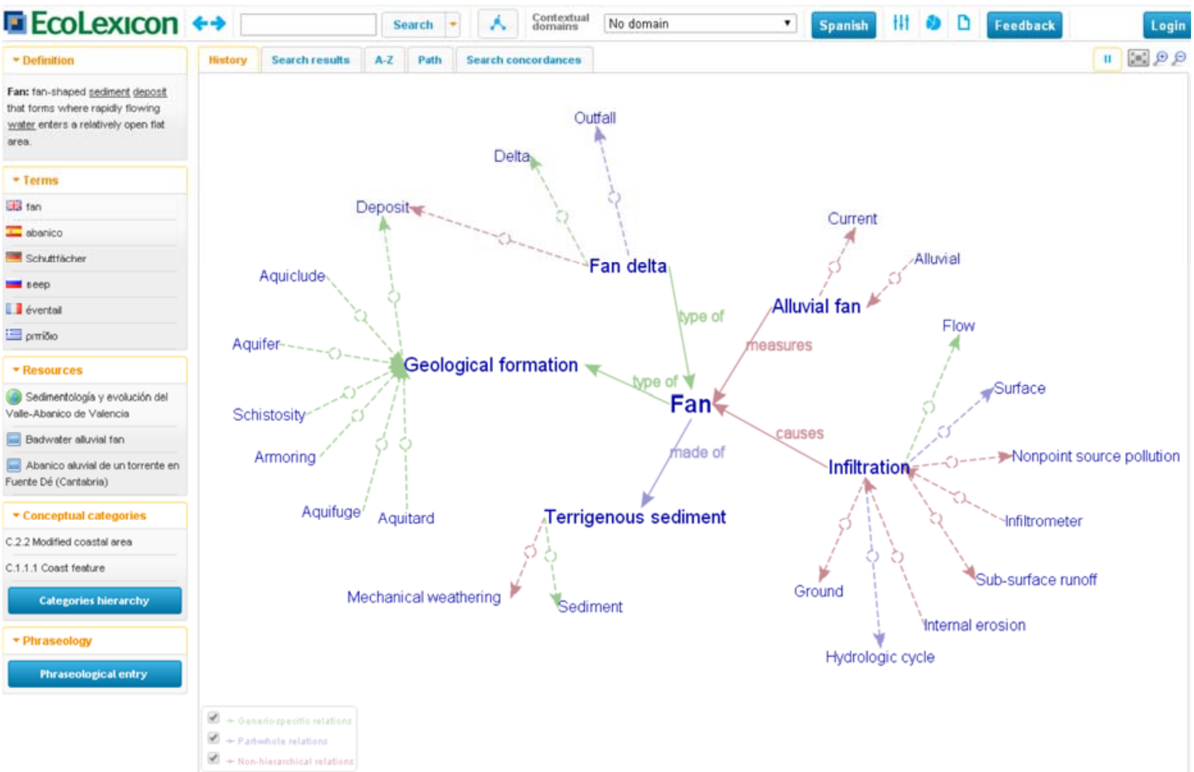

Figure 1: EcoLexicon user interface

Each definition makes category membership explicit, reflects a concept's relations with other concepts, and specifies essential attributes and features (León-Araúz, Faber & Montero-Martínez, 2012: 153-154). Accordingly, the definition is the linguistic codification of the relational structure shown in the concept map, at the center of the screen. Although users can configure the map to their needs, the standard representation mode (see Figure 1) shows a multi-level semantic network whose concepts are all linked in some way to the search concept, which is at its center.

A specialized corpus was specifically compiled for EcoLexicon in order to extract linguistic and conceptual knowledge. Currently, the corpus has over 50 million words and each of its texts has been tagged according to a set of XML-based metadata, which contain information about the language of the text, the author, date of publication, target reader, contextual domain, keywords, etc. This was done in order to provide users with a direct and flexible way of accessing the corpus. It also allows them to constrain corpus queries based on pragmatic factors, such as contextual domains or target reader. In this way, users can compare the use of the same term in different contexts. The corpus was first made available in the Search concordances tab (center area menu just above the concept map in Figure 1). However, currently, the English EcoLexicon Corpus (23 million words) is also hosted and freely available in Sketch Engine Open Corpora[4].

To fully exploit the contents and components of EcoLexicon for purposes of translation, we developed EcoLexiCAT. A terminological knowledge base (TKB) such as EcoLexicon provides a great amount of interconnected information in many different formats. However, in the professional translation workflow, especially when the source text has a high term density, searching in EcoLexicon, together with other resources, might cause high search and comprehension costs (see Section 1). EcoLexiCAT provides all this knowledge as an integral part of the translation workflow, where it is presented according to a specific context and during a specific phase of the translation process (see Section 4).

3.3 BabelNet and Babelfy

The multilingual encyclopedic dictionary and semantic network BabelNet[5] was created by linking Wikipedia to WordNet (Navigli & Ponzetto, 2012: 218). It connects concepts and named entities in a network of semantic relations, made up of about 14 million entries, called Babel synsets. Each Babel synset represents a given meaning and contains all the synonyms expressing that meaning in a range of different languages. Wikipedia and WordNet are integrated through automatic mapping and by filling in lexical gaps in resource-poor languages with MT.

[4] https://the.sketchengine.co.uk/open/

[5] babelnet.org

BabelNet is an enormous information resource that can be accessed through an open API, and was considered to be a valuable addition to EcoLexiCAT in those cases where EcoLexicon, a manually-built resource, did not include sufficient information regarding general language issues or for texts that combine environmental issues with other domains of expertise. Furthermore, the BabelNet researchers created their own algorithm, called Babelfy[6] for the disambiguation of polysemic words when found in the context of a particular text (Moro, Raganato & Navigli, 2014; Moro, Cecconi & Navigli, 2014).

Babelfy is a unified multilingual, graph-based approach to entity linking (the disambiguation of named entities) and word-sense disambiguation (the disambiguation of common nouns, verbs and adjectives). When presented with an input segment, the system extracts all the linkable fragments and lists the possible meanings of each of them according to the semantic networks of BabelNet. Evidently, this is of great help when dealing with polysemic terms. In EcoLexiCAT, the source text is disambiguated through Babelfy before matching the terms with BabelNet.

3.4 Sketch Engine

Sketch Engine (Kilgarrif et al., 2004) is an online corpus query system with a very efficient search engine and a statistical component for enhanced precision. It contains over 300 corpora in over 60 languages and allows end users to create their own corpora as well. One very interesting module is information extraction through word sketches. Word sketches are summaries of collocational information of a search term, where the term is analyzed according to the verbs, modifiers and other usual constructions that accompany it in real texts. Word sketches are created through sketch grammars that launch specific queries to a corpus. End users can create their own grammars for word sketches and therefore adapt the tool to their specific needs.

With an account, users have access to pre-loaded corpora and a corpus compiler called WebBootCaT. They can download corpora, add new documents to a corpus, extract domain keywords, view texts, and generate concordances, wordlists, frequency lists, collocations, and word sketches. Sketch Engine also hosts a set of freely available open corpora that can be queried with full Sketch Engine functionalities. This option, where the EcoLexicon corpus can be uploaded and afterwards freely accessed, made it a perfect option as a corpus query system for EcoLexiCAT.

