Towards a Tool for Extracting Specialized Argument Structures*

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

[1] *Universidad de Granada, calle Buensuceso, 11, 18002, España*
[2] *Universidad de Granada, avenida del Hospicio, 1, 18010, España*
[3] *Universidad de Granada, calle Periodista Daniel Saucedo Aranda S/N 18071 España*
[4] *Aix Marseille Université, CNRS, LIS, Marseille, France*

Abstract

This contribution presents the design and development of MarcoTAO, 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, MarcoTAO 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, MarcoTAO addresses the limitations of previous tools by streamlining corpus analysis and making phraseological studies more accessible to academia.

Keywords

paper template, paper formatting, CEUR-WS

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 2005, Granger and Meunier 2008, Faber 2012). Indeed, the production of texts relies on managing structured lexico-grammatical constructs, such as phraseology (Corpas Pastor, 2008; Tutin, 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 and Tutin, 2018), or verb-noun combinations (*to eject lava*) (Buendía Castro and Sánchez-Cárdenas, 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, 2013; Buendía and Sánchez Cárdenas 2016) that form an argumental structure. For instance, in Environmental Sciences, verbs such as *drive, encourage* or *provoke* are often used to describe the causes of *deforestation*. Interestingly, translating these verbs into other languages requires not only knowing the target

* *4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece.*
[1*] Corresponding author.
[†] These authors contributed equally.
bsc@ugr.es (B. Sánchez); prienda@correo.ugr.es (P. Rienda); nmedina@ugr.es (N. Media); carlos.ramisch@lis-lab.fr (C. Ramisch)
0002-1904-675X (B. Sánchez); 0002-6013-732X (N. Media); 0001-7466-9039 (C. Ramish)

language but also how scientist in this domain express such a process. As a result, translating just the verb, rather than the whole semantic structure, might lead to non-idiomatic sentences. Such information is typically absent from general dictionaries and even from most terminological resources, but it can be found in specialized corpora. By adopting this approach, we are more likely to identify appropriate equivalents to express the consequences of deforestation, such as *agravar, ocasionar* or *alterar* in Spanish.

Interestingly, TAN does not always produce satisfactory results for this issue, as it tends to translate verbs into their formal equivalents (*provoque/provocar; cause/causar*) rather than domain-specific ones. This might be partly due to the fact that general machine translation tools are trained 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. This is known as the "Digital Linguistic Bias" (DLB) (Muñoz-Basols et al., 2024). As a result,

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. 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 TAN and IA. 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 triples form corpora in the form of "noun-verb-noun" structures, called triples, reflecting the argumental structure of a given concept across several languages. For instance: [**Soybean expansion**] in southern Brazil [**contributed**] to [**deforestation**] by stimulating migration to agricultural frontier regions.

In order to simply this complex task, we designed a tool prototype that automates the extraction of [noun-verb-noun] structures from specialized corpora in multiple languages. It is an easy-to-use web interface designed to help researchers and linguists analyze and this type of linguistic information more efficiently.

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

Nevertheless, the process of triple extractions can be achieved by employing a range of corpus tools being able to identify argument structures. Previous research (Sánchez Cárdenas 2024) compared the performance of 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. Our study concluded that both tools had challenges in terms of noise in the retrieved triples. Sketch Engine had a higher percentage of noise (90.9%), while MWEtoolkit had a relatively lower percentage (34.4%). Additionally, MWEtoolkit achieved a higher percentage of accurate triples (65.5%) compared to Sketch Engine (9.1%).

In Section 2, we describe the protocol for extracting argumental 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 argumental structures

In previous research (Sánchez Cárdenas & Ramsich 2019; Sánchez Cárdenas 2024), we employed MWEtoolkit[2], a computational tool for the identification of multiword expressions in corpora (Ramisch 2015, 2023) in order to isolate triples [noun1-verb-noun2] representing argumental

[2] http://mwetoolkit.sourceforge.net

structures. For this endeavor, queries were designed through Python scripts in order to process and query the corpora, extract candidates and sort the results.

2.1. Processing the corpora

During preprocessing phase, texts were automatically converted to UTF-8. They were processed and analyzed using UDPipe, a natural language processing tool. It performed the following tasks: tokenizing sentences into words, tagging words with their part-of-speech (POS) using the Universal Dependencies tagset, assigning lemmas, and generating syntactic dependency trees to map relations between words.

