Journal of Digital Terminology and Lexicography - ISSN:3103-3601
Vol. 2, No. 1 (2026) DOI: 10.25430/pupj.jdtl.1773243626

Requirements for Constructing a Tool for the Extraction of Phraseological Structures

Beatriz Sánchez-Cárdenas[1*], Pablo Rienda[2], Nuria Medina-Medina[3] and Carlos Ramisch[4]

[1]Department of Translation and Interpreting, Universidad de Granada, calle Buensuceso, 11, Granada, 18002, España.
[2]Universidad de Granada, calle Periodista Daniel Saucedo Aranda S/N, Granada, 18071, España.
[3]Department of Computer Languages and Information Systems, Universidad de Granada, calle Periodista Daniel Saucedo Aranda S/N, Granada, 18071, España.
[4]Aix Marseille Univ, CNRS, LIS, Marseille, France.

*Corresponding author(s). E-mail(s): bsc@ugr.es
Contributing authors: prienda@correo.ugr.es; nmedina@ugr.es; carlos.ramisch@lis-lab.fr

Abstract

This contribution presents the design and development of a web-based prototype for the extraction and analysis of specialized argument structures in multilingual corpora. The tool encapsulates complex command-line scripts into a user-friendly interface, allowing researchers to load, parse, and index corpora, search for noun-verb-noun triples, and organize results into lexical clusters. By leveraging distributional semantics models like word2vec, the tool refines clusters by filtering irrelevant terms and enriching them with semantically related ones. The prototype supports cross-platform accessibility, ensures centralized server-side storage, and provides scalable functionality for future extensions. Currently in the testing phase, it addresses the limitations of previous tools by streamlining corpus analysis and making phraseological studies more accessible to academia.

Keywords: Verb-Noun Collocation, Lexical Clusters, Triples Extraction

1 Introduction

Until recently, the study of specialized language tended to focus on terms. However, at the beginning of the century, researchers realized that the description of any specialized domain should go beyond noun description and consider other information such as phraseological structures, which are essential to write or translate scientific texts (L'Homme, 2015; Granger & Meunier, 2008; Faber, 2012). Indeed, the production of texts relies on managing structured lexico-grammatical constructs, such as phraseology (Corpas Pastor, 2008; Tutin & Grossmann, 2014; Vezzani, 2023).

Obviously, phraseological units occur not only in general language but also in scientific discourse. These include metatextual expressions (*formulating a hypothesis*), interpersonal markers (*as it is well known*), logical connectors (*therefore*), attitude expressions (*defending a position*), modal markers (*to some extent*) (Jacques & Tutin, 2018), or verb-noun combinations (*to eject lava*) (Sánchez-Cárdenas & Buendía Castro, 2012).

This study examines verb-noun collocations in specialized discourses. By "verb-noun collocation", we refer to combinations of a verb and a noun (Buendía Castro, 2013; Buendía Castro & Sánchez-Cárdenas, 2016) that form an argument structure. For instance, in environmental sciences, verbs such as *drive*, *encourage* or *provoke* are often used to link *deforestation* to its causes. Interestingly, finding the equivalent expression in the target language requires not only knowing the language but also knowing how scientists express processes in each domain. As a result, translating the verb and the nouns or terms separately, rather than the whole semantic structure, might lead to non-idiomatic sentences. For instance, translating "la ganadería extensiva ayuda a prevenir los incendios" as "extensive livestock farming helps to prevent wildfires" might not be the most idiomatic way to express that idea in English. A more accurate choice would be to use verbs such as *to reduce the risk*, *to mitigate* or *to inhibit*. However, the verb *prevenir* might be translated by other verbs according to each specialized domain (*mitigate symptoms*, *reduce the risk*, *protect against something*) or event by its cognate (*to prevent dysbiosis*). Indeed, specialized languages follow their own lexical preferences, which differ from those of general language. This type of information is usually absent from most terminological resources (L'Homme, 2020), but it can be found in comparable corpora and in some specialized dictionaries, for instance those created at the Observatoire Linguistique Sens-Texte (such as DiCoInfo, DiCoEnviro or JuriDiCo). These dictionaries describe lexical relations, actantial structures, and semantic patterns.

In this sense, our approach can be related to Explanatory Combinatorial Lexicology (Mel'čuk, 1998), which places collocational relations at the core of lexical description. We share with this framework the assumption that recurrent combinatorial patterns constitute a central source of linguistic and conceptual information. This view is also reflected in more recent terminological work on specialized collocations and combinatorial behavior (L'Homme, 2020), which emphasizes the importance of systematically describing verb–noun relations in specialized discourse.

Interestingly, Neuronal Machine Translation (NMT) does not always produce satisfactory results for this issue, as it tends to translate verbs into their formal equivalents in general language (*provoque/provocar*; *cause/causar*) rather than domain-specific ones. This might be partly due to the fact that general machine translation tools are mostly trained, for now, with English corpora and do not discriminate different genre (blogs vs scientific papers), resulting in non-idiomatic texts with a plain style and standardized language when translating to other languages than English. This is known as the "language modelling bias" or the "digital linguistic bias" (Muñoz-Basols et al., 2024).

The implications of this go beyond the preservation of linguistic heritage and the specificity of each language and culture. Indeed, it can also lead to a loss of nuance, terminological imprecision or semantic inaccuracy. For instance, in NLP we say "to fine-tune a model", which in French is "affiner un modèle" and not "peaufiner/faire un réglage fin" as suggested by Google Translate. Extracting the linguistic structures from comparable corpora is one potential solution to create linguistic tools that help mitigate the standardization that comes with the use of NMT and AI-based text generation. However, analyzing concordances manually is a rather inefficient strategy. A more efficient alternative is to run complex queries that are capable of modeling lexico-grammatical co-occurrence patterns that approximate predicate-argument structures. Yet, this is also time-consuming and demands specific skills that most scientific translators or writers lack.

In this perspective, we developed a methodology to extract such phraseological units from corpora in the form of [noun-verb-noun] combinations, called triples, reflecting the argument structure of a given concept across several languages. For example: *[**Soybean expansion**] in southern Brazil [**contributed**] to [**deforestation**] by stimulating migration to agricultural frontier regions.*

The process of triple extractions can be achieved by employing a range of corpus tools capable of identifying argument structures. Previous research (Sánchez-Cárdenas, 2024) compared the performance of two computational tools – MWEtoolkit and Sketch Engine – in facilitating this specific task. The key difference between these two tools lies in the specific purpose for which they were originally designed. The study concluded that both tools struggled with noise in the extracted triples. Sketch Engine produced a high level of noise (90.9%), whereas MWEtoolkit performed better with 34.4% of noise. Conversely, MWEtoolkit retrieved a much higher percentage of accurate triples (65.5%) than Sketch Engine (9.1%).

In order to carry out these extractions, we designed a tool prototype that automates the extraction of [noun-verb-noun] structures from specialized corpora in multiple languages. It offers a web interface designed to help researchers and linguists analyze and manipulate this type of linguistic information more efficiently. This article presents our project, covering both the corpus-based methodology used to extract the candidate units and the corresponding web interface of the tool.

Although similar projects and initiatives exist (Baroni & Bernardini, 2004; Orliac, 2006; Vezzani, 2023), to the best of our knowledge, none offers the possibility to extract argument structures in the form of triples from specialized corpora across languages.

In Section 2, we describe the protocol for extracting argument structures in the form of triples from specialized corpora. Section 3 explains the features of a web-based tool prototype created to simplify these searches and includes relevant screenshots for illustration.

2 Retrieving Triples to Represent Argument Structures

In previous research (Sánchez-Cárdenas & Ramisch, 2019), we employed MWEtoolkit, a computational tool for the identification of multiword expressions in corpora (Ramisch, 2015, 2023), to isolate triples [noun1-verb-noun2] representing argument structures. For this endeavor, queries were designed through Python scripts to process and query the corpora, extract candidates, and sort the results.

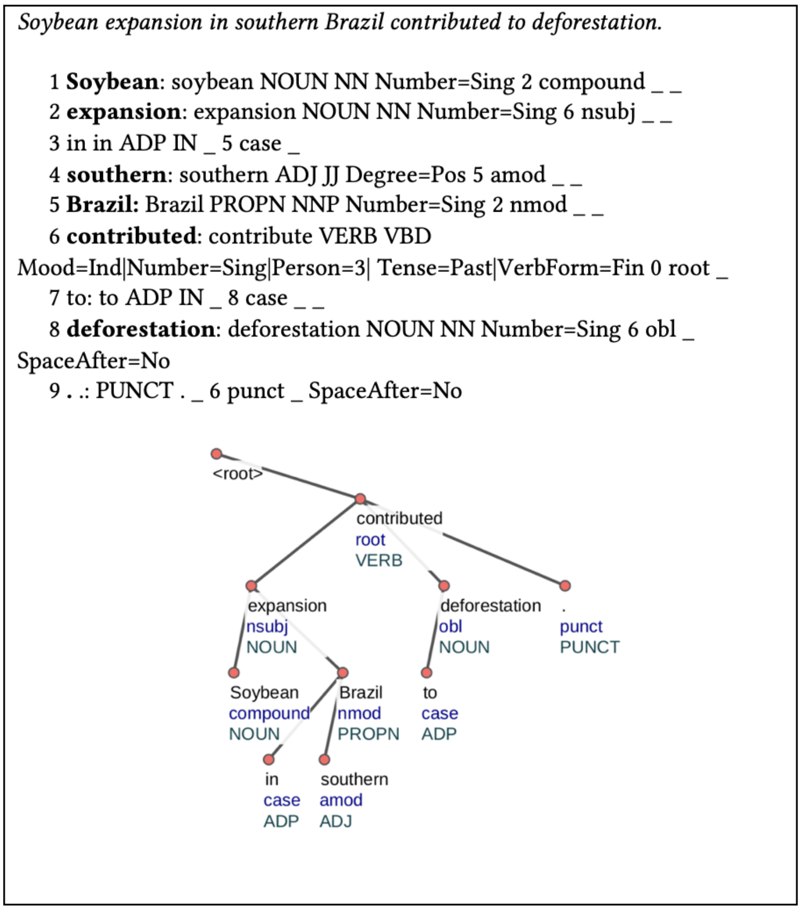
The experiments reported in this paper are based on two comparable specialized corpora in English and Spanish, compiled within the domain of deforestation. The English corpus contains approximately 1,105,718 tokens, while the Spanish corpus comprises around 971,788 tokens. The corpora are comparable, which means that they were designed to reflect similar thematic and specialization levels. Both corpora include different kinds of specialized text types, such as scientific articles, encyclopedia entries and specialized news reports.

2.1 Processing the Corpus

During the preprocessing phase, texts were automatically cleaned and converted to UTF-8 encoding. They were then processed and analyzed using UDPipe, a multilingual natural language processing tool for the analysis of texts (Straka, 2018). UDPipe performed the following tasks: splitting the text into sentences, tokenizing sentences, tagging words with their part-of-speech tags using the Universal Dependencies tagset, assigning lemmas, and generating syntactic dependency trees to map relations between words, as shown below in the example in Figure 1.

2.2 Querying the Corpus

- **Step 1: Regular expression queries**
  Using MWEtoolkit, queries were designed as multi-level regular expressions to extract [noun1-verb-noun2] triples, such as [volcano-eject-lava] (Sánchez-Cárdenas & Ramisch, 2019). The queries allow extracting lemmas (word occurrences neutralized for inflection) occurring within given sequences of automatically assigned parts of speech. These searches capture argument structures, but also irrelevant triples such as [volcano-see-lava], which required manual filtering. Searches were encapsulated in shell scripts for more efficient execution. They require specifying the target corpus and at least one element of the triple ([noun1], [verb] or [noun2]).
- **Step 2: Search strategy**
  In order to test the validity of the queries, a pilot study was conducted. Initial queries were constructed using terms within the domain of environmental sciences, derived from the terminological knowledge base MarcoTAO. Specifically, the semantic relations between concepts



Soybean expansion in southern Brazil contributed to deforestation.

1 **Soybean**: soybean NOUN NN Number=Sing 2 compound _ _
2 **expansion**: expansion NOUN NN Number=Sing 6 nsubj _ _
3 in in ADP IN _ 5 case _
4 **southern**: southern ADJ JJ Degree=Pos 5 amod _ _
5 **Brazil:** Brazil PROPN NNP Number=Sing 2 nmod _ _
6 **contributed**: contribute VERB VBD Mood=Ind|Number=Sing|Person=3| Tense=Past|VerbForm=Fin 0 root _
7 to: to ADP IN _ 8 case _ _
8 **deforestation**: deforestation NOUN NN Number=Sing 6 obl _ SpaceAfter=No
9 **.** .: PUNCT . _ 6 punct _ SpaceAfter=No

Fig. 1 Dependency tree example

were used to identify verbs lexicalizing those semantic relations. For instance, the query pattern [volcano-ANY-lava], with ANY representing an unspecified element, retrieved verbs like *eject, emit* or *spew.*

These verbs were then reused to identify additional nouns that could occupy the noun1 or noun2 positions. With each iteration, one of the three elements (noun1, verb, or noun2) was underspecified, while the others were specified based on previous query results. This iterative process gradually expanded the representativity of the results, covering a broader range of phraseological patterns in the domain (Table 1).

- **Step 3: Filtering and sorting search results**

Triples were automatically scored using pointwise mutual information, calculated from co-occurrence frequencies in the specialized corpus. Results were then sorted in descending order of relevance (according to this score), with the most relevant triples prioritized for further analysis. Another script performed this step, which involves concatenating the search results, calculating the relevance score, and ranking the triples. The output was stored in a textual file in tab-separated format (TSV), compatible with spreadsheet editors, for manual review. Finally, the triples were manually assessed by the authors using a numerical code from 0 to 3 (Table 2):

(0) Incorrect: The extracted triple does not accurately represent the argument structure of the term.
(1) Partially correct: The extraction is not entirely accurate, but it provides useful insights for enhancing future extractions.
(2) Almost correct: The extraction is accurate, but some essential element is missing.
(3) Correct: The extracted triple is wholly accurate and requires no further information.

Table 1 Examples of iterative searches with *volcano*

Argument 1	verbe	Argument 2	Résultat
volcano	**ANY (unknown)**	lava	eject|emit|erupt|expel| produce| spew |create
volcano	eject|emit|erupt|expel| produce|spew|create|produce	**ANY (unknown)**	lava|gas|smoke|ash| cloud| rock|dust| steam|continent| land | crust
ANY (unknown)	create|produce	continent| land	eruption |hot spot | tectonic activity | earthquake| continental collision

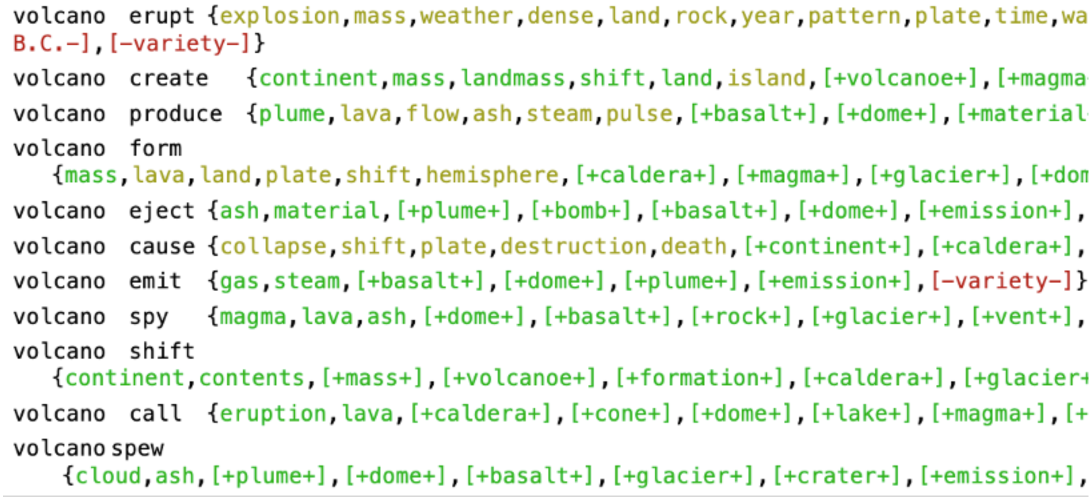
Table 2 Examples of triples coding

Triple extracted	Code	Explanation
deforestation-produce-Table	0	Error in extraction.
highway-lead-deforestation	1	The project of building new highways led to deforestation but this information was missing in the concordance.
deforestation-lead-shrinkage	2	Since complex nouns were not extracted, *forest shrinkage* was completed manually.
pasture-affect-deforestation	3	Accurate

2.3 Distributional Clustering of Triples

The final step involves the grouping and distributional clustering of triples marked as 2 or 3. Triples annotated as 0 and 1 were discarded at this stage. A specific script organizes the selected triples into clusters, grouping similar triples into linguistic schemas based on shared patterns, and extending their coverage based on distributional representations. The resulting patterns show common phraseological structures in the domain.