4. EcoLexiCAT: a terminology-enhanced translation tool

When users start a new project in EcoLexiCAT, they first access the project settings interface in Figure 2, where they can do the following: (1) name the project; (2) choose

[6] babelfy.org

directionality (so far, English–Spanish or Spanish–English); (3) select a particular domain within the environment—these are in accordance with the domains according to which EcoLexicon is organized and are included in this first step as a way to classify projects and TMs for later reuse; (4) choose between general and patent segmentation rules, for the source text to be segmented accordingly; (5) optionally add an MT provider for post-editing—MyMemory is freely available, but others (e.g. Moses, DeepLingo, IP Translator) can also be added if users have an account with them; (6) optionally add users' own TMs and/or glossaries—otherwise a collective TM stored in the system will be used; and (7) upload the source text. These steps, except for (3), are default options in MateCat.

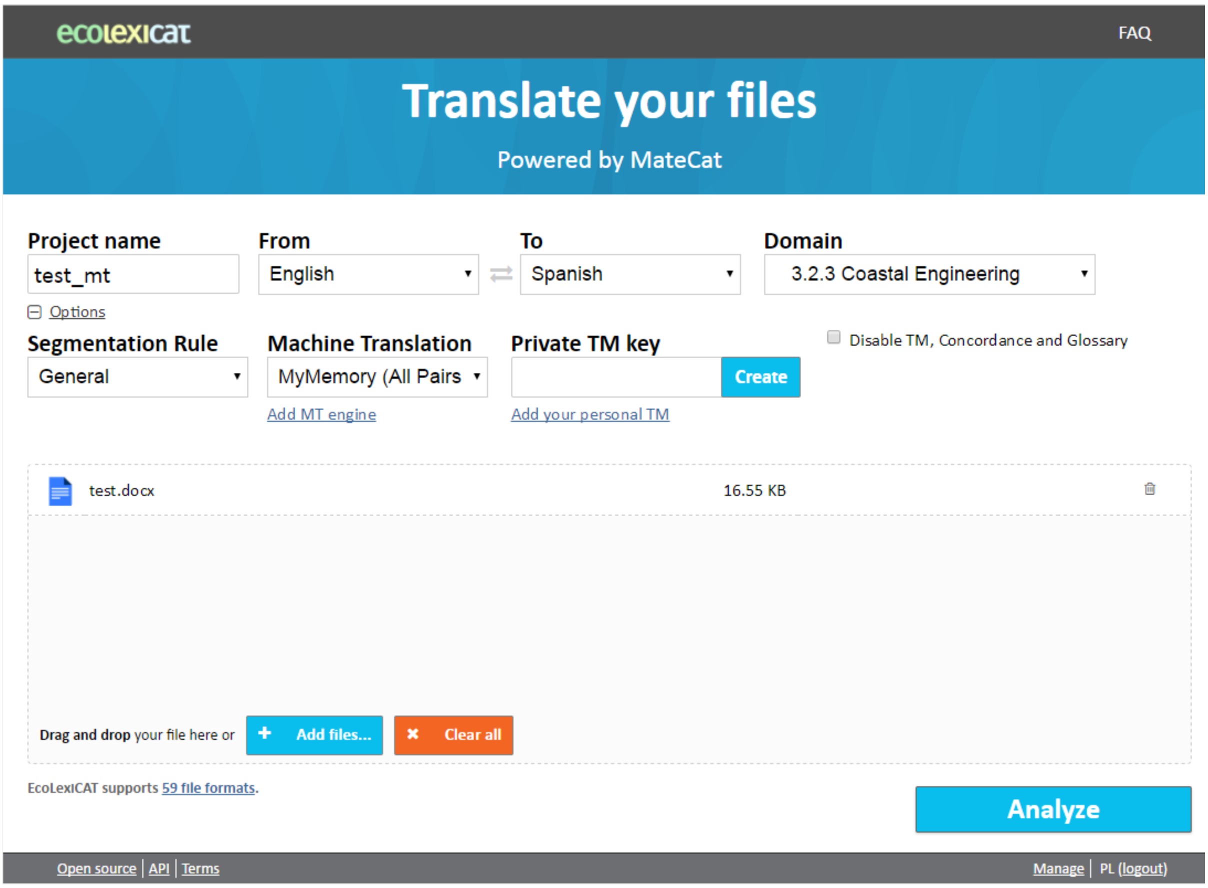

Figure 2: Project settings in EcoLexiCAT

Once the source text is processed and converted into a bilingual format (XLIFF), users can access the main interface (Figure 3), which is divided into two main sections. The left-hand section is where the three external resources (i.e. EcoLexicon, BabelNet/Babelfy and Sketch Engine) provide the terminological enhancement of the translation process. The right-hand section is where the target text is produced, an editor where the source text appears split into different segments. In the right upper part of the editor, users may download the target or the source text in their original format, and export the bilingual file in SDLXLIFF (SDL Trados Studio's native format) or the whole project in OmegaT's native format, another desktop open source CAT tool. This, together with the possibility of downloading the TM and the glossary created during the project, ensures the interoperability of different formats across different CAT tools, an issue that professional translators must often deal with.

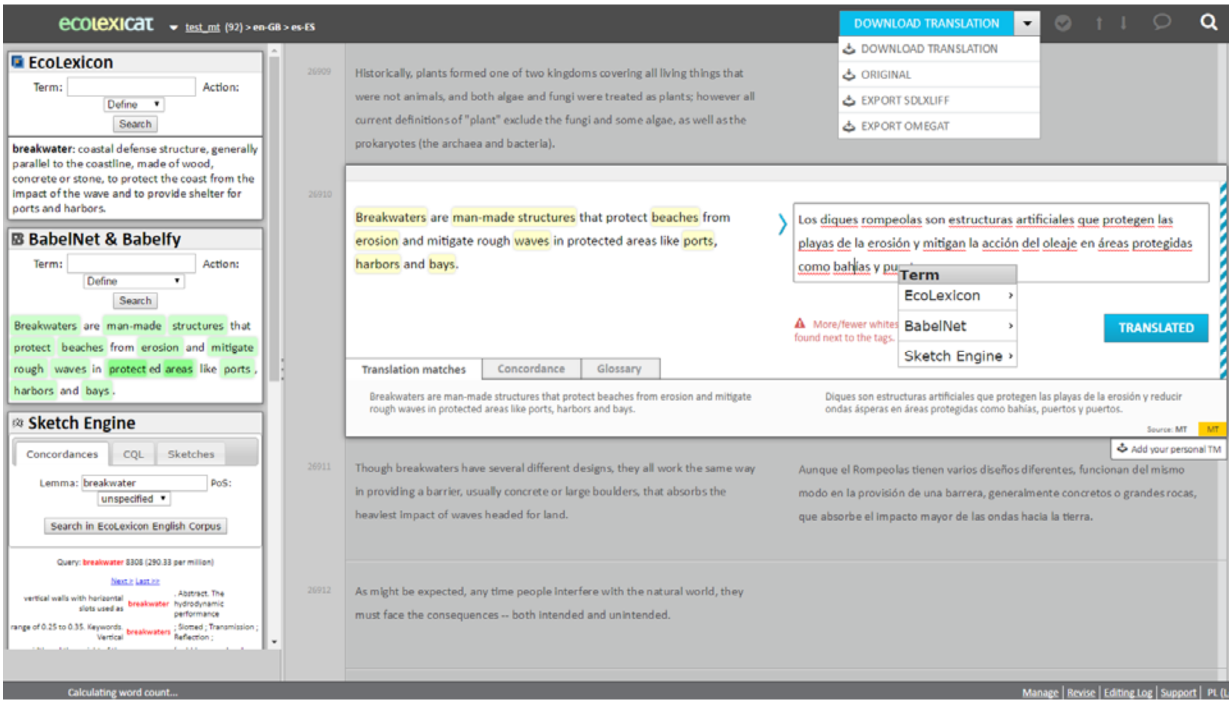

Figure 3: Main user's interface of EcoLexiCAT

Figure 4 shows a segment within the editor. This editor offers the usual editing features of any CAT tool. Users can split or merge segments, copy the source text in the target segment, benefit from a QA system that detects missing spaces or tags, create on-the-fly glossary entries, search for concordances within the TM and get suggestions from previously stored segments in the TM or, if added, from an MT engine. Once a segment is confirmed, it is stored both in the users' TM and in a collective TM from which other users can benefit. This converts the tool into a collaborative environment.

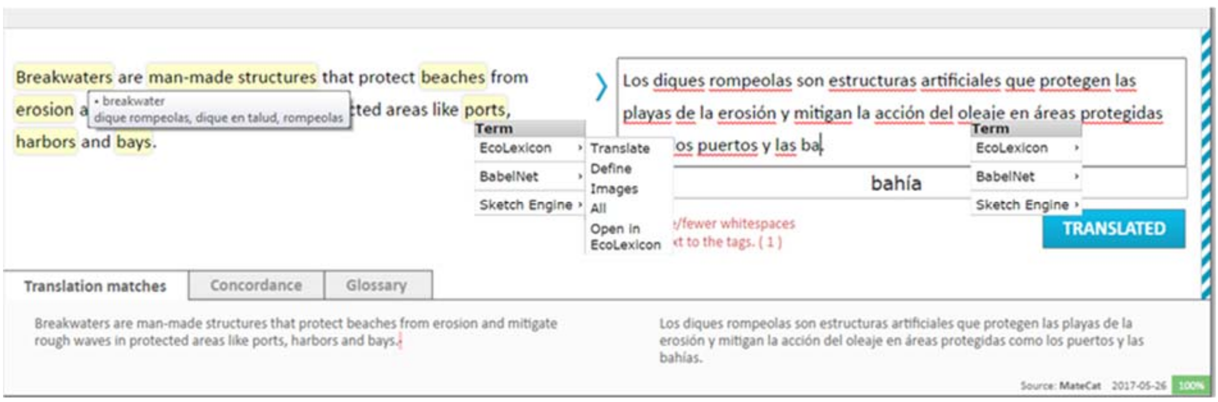

Figure 4: EcoLexiCAT editor

However, the difference between an ordinary CAT tool and EcoLexiCAT is that the EcoLexiCAT is a terminology-enhanced translation tool. This means that the editor interacts with external terminological resources that can assist the translator during the different phases of the translation workflow. First of all, the source segment is enriched with information from EcoLexicon. This is done by lemmatizing all the words in the segment and matching them against the term entries in the TKB.

All matching terms are highlighted in yellow, and users can interact with them in three ways: (1) if they hover the mouse over them, all possible translations (equivalent terms and synonyms) are displayed in an emerging box; (2) if they click on any of them, the EcoLexicon box of the left-hand side shows both the translations and the definition; and (3) if they right-click on any of them, a scroll-down menu gives access to all the different options provided by each of the resources of the left-hand section (see Figures 5–11).

For instance, in the case of EcoLexicon, these options correspond to the data categories in the TKB that usually serve for text comprehension: translations, synonyms, definitions, and images. Also from this menu, a new tab can be opened in the browser to access the EcoLexicon TKB for a more detailed analysis of the conceptual networks.

In turn, the target segment is enriched with a predictive typing feature. As soon as users start typing a word that has been matched as the translation of one of the terms in the source segment, all possible translations are shown in a drop-down list. In addition, as in the source segment, users can right-click on any term they type in the target segment and send queries to the three resources in the opposite language directionality. This is especially relevant in the case of corpus queries, since this is the resource that will usually be most useful during the text production phase.

Thus, the external resources of EcoLexiCAT interact with the segments in the editor during the different phases of the translation process, since they are terminologically enhanced for both source text comprehension and target text production tasks.

In Figures 5–11, a detailed view of the external resource boxes is provided. Figure 5 shows the EcoLexicon box as it appears when all features (i.e. translations, definition, images) are requested from any of the modules where the scroll-down menu may be activated (i.e. the EcoLexicon box itself, the BabelNet & Babelfy box, the source segment or the target segment). Users can also choose to visualize these features separately.

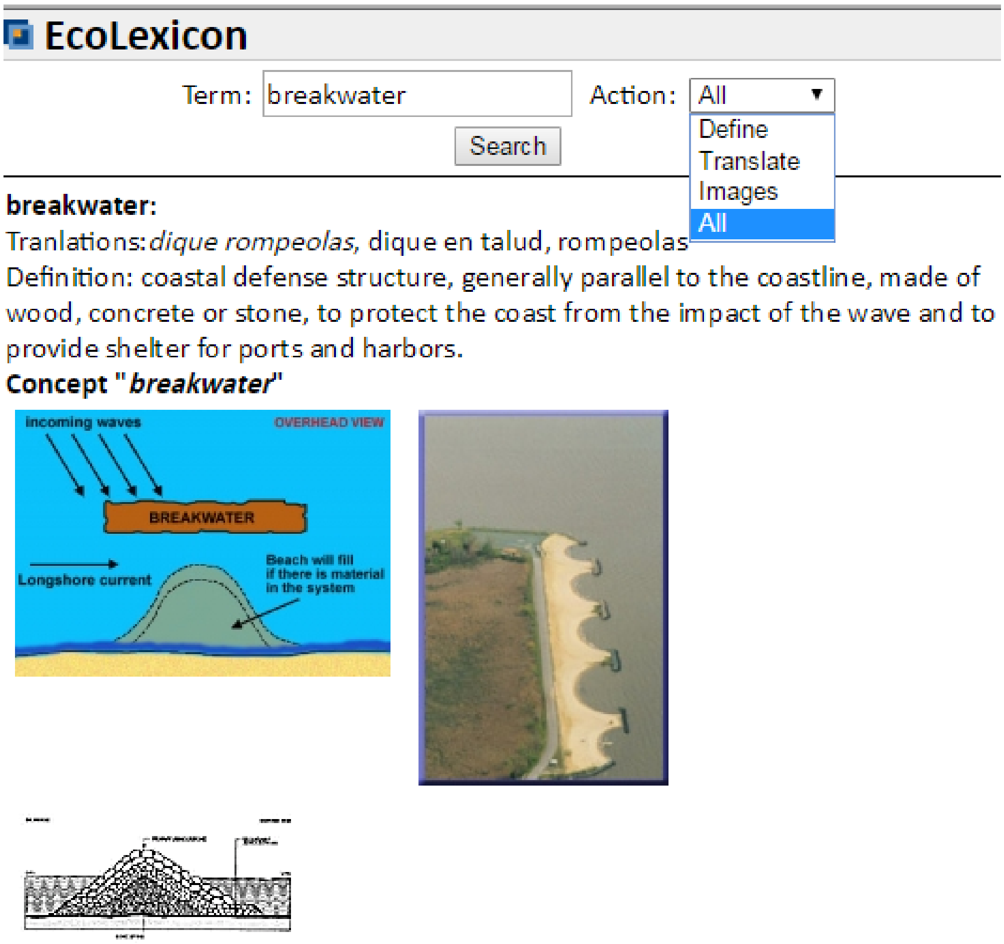

Figure 5: EcoLexicon box in EcoLexiCAT

Below the EcoLexicon box, users can find the BabelNet & Babelfy box (Figure 6), where the source text is also matched against the BabelNet network previously disambiguated by the Babelfy algorithm. This enables the system to propose statistically relevant candidate translations, which is a significant advantage taking into account that BabelNet covers any specialized or general domain and ambiguity can be frequently encountered. Furthermore, it helps the system to arrange definitions or images in the most plausible order. For instance, in Figure 6, while the first three definitions can be useful for EcoLexiCAT users, the fourth clearly belongs to a different domain and shows a different sense of the term *erosion*.

In this box, all matched terms are highlighted in green and behave in the same manner as the terms in the source segment with regard to EcoLexicon: (1) if users hover the mouse over them, all possible translations (equivalent terms and synonyms) are displayed in an emerging box; (2) if they click on any of them, the BabelNet & Babelfy box on the left-hand side shows both the translations and the definition; and (3) if they right-click on any of them, a scroll-down menu gives access to all the different options provided by each of the resources of the left-hand section. In the case of

BabelNet, these options correspond to the data categories that have been considered most interesting for translators: definitions, translations, compound words and images. Also, from the definitions option, a new tab can be opened in the browser to access the semantic networks in BabelNet.

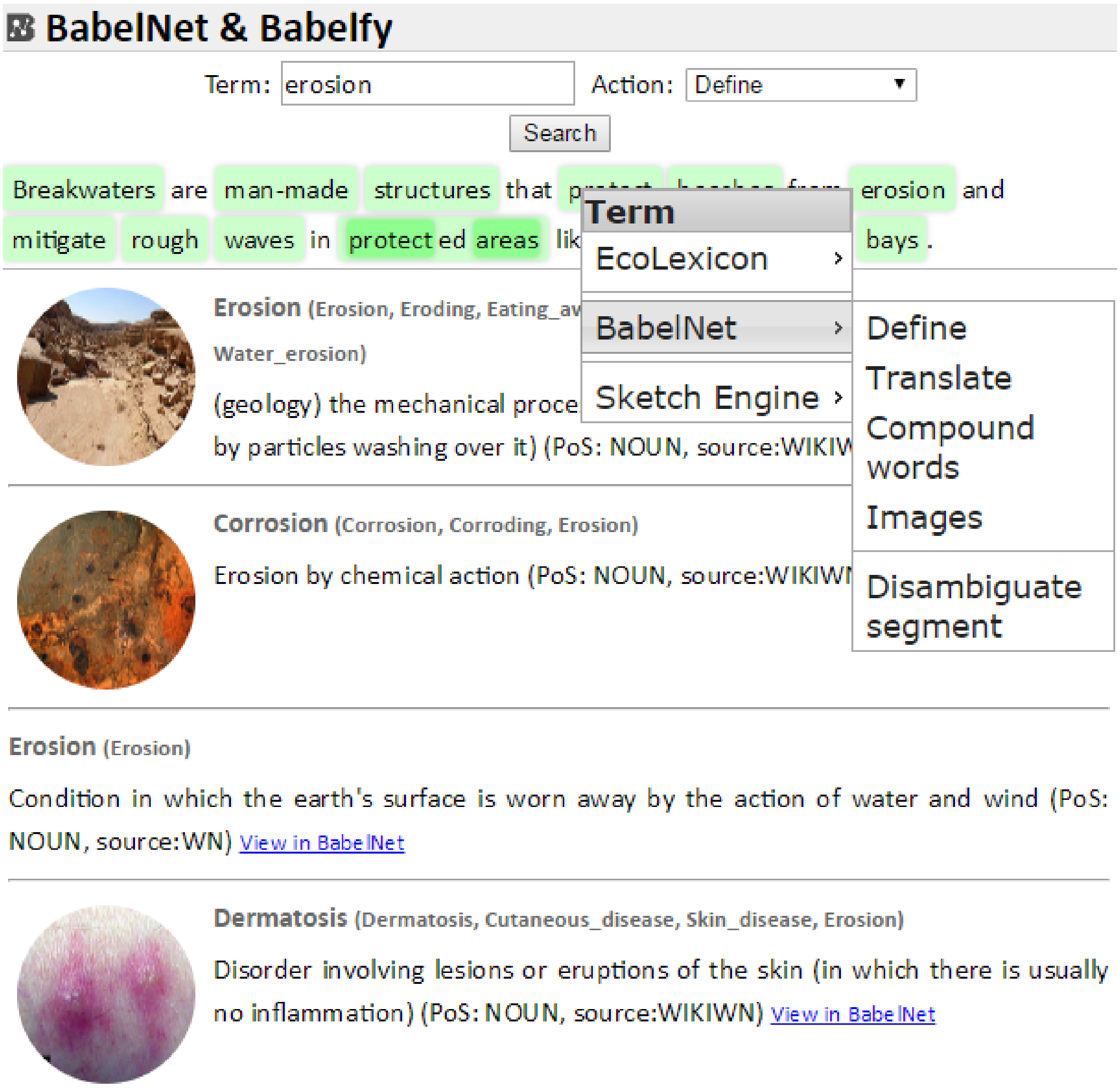

Figure 6: BabelNet and Babelfy box in EcoLexiCAT – Definitions

This box is particularly interesting for terms that are not available in EcoLexicon. This may occur when entries in EcoLexicon have not yet been included (it is a developing resource), when general language issues arise or when the source text combines environmental terms with terms from other specialized domains. Nevertheless, users should be cautious because the Babelfy algorithm may fail or produce candidate translations that do not account for domain specificity. Being an automatically built resource based on the synsets of WordNet (a general language lexical database), BabelNet often offers a set of concepts with different levels of granularity under the same entry.