2.2. Querying the corpora

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]. These searches captured argumental structures, but also irrelevant triples such as [volcano-see-lava], which required manual filtering. To streamline the process, searches were encapsulated in shell scripts for easier and more efficient execution.

Future research will enhance triple extraction by addressing current challenges and incorporating new strategies. Improvements include handling complex nouns, argumental structures with more than two complements, negations or phrasal verbs.

Step 2: Search strategies

In order to test the validity of the scripts, a pilot study was conducted.

Initial queries were constructed using seed terms extracted from the corpora. In previous pilot studies within the domain of environmental sciences, these seeds terms were derived from EcoLexicon[3], a knowledge data base. Specifically, the semantic relations between concepts were used to identify verbs lexicalizing those semantic relation. For instance, the query pattern [volcano-?-lava] 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.

Step 3: Filtering and sorting Results

Triples were automatically ranked by relevance using pointwise mutual information (PMI), calculated from co-occurrence frequencies. Results were sorted in descending order of relevance, with the most significant triples prioritized for further analysis. Encapsulated scripts simplified queries, and output was stored in tsv files for manual review. Finally, Triples were gathered into a single tsv file and manually ranked using a code from 0 to 4: 0 (accurate), 1 (acceptable with minor manual modifications), 2 (irrelevant but potentially useful for refining future searches), and 4 (incorrect).

2.3. Distributional clustering of triples

The final step involves the distributional clustering of triples marked as 0 or 1. A specific script organizes these results into clusters, grouping similar triples into linguistic schemas based on shared patterns. For instance: (volcano, expel, {lava, magma, rock}) or ({volcano, crater}, expel, lava). These patterns show common phraseological structures in the domain.

---

[3] http://ecolexicon.ugr.es/es/index.htm

To that end, we use distributional semantics via Word2vec, where words in the triples are represented as vectors based on their co-occurrence context in the corpus. Triples are automatically grouped using a semantic similarity measure based on word embeddings (Pilehvar & Camacho-Collados 2021) with Gensim software.

The system removes words that are infrequent in the patterns and includes words that are semantically close to the group. A word is added only if its average similarity to the group is above a threshold, usually 0.3.

As a result, all possible combinations of nous and verbs that appear in each position of the triples are generated. This process results in the recurring phraseological-verb-nous argument structures of the analyzed concept. This lexical clustering organizes the extracted triples in a way that is useful from a terminological point of view. In fact, it highlights productive lexical patterns relevant for domain-specific phraseology. This could contribute to the development of terminological resources and improve translation quality and scientific text production.

Figure 1 shows the raw output for the clustering of VOLCANO. Needless to say, this information requires some manual refinement. Table 2 presents the lexical cluster of DEFORESTATION after manual treatment.

```
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+],
```

Figure 1. Lexical clustering of VOLCANO (EN)

Noun phrase 1	Verb	Noun phrase 2
{transportation_cost, transport costs, technology, fallow, forest_cover, technological_change,	increase	deforestation
{cattle ranching, population concentration, production, introduction of new crop, timber, technological progress}	lead to	deforestation
{forest clearing, infrastructure project, forestry, cocoa, technological change,	accelerate	deforestation
{highway construction, technological_change, crop production, soybean expansion, agricultural	lead_to	deforestation
{crop,development,technological_change,technological	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

Table 1. Lexical clustering of DEFORESTATION (EN) after manual analysis

This kind of information is highly relevant for encoding texts in a specialized domain. In fact, when comparing these lexical schemas with those referring to the same concept in other languages, differences in the lexical schemes across languages become evident. This reveals not only linguistic differences, but also conceptual and cultural dysmorphism. This kind of information is important for improving terminological resources and translation tools, but also to understand how a concept is conceptualized across languages and cultures.

3. Design of MarcoTAO: towards a web user interface[4]

However, this protocol cannot be yet 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 the academia, we developed a web interface, currently in the prototype phase, that encapsulates the existing scripts for all the phrases described above. The MarcoTAO prototype is capable of: Loading, parsing and indexing corpora; Searching for "noun-verb-noun" triples in the indexed corpora; Grouping search results according to similar annotations; Creating lexical clustering; Visualizing and storing the results. The interface is accessible, user-friendly and compatible with various operating browsers. The interface shown in Figure 2 allows users to design triple searches starting with two elements. For instance, the query [*deforestation – provoke –* ?] generates results such as floods, droughts, or desertification. Figure 4 shows a screen shot of the search interface and figure 6 illustrates the lexical clusters obtained when analyzing the concept CLIMA in Spanish.

Additionally, users can perform bulk searches using word lists. This feature enables the inclusion of denominative variants of a term (e.g., *deforestation, logging, forest loss*) or verbs that express the same concept (e.g., *cause, provoke, generate*). A key advantage of the tool is its ability to seamlessly incorporate the results of one search into subsequent queries.

Concerning the technical details, the MarcoTAO prototype uses a client-server architecture. All data, including user credentials, project information, and analysis results, are stored in a MySQL database on the server.

The frontend uses standard web technologies such as HTML, CSS, and JavaScript. These allow the interface to change dynamically depending on the user, the selected project, or the current analysis step. On the server side, PHP handles the backend logic. It also manages the connection with the database and prepares the content before it is sent to the client.

The application runs scripts in Python by generating shell commands from the backend. These commands start Python scripts on the server. They perform tasks such as preprocessing the corpus, extracting triples, filtering results, and creating clusters. The backend prepares the input, launches the scripts, and collects the output. Results are saved in structured formats like TSV or JSON. The frontend then displays these results visually. This allows users to perform complex analysis without needing technical expertise.

One main feature of the application is that users can run scripts from any operating system. Since all scripts run on the server, users only need a web browser. The interface shows the output in a clear and user-friendly way. Another advantage is that users do not need to install anything. All dependencies and configurations are stored on the server. This makes the tool easy to access and maintain.

In order to illustrate the functionalities of MarcoTAO, the following figures illustrate key steps in the workflow, from queries of triples (Figure 4) and triple extraction (Figure 5) to the generation of lexical clusters based on distributional similarity (Figure 6).

[4] The screenshots included 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.

Figure 2. Welcome screen of the MarcoTAO prototype

Figure 3. Concept project management screen of MarcoTAO

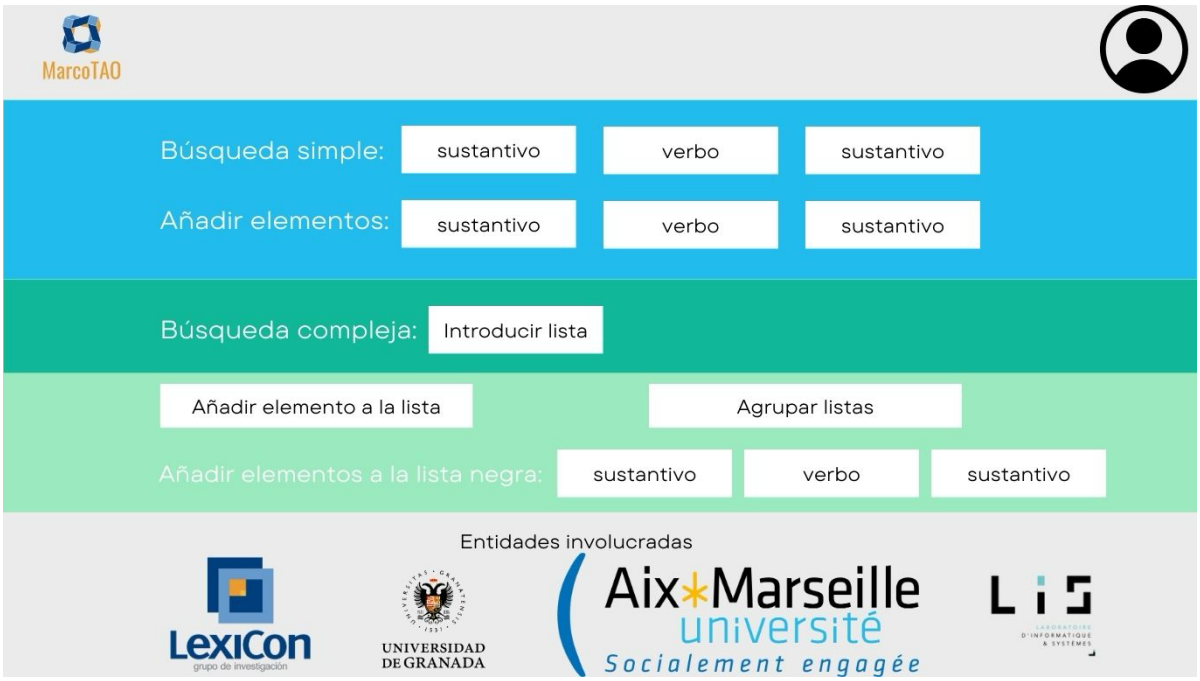

Figure 4. Triples search interface

41	deforestation			burn		fuel				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 thisArti 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

Figure 5. List of extracted triples related to DEFORESTATION (EN) in the prototype

Figure 6. Automatically clustered triples related to CLIMA (ES) based on distributional similarity

Acknowledgements

This research was carried out as part of the project PID2020-118369GB-I00, Transversal integration of culture into an environmental terminological knowledge base (TRANSCULTURE), funded by the Spanish Ministry of Science and Innovation.

Declaration on Generative AI

During the preparation of this work, the authors used X-GPT-4 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Baroni, Marco, and Bernardini, Silvia. BootCaT: Bootstrapping Corpora and Terms from the Web. Proceedings of LREC 2004.

[2] Buendía-Castro, Miriam, and Beatriz Sánchez-Cárdenas. "Using Argument Structure to Disambiguate Verb Meaning." In *Proceedings of the XVII EURALEX International Congress*, eds. T. Margalitadze and G. Meladze, Tbilisi: Ivane Javakhishvili Tbilisi University Press. 482–490, (2016).

[3] Buendía-Castro, Miriam. Phraseology in Specialized Language and its Representation in Environmental Knowledge Resources. PhD Thesis, Universidad de Granada, (2013).

[4] Corpas Pastor, Gloria. *Investigar con corpus en traducción: los retos de un nuevo paradigma* (Vol. 49). Peter Lang (2008).

[5] Faber, Pamela, ed. *A Cognitive Linguistics View of Terminology and Specialized Language.* Vol. 20, Walter de Gruyter, (2012). https://doi.org/10.1515/9783110277203

[6] Granger, Sylviane, and Fanny Meunier, eds. *Phraseology: An Interdisciplinary Perspective.* John Benjamins Publishing, (2008).

[7] Jacques, Marie-Paule, and Agnès Tutin. *Lexique transversal et formules discursives des sciences humaines.* ISTE Group, (2018).

[8] L'Homme, Marie-Claude. "Predicative Lexical Units in Terminology." In *Language Production, Cognition, and the Lexicon*, eds. N. Gala, R. Rapp, and G. Bel-Enguix, Berlin: Springer, pp. 75–93, (2015).

[9] Muñoz-Basols, Javier, María del Mar Palomares, and Francisco Moreno Fernández. “El Sesgo Lingüístico Digital (SLD) en la inteligencia artificial: implicaciones para los modelos de lenguaje masivos en español.” Lengua y Sociedad 23.2, (2024). 623-648. Doi: https://orcid.org/0000-0002-3136-4443.

[10] Orliac, Brigitte. “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 (2006): 261-280.

[11] Pilehvar, M.T., Camacho-Collados, J. Word Embeddings. In: *Embeddings in Natural Language Processing. Synthesis Lectures on Human Language Technologies*. 2021. Springer, Cham. https://doi.org/10.1007/978-3-031-02177-0_3

[12] Ramisch, Carlos. "Multiword Expressions Acquisition: A Generic and Open Framework", *Theory and Applications of Natural Language Processing* series, XIV, Springer, ISBN 978-3-319-09206-5, 230. (2015).

[13] Ramisch, Carlos. "Multiword expressions in computational linguistics: down the rabbit hole and through the looking glass", Habilitation à diriger des recherches, Aix Marseille University, Marseille, France, 2023.

[14] Sánchez Cárdenas Beatriz. “Extracting Semantic Frames from Specialized Corpora for Lexicographic Purposes”, *Círculo de Lingüística Aplicada a la Comunicación*, 99, 163-177, 2024. https://doi.org/10.5209/clac.90626

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

[16] Sánchez-Cárdenas, Beatriz, and Miriam Buendía-Castro. “Inclusion of Verbal Syntagmatic Patterns in Specialized Dictionaries: The Case of EcoLexicon.” In *Proceedings of the 15th EURALEX International Congress*, eds. R. V. Fjeld and J. M. Torjusen, Oslo: EURALEX, pp. 554–562, (2012).

[17] Tutin, Agnès. *L'écrit scientifique: du lexique au discours*. Eds. Francis Grossmann, and Presses Universitaires de Rennes. Presses universitaires de Rennes, (2014).

[18] Vezzani, Federica. “Vers une méthodologie pour l’extraction et la classification automatiques des collocations terminologiques verbales en langue médicale”. *Phraséologie et terminologie*, 480, 259. (2023).