- **Step 1: Grouping Triples into Clusters**
  First, our tool groups entries of the form [noun1-verb-noun2] by retrieving the lemmas that co-occur in several triples of the initial input list. In other words, we generate all possible cluster keys from a triple (i.e. [noun1-verb], [verb-noun2] and [noun1-noun2]). All third elements that co-occur with a given cluster key pair are collected as its members. For example, as shown in Figure 2, if the triples [crater, expel, lava], [volcano, expel, lava], [volcano, expel, magma] and [volcano, expel, rock] appear in the input list, the cluster key pair [volcano, expel] defines one cluster whose members in position [noun2] are {lava, magma, rock}. The cluster key pair [expel, lava] defines another cluster, with {crater, volcano} as its members occupying position [noun1]. At this stage, distributional representations are not yet involved: clustering is entirely determined by which third elements were observed together with each pair in the list of triples. Also, we did not try to create complex clusters such as [{volcano, crater}, expel, {lava, magma, rock}], since it is hard to ensure the compatibility of members based on a single shared lemma (here, *expel*).
- **Step 2: Creation of word embeddings**
  In order to enrich and filter these initial clusters, our tool relies on word embeddings (Pilehvar & Camacho-Collados, 2021). They are mathematical representations where each word is associated with a vector: a (fixed-length) sequence of numbers. The length of the sequence corresponds to the number of "dimensions" *D* of the vector. This sequence of *D* numbers can be seen as a "coordinate" or a "vector" in an abstract mathematical space, hence their interpreta-



```
volcano  erupt {explosion,mass,weather,dense,land,rock,year,pattern,plate,time,wa
B.C.-],[-variety-]}
volcano  create   {continent,mass,landmass,shift,land,island,[+volcanoe+],[+magma
volcano  produce  {plume,lava,flow,ash,steam,pulse,[+basalt+],[+dome+],[+material
volcano  form
   {mass,lava,land,plate,shift,hemisphere,[+caldera+],[+magma+],[+glacier+],[+dom
volcano  eject {ash,material,[+plume+],[+bomb+],[+basalt+],[+dome+],[+emission+],
volcano  cause {collapse,shift,plate,destruction,death,[+continent+],[+caldera+],
volcano  emit  {gas,steam,[+basalt+],[+dome+],[+plume+],[+emission+],[-variety-]}
volcano  spy   {magma,lava,ash,[+dome+],[+basalt+],[+rock+],[+glacier+],[+vent+],
volcano  shift
   {continent,contents,[+mass+],[+volcanoe+],[+formation+],[+caldera+],[+glacier+
volcano  call  {eruption,lava,[+caldera+],[+cone+],[+dome+],[+lake+],[+magma+],[+
volcano spew
    {cloud,ash,[+plume+],[+dome+],[+basalt+],[+glacier+],[+crater+],[+emission+],
```

Fig. 2 Lexical cluster of *volcano*

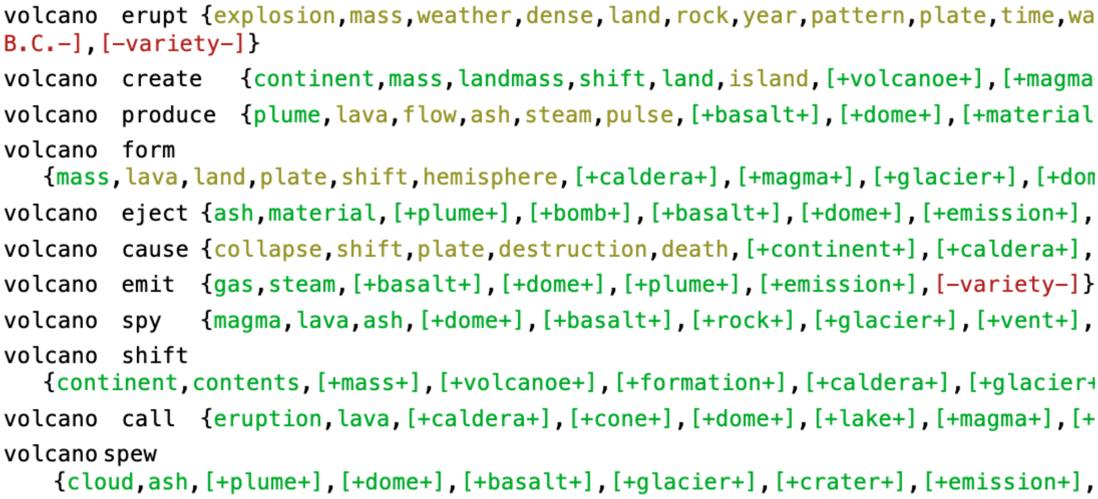
tion as being “embedded” in a *D*-dimensional space. In our experiments, we use *D*=400, that is, each word is represented as a sequence of 400 numbers.

We create word embeddings using *word2vec* (Mikolov et al., 2013), which indirectly relies on Harris’ (1971) distributional hypothesis. Given a specialized corpus, *word2vec* looks at how words appear in context — for example, if *volcano* and *crater* often occur near words like *erupt*, *magma*, and *ash*, they are probably semantically related. Words occurring in similar contexts will have similar vectors, and their proximity in the vector space reflects semantic relatedness. This idea is implemented as a neural network predicting context words (e.*g. erupt, magma,* and *ash*) from the input target words (*volcano* and *crater*). *Word2vec* word embeddings are extracted from the parameters of this neural network. *Word2vec* generates vectors for all words occurring at least twice in the corpus, not only those present in our triples. Thus, the embedding model provides a way to objectively quantify lexical similarity beyond the words that happened to co-occur in the raw triples. In our tool, the whole procedure of embedding creation is encapsulated within the *gensim* library. We provide the preprocessed specialized corpus as input to *gensim*, where each word is represented as a concatenated lemma/POS pair. We obtain a list of 400-dimensional word embeddings as output.

- **Step 3: Distributional expansion**

  Once the clusters (step 1) and word embeddings (step 2) have been built, we can enrich and refine the former with the help of the latter. For each observed member of a cluster (e.g. *lava* in the [*volcano*, *expel*] cluster), we look up its nearest neighbors in the embedding space according to cosine similarity. As a simplified example, suppose *volcano* is represented by the 3-dimensional vector (0.8, 0.2, 0.3) and *lava* by (0.7, 0.1, 0.6). To measure how close they are, cosine similarity checks whether both vectors have similar values in the same positions (e.g. 0.8 and 0.7 in position 1). In this toy example, cosine similarity is 0.93, which is high considering that the scale of possible similarity scores ranges from −1 to 1. This is interpreted as follows: *volcano* and *lava* share many contexts in the specialized corpus, so the model places them close together. If they co-occurred less often, their cosine similarity score would be lower.

  We generate the top 10 nearest neighbors to each cluster member according to cosine similarity. This window provides a sufficiently broad contextual spectrum to capture elements of the argument structure, while offering a good balance between coverage and precision, according to preliminary tests. Neighbors can in principle be any words in the vocabulary, not just terms that already occurred in other clusters. Neighbors are filtered so that they match the expected part of speech (i.e. if the cluster member is a noun, we only retrieve neighbor nouns).



```
volcano  erupt {explosion,mass,weather,dense,land,rock,year,pattern,plate,time,wa
B.C.-],[-variety-]}
volcano  create   {continent,mass,landmass,shift,land,island,[+volcanoe+],[+magma
volcano  produce  {plume,lava,flow,ash,steam,pulse,[+basalt+],[+dome+],[+material
volcano  form
   {mass,lava,land,plate,shift,hemisphere,[+caldera+],[+magma+],[+glacier+],[+dom
volcano  eject {ash,material,[+plume+],[+bomb+],[+basalt+],[+dome+],[+emission+],
volcano  cause {collapse,shift,plate,destruction,death,[+continent+],[+caldera+],
volcano  emit  {gas,steam,[+basalt+],[+dome+],[+plume+],[+emission+],[-variety-]}
volcano  spy   {magma,lava,ash,[+dome+],[+basalt+],[+rock+],[+glacier+],[+vent+],
volcano  shift
   {continent,contents,[+mass+],[+volcanoe+],[+formation+],[+caldera+],[+glacier+
volcano  call  {eruption,lava,[+caldera+],[+cone+],[+dome+],[+lake+],[+magma+],[+
volcano spew
    {cloud,ash,[+plume+],[+dome+],[+basalt+],[+glacier+],[+crater+],[+emission+],
```

Fig. 3 Lexical clustering of VOLCANO (EN)

Each neighbor is assigned a weight based on its cosine similarity to the cluster key pair (here [*volcano*, *expel*]). Only neighbors not already present in the cluster, matching the correct part of speech, and whose similarity is above a predefined threshold (0.3) are kept. In this way, new terms such as *plume* or *basalt* are added to the cluster even though they were not seen in the original triples, provided they are sufficiently close in the embedding space. Such terms are marked in green and added to the cluster surrounded by {+plus+} signs, as illustrated in Figure 2.

- **Step 4: Distributional filtering**
  The original clusters may contain spurious or less relevant elements, and embeddings can help identify them automatically. We calculate the cosine similarity between each member of a cluster and all other members. Elements considered as dissimilar with the rest of the cluster will have a similarity lower than the 0.3 threshold. For instance, a borderline case such as *rock*, with a similarity just below the threshold, would be flagged in yellow as a "probable removal."
  Finally, words explicitly listed in a provided blacklist (for example *variety* in Figure 3) are always shown in red and excluded, regardless of their score. This scoring and colour coding scheme provides a compact visual summary of which elements in each cluster are probably relevant and originally present in the triples (green), newly suggested via distributional expansion (in green and surrounded by {+plus+} signs), doubtful (in yellow), or explicitly spurious (in red and surrounded by {-minus-} signs).
- **Step 5: Manual refinement**
  The resulting clusters are the result of a fully automatic procedure and still require manual analyses. After the analysis of the automatic clusters, experts can create phraseological tables such as the one illustrated in Table 3. From the comparison of such tables across languages, knowledge of the argument structure and translation patterns in different languages will emerge.

Viewing terms as a mapping of argument-structure configurations rather than a list of isolated lexical items sheds light on how they behave in the syntagmatic axis in each domain, language and culture. These patterns illustrate not only linguistic uses but also how each domain encodes semantic relations such as agency or causality.

Interestingly, certain verbs recur across different rows, such as *lead to* which appears three times. However, each entry corresponds to a distinct N–V–N configuration. Although the verb is identical, the noun phrases that fill the argument slots differ across rows, representing different causal pathways within the domain. Collapsing the different rows around one verb would obscure the specific patterns combi-

Table 3 Lexical clustering of DEFORESTATION (EN) after manual analysis

Noun phrase 1	**Verb**	**Noun phrase 2**
{transportation cost, technology, fallow, forest cover, technological change]	increase	deforestation
{cattle ranching, population concentration, production, new crop, timber, technological progress}	lead to	deforestation
{forest clearing, infrastructure project, forestry, cocoa, technological change}	accelerate	deforestation
{crop, development, technological change}	promote	deforestation
{agricultural land, acquisition of land, development project, cropland, pasture}	drive	deforestation
technological change	lead to	{replanting, loss in forest cover, deforestation, reforestation}
{banana production, production, technological change, technology}	affect	deforestation
{pasture technology, infrastructure, multiple effect, progress}	stimulate	deforestation

nation. For instance, when *technological change* is the Agent of *lead to*, typical Patients are *replanting*, *loss in forest covert* or *deforestation.*

When lexical units are described according to the patterns in their syntagmatic axis, differences in the lexical schemes across languages arise. A similar study conducted in Spanish (Sánchez-Cárdenas, 2024) showed equivalent patters such as [sobreexplotación agrarian-provocar-desertización], [pérdida de bosque-acelerar-inundaciones], and [agricultura intensive, impulsar, pérdida forestal].

In order to express processes of creation, generation, or intensification related to deforestation, the Spanish corpus uses verbs such as *provocar, ocasionar, originar, causar, producir, impulsar, acelerar*, and *generar.* For processes of destruction or degradation, it employs verbs like *destruir, arrasar, eliminar, afectar,* and *alterar.* In the English corpus, the same types of processes are lexicalized through verbs such as *lead to, drive, trigger, stimulate, encourage, spur,* and *increase* for creation or intensification, and *destroy, degrade, remove, affect*, and *aggravate* for destruction or deterioration.

This shows how each language selects different predicates to encode equivalent processes within the domain of deforestation, which reveals not only linguistic differences, but also deeper conceptual and cultural dysmorphism.

Furthermore, the interest of the information obtained through corpus-based triples goes far beyond what a bilingual dictionary can offer. If we look up the English verbs that appear in the corpus such as *drive, trigger,* or *spur,* most bilingual dictionaries provide Spanish equivalents like *conducir, activar,* or *estimular.* However, these are not the verbs actually used by the Spanish speakers specialized on this domain. Instead, verbs such as *provocar, originar,* or *causar* are more natural in this context. This mismatch demonstrates that bilingual dictionaries fail to capture the real phraseological behaviour of specialized language. We can therefore conclude that the analysis of argument structure patterns in the form of triples reveals the true functional equivalents of verbs across languages, as opposed to the semantic correspondences provided by dictionaries.

In sum, this kind of information can be useful for improving terminological resources and translation tools, but also to understand how a concept is conceptualized and lexicalized across languages and cultures.

3 From Command-Line Scripts to a Web-Based Prototype: Design and Management of the Tool

Designing and managing terminological resources requires a clear definition of scope, domain, languages, and user groups, adherence to core terminological principles, and systematic structuring to ensure consistency, interoperability, and long-term usability. The original scripts fell short in usability, which motivated the transition to a web-based prototype.

3.1 User needs

From a user-oriented perspective, MarcoTAO is conceived as a functional resource whose relevance depends on specific user profiles and needs. Following the functional theory of lexicography (Bergenholtz & Tarp, 2010), it is conceived as a resource whose adequacy depends on specific user profiles and usage situations. The tool addresses different cognitive and communicative needs, depending on whether the user is a researcher, a professional translator or a student. The main user profiles that could benefit from this resource are researchers, professional translators, and students in advanced translation or linguistic programs.

First, researchers in linguistics, terminology or lexicography constitute the primary target users. For this group, the tool supports corpus exploration and the extraction of lexicographic data in the form of argument structures, which identify recurrent phraseological patterns across languages. In particular, clustering lexical triples is relevant for researchers whose objective is to have an in-depth understanding of specialized language and underlying conceptual patterns.

Secondly, professional translators represent a different user profile, since their needs are mostly communicative and often characterized by time constraints. However, this tool is not intended to function as a fast answer resource. Rather, it supports the preparatory phase of a specialized translation that needs an extensive documentation on phraseological recurrent patterns, which can subsequently feed the database of the translator.

Lastly, students in advanced translation or linguistic programs constitute a third relevant user profile. In an educational context, the tool can be used to support guided exploration of specialized corpora, helping students identify differences in recurring patterns across languages. Nevertheless, the use of this tool requires supervision and a certain learning investment.

3.2 From Scripts to Prototype

MarcoTAO was initially developed as command-line scripts running on user machines, which limited usability and sharing. The main goal of the prototype was to centralize execution and data storage on a server, accessible from any device with an Internet connection and browser. This ensured cross-platform compatibility, extensibility for future scripts, and simplified maintenance. The prototype offers an intuitive and accessible interface, and supports key tasks such as corpus parsing and indexing, triple-based searches, clustering, and project-specific user management. A structured database and a secure file system safeguard user-generated content.

3.3 Development Methodology

The prototype followed a user-centered, iterative design approach (Wallach & Scholz, 2012; ISO 9241-210, 2010), delivering progressively refined versions informed by feedback (Amos et al., 2011). This methodology, suited to contexts where end users cannot provide exhaustive requirements (Maida & Pacienzia, 2015), proved effective for the project's multidisciplinary team of developers and specialists in terminology, semantics, and knowledge representation, supported by the "Laboratoire d'information et des systems" (LIS) computing infrastructure. Iterative prototyping facilitated communication across disciplines, reducing epistemological gaps and aligning design with cognitive and communicative needs.

Co-design (Schewe & Thalheim, 2005) further ensured active participation of experts, fostering mutual understanding and iterative refinement of both interface and functionalities. This participatory approach not only improved usability but also supported long-term adoption (Sanders & Stappers, 2008). The phases of development followed standard prototyping cycles — investigation, requirements, rapid design, iterative testing, and deployment (Sharma, 2022) — with clear versioning goals and evaluation checkpoints to mitigate risks of undefined iteration or scope creep.

3.4 Functional and Non-functional Requirements

The system architecture was organized into six modules: user, project, concept, language, corpus, and search. Administrators can manage accounts, while authenticated users can create projects, associate concepts (e.g. DEFORESTATION), upload and process corpora, and perform triple-based searches. Features include clustering, exporting results, and reusable formats, all within an integrated interface.

Non-functional requirements include compliance with data protection regulations, performance and availability, integrity via backup mechanisms, cross-platform compatibility, and robust error handling with user feedback.

3.5 Design and Architecture

The design process progressed from sketches to digital mock-ups, with emphasis on usability (e.g., collapsible sidebar menus, profile images). The architecture follows a client–server model (Oluwatosin, 2014): browsers send HTTP requests, the server executes PHP scripts, queries a MySQL database, and integrates Python scripts for corpus processing, triple extraction, filtering, and clustering. Results are stored in structured formats (TSV, JSON) and displayed on the frontend, ensuring consistency and scalability. Running all scripts on the server allows users to work from any device without installing dependencies.

3.6 Accessibility and Usability

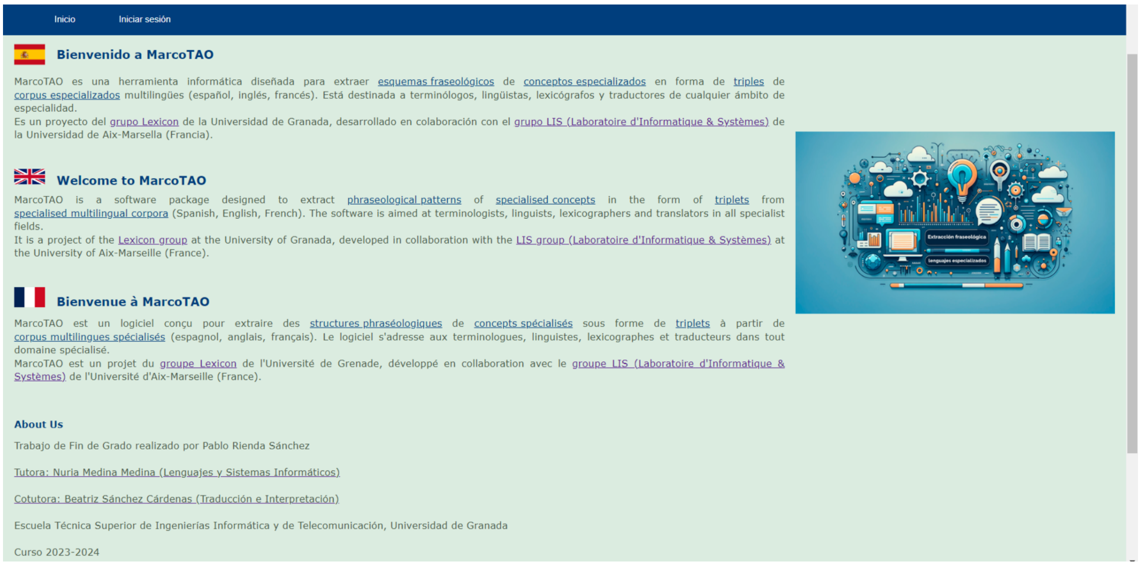
MarcoTAO was developed in line with W3C accessibility (Caldwell et al., 2008) and usability principles (Bruno et al., 2005; Nielsen & Loranger, 2006). Accessibility features include semantic HTML, alternative text for images, and avoidance of flashing content. Usability is ensured through intuitive navigation, consistent structure, clear visual hierarchy, and effective error feedback. Together, these principles increase inclusivity, productivity, and data quality, while responsive design ensures full functionality across devices and screen sizes.

4 Design of MarcoTAO: Towards a Web User Interface[1]

The analysis protocol described in previous sections cannot yet be widely used by other researchers, since it is composed of several command-line scripts that must be executed separately. The whole process is error prone and lacks user-friendliness. To address these limitations and make the whole process available to academia, we developed a web interface, currently in the prototype phase, that encapsulates the existing scripts for all the phrases described above. Tutorials are made available to users to support them in completing each task.

During the development of the MarcoTAO web prototype, several technical challenges were addressed that proved decisive for its correct functioning. The most relevant concerned script installation and configuration, visualization of processed results, management of directories and permissions, and the processing of large corpora.

[1] The included screenshots aim to demonstrate the functionalities of the prototype. At this stage, the linguistic content shown is illustrative, and does not reflect the final output quality expected.



Bienvenido a MarcoTAO

MarcoTAO es una herramienta informática diseñada para extraer esquemas fraseológicos de conceptos especializados en forma de triples de corpus especializados multilingües (español, inglés, francés). Está destinada a terminólogos, lingüistas, lexicógrafos y traductores de cualquier ámbito de especialidad.
Es un proyecto del grupo Lexicon de la Universidad de Granada, desarrollado en colaboración con el grupo LIS (Laboratoire d'Informatique & Systèmes) de la Universidad de Aix-Marsella (Francia).

Welcome to MarcoTAO

MarcoTAO is a software package designed to extract phraseological patterns of specialised concepts in the form of triplets from specialised multilingual corpora (Spanish, English, French). The software is aimed at terminologists, linguists, lexicographers and translators in all specialist fields.
It is a project of the Lexicon group at the University of Granada, developed in collaboration with the LIS group (Laboratoire d'Informatique & Systèmes) at the University of Aix-Marseille (France).

Bienvenue à MarcoTAO

MarcoTAO est un logiciel conçu pour extraire des structures phraséologiques de concepts spécialisés sous forme de triplets à partir de corpus multilingues spécialisés (espagnol, anglais, français). Le logiciel s'adresse aux terminologues, linguistes, lexicographes et traducteurs dans tout domaine spécialisé.
MarcoTAO est un projet du groupe Lexicon de l'Université de Grenade, développé en collaboration avec le groupe LIS (Laboratoire d'Informatique & Systèmes) de l'Université d'Aix-Marseille (France).

About Us

Trabajo de Fin de Grado realizado por Pablo Rienda Sánchez

Tutora: Nuria Medina Medina (Lenguajes y Sistemas Informáticos)

Cotutora: Beatriz Sánchez Cárdenas (Traducción e Interpretación)

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación, Universidad de Granada

Curso 2023-2024

Fig. 4 Welcome screen of MarcoTAO prototype

First, script installation was complex due to dependencies on specific Python libraries. Version heterogeneity and server incompatibilities required a rigorous validation procedure. A testing environment on an Ubuntu-based virtual machine replicated the execution setting, enabling early detection of conflicts and systematic documentation. This process produced a technical installation manual and, subsequently, an end-user guide for researchers. Ensuring correct visualization of results required a mechanism combining selective read permissions with dynamic navigation, so that users could access only outputs linked to their projects, thus safeguarding data security and privacy.

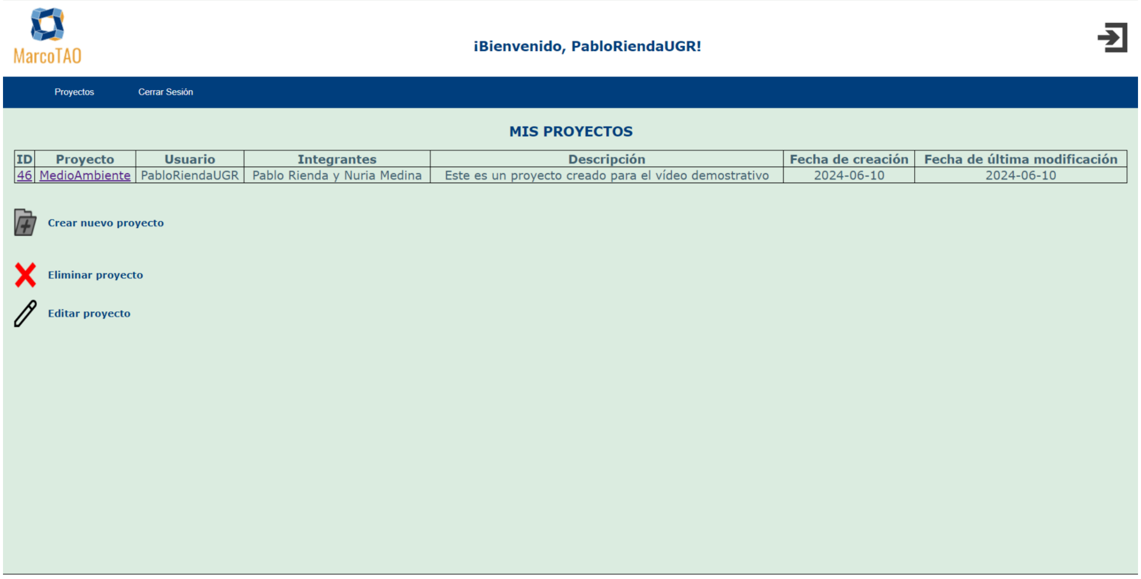
The hierarchical creation of user directories also posed challenges. A tree-shaped structure was implemented: each new user was assigned a root directory, with nested subdirectories for projects, concepts, corpora, and generated files. This strategy demanded careful management of permissions, to prevent cross-user access and preserve data integrity.

Finally, processing large corpora often caused excessive execution times, particularly during indexing. A "divide and conquer" strategy segmented corpora into smaller subsets linked to the same concept. This reduced execution time and improved efficiency. An aggregation script then consolidated the partial results into a unified table, enabling proper clustering.

The MarcoTAO prototype is organized into several interactive web pages. The welcome page of the application (Figure 4) includes a header with the name and logo of the tool. The main content area presents general information about the tool in multiple languages, together with an "About Us" section describing the academic context of the project. Key terms in the multilingual text display short definitions as tooltips. The login page provides a basic authentication form, redirecting users to a personalized dashboard. Administrators have access to additional account-management functions — registration, password updates, and deletion — validated with JavaScript to ensure correct input (e.g., passwords with at least eight characters).

The project overview page (Figure 5) displays a list of existing projects, with details such as concept studied, project description, creation or modification date. Users can create, edit, or delete projects using collapsible forms triggered by clicking on corresponding icons. JavaScript handles form validation and ensures that required fields are completed before submission.

In the current version of the prototype, users can only view and manage their own projects, since each project is associated with a single user account and cannot be collaboratively edited. While it is technically possible to share access by using a common user account, multiuser collaboration is not supported in the current implementation and would require further development, particularly to address issues related to simultaneous access and data consistency.



¡Bienvenido, PabloRiendaUGR!

MIS PROYECTOS

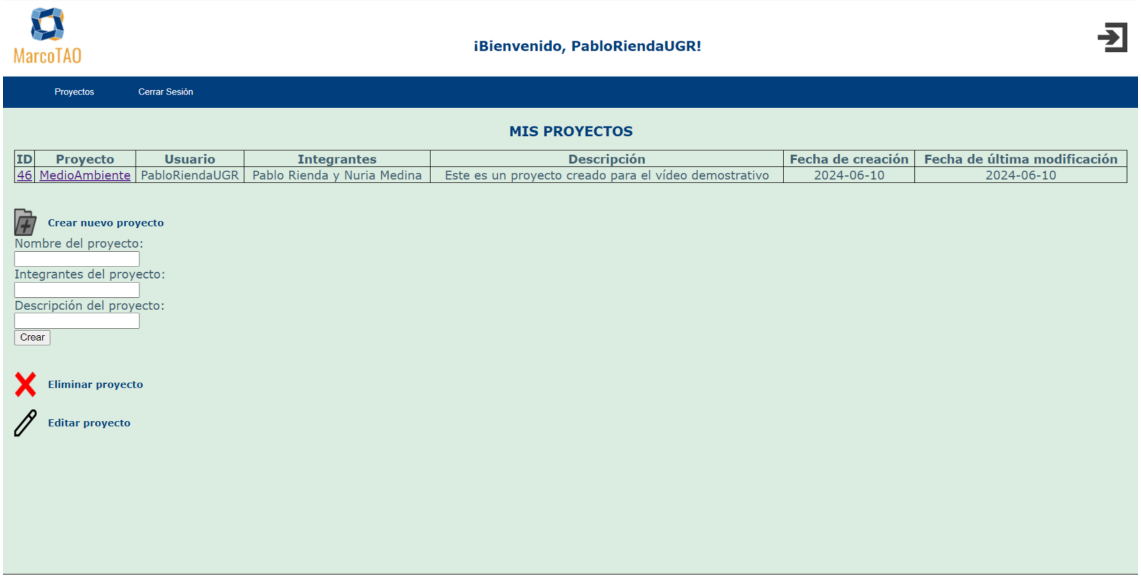
ID	Proyecto	Usuario	Integrantes	Descripción	Fecha de creación	Fecha de última modificación
46	MedioAmbiente	PabloRiendaUGR	Pablo Rienda y Nuria Medina	Este es un proyecto creado para el vídeo demostrativo	2024-06-10	2024-06-10

Crear nuevo proyecto

Eliminar proyecto

Editar proyecto

Fig. 5 Concept project management screen of the prototype tool



¡Bienvenido, PabloRiendaUGR!

MIS PROYECTOS

ID	Proyecto	Usuario	Integrantes	Descripción	Fecha de creación	Fecha de última modificación
46	MedioAmbiente	PabloRiendaUGR	Pablo Rienda y Nuria Medina	Este es un proyecto creado para el vídeo demostrativo	2024-06-10	2024-06-10

Crear nuevo proyecto

Nombre del proyecto:

Integrantes del proyecto:

Descripción del proyecto:

Eliminar proyecto

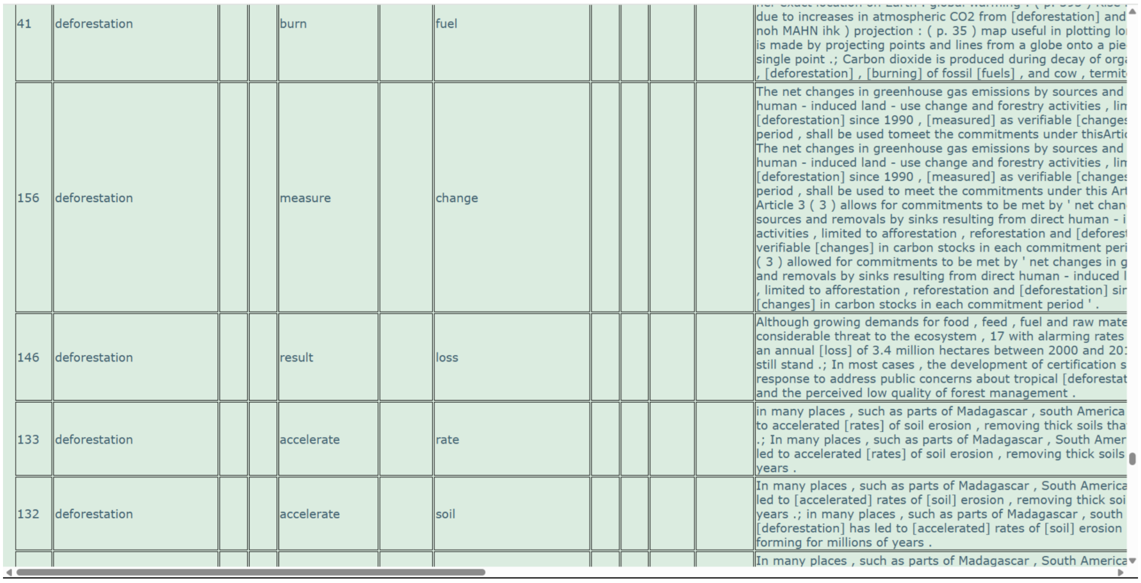
Editar proyecto

Fig. 6 Triple search interface

Each project has its own dedicated page, where users can view and manage associated concepts. Selecting a concept opens another view that displays its corpora. Clicking on a corpus element reveals the text in a scrollable area with size constraints for readability. Users can also upload new corpora oredit existing ones using provided interfaces. JavaScript is used to validate inputs such as language codes and file formats. The application also includes a section for uploading processed corpora.

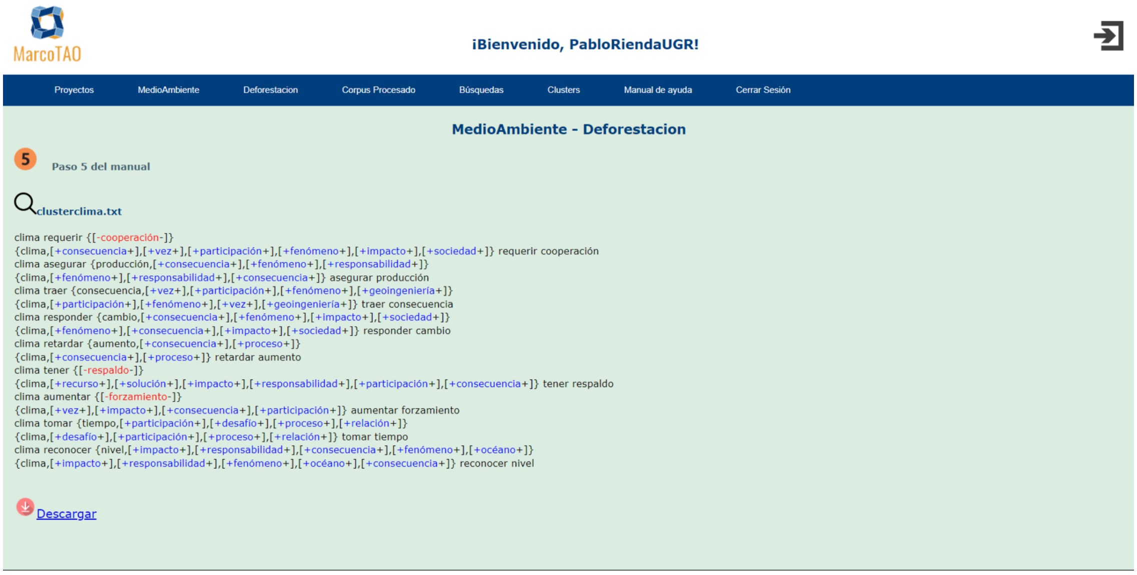
Beyond this interface, users can upload raw text corpora, which are then parsed and indexed. Users are then redirected to the search interface (Figure 6), where queries are based on the [noun–verb–noun] structure. In these queries, two positions must be specified, while one can be underspecified with a wildcard. It is possible to write the lexical units, or to upload them from a predefined list. It is also possible to define elements in a blacklist or to access items stored from previous results.

The search results page (Figure 7) allows users to view and download the triples generated by the search scripts, together with the originating concordance.



41	deforestation			burn		fuel			her exact location on Earth : global warming : (p. 393) Rise due to increases in atmospheric CO2 from [deforestation] and noh MAHN ihk) projection : (p. 35) map useful in plotting lo is made by projecting points and lines from a globe onto a pie single point .; Carbon dioxide is produced during decay of orga , [deforestation] , [burning] of fossil [fuels] , and cow , termit
156	deforestation			measure		change			The net changes in greenhouse gas emissions by sources and human - induced land - use change and forestry activities , lin [deforestation] since 1990 , [measured] as verifiable [changes period , shall be used tomeet the commitments under thisArtic The net changes in greenhouse gas emissions by sources and human - induced land - use change and forestry activities , lin [deforestation] since 1990 , [measured] as verifiable [changes period , shall be used to meet the commitments under this Art Article 3 (3) allows for commitments to be met by ' net chan sources and removals by sinks resulting from direct human - i activities , limited to afforestation , reforestation and [deforest verifiable [changes] in carbon stocks in each commitment peri (3) allowed for commitments to be met by ' net changes in g and removals by sinks resulting from direct human - induced l , limited to afforestation , reforestation and [deforestation] sir [changes] in carbon stocks in each commitment period ' .
146	deforestation			result		loss			Although growing demands for food , feed , fuel and raw mate considerable threat to the ecosystem , 17 with alarming rates an annual [loss] of 3.4 million hectares between 2000 and 20: still stand .; In most cases , the development of certification s response to address public concerns about tropical [deforestat and the perceived low quality of forest management .
133	deforestation			accelerate		rate			in many places , such as parts of Madagascar , south America to accelerated [rates] of soil erosion , removing thick soils tha .; In many places , such as parts of Madagascar , South Amer led to accelerated [rates] of soil erosion , removing thick soils years .
132	deforestation			accelerate		soil			In many places , such as parts of Madagascar , South America led to [accelerated] rates of [soil] erosion , removing thick soi years .; in many places , such as parts of Madagascar , south [deforestation] has led to [accelerated] rates of [soil] erosion forming for millions of years .
									In many places , such as parts of Madagascar , South America

Fig. 7 List of extracted triples related to DEFORESTATION (EN) in the prototype



¡Bienvenido, PabloRiendaUGR!

MedioAmbiente - Deforestacion

Paso 5 del manual

clusterclima.txt

```
clima requerir {[-cooperación-]}
{clima,[+consecuencia+],[+vez+],[+participación+],[+fenómeno+],[+impacto+],[+sociedad+]} requerir cooperación
clima asegurar {producción,[+consecuencia+],[+fenómeno+],[+responsabilidad+]}
{clima,[+fenómeno+],[+responsabilidad+],[+consecuencia+]} asegurar producción
clima traer {consecuencia,[+vez+],[+participación+],[+fenómeno+],[+geoingeniería+]}
{clima,[+participación+],[+fenómeno+],[+vez+],[+geoingeniería+]} traer consecuencia
clima responder {cambio,[+consecuencia+],[+fenómeno+],[+impacto+],[+sociedad+]}
{clima,[+fenómeno+],[+consecuencia+],[+impacto+],[+sociedad+]} responder cambio
clima retardar {aumento,[+consecuencia+],[+proceso+]}
{clima,[+consecuencia+],[+proceso+]} retardar aumento
clima tener {[-respaldo-]}
{clima,[+recurso+],[+solución+],[+impacto+],[+responsabilidad+],[+participación+],[+consecuencia+]} tener respaldo
clima aumentar {[-forzamiento-]}
{clima,[+vez+],[+impacto+],[+consecuencia+],[+participación+]} aumentar forzamiento
clima tomar {tiempo,[+participación+],[+desafío+],[+proceso+],[+relación+]}
{clima,[+desafío+],[+participación+],[+proceso+],[+relación+]} tomar tiempo
clima reconocer {nivel,[+impacto+],[+responsabilidad+],[+consecuencia+],[+fenómeno+],[+océano+]}
{clima,[+impacto+],[+responsabilidad+],[+fenómeno+],[+océano+],[+consecuencia+]} reconocer nivel
```

Descargar

Fig. 8 Automatically clustered triples related to CLIMA (ES) based on distributional similarity

Search results are then grouped into lexical clusters, which can be visualized and stored. The cluster visualization page (Figure 8) provides an interface for interpreting grouped search results through clustering techniques. Each cluster is displayed with a color code scheme in order to improve readability and semantic differentiation. Users can inspect both the list of generated clusters and the specific contents assigned to each group.

In addition to the graphical interface, the application includes a command-line script system designed to process corpora and extract linguistic data. These scripts are executed via SSH access to the server, following a sequential workflow.

This combination of user-friendly design and modular architecture enables both ease of use and advanced functionality, supporting reusability and customization across different linguistic domains and research contexts.

The proposed analyses aim to identify cross-linguistic differences in specialized events, rather than isolated lexical correspondences. Focusing on DEFORESTATION, the extraction of recurrent argument structures reveals different semantic dimensions of the same concept, such as causation, agency, impact or mitigation. For instance, Spanish frequently lexicalizes DEFORESTATION through verbs expressing direct causation or damage (*provocar deforestación, causar la deforestación, arrasar bosques*), typically following patterns such as [*expansión agrícola – provocar – deforestación*] or [*actividades humanas – destruir – bosques*]. By contrast, English shows a wider use of constructions encoding processes and indirect causation (*lead to deforestation, contribute to deforestation, drive deforestation*), as well as desagentivised patterns in which the event itself functions as subject, for example agricultural expansion (*contribute to deforestation; deforestation – affect – biodiversity*). These differences emerge from recurrent patterns observed across comparable specialized corpora.

From a translation perspective, such contrasts are relevant because they reflect lexical and constructional preferences that directly affect accuracy and idiomaticity. A literal transfer of a Spanish causative construction (*la agricultura intensiva provoca la deforestación*) into English may result in a less idiomatic rendering if process-oriented patterns (*intensive agriculture contributes to deforestation*) are preferred in the target discourse. Making these recurrent patterns explicit helps translators anticipate shifts in causality and agency that are important in order to produce natural and idiomatic target texts.

5 Concluding Remarks

This article has presented the development of the MarcoTAO prototype, a web-based environment that evolved from a set of command-line scripts into an accessible and extensible tool for linguistic research. Guided by iterative, user-centered, and co-design methodologies, the system integrates principles of accessibility, usability, and interoperability while addressing key technical challenges in deployment, visualization, and large-corpus processing. The resulting prototype offers an intuitive interface and robust architecture that enable users to upload, parse, and index corpora, perform triple-based queries, and cluster results for further analysis.

A previous comparison with Sketch Engine suggested that MarcoTAO may produce extractions with less noise and greater precision, while exploratory tests with ChatGPT highlighted the potential of LLM-based prompting strategies for similar tasks without corpus analysis (Sánchez-Cárdenas, 2024). Together, these findings underscore the complementary strengths of corpus-driven and LLM-driven approaches, and point toward future research avenues that combine both methods to enhance translation, terminology management, and linguistic analysis.

Future research will enhance triple extraction by addressing current challenges and incorporating new strategies. From a linguistic point of view, improvements include handling complex nouns, argument structures with more than two complements, negations, or phrasal verbs.

Concerning the prototype development, one of the main tasks planned for future development is migrating the application to a private server, a transition for which the system has been technically prepared. Another important line of work is the continuous improvement of web accessibility. Although the current implementation of the prototype MarcoTAO meets level AA compliance in the TAW test (Fundacion CTIC, n.d.), the long-term goal is to reach AAA certification. This would involve refining various elements of the interface and interaction design to ensure broader accessibility for users with diverse needs.

A major technical improvement involves enabling server-side script execution directly from the web interface. This functionality would simplify the workflow, allowing users — especially those with less technical expertise — to process and analyze data without relying on a terminal connection. To make this possible, the application will need to be hosted on a more robust server with increased storage capacity and optimized script management. In addition, some additional scripts not covered in this prototype will be implemented in future development stages.

In terms of usability, the current system would benefit from formal usability testing with real users. This will allow developers to identify and address specific issues using the five usability attributes proposed by Nielsen and Loranger (2006): learnability, efficiency, memorability, errors, and satisfaction.

Evaluating the application along these dimensions will help improve its interface and user experience, ensuring that the platform is both powerful and user-friendly.

Declaration on Generative AI

During the preparation of this work, the authors used X-GPT-4 in order to improve readability and avoid redundance, check APA style in references and grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of this publication.

Acknowledgements

This work has been carried out within the framework of the project Transversal integration of culture in an environmental terminology knowledge base (TRANSCULTURE, PID2020-118369GB-I00), funded by the Spanish Ministry of Science and Innovation.

This work received support from the CA21167 COST action UniDive, funded by the European Union via COST (European Cooperation in Science and Technology), and from the French Agence Nationale pour la Recherche, via the SELEXINI project (ANR-21-CE23-0033-01).

References

Amos, D., Lesea, A., & Richter, R. (2011). *FPGA-based Prototyping Methodology Manual: Best Practices in Design-for-Prototyping*. Synopsis Press.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (pp. 1313–1316*)*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2004/pdf/509.pdf

Bergenholtz, H., & Tarp, S. (2010). Chapter 1. LSP lexicography or terminography? The lexicographer's point of view. In P. A. Fuertes-Olivera (Ed.), *Specialised Dictionaries for Learners* (pp. 27–38). De Gruyter. https://doi.org/10.1515/9783110231335.1.27

Bruno, V., Tam, A., & Thom, J. (2005). Characteristics of web applications that affect usability: A review. In *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future* (pp. 1–4). Computer-Human Interaction Special Interest Group (CHISIG) of Australia. https://dl.acm.org/doi/10.5555/1108368.1108445

Buendía Castro, M. (2013). *Phraseology in specialized language and its representation in environmental knowledge resources* (Doctoral dissertation, Universidad de Granada).

Buendía Castro, M., & Sánchez-Cárdenas, B. (2016). Using argument structure to disambiguate verb meaning. In T. Margalitadze & G. Meladze (Eds.), *Proceedings of the 17th EURALEX International Congress* (pp. 482–490). Ivane Javakhishvili Tbilisi University Press. https://euralex.org/publications/using-argument-structure-to-disambiguate-verb-meaning/

Caldwell, B., Cooper, M., Reid, L. G., Vanderheiden, G., Chisholm, W., Slatin, J., & White, J. (Eds.) (2008). *Web content accessibility guidelines (WCAG) 2.0.* WWW Consortium (W3C). https://www.w3.org/TR/WCAG20/

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: Los retos de un nuevo paradigma*. Peter Lang. https://www.peterlang.com/document/1105581

Faber, P. (Ed.). (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. De Gruyter Mouton. https://doi.org/10.1515/9783110277203

Fundación CTIC. (n.d.). *TAW: Test de accesibilidad web*. Retrieved July 18, 2025, from https://www.tawdis.net

Granger, S., & Meunier, F. (Eds.). (2008). *Phraseology: An Interdisciplinary Perspective*. John Benjamins Publishing Company. https://doi.org/10.1075/z.139

Harris, Z. S. (1971). *Structures Mathématiques du Langage*. Dunod.

ISO 9241-210 (2010). *Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems*. International Organization for Standardization. https://www.iso.org/standard/52075.html

Jacques, M.-P., & Tutin, A. (Dir.). (2018). *Lexique Transversal et Formules Discursives des Sciences Humaines*. ISTE Éditions.

L'Homme, M.-C. (2015). Predicative lexical units in terminology. In N. Gala, R. Rapp, & G. Bel-Enguix (Eds.), *Language Production, Cognition, and the Lexicon* (pp. 75–93). Springer. https://doi.org/10.1007/978-3-319-08043-7_6

L'Homme, M.-C. (2020). *Lexical Semantics for Terminology: An Introduction*. John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.20

Maida, E. G., & Pacienzia, J. (2015). *Metodologías de desarrollo de software*. (Bachelor's thesis, Universidad Católica Argentina). https://repositorio.uca.edu.ar/handle/123456789/522

Mel'čuk, I. (1998). Collocations and lexical functions. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications* (pp. 23–53). Clarendon Press Oxford. https://doi.org/10.1093/oso/9780198294252.003.0002

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781

Muñoz-Basols, J., Palomares Marín, M. del M., & Moreno Fernández, F. (2024). El sesgo lingüístico digital (SLD) en la inteligencia artificial: Implicaciones para los modelos de lenguaje masivos en español. *Lengua y Sociedad*, 23(2), 623–647. https://doi.org/10.15381/lengsoc.v23i2.28665

Nielsen, J., & Loranger, H. (2006). *Prioritizing Web Usability*. New Riders.

Oluwatosin, H. S. (2014). Client-server model. *IOSR Journal of Computer Engineering*, 16(1), 67–71.

Orliac, B. (2006). Colex: Un outil d'extraction de collocations spécialisées basé sur les fonctions lexicales. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 12(2), 261–280. https://doi.org/10.1075/term.12.2.06orl

Pilehvar, M. T., & Camacho-Collados, J. (2021). Word embeddings. In *Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning* (pp. 25–40). Springer. https://doi.org/10.1007/978-3-031-02177-0_3

Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer. https://doi.org/10.1007/978-3-319-09207-2

Ramisch, C. (2023). *Multiword Expressions in Computational Linguistics: Down the Rabbit Hole and Through* *the Looking Glass* (Habilitation thesis, Aix-Marseille University).

Sánchez-Cárdenas, B. (2024). Extracting semantic frames from specialized corpora for lexicographic purposes. *Círculo de Lingüística Aplicada a la Comunicación*, 99, 163–177. https://doi.org/10.5209/clac.90626

Sánchez-Cárdenas, B., & Buendía Castro, M. (2012). Inclusion of verbal syntagmatic patterns in specialized dictionaries: The case of EcoLexicon. In R. V. Fjeld & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 554–562). Department of Linguistics and Scandinavian Studies, University of Oslo. https://euralex.org/publications/inclusion-of-verbal-syntagmatic-patterns-in-specialized-dictionaries-the-case-of-ecolexicon/

Sánchez-Cárdenas, B., & Ramisch, C. (2019). Eliciting specialized frames from corpora using argument-structure extraction techniques. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 25(1), 1–31. https://doi.org/10.1075/term.00026.san

Sanders, E. B.-N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *CoDesign*, 4(1), 5–18. https://doi.org/10.1080/15710880701875068

Schewe, K.-D., & Thalheim, B. (2005). The co-design approach to Web information systems development. *International Journal of Web Information Systems*, 1(1), 5–14. https://doi.org/10.1108/17440080580000078

Sharma, P. (2022). *Los 9 mejores modelos de desarrollo de software para elegir: fases y aplicaciones*. Cynoteck Technology Solutions. https://www.cynoteck.com/blog-post/top-software-development-models-to-choose-from

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In D. Zeman & J. Hajič (Eds.), *Proceedings of the CoNLL 2018 Shared Task*: Multilingual Parsing from Raw Text to Universal Dependencies (pp. 197–207). Association for Computational Linguistics. https://doi.org/10.18653/v1/K18-2020

Tutin, A., & Grossmann F. (Eds.). (2014). *L'écrit scientifique: Du lexique au discours. Autour de Scientext*. Presses Universitaires de Rennes.

Vezzani, F. (2023). Vers une méthodologie pour l'extraction et la classification automatiques des collocations terminologiques verbales en langue médicale. In P. Frassi (Ed.), *Phraséologie et terminologie* (pp. 259–278). De Gruyter. https://doi.org/10.1515/9783110749854-013

Wallach, D., & Scholz, S. C. (2012). User-centered design: Why and how to put users first in software development. In A. Maedche, A. Botzenhardt, & L. Neer (Eds.), *Software for People: Fundamentals, Trends and Best Practices* (pp. 11–38). Springer. https://doi.org/10.1007/978-3-642-31371-4_2