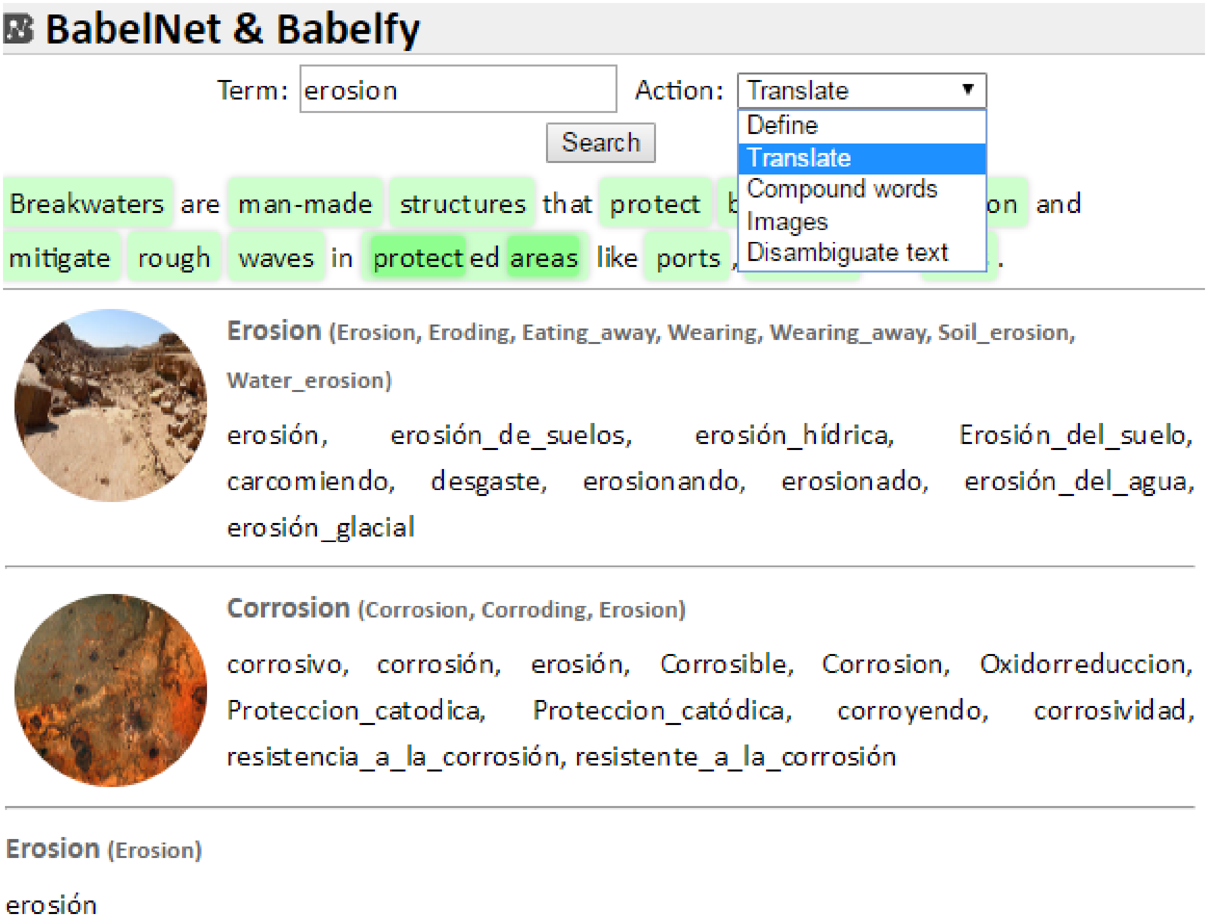

Figure 7: BabelNet & Babelfy box in EcoLexiCAT – Translations.

For instance, in Figure 7, the candidate translations go beyond equivalence, since some of the terms are hyponyms or derivatives of *erosion*. Nonetheless, when used with caution, these results can help to expand user knowledge of the semantic network of the domain

However, for this purpose, and especially for text production tasks, there is another option in this box, namely compound words. Figure 8 shows different compound terms of *erosion*, whether it acts as the head (e.g. *beach erosion*) or the modifier of the compound (e.g. *erosion control*). All of them can be clicked to access their definitions. In this way, users can browse the resource through different interconnected concepts and terms and gain a better understanding of the domain. Finally, images are the last option available from BabelNet (Figure 9). They can be very useful when understanding and translating complex concepts, such as processes or parts of entities, and can complement the images offered by EcoLexicon.

Figure 8: BabelNet & Babelfy box in EcoLexiCAT – Compound words

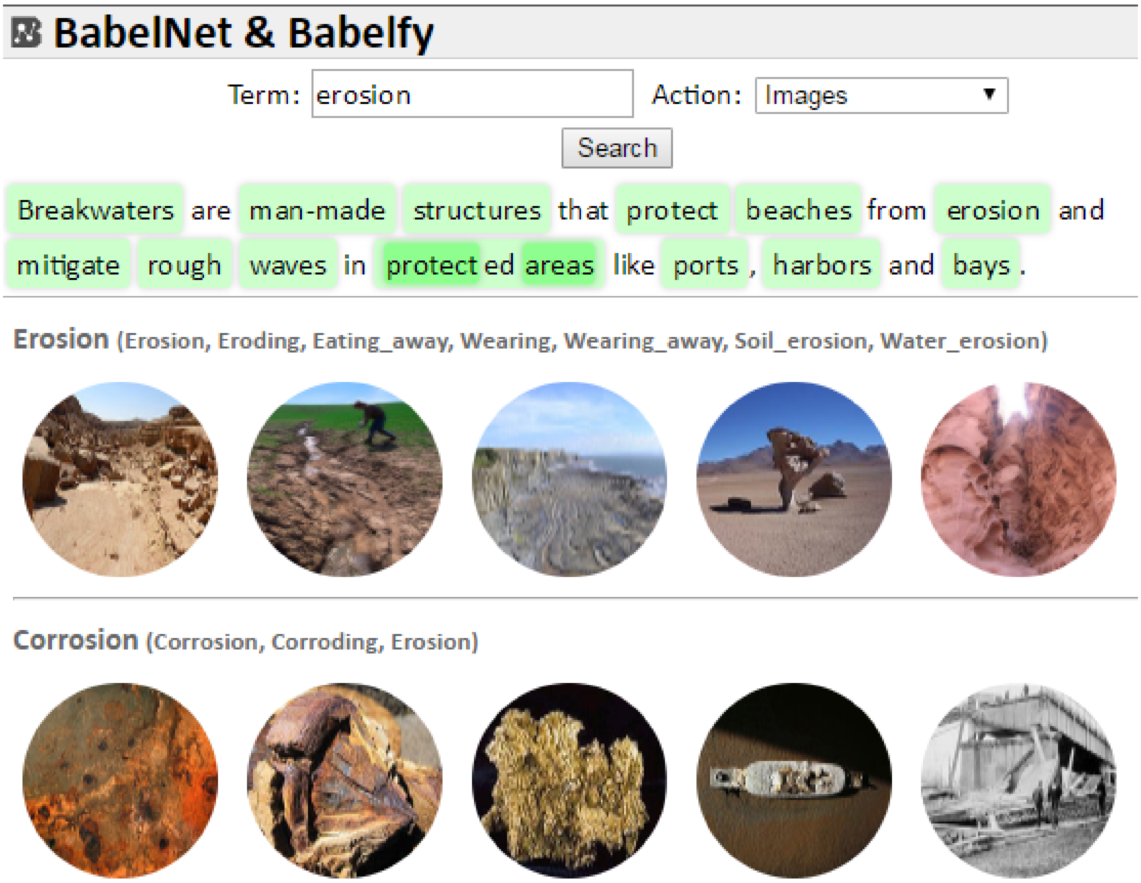

Figure 9: BabelNet & Babelfy box in EcoLexiCAT – Images

Below the BabelNet & Babelfy box, the Sketch Engine box appears (Figure 10). This box can be used to select a term from both the source and target segments and analyze its behavior in the EcoLexicon Corpus. So far, only the EcoLexicon English Corpus is hosted in Sketch Engine Open Corpora. The EcoLexicon Spanish Corpus is still in the compilation phase but will be made available in the near future.

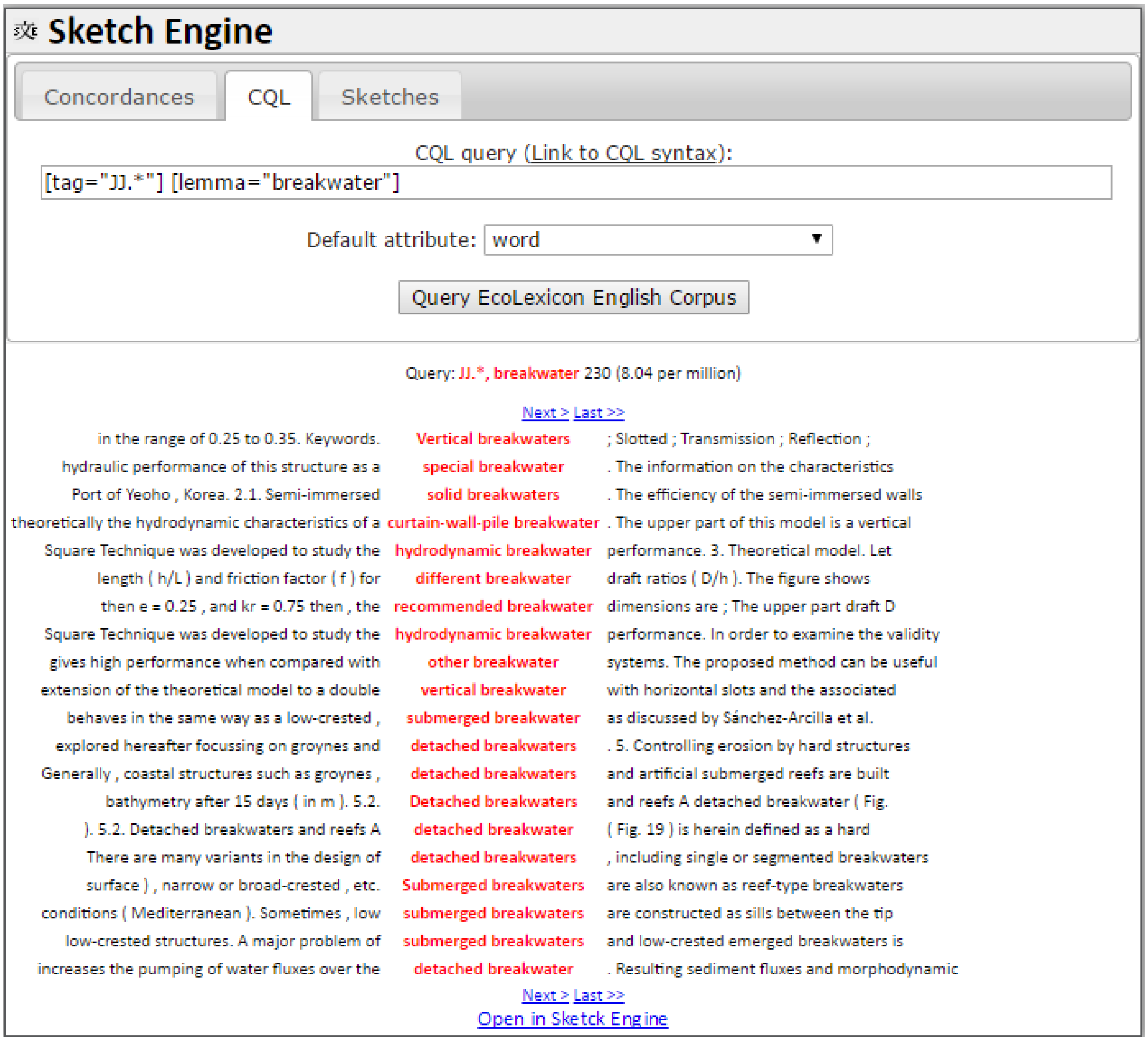

Figure 10: Sketch Engine box in EcoLexiCAT – CQL queries

The corpora can be queried through basic or CQL queries (Figure 10) as well as through word sketches (Figure 11). The output of the queries can be opened in a new tab that sends users to the website of Sketch Engine Open Corpora for a more detailed analysis. In this way, they can use all the functionalities of the tool (e.g. Context, Word list, Thesaurus, Sketch Diff, etc.) and make more specific queries filtered by the features according to which the corpus is tagged (i.e. year, genre, contextual domain, user type and linguistic variant). As previously mentioned, this information can be very useful during the text production phase (e.g. searching for modifiers or verbs that collocate with a particular noun, looking for synonyms or frequent syntactic structures, etc.). However, corpora can also help translators to understand how concepts interrelate with each other within the domain. For this reason, corpus queries are enabled from both source and target segments.

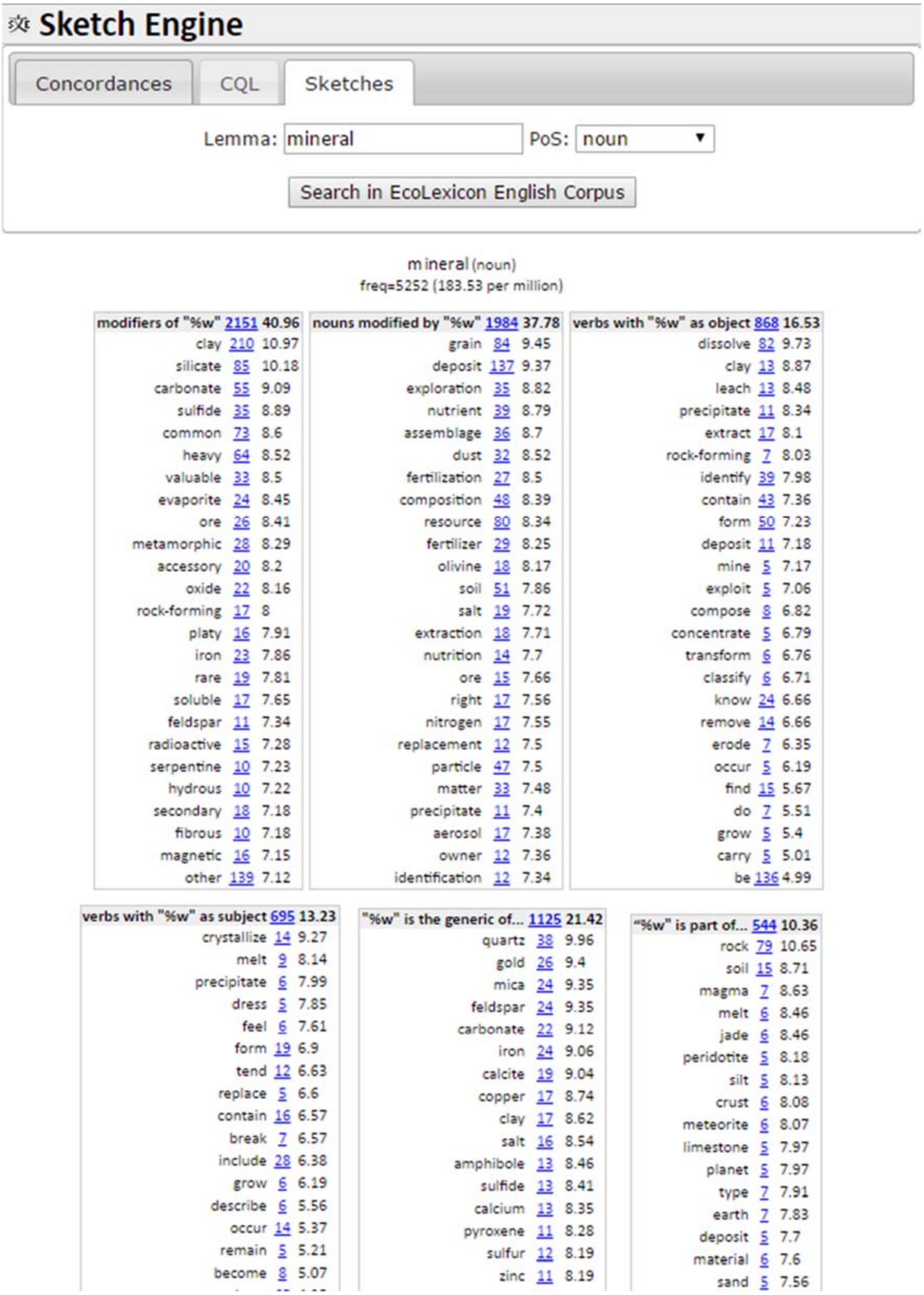

Figure 11: Sketch Engine box in EcoLexiCAT – Word Sketches

For instance, with the CQL query in Figure 10, users can not only access the adjectives that modify the term *breakwater* but also infer that breakwaters are usually classified according to position, material, function, etc. Furthermore, in Figure 11, Sketch Engine's default word sketches (e.g. modifiers and verbs) are combined with a series of customized word sketches (León-Araúz et al., 2016) especially focused on the comprehension phase, since they are based on semantic relations and thus provide knowledge rich contexts (Meyer, 2001). In Figure 11, the customized word sketches of the relations *is_the_generic_of* and *is_part_of* are shown for the term *mineral*. In this way, users can have quick access to part of the conceptual network of all concepts sufficiently represented in the corpus.

Finally, there are two other features powered by MateCat that can be of interest to professional translators, as well as to lecturers and researchers. As soon as a segment is confirmed, users can open their editing log (Figure 12) and monitor their own performance. This includes different types of information on each segment, such as: (1) the time invested in post-editing it; (2) the suggestion source, whether it comes from MT or TMs; (3) the matching percentage between the source segment and the suggestion; and (4) the post-editing effort and tracked changes of the final target segment. These data help to raise user awareness regarding their strengths and weaknesses as professional translators as well as those of the tool. For this reason, the editing log can also be exploited by Translation lecturers and researchers who are interested in assessing both the work of students and/or the performance of the tool.

Editing Details

Secs/Word	Job ID	Segment ID	Words	Suggestion source	Match percentage	Time-to-edit	PE Effort
6.4	92	26910	21.00	Machine Translation	85%	02m:13s	33%

Segment	Breakwaters are man-made structures that protect beaches from erosion and mitigate rough waves in protected areas like ports, harbors and bays.
Suggestion	Diques son estructuras artificiales que protegen las playas de la erosión y reducir ondas ásperas en áreas protegidas como bahías, puertos y puertos.
Translation	Los diques rompeolas son estructuras artificiales que protegen las playas de la erosión y mitigan la acción del oleaje en áreas protegidas como los puertos y las bahías.
Diff View	~~Diques~~ Los diques rompeolas son estructuras artificiales que protegen las playas de la erosión y ~~reducir ondas ásperas~~ mitigan la acción del oleaje en áreas protegidas como ~~bahías,~~ los puertos y ~~puertos.~~ las bahías.

Figure 12: Editing log in EcoLexiCAT

In this line, the revision panel (Figure 13) helps to perform the last phase of the translation workflow. Revisers can approve or correct all target segments. If corrected, the changes are tracked in the target cell, and revisers can use a metric for translation quality evaluation commonly used in the industry. This metric is based on different error types (i.e. tag issues, translation errors, terminology and translation consistency, language quality and style) and degrees (i.e. enhancement and error). At the end, users can generate a quality report that automatically scores the overall quality of the translation based on the issues highlighted by the revisers. Therefore, this feature can also be used by Translation lecturers if they want to grade their students' work in a systematic way.

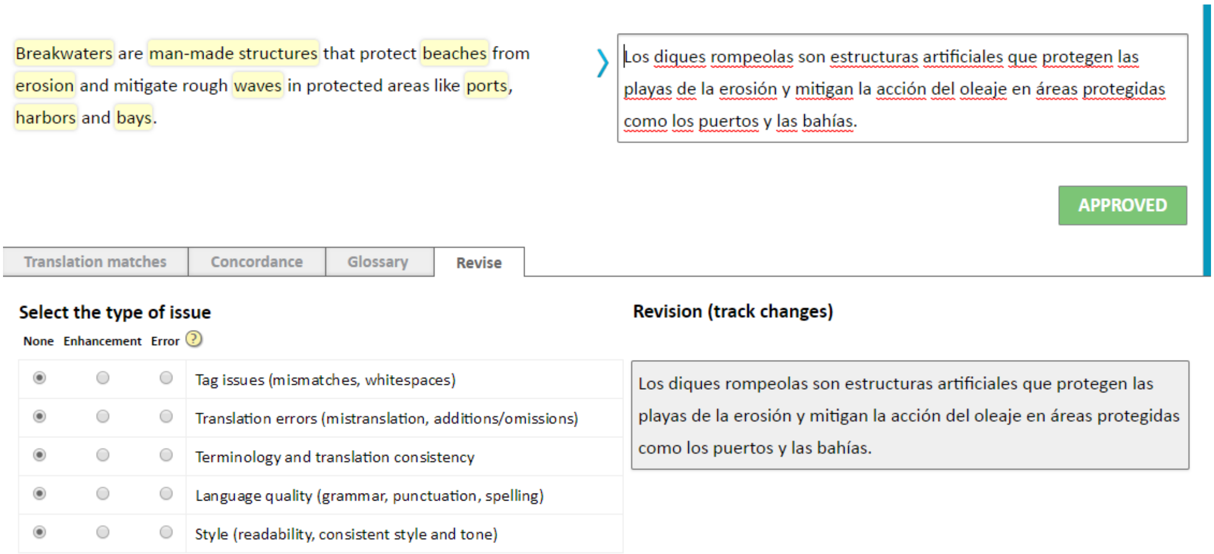

Figure 13: Revision in EcoLexiCAT

5. Conclusions and future work

In this paper we have presented the first version of EcoLexiCAT, a terminology-enhanced tool that enriches both source and target segments with terminological information from three external resources in an interactive environment. The tool has been designed to meet the expectations of professional translators regarding terminology management. However, it still needs to be evaluated by prospective users. A study comparing the performance of EcoLexiCAT users versus non EcoLexiCAT users will thus be carried out in the near future.

However, there are still other features that will be added to the tool before starting the evaluation process. For instance, EcoLexiCAT will also be enriched with other external resources. Part of the Inter-Active Terminology for Europe[7] (IATE), EU's multilingual term base, has been recently downloaded and stored in a database to interact with EcoLexiCAT as a fourth external resource. The IATE dump will cover the entries in English and Spanish belonging to environment-related domains.

Furthermore, EcoLexicon is currently being linked to other encyclopedic (i.e. DBpedia) and environmental resources (i.e. GEMET, AGROVOC) by means of Linked Data. Once the TKB is fully integrated into the Linguistic Linked Open Data, EcoLexiCAT will also benefit from reliably disambiguated encyclopedic and specialized term entries.

[7] http://iate.europa.eu/tbxPageDownload.do

In the same line, we plan to add another box enabling users to customize for each project a resource console based on the URLs of the resources that they usually consult, such as WordReference, TERMIUM Plus, MetaGlossary, Linguee, etc. This will work as the SDL Trados Studio plug-in Web Lookup or the MemoQ web search feature. Two other features from EcoLexicon will also be added once they are ready. These are the EcoLexicon Spanish Corpus and phraseological patterns from a new module that is currently under construction.

Finally, when all of these features are included in the tool, EcoLexiCAT will be made freely available for any user interested in translating English or Spanish environmental texts. Users will only need to register and indicate their educational background, translation experience and the purpose for which they will be using the tool. This will help us analyze user profiles and behaviour when interacting with the tool. Moreover, it will allow us to classify the resources generated (i.e. TMs), which can be used as a parallel corpus, thus enriching both the tool and the EcoLexicon Corpus.

6. Acknowledgements

This research was carried out as part of project FF2014-52740-P, *Cognitive and Neurological Bases for Terminology-enhanced Translation* (CONTENT), funded by the Spanish Ministry of Economy and Competitiveness.

7. References

Bertoldi, N., Cettolo, M. & Federico, M. (2013). Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of the MT Summit XIV*, Nice, France, September, pp 35–42.

Durán Muñoz, I. (2010). Specialized lexicographical resources: a survey of translators' needs. In S. Granger & M. Paquot (eds) (2010). *eLexicography in the 21st century: New Challenges, new applications. Proceedings of ELEX2009.* Cahiers du Cental. Vol. 7. Louvain-La-Neuve: Presses Universitaires de Louvain, pp. 55 – 66.

Durán Muñoz, I. (2012). Meeting translators'needs: translation-oriented terminological management and applications. *The Journal of Specialised Translation*, 18, pp. 77–92.

Faber, P., León-Araúz, P. & Reimerink, A. (2011). Knowledge representation in EcoLexicon. In N. Talaván, E. Martín Monje & F. Palazón (eds.) *Technological Innovation in the Teaching and Processing of LSPs: Proceedings of TISLID*, 10. Madrid: Universidad Nacional de Educación a Distancia, pp 367–385.

Faber, P. (ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/New York: Mouton de Gruyter.

Faber, P. (2015) Frames as a framework for terminology. In H. J. Kockaert & F. Steurs (eds.) *Handbook of Terminology*, 1. John Benjamins Publishing Company, pp. 14–33.

Faber, P., León-Araúz, P. & Reimerink, A. (2014). Representing environmental knowledge in EcoLexicon. In *Languages for Specific Purposes in the Digital Era. Educational Linguistics*, 19. Springer, pp 267–301.

Faber, P., León-Araúz, P. & Reimerink, A. (2016) EcoLexicon: new features and challenges. In I. Kernerman, I. Kosem Trojina, S. Krek, & L. Trap-Jensen (eds.) *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, Portorož, pp. 73–80.

Federico, M. et al. (2014). The MateCat Tool. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, Dublin, Ireland, August 23-29, pp 129–132.

Fillmore, C. J. (1982). Frame Semantics. In The Linguistic Society of Korea (ed.) *Linguistics in the Morning Calm*. Seoul: Hanshin, pp. 111–137.

Fillmore, C. J. & Atkins, B. T. S. (1992). Toward a Frame-based Lexicon: The Semantics of RISK and Its Neighbors. In A. Lehrer & E. Kittay (eds.) *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*. Hillsdale NJ: Erlbaum, pp. 75–102.

Gornostay, T. (2014). Dreams of better terminology tools. *Multilingual Magazine* April/May, pp. 44–45.

International Organization for Standarization (ISO) (1999). *ISO 12620. Computer applications in terminology - Data categories*. Ginebra: ISO.

Kilgarriff, A, Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress*. Lorient: EURALEX, pp. 105–116.

León Aráuz, P., Faber, P. & Montero Martínez, S. (2012). Specialized Language Semantics. In P. Faber (ed.) *A cognitive linguistics view of terminology and specialized language.*, 20. Berlin, Boston: De Gruyter Mouton, pp. 95–175.

Meyer, I. (2001). Extracting Knowledge-Rich Contexts for Terminography: A conceptual and methodological framework. In Bourigault, Jacquemin, L'Homme (eds.), *Recent Advances in Computational Terminology*, pp. 279–302.

Moro, A, Raganato, A., Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics* (TACL), 2, pp. 231–244.

Moro, A., Cecconi, F., Navigli, R. (2014) Multilingual Word Sense Disambiguation and Entity Linking for Everybody (2014). *Proc. of the 13th International Semantic Web Conference, Posters and Demonstrations* (ISWC 2014), pp. 25–28, Riva del Garda, Italy, 19-23 October 2014

Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pp. 217–250.

Nielsen, S. (2008). The Effect of Lexicographical Information Costs on Dictionary Making and Use. *Lexikos*, 18, pp. 170–189.

Tarp, S. (2013). What should we demand from an online dictionary for specialized translation? *Lexicographica - International Annual for Lexicography*, 29(1), pp. 146–162.

Tudhope D., Koch T. & Heery R. (2006). *Terminology Services and Technology: JISC state of the art review.* Available at: http://www.ukoln.ac.uk/terminology/JISC-review2006.html