EDICIONES COMPLUTENSE

Círculo de Lingüística Aplicada a la Comunicación
ISSN-e: 1988-2696

ARTÍCULOS

Extracting Semantic Frames from Specialized Corpora for Lexicographic Purposes

Beatriz Sánchez-Cárdenas
Universidad de Granada. Grupo LexiCon

https://dx.doi.org/10.5209/clac.90626 Recibido: 22 de julio de 2023 • Aceptado: 13 de enero de 2024

Abstract: This paper focuses on Frame-based Terminology (FBT), a theoretical framework based on cognitive premises that underpins the representation of specialized objects and events. It demonstrates that specialized semantic frames can be inferred from lexical schemas and semantic annotation. This research used corpus techniques to collect data for the representation of multilingual semantic frames. It focuses on the subdomain of deforestation within the general domain of Environmental Science. The methodology involves the encapsulation of corpus queries and semantic annotation to guide the creation of specialized frames associated with the DEFORESTATION event. Although the creation of frame-based terminological resources can be time-consuming, our results show that it is facilitated by the use of e-tools that extract the arguments of specialized predicates, which are often immersed in complex syntactic constructions.
Contents: 1. Introduction. 2. Semantic Frames of Specialized Concepts. 3. Methodology for Extracting Semantic Frames in Complex Concepts. 4. Semantic Frame Induction from Lexical Patterns. 5. Results. 6. Conclusions.

Cómo citar: Sánchez-Cárdenas, B. (2024). Extracting Semantic Frames from Specialized Corpora for Lexicographic Purposes. *Círculo de Lingüística Aplicada a la Comunicación* 99 (2024) 163-177. https://dx.doi.org/10.5209/clac.90626

1. Introducción

In recent years, cognitive approaches have attracted considerable attention in Terminology research (Faber & L'Homme, 2022). These theoretical approaches aspire to reflect the way that specialized knowledge units are organized in the mind. Frame-based Terminology (FBT) (Faber 2012, 2016, 2022) is a cognitive text-based theory that creates specialized frames, based on the data extracted from corpora. This includes the automatic extraction of predicate-argument structure.

Frame Semantics, as proposed by Fillmore in 2006 (Fillmore et al., 2003, Fillmore 2006), introduces a knowledge representation model that characterizes concepts based on their position within a comprehensive conceptual system. The central premise underlying this approach is that language acts as a reflection of our cognitive processes, thereby enabling the depiction of any language through cognitive structures, commonly referred to as *frames*. A frame essentially serves as a blueprint, outlining a particular situation in a schematic manner. To effectively define a frame, the identification of key participants or frame elements within each schematized situation becomes imperative. Among these frame elements, the ones crucial for precisely delineating the frame's meaning are referred to as *core frame elements*.

FrameNet (https://framenet.icsi.berkeley.edu) constitutes a valuable resource that applies Frame Semana tics (Fillmore et al., 2003, Ruppenhofer et al., 2016) to depict the intricacies of English language. By employing Frame Semantics, FrameNet enhances our understanding of English by revealing its underlying cognitive structures and the connections between concepts.

FBT applies basic premises of Frame Semantics to the study of the conceptual organization that underlies specialized domains (Faber, 2012, 2015). According to this model, specialized semantic frames model language independent descriptions of terms cognitive structure (Faber, 2015). Specialized semantic frames provide a way to cluster related lexical items that account for language-independent dimensions of specialized knowledge. For instance, in the domain of Environmental Science, the DEFORESTATION event is lexicalized in English by units related to its causes (e.g. *logging, agriculture, urbanization*) and consequences (e.g. *erosion, greenhouse effect, soil pollution*). Although lexical patterns evidently vary across languages, we claim that there is a core set of items that is common to all languages (Wierzbicka, 1999).

The advantages of semantic frames for terminological and translation purposes are numerous. In fact, they provide a rich and structured framework to organize concepts in a specialized domain. Such a representation

allows the inference of the cognitive structures underlying scientific texts. Getting to know the frame structures of these concepts and their linguistic correlates is useful both for translators, to understand a concept, and for non-native speaker experts, to produce texts in a foreign language. However, as Faber and L'Homme (2022: 2) write, «translators often point out that the focus on concepts and knowledge organization of many terminological resources does not allow them to address all the questions raised by the translation of specialized texts.»

Lexicographers often say that representing the behavior of terms in different languages can be crucial in order to produce texts that reflect the lexical preferences of a word and its semantic prosody. As a matter of fact, «terminological resources need to account for the linguistic and textual behavior of terms in addition to providing information about the kind of knowledge they convey» (Faber and L'Homme, ibid). Consequently, the description of the conceptual, linguistic and communicative dimension of terms should lead to the creation of new multilingual lexicographic resources and, ultimately, to the improvement of translation tools.

Given that conceptual frames are formed by complex argument structures, their creation requires both linguistic and domain expertise, as well as tools for performing corpus-based searches (L'Homme et al., 2014, Hermann et al., 2014). When building entries of frame-based resources, lexicographers need to access specialized corpora, so that they can perform complex searches, involving verbs and their arguments. Nevertheless, constructing lexical resources that model the frame structure of domain concepts is not only challenging but time-consuming.

Not surprisingly, to the best of our knowledge, there is still no methodology for the creation of semantic frames in specialized language. Ideally, computer tools could support and facilitate corpus analysis to confirm and generalize linguistic introspection. However, concordances are not sufficient since it is necessary to run complex queries that are capable of modeling morphosyntactic and syntactic co-occurrence patterns that approximate predicate-argument structure. Moreover, since variability can influence the results of corpus queries, it is also important to consider complex phenomena such as verbal alternation or complex nominals. Natural language processing techniques and linguistic analysis can help to overcome some of these difficulties.

This article presents a protocol for the construction of specialized frames, composed of the following steps: (i) design and application of complex corpus queries for lexico-grammatical patterns in the form of triples (noun-verb-noun); (ii) systematic annotation of these triples by lexicographers; (iii) semi-automatic grouping of these triples and manual frame construction.

As a case study, we focus on the concept of DEFORESTATION. From an environmental point of view, DEFORESTATION is one of the main ecological issues worldwide. From a terminological perspective, the concept of DEFORESTATION raises questions such as which entities and processes cause this event and what are its consequences for the ecosystem. Some of the terminological resources dealing with deforestation are EcoLexiCon[1], DiCoEnviro[2] and Gemet[3]. This paper explains how the information in all of them could be improved. Our approach tackles this problem from the perspective of Frame-based Terminology with the goal of enriching these resources. The corpus used for this research was extracted from an English environmental corpus of more than 23 million words, and consists of a sub-corpus of 1,257,216 words composed of texts about the DEFORESTATION event.

2. Semantic frames of specialized concepts

Although translators have a wide range of resources available (general and specialized dictionaries, multilingual glossaries, terminology databases, translation memories, and parallel and comparable corpora), none of them describes the semantic structure of specialized concepts and their correspondence with linguistic structures. One way to fill this gap is to create resources that describe the semantic frames, in other words, the semantic relationships between specialized concepts, which are language-independent, coupled with their lexicalizations in each language.

To elaborate this type of description based on the observation of texts, the study of concordances is not sufficient, semantic frames must be systematically constructed, starting from the analysis of specialized corpora and arriving at generalizations with a high level of interactivity. This requires linguistic expertise, as well as the use of appropriate computer tools to perform corpus queries and extractions. Such tools can support, enhance and facilitate corpus analysis to confirm and generalize linguistic introspection.

According to Cognitive Linguistics (Lakoff and Johnson, 1997; Rosch and Lloyd, 1978), the structure of language and cognition stem from the psychosensory perception of our environment. In Frame Semantics (Fillmore, 2006), a frame represents a concept or event in a schematic way. To define a frame, it is necessary to identify the main participants in each situation. For example, the ATTACK frame has three basic elements: the AGGRESSOR, the WEAPON and the VICTIM. There may be other non-essential frame elements such as PLACE or MANNER.

EcoLexicon (http://ecolexicon.ugr.es) (Faber et al., 2011; Faber and Buendía-Castro, 2014; Faber et al., 2016) is a multilingual terminological knowledge base (TKB) on environmental sciences developed by the LexiCon research group (http://lexicon.ugr.es) at the University of Granada (Spain). It currently has 4 578 concepts and 24 587 terms in English, Spanish, German, French, Russian and Modern Greek.

---

1 http://ecolexicon.ugr.es

2 http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi

3 https://www.eionet.europa.eu/gemet/en/themes/

EcoLexicon is the practical application of Frame-Based Terminology (FBT) (Faber et al., 2006; Faber, 2012, 2015), a theory on the representation of specialized knowledge that employs certain aspects of Frame Semantics (Fillmore, 1982; Fillmore and Atkins, 1992) to structure specialized domains and create language independent representations. FBT focuses on conceptual organization, the multidimensional nature of specialized knowledge units, and the extraction of semantic and syntactic information through the use of multilingual corpora. The target users of this TKB include language professionals and environmental experts as well as the general public. The representations offered in EcoLexicon are designed for translators, technical writers and environmental experts who need to gain a better understanding of specialized environmental concepts in order to write or translate specialized or semi-specialized texts.

Theoretically, semantic frames are only valid for one culture and are therefore not universal. However, they can be generalized for a group of cultures with common characteristics (e.g. Western culture). The same is true for specialized language, where a semantic frame is shared by all experts in a domain. For example, the concept HURRICANE, represented by the terms *huracán*, *hurricane*, *ouragan*, in Spanish, English and French respectively, should be described according to the entities causing that meteorological event and its consequences on other entities.

For instance, the concepts HURRICANE and TORNADO, lexicalized in any language have a set of common features, namely its causes (*atmospheric conditions*), consequences (*floods, waves, debris flow*) and typical actions (*rotate, turn, strike*). Accordingly, they share the same semantic frame, whose description should at least contain the actors and verbs involved in the conceptual and linguistic realization of the concept. A terminologist's work adopting this point of view should consist of extracting such linguistic structures from specialized texts and to represent them in semantic frames.

Since semantic frames are made up of complex argument structures, their creation requires both linguistic and domain knowledge, as well as tools to extract such information from the corpus (L'Homme et al., 2014). To learn the semantic networks and the linguistic structures of specialized concepts, it is necessary to perform complex queries, capable of modelling morphosyntactic and syntactic cooccurrence patterns that reflect the structure of the predicates. Computer tools can support, enhance and facilitate corpus analysis to confirm linguistic introspection and lead to objective generalizations.

3. Methodology for extracting semantic frames in complex concepts

Our starting point is based on the hypothesis that certain recurrent lexical-semantic patterns play a significant role in the construction of semantic frames (Sánchez Cárdenas and Ramisch, 2019). These patterns take the form of argument structures referred to as «triples». Triples typically consist of a predicate (v) accompanied by its corresponding arguments (n1 and n2), thus a subject and an optional object. Formally, they comprise a nominal phrase, followed by a verb or verbal phrase, and concluding with one or more nominal phrases (e.g., *deforestation, accelerate, global warming*). Following the protocol outlined in a previous study (Sánchez Cárdenas and Ramisch, 2019), we utilized the MWEtoolkit computational tool, originally designed for extracting polylexical words (Ramisch, 2015), to extract triples as argument structures from the corpus. Subsequently, these triples were annotated according to several typologies of semantic categories designed for verbs and nouns. By assigning semantic labels, we were able to cluster similar triples to extract recurrent lexical-semantic patterns, thereby facilitating the emergence of semantic frames.

In this study, we apply this methodology to construct semantic frames for complex concepts, such as DEFORESTATION. Additionally, we have developed a computer tool to enhance and streamline the entire process. To illustrate this methodology, we have analyzed an English subcorpora comprising 1,257,216 words, with a primary focus on deforestation. These texts were manually selected and encompass scientific articles, encyclopedic entries, and specialized news items that address deforestation with a medium-high level of specialization.

In the subsequent section, we provide a detailed explanation of our methodology, which includes the extraction of triples, the annotation process, and the generalization of frames.

3.1. Lexical patterns extraction techniques

Our objective was to extract pertinent lexical patterns, represented as triples, from our corpus. A triple consists of an initial nominal phrase, followed by a verb or verbal locution, and concluded with an optional second nominal phrase, such as [*extensive farming, lead to, deforestation*]. These triples were obtained by executing queries in which at least one element of the triple was left underspecified (ANY). Various Python scripts were employed to extract triples forming argument structures from the corpus using MWEtoolkit (Sánchez Cárdenas and Ramisch, 2019). For instance, the triple [extensive farming ANY deforestation] was retrieved from the sentences such as (1) and (2).

(1) The expansion of **intensive livestock production systems** has resulted in widespread **deforestation** in areas such as Southeast Asia and Central Africa.

(2) The United Nations Food and Agriculture Organization (FAO) report emphasizes that the conversion of forests to accommodate **large-scale intensive farming** operations, such as palm oil plantations and industrial cattle ranching, **has contributed** significantly **to deforestation** rates in various tropical regions.

One limitation of this methodology is the presence of noise and silence in the search results. Although certain extracted triples were deemed irrelevant for various reasons (see 4.3 for details), there were instances where the extraction of predicate-argument structures, represented as triples, failed to capture valuable information. This inadequacy can be attributed to several factors. Firstly, certain predicates predominantly manifest as nominal rather than verbal forms (e.g., *erosion* rather than *erode*). Additionally, some predicates may exhibit a different number of arguments, either exceeding or falling short of the expected two arguments (3).

(3) [Soybean expansion] in southern Brazil [contributed] to [deforestation] by [stimulating migration to agricultural frontier regions].

Despite these efforts, it is important to acknowledge that predicate-argument structures represented as simplified triples [n1, v, n2] might not capture all pertinent information. For example, certain predicates are primarily nominal rather than verbal in nature (e.g., *erosion* rather than *to erode*), and some predicates can possess more or fewer than the expected two arguments (e.g., «[Soybean expansion] in southern Brazil [contributed] to [deforestation] by [stimulating migration to agricultural frontier regions]»). Nevertheless, by conducting multiple queries using various variants of the target terms, we anticipate compensating for this limitation and achieving a comprehensive coverage of the phraseological patterns under investigation (cf. 4.1).

In a parallel way, terminological variation (Freixa, 2022) can also contribute to the occurrence of silence. For instance, the concept of DEFORESTATION is sometimes expressed using nominal compounds such as *forest loss* or *reduction of forest*. In previous studies (reference hidden for anonymous purposes), the inability to include nominal compounds resulted in the silencing of meaningful patterns.

To address the issue of silence, our approach aims to alleviate this limitation by incorporating prior searches, which integrate their results into the search patterns. Thus, we introduced the capability to perform searches using a comprehensive list of lexical units in specified positions: list_n1, list_verbs, and list_n2. These lists were derived from the corpus through the described searches (see 3.2. and 3.3) and saved as .txt files, which were subsequently utilized for the triples queries.

This strategy expands the search spectrum of the triples, enhancing the chances of capturing relevant information. In the upcoming sections, we provide a detailed explanation of the methodology employed to conduct these searches, with the objective of achieving comprehensive coverage of the semantic frame associated with the analyzed concept. Our aim is to encompass as much of the concept's semantic frame as possible within our study.

3.2. Verbs associated to the event DEFORESTATION

In our study, a query such as [*n1=list-complex* + *v=ANY* + *n2=list-deforestation*] allows us to uncover unknown verbs. In this triple, we specify already known nouns such as [*technological change*, *reduce*, *deforestation*] in the «list-complex» and [*population growth*, *be associated to*, *forest loss*] in «list-deforestation». They are utilized as a first step to identify verbs that have at least one of those nouns as an argument, such as *to reduce* or *to be associated with*, and that emerge as a result of this query.

To perform these types of preliminary searches that cover a broad range, we carry out multiple extractions of events, actions, and entities that are affected by or serve as triggers for DEFORESTATION. To accomplish this, we employ several CQL searches in Sketch Engine. For instance, Figure 1 illustrates the results of CQL search CQL [lemma="deforestation"] []{0,5} [tag="V.*"&lemma!="be|do|have"] performed to retrieve from the corpus verbs associated with the term *deforestation*.

Figure 1. Concordances of the CQL query [lemma="deforestation"][]{0,5}[tag="V."&lemma!="be|do|have"]*

Left context	KWIC	Right context
ge climate change global warming climate policy	**Deforestation affects**	the global climate both by releasing the carbon
the physical properties of the planetary surface.	**Deforestation exerts**	a warming influence by (i) adding CO2 to the at
ture changes. Previous studies have shown that	**deforestation in the tropics would decrease**	evapo- transpiration rates and increase sensible
han the global average in our model. Temperate	**deforestation produces**	local cooling due to dominant albedo change an
el climate system. Cooling biophysical effects of	**deforestation dominate**	the climate response in the Boreal deforestation
degree, towards the west (Fig. 2B). The greatest	**deforestation seems**	to have taken place in the vicinity of the major p
degree, towards the west (Fig. 2B). The greatest	**deforestation seems to have taken**	place in the vicinity of the major population cent
ion types showed a mitigating effect on potential	**deforestation as driven**	by transportation networks, with Federal Strict P
h tropical deforestation rates (16). Rapid tropical	**deforestation leads**	to a radical exchange between natural capital a
ssociated with west Africa, for example, in which	**deforestation is linked**	to the funding of civil wars and armed conflicts.
ion of greenhouse gases in the atmosphere and	**deforestation has caused**	, nevertheless, a dramatic rise in the average of
wn the trees near enough everything involved in	**deforestation produces**	CO2 the humans do the machinery the trees ev

In addition to simple verbs (e.g. *disrupt*, *accelerate*, *cause*), we also extracted from the corpora a list of verbal locutions that often co-occur with terms related to deforestation (e.g. *lead to*, *result in*, and *depend on*) as shown below:

lead to, depend on, result from, occur within, contribute to, increase by, influence by, occur in, drive by, clear, conserve, degrade, protect, destroy, restore, log, preserve, maintain, remove, survive, increase, stimulate, encourage, spur, aggravate, decrease, promote, drive, accelerate, reduce, lead, displace, affect, aggravate...

The retrieved verbs were incorporated into the list-verbs file with the specific aim of being utilized in triple searches in combination with the lists of nouns described below.

3.3. Nominal compounds associated to the event DEFORESTATION

A preliminary version of our tool allowed to locate triples where *n1* and *n2* were single nouns. However, we realized that it was not very useful to extract triples containing only the head nouns (e.g. *warming*, *loss*, and *effect*). Indeed, one characteristic of the DEFORESTATION event is that it involves participants which are lexicalized as complex noun compounds such as *global warming*, *forest loss* or *greenhouse effect*. We will refer to these participants simply as *complex nominals,* to avoid discussing whether they are terms.

Therefore, as a preliminary extraction step, we run a query to extract recurrent noun phrases from the corpus, expressed as a regular expression. In the preliminary version of our tool, we initially focused on locating triples (n1 v n2) where n1 and n2 consisted of single nouns. However, we soon realized that extracting triples containing only the head nouns (e.g., *warming, loss, effect*) did not provide us with sufficient information for our analysis. This limitation led us to recognize the importance of capturing the complexity of the DEFORESTATION event, which involves participants lexicalized as compound nouns such as *global warming, forest loss*, or *greenhouse effect*. In order to incorporate these elements, we designed a preliminary extraction step that involved running this query on MWEtoolkit (Ramisch, 2015) to identify recurrent noun phrases in the corpus. This query was formulated as a regular expression pattern over parts of speech: (ADJ|(NOUN ADP?)){1,4} NOUN. As a result, nouns (NOUN) preceded by a sequence of 1 to 4 modifiers ({1,4}) were retrieved from the corpora, that could be either an adjective (ADJ) or another noun. In cases where the modifier was a noun, it could be optionally (?) followed by a preposition (ADP).

Through the implementation of this approach, our objective was to enhance the relevance and depth of our triple extraction process. The applied pattern successfully extracted combinations such as *carbon dioxide*, *environmental harm, international environmental law*, and *victim of environmental harm*. To further refine the selection, we utilized MWEtookint to calculate the T-score association measure for each extracted combination, following the approach outlined by Evert (2004). The combinations were ranked in descending order based on their T-scores, and a manual selection was made to identify those that appeared most relevant for modeling the deforestation event.

The resulting list, referred to as «list-noun1» henceforth, consists of 94 recurrent simple nouns and complex nominals that were used in our subsequent queries. We present below some illustrative examples.

Nouns associated with DEFORESTATION
agricultural land
arable land
banana production
cattle ranching
climate change
commercial ranching
cropland
damage to ecosystem
demand for energy
demand for food
diversity decline
environmental degradation
extinction
fire
global warming
growth in demand for food growth in demand energy
intensification of land use increased population
land degradation
technological change
temperature change
highway construction
livestock
population growth
soybean farm
species decline
transportation cost
unsustainable development

3.4. Terminological variation of deforestation

Terminological variation refers to the existence of different terms used to describe the same concept. For example, terms like *climate change, global warming*, and *climate crisis* all refer to the same environmental issue. Notably, term variation serves as a cognitive device that conveys information about the concept's features and its relationships with other concepts. Additionally, it serves as a communicative strategy to avoid repetition, tailor the text to the audience, and generalize the content (Sabela, 2022, p. 438). Terminological variation can arise from various factors, including diachronical, cognitive, communicative, linguistic, or contextual influences (Dury, 2022; Fernández-Silva, 2022; Freixa, 2022). In line with this, our corpus analysis reveals that DEFORESTATION can be expressed through monolexical and polylexical units.

Therefore, we extracted from the corpus a second list of terms corresponding to the usual denominations of this concept in English, such as *forest shrinkage, forest loss, forest scarcity, scarcity of forest,* and *land clearing* as shown below. This list will be referred to as «list-noun2». To this end, we performed extractions with Sketch Engine, essentially using the function «Thesaurus» with words such as *deforestation, tree, forest*.

terminological variation in DEFORESTATION
forest shrinkage, forest clearing, clearing of forest, forest loss, loss of forest, destruction of forest, forest scarcity, scarcity of forest, degradation of forest, forest cover reduction, reduction of forest, logging, forest clearing, land clearing

Hence, we compiled a list of terms representing the commonly used designations for this concept in English. This list encompasses compound nouns such as *forest shrinkage*, *forest loss, forest scarcity, scarcity of fores*t, and *land clearing*, as illustrated above these lines.

There is room for debate regarding whether all these terms refer to the same concept. While some focus on the action of tree elimination (*logging*), others emphasize the process (*forest clearing*), and some highlight the outcome of this action (*forest loss*). However, they all revolve around the process through which a forested area transitions into treeless land, often for purposes like agriculture or livestock.

Our study also acknowledges natural causes of forest mass reduction, such as wildfires or controlled logging for timber extraction. Although these actions result in tree removal, we consider that they should not classified as DEFORESTATION, since the forest undergoes self-regeneration afterwards.

4. Semantic Frame Induction from Lexical Patterns

In a previous study (Sánchez Cárdenas y Ramisch, 2019), we developed a set of Python scripts to extract «noun-verb-noun» argument structures from specialized corpora in multiple languages. Although these scripts may require future improvements and extensions, they effectively extract dozens of triples. Once annotated using the protocol described in a previous publication (reference hidden for anonymity, SJR Q2 journal), these triples are automatically grouped based on their semantic similarity, revealing the underlying semantic framework.

However, the current state of our search system is limited to a collection of isolated Python scripts that need programming skills to operate. Consequently, sharing the system with other users is challenging. Moreover, the system is rudimentary, slow, and cumbersome, regardless of the user's computer proficiency.

In order to address these challenges and enable other experts to utilize our protocol analysis, we are developing MarcoTAO, a prototype tool under development that is not currently available to the public. This web interface is built upon the MWEtoolkit computational tool, originally designed for constructing lexicons of multiword expressions (Ramisch, 2015). The query engine in the tool is designed to handle complex patterns. Accessible free of charge, the platform empowers users with linguistic expertise to perform the necessary searches and access the semantic frames of the concepts.

MarcoTAO enables the extraction and processing of triples through two distinct approaches. Firstly, the previously extracted triples can be clustered using word embeddings, a technique that represents words or phrases as numerical vectors. This allows for comparing similarities and grouping them based on semantic relationships. Secondly, the triples can undergo semantic annotation, where additional metadata is added to generate conceptual schemas. These schemas are created based on the similarity of semantic annotations.

4.1. Comparing triple extraction with MWE toolkit and Sketch Engine

The process of triple extractions can be achieved by employing a range of corpus tools being able to identify argument structures. In our research, we assess and compare 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. MWEtoolkit enables the identification of multi-word expressions within a text, while Sketch Engine facilitates in-depth analysis of a large text corpus, identifying patterns and frequency data among other features.

Since neither tool is specifically engineered for triples extraction tasks, we have adapted both of them for the purpose of triples extraction. In order to analyze which tool is more suitable for triples extraction, a small-

scale analysis was conducted. For this purpose, a simple search was performed using both tools with our corpus. The search focused on identifying triples related to DEFORESTATION, such as the relationship between the cause of deforestation, environmental impact, and mitigation measures. Two different searches were performed using MWEtoolkit and Sketch Engine to extract relevant information.

The following search pattern was utilized with MWEtoolkit with *deforestation* both the preverbal and postverbal position:

```
./search-pattern.sh "[${patn1} pos~/${deforestation}/] []{repeat={0,3} ignore=true}
[${patv} pos~/${verbalPOS}/] []{repeat={0,3} ignore=true} [${patn2}
pos~/${nominalPOS}/]" $corpus.
```

This search aimed to capture specific patterns in the corpus related to *deforestation*. It looked for a nominal phrase referring to deforestation (${patn1}), followed by zero to three other elements (ignoring any POS), then a verbal phrase (${patv}), followed by zero to three additional elements (ignoring any POS), and finally a nominal phrase (${patn2}). This pattern allowed the extraction of relevant triples related to deforestation.

In Sketch Engine, this CQL search query was used, also with *deforestation* both the preverbal and postverbal position:

```
1:[lemma="deforestation"] [tag!="V."]{0,5} 2:[tag="VV."] [tag!="V."]{0,5}
3:[tag="N."] within <s/>
```

This query aimed to capture the context surrounding the keywords ‘volcano’. It looked for a lemma of «deforestation» followed by zero to five words that are not verbs (${tag!=”V.”}), then a verb in any form that starts with ‘VV’ (${tag=”VV.”}), followed by zero to five words that are not verbs, and finally a noun (${tag=”N.*”}), all within the same sentence. This query allowed the extraction of triples related to volcanoes and their associated actions or characteristics.

Subsequently, we have carried out an analysis of the extracted triples with both tools to ensure systematic evaluation, each line was annotated with a distinct code as follows:

(1) Incorrect: The extracted triple does not accurately represent the information.
(2) Partially correct: The extraction is not entirely accurate but provides useful insight for enhancing future extractions.
(3) Almost correct: The extraction is accurate, but additional minor information is necessary for completeness.
(4) Correct: The extracted triple is wholly accurate and requires no further information.

Tables 1 and 2 show the results obtained in terms of noise and precision.

Table 1. Noise and accuracy of triples retrieved with Sketch Engine

	Sketch Engine: 999 triples retrieved			
	Code	**Number of results**	**Total number of triples**	**Percentage**
noise	0	888	908	90.9%
	1	20		
accuracy	2	46	91	9.1%
	3	45		

Table 2. Noise and accuracy of triples retrieved with MWEtoolkit

	MWEtookit: 195 triples retrieved			
	Code	**Number of results**	**Total number of triples**	**Percentage**
noise	0	2	67	34.4%
	1	65		
accuracy	2	82	128	65.5%
	3	46		

As shown in Table 1, Sketch Engine retrieved a total of 999 triples. However, the majority of these triples (90.9%) were classified as noise (Code 0), indicating that they were not directly related to the target triples. On the other hand, only a small portion (9.1%) of the retrieved triples were classified as accurate (Code 2).

Regarding MWEtoolkit (Table 2), a total of 195 triples were retrieved. Among these triples, 34.4% were classified as noise (Code 0), indicating some unrelated results. The majority (65.5%) of the retrieved triples were classified as accurate (Code 2).

From the analysis of these results, it can be observed 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%).

Accordingly, we chose to employ MWEtoolkit to extract lexico-patterns in this study.

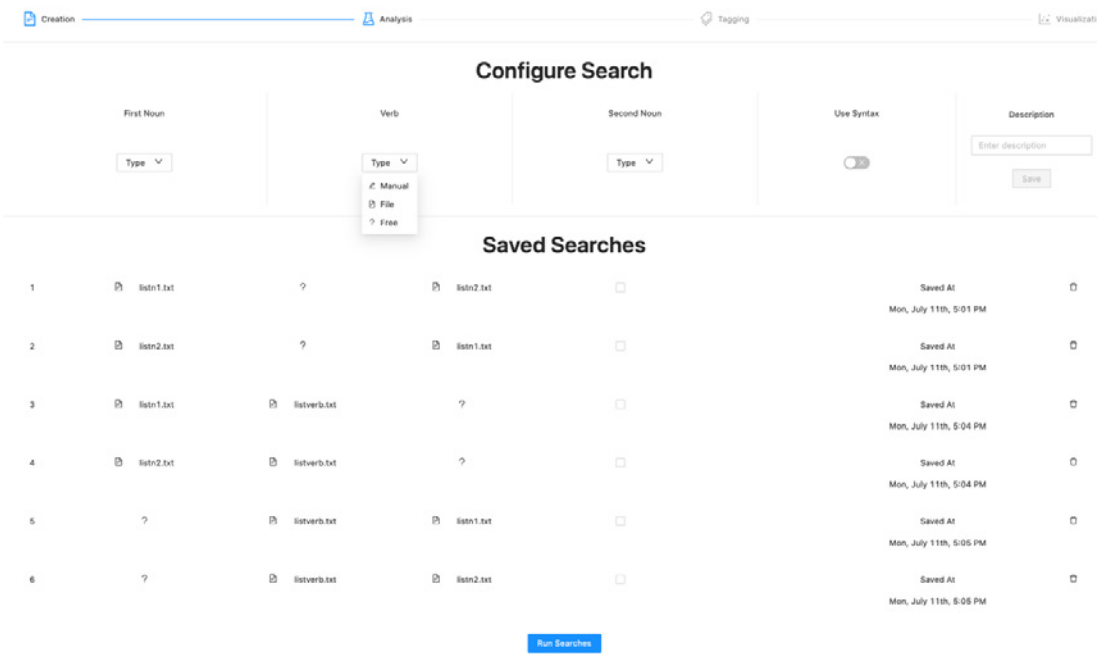
4.2. Lexico-patterns Extraction with MWEtoolkit

The process of pattern extraction using MWEtoolkit begins with corpus loading, processing, parsing and indexing the corpus files in txt format. Subsequently, queries designed to retrieve lexical patterns associated to DEFORESTATION are defined. Table 3 outlines the queries formulated for this project. The predefined word lists (called listn1, listn2 and listverbs) described in section 3 enable simultaneous exploration across multiple queries.

Table 3. Queries designed to retrieve lexical patterns

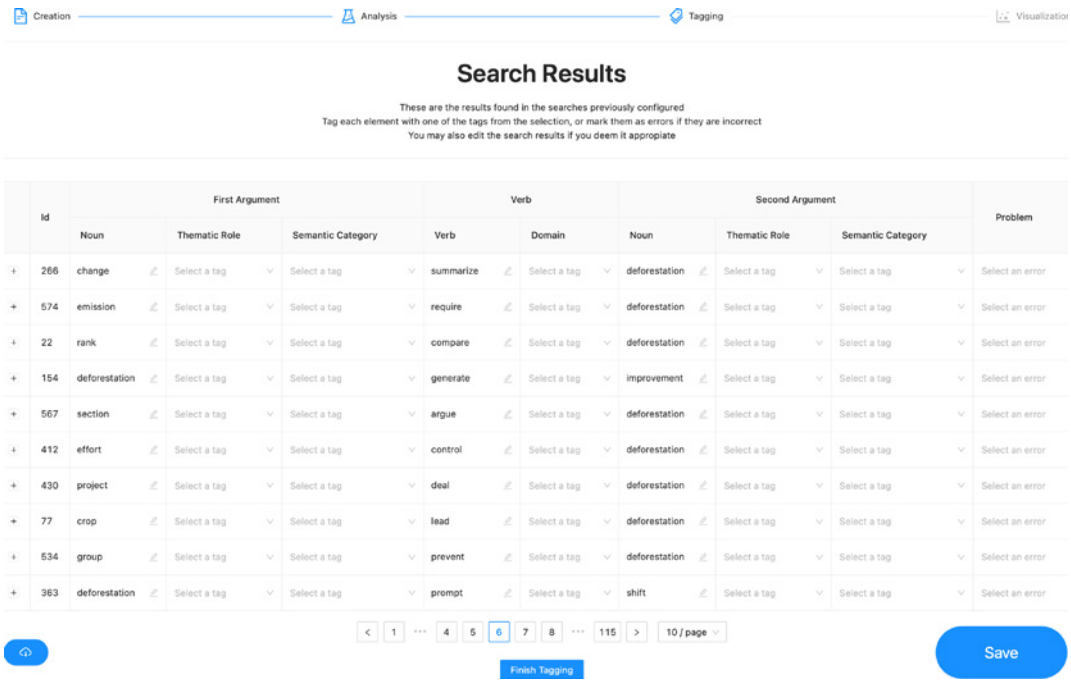
Noun Phrase 1 + verb + Noun Phrase 2
listn1.txt ANY listn2.txt
listn2.txt ANY listn1.txt
listn1.txt listverb.txt ANY
list2.txt listverb.txt ANY
ANY listverb.txt listn1.txt
ANY listverb.txt listn2.txt

Figure 2. Searches performed using the platform MarcoTAO



Through this approach, we conducted comprehensive searches encompassing two identified elements associated with DEFORESTATION, encapsulated in these lists of nouns and verbs, and one unidentified element, with the objective of including a broad spectrum of possible queries. As a result, we obtained 598 triples, represented in Figure 3. Once this was done, the semantic labelling phase took place (section 4.4).

Figure 3. Example of triples retrieved



4.3. Triple extraction error analysis

An error analysis was conducted to improve the protocol for future research. Identifying the kind of problem in the incorrect triples leads to new research avenues and directions.

The rate of errors in the triples retrieved is mostly due to the fact that DEFORESTATION is a highly complex event that entails a large variety of interrelated participants. As shown below, some complex structures are difficult to identify automatically. These drawbacks will be addressed in future research.

1. Complex syntactic structures with three arguments such as an Agent (*deforestation*), a Patient (*acres of once production land*) and a Result (*desert*) where we obtained the misleading triple [*land*, *turn into*, *deforestation*]:

   – *Each year, **millions of acres of once productive land** are turned into **desert** through **overgrazing and deforestation**.*

2. Deep semantic structures with a positive verb that hides a negation are problematic. For instance, the occurrence below retrieved the triple [deforestation, leave, tree], incorrect not only because of the negation but also due to the fact that this structure has three arguments:

   – *Deforestation **leaves fewer** trees to absorb carbon dioxide*.

3. Causal structures are difficult to identify since it requires to extract deep structure information. In the example below the triple [*activist, decry, deforestation*] misses the most relevant information. Future searches should contain knowledge-rich patterns, such as Sketch Grammars (León et al., 2016).

   – *Environmental activists decried the apparent accelerating pace of deforestation in the twentieth century **because of** the potential loss of wildlife and plant habitat and the negative effects on biodiversity.*

4. Coordination of several nouns in one phrase. Currently only the first one of these nouns [e.g. *deforestation*] is detected by our MWEtoolkit scripts:

   – *The progressive conversion of the coastal land to alternative uses has been documented to cause **deforestation, pollution of marine and inland waters, coral reef destruction, coastal erosion and flood.***

5. Phrasal verbs are not yet correctly identified. This could be solved by detecting these verbal forms prior to running the searches:

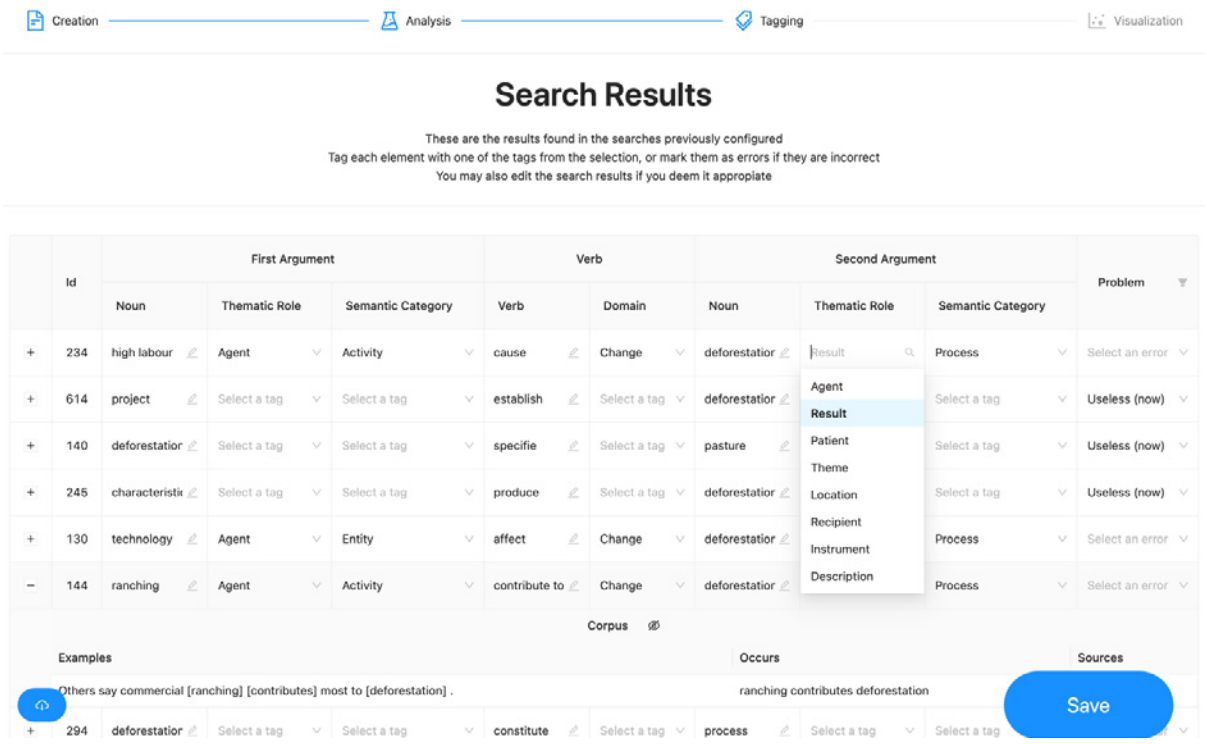
   – *Cropper et al. (1999**) found** population pressure, road density and proximity to the capital city **as (found as=are)** the major factors responsible for deforestation in Thailand.*

Most of these issues are among the greatest challenges in NLP today. The next section explores new extraction techniques in order to overcome some of these drawbacks in future research.

4.4. Semantic Annotation of Lexico-patterns

The selected lexico-patterns retrieved in the form of triples were manually annotated using the platform interface as shown in Figure 4.

Figure 4. Semantic annotation of triples with MarcoTAO



The semantic annotation process was conducted in three stages. Firstly, verbs were classified into lexical domains, which are semantic categories based on the nature of the verb, such as ACTION, CHANGE or EXISTENCE (Faber & Mairal, 1999; Mairal & Faber 2002). Then, verb arguments were ranked according to their semantic class, such as LANDFORM, ACTION or FLORA, using a typology of environmental sciences, currently under development (Sánchez Cárdenas and Ramisch, 2019; Gil-Berrozpe et al., 2019). Finally, verb arguments were assigned a thematic role from a closed inventory, including roles such as Agent, Theme or Result (Rojas, 2022; Sánchez Cárdenas 2022; Sánchez Cárdenas and Ramisch 2019).

The three-level annotation (verb lexical domains, noun semantic classes, and noun thematic roles) serves two main purposes. Firstly, it helps uncover recurring lexico-grammatical patterns in the corpora. For instance, the structure «ACTION increase/stimulate PROCESS» is a common pattern, where the first noun functions as the Agent and the second noun as the Patient. Within this structure, some nouns categorized as ACTION include *banana production, commercial ranching*, and *soybean expansion*, while the PROCESS is represented by *deforestation, land clearing*, or *forest erosion*. In Table 4, a summary of the semantic annotation of several triple structures is offered.

As a conclusion, semantic annotation of lexical patterns in the form of noun-verb-noun combinations allows the inference of recurrent conceptual schemes lexicalized that would be difficult to infer otherwise. In other words, this annotation decomposes specialized frame creation into systematic steps that are individually more tractable than the creation of the whole frame at once based on extracted triples. For instance, there are three main conceptual dimensions activated by this concept; (1) DEFORESTATION begins to exist, (2) DEFORESTATION is intensified and (3) DEFORESTATION is seen as a direction towards which leads several processes and actions.

5. Results

Based on the assumption that similar phraseological patterns reveal semantic similarity, the triples were automatically grouped into semantic classes using a script that gathers triples sharing the same annotation. For instance, rows 3 and 4 of Table 4 were grouped, as the lexical domain of the verb, as well as the semantic classes and roles of the nouns in their arguments coincide. This information was then analyzed for the construction of the specialized frame in the next subsection.

Table 4. Example of the lexical patterns semantic annotation

Noun1	N1 role	N1 class	Verb	lexical domain	Noun2	N2 role	N2 class
demand for land	Patient	process>action	**spur**	CHANGE	**deforestation**	Agent	attribute
deforestation	Agent	process>change	**intensify**	CHANGE	**natural_flood**	Patient	process>loss
forest_clearing	Agent	process>loss	**contribute_to**	CHANGE	**change in biodiversity**	Patient	process>loss
forest_clearing	Agent	process>loss	**contribute to**	CHANGE	**climate_change**	Patient	process>loss
forest_degradation	Cause	process>change	**be a precursor of**	EXISTENCE	**deforestation**	Result	process>loss
forest_scarcity	Theme	attribute	**drive by**	MOVEMENT	**land_price**	Agent	process>action
forest_scarcity	Agent	attribute	**lead to**	MOVEMENT	**higher land price**	Result	process>action
technological_change	Cause	process>change	**promote**	EXISTENCE	**deforestation**	Result	process>loss
technological_change	Agent	process>change	**affect**	CHANGE	**forest_clearing**	Patient	process>loss

5.1. The semantic frames induction of DEFORESTATION

Specialized semantic frames serve as a mechanism to group related lexical structures, thereby accounting for dimensions of specialized knowledge that are independent of any specific language. With this goal in mind, the previously annotated triples were systematically grouped based on their similarity, guided by the three-layer annotation. A Python script was used in order to do an automatic grouping that assembled all verbs sharing identical annotated lexical domains, thematic roles, and semantic categories. Consequently, the resulting lexical schemas encapsulate not only the lexical patterns of DEFORESTATION but also effectively mirror the conceptual structure of the concept. The aggregated annotation reveals a minimum of three activated frames or conceptual dimensions of DEFORESTATION within scientific texts: EXISTENCE, CHANGE, and MOVEMENT.

The conceptual structure of DEFORESTATION in the EXISTENCE lexical domain is displayed in Table 5. It provides a structured understanding of the various roles, actions, and outcomes associated with this phenomenon. The *deforestation* event is seen as the result of a cause lexicalized by several actions (*banana plantation, labor, unsustainable development*), as the result of a change process (*maize boom price, loss of biodiversity, migration*) or as a theme that occur at certain places (*tropical area*). The semantic categories in the table provide additional insights into the relationships between the arguments, verbs, and thematic roles.

Table 5. Lexical domain of EXISTENCE in DEFORESTATION

LEXICAL DOMAIN	ARGUMENT 1			VERB	ARGUMENT 2		
	SYNTAX	THEMATIC ROLE	SEMANTIC CATEGORY		SYNTAX	THEMATIC ROLE	SEMANTIC CATEGORY
EXISTENCE	Subject	CAUSE	ARTIFICIAL PLACE	*associate to*	Direct Object	RESULT	PROCESS>ACTION
			ATTRIBUTE	*provoke, promote*			
			LANDFORM	*tend to, cause*			PROCESS>CHANGE
			METHOD	*produce*			
			PROCESS>ACTION	*cause*			PROCESS>LOSS
			PROCESS>CHANGE	*provoke, promote, cause*			
		RESULT	ATTRIBUTE	*be a precursor of, promote, result in, encourage, determine*	Direct Object	CAUSE	PROCESS>ACTION
			FLORA	*result_from, result from*	Passive Subject		
			PROCESS>ACTION	*promote*	Direct Object		PROCESS>LOSS
			PROCESS>CHANGE	*result_from*	Passive Subject		
			PROCESS>LOSS	*depend_on, promote, explain_by, encourage, result, generate*	Direct Object/Passive Subject		
		THEME	LANDFORM	*result from*	Circumstantial	LOCATION	PROCESS>CHANGE
			PROCESS>CHANGE	*occur_in*			
			PROCESS>CHANGE	*occur_in*		THEME	PROCESS>LOSS
			PROCESS>LOSS	*correlated with*		LOCATION	
			PROCESS>LOSS	*occur_within*		THEME	

This kind of description aims to reflect the deep structure of the concept. Hence it does not consider syntactic notions. However, they could be easily added as shown in Table 5, where examples of possible syntactic functions for each argument are provided.

Similarly, Tables 6 and 7 represent the conceptual structures within the domains of CHANGE and MOVEMENT.

Table 6. Lexical domain of CHANGE in DEFORESTATION

LEXICAL DOMAIN	ARGUMENT 1		VERB	ARGUMENT 2	
	THEMATIC ROLE	SEMANTIC CATEGORY		THEMATIC ROLE	SEMANTIC CATEGORY
CHANGE	AGENT	ARTIFICIAL OBJECT	*increase*	PATIENT	LANDFORM
			increase		LOCATION
			stimulate	RESULT	
			stimulate		PROCESS>ACTION
		COGNITIVE CATEGORY	*influence*	PATIENT	
			stimulate	RESULT	
		FAUNA	*increase*	PATIENT	
		FLORA	*induce*		
		LANDFORM	*increase, contribute_to*		PROCESS> CHANGE
		MAGNITUDE	*increase*		
		METHOD	*increase, affect*		
		NATURAL PLACE	*increase*		
		PROCESS>ACTION	*affect, aggravate*		
			contribute to		PROCESS>ELIMINATION
			increase, contribute_to, contribute to		PROCESS>LOSS
		PROCESS>CHANGE	*stimulate*	PATIENT	PROCESS>LOSS
			contribute_to, slow, intensify, increase, trigger, induce, influence_by, affect, spur, exacerbate	RESULT	PROCESS>ACTION
			stimulate		PROCESS>LOSS
		PROCESS>LOSS	*reduce*		
	THEME	PROCESS>ACTION	*contribute to*	THEME	PROCESS>CHANGE

Table 7. Lexical domain of MOVEMENT in DEFORESTATION

LEXICAL DOMAIN	ARGUMENT 1		VERB	ARGUMENT 2	
	THEMATIC ROLE	SEMANTIC CATEGORY		THEMATIC ROLE	SEMANTIC CATEGORY
MOVEMENT	AGENT	ARTIFICIAL OBJECT	*associate with*	THEME	PROCESS>ACTION
		ARTIFICIAL PLACE	*accelerate*	RESULT	
		ATTRIBUTE	*lead_to*	RESULT	
		FLORA	*lead*	THEME	
			accelerate	THEME	PROCESS>ADDITION
		LANDFORM	*lead to*	RESULT	PROCESS>CHANGE
			lead	THEME	
		PROCESS>ACTION	*drive*	RESULT	ATTRIBUTE
			lead to, lead		PROCESS>LOSS
			lead_to, drive, lead, lead to	THEME	ATTRIBUTE
			accelerate, drive		PROCESS>LOSS
		PROCESS>CHANGE	*accelerate*	RESULT	PROCESS>LOSS
			lead_to, lead to	THEME	PROCESS>ACTION
			drive		PROCESS>LOSS
		PROCESS>LOSS	*lead_to, accelerate, drive*	RESULT	PROCESS>LOSS
	THEME	ATTRIBUTE	*lead_to, lead to*	AGENT	PROCESS>ACTION
		PROCESS>ACTION	*drive by*	AGENT	PROCESS>LOSS
		PROCESS>CHANGE	*drive_by*		

5.2. Lexicographic applications

From a text production perspective, the type of representation explained in 5.1. serves as a scheme that uncovers the correspondence between semantic and lexical structures. Hence, not only it deepens our understanding of the concept described but it also fosters the generation of more precise and effective discourse, specifically tailored to a given specialized domain.

The information of this kind could be particularly beneficial for creating lexicographic definitions. Indeed, this representation could help terminologists and lexicographers to craft new entries that could enable a better understanding of the concept, and a more accurate and effective discourse production according to a certain specialized domain.

In order to illustrate this, we present in Tables 3-5 some the definitions that can be inferred from the relationships and patterns outlined in Tables 8-10. Consequently, when using the data from the semantic frame schemes, it becomes feasible to produce clear, yet succinct definitions that capture the fundamental nature of the concepts and their interrelations with other concepts.

Such definitions could represent an improvement in lexicographic resources by providing context-specific, nuanced explanations of deforestation, since they enhance the accuracy, specificity, and comprehensiveness of our understanding of the concept studied, enabling users to grasp the intricacies of deforestation more effectively.

Table 8. Definition of DEFORESTATION *in* the lexical domain of EXISTENCE

EXISTENCE Definition: An AGENT (natural or artificial process) destroys or protects a PATIENT (flora: tree), causing a RESULT (event: deforestation) which in turn produces another RESULT (event: acid rain). – Landslides increase deforestation. – Deforestation increases aridification. – Aridification reduces the habitat of numerous species.

Table 9. Definition of DEFORESTATION *in* the lexical domain of CHANGE

CHANGE Definition: An AGENT (human action | natural disaster | natural process | artificial process) generates a change in a PATIENT (flora: tree) that affects a THEME (Event: deforestation) or reverses it. – Agricultural expansion triggers deforestation – Technology can minimize the impact of deforestation

Table 10. Definition of DEFORESTATION *in* the lexical domain of MOVEMENT

MOVEMENT Definition: An AGENT (natural or artificial process) leads to a RESULT (Event: deforestation). – The intensification of cattle grazing activities can lead to deforestation. – Erosion processes are accelerated by the deforestation of riverbanks.

6. Conclusions

This work not only contributes to our understanding of Frame-based Terminology but also has the potential to streamline and enhance the creation of lexicographic resources. Our investigation into the conceptual structure of DEFORESTATION, as reflected through the analysis of corpus triples extraction, reveals a complex scenario encompassing numerous interdependent participants. The semantic frame of the DEFORESTATION event represents a process triggered by either human activity or natural events. It is evident from the results that human activities contribute to this process directly (through activities like *tree-cutting*) or indirectly (through greenhouse gas emissions instigating natural disasters that subsequently lead to *soil erosion*). The ramifications of deforestation permeate the entire ecosystem and profoundly impact living organisms across many levels.

Interestingly, despite deforestation being a global environmental concern requiring urgent prevention and reversal, scientific texts in our corpora insufficiently address what actions should be taken towards forest protection and sustainable forest management. Future research is necessary to ascertain whether this lacuna reflects a biased corpus or indicative of the scientific community's inadequate attention towards this issue.

The functionalities of the scripts developed and the future resource MarcoTAO include extensive corpus analysis, advanced query capabilities, semantic annotation, and automatic generation of conceptual schemas. The platform offers a user-friendly and accessible web interface, designed to aid researchers and practitioners in efficiently exploring and comprehending semantic frames. Our future vision for this resource involves making it accessible to the public and its capability to analyze corpora across various disciplines, such as Literature, Sociology, or Psychology. By doing so, we aim to unearth and analyze the semantic structures

embedded within texts and discourses in these fields. This would facilitate an interdisciplinary understanding of the intricate relationship between language and conceptual frames, thereby benefiting a wide array of scholarly field.

The current study has also highlighted linguistic structures overlooked by our automatic extraction procedures. Future research efforts should also aim to replicate this procedure in different languages, establishing interlinguistic correspondences. These research directions align with our ultimate objective of enriching terminological resources with more pertinent conceptual and linguistic information.

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. My deepest gratitude to Carlos Ramisch, researcher and developer of the MWEtoolkit program, for his invaluable support and dedication to the background work that has been fundamental to this research.

References

Dury, Pascaline (2022). Diachronic variation. In Pamela Faber and Marie-Claude L'Homme (eds.), *Theoretical Perspectives on Terminology Explaining terms, concepts and specialized knowledge*, John Benjamins Publishing Company Amsterdam / Philadelphia, pp. 421–434. 10.1075/tlrp.23.19dur

Evert Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis. Institut für maschinelle Sprachverbeitung, University of Stuttgart, Germany. 10.18419/opus-2556

Faber Pamela, and Ricardo Mairal (1999). *Constructing a Lexicon of English Verbs*, New York, Walter de Gruyter. doi.org/10.1515/9783110800623

Faber, Pamela (Ed.) 2012. *A cognitive linguistics view of terminology and specialized language* (Vol. 20). Walter de Gruyter. 10.1515/9783110277203

Faber, Pamela, and Marie-Claude L'Homme. Lexical semantic approaches to terminology. *Terminology* 20.2 (2014): 143–150.

Faber, Pamela (2015). Frames as a framework for terminology. *Handbook of Terminology*, 1(14). In Kockaert, H.J. and Steurs, F. (Eds.), 1:14–33. John Benjamins Publishing Company. 10.1075/term.20.2.01int.

Fernández-Silva, Sabela (2022). Cognitive approaches to the study of term variation. In Pamela Faber & Marie-Claude L'Homme (Eds.) *Theoretical Perspectives on Terminology Explaining terms, concepts and specialized knowledge*, John Benjamins Publishing Company Amsterdam / Philadelphia, pp. 435–456. https://doi.org/10.1075/tlrp.23.20fer

Fillmore, Charles, Christopher Johnson and Miriam Petruck (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3), pp. 235–250. 10.1093/ijl/16.3.235.

Freixa, Judith (2022). Causes of terminological variation. In Pamela Faber and Marie-Claude L'Homme (Eds.), *Theoretical Perspectives on Terminology Explaining terms, concepts and specialized knowledge*, John Benjamins Publishing Company Amsterdam / Philadelphia, pp. 399–420. 10.1075/term.00071.ben

Gil-Berrozpe, Juan Carlos, León-Araúz and Pamela Faber (2019). Ontological Knowledge Enhancement in EcoLexicon. In Kosem, I., Zingano-Kuhn, T., Correia, M., Ferreira, J.P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. and Tiberius, C. (Eds.), *Proceedings of the eLex 2019 conference: Electronic lexicography in the 21st century*, pp. 177–197. Brno: Lexical Computing CZ, s.r.o

Hadouche, Fadila, Guy Lapalme and Marie-Claude L'Homme (2011). Attribution de rôles sémantiques à des actants. *Proceedings of TALN 2011*.

Hermann, Karl Moritz, D. Das, J. Weston, and K. Ganchev (2014). Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers*) pp. 1448–1458.

L'Homme, Marie-Claude, Benoît Robichaud, and Carlos Subirats Rüggeberg (2014). Discovering frames in specialized domains. *LREC* 201, pp. 1364–1371.

L'Homme, M. C. 2015. Predicative Lexical Units in Terminology. In Gala, N., R. Rapp and G. Bel-Enguix (Eds), *Language Production, Cognition, and the Lexicon*. Berlin: Springer, pp. 75–93.

Laufer, Batio and Tamar Levitzky-Aviad (2006). Examining the Effectiveness of 'Bilingual Dictionary Plus' — a Dictionary for Production in a Foreign Language. *International Journal of Lexicography* 19(2), pp. 135–155.

León Araúz, Pilar, Antonio San Martín and Pamela Faber (2016). Pattern-based Word Sketches for the ExtracP tion of Semantic Relations. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pp. 73–82. Osaka, Japan: COLING 2016.

Mairal Usón, Ricardo, and Pamela Faber (2002). Functional Grammar and lexical templates. In A. Machtelt Bolkestein, Casper de Groot and J. Lachlan Mackenzie, *New perspectives on argument structure in Functional Grammar*, Mouton de Gruyter, pp 39–94.

Ramisch, Carlos (2015). «Multiword Expressions Acquisition: A Generic and Open Framework». *Theory and Applications of Natural Language Processing series*, XIV. Springer. 10.1007/978-3-319-09207-2

Rojas-Garcia, Juan(2022). Representation of Hydronyms in Terminological Knowledge Bases on the Environment. PhD Thesis. University of Granada.

Sánchez-Cárdenas, Beatriz (2022) *Le syntagme verbal en français. Paradigmes, analyses et représentations*, Serie Estudios Franceses-Lingüística, Editorial Comares

Sánchez Cárdenas, Beatriz, and Carlos Ramisch (2019). *Framing specialized concepts through automatic extraction and semantic annotation: the deforestation event*. Proceedings of the 12th International Conference of The Asian Association for Lexicography «Lexicography in the Digital World», Krabi, Thailand.

Sánchez Cárdenas, Beatriz, and Carlos Ramisch (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. doi: https://doi.org/10.1075/term.00026.san.

Straka, Milan, and Jana Straková (2017). «Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe». *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada.

Straka, Milan, Jan Hajič, and Jana Straková (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, pp. 4290–4297.

Wierzbicka, Anne (1999). Emotional universals. *Language design: journal of theoretical and experimental linguistics*, 2, pp. 23–69.

Declaración de contribución de autoría

Conceptualización, curación de datos, análisis formal, investigación, metodología, administración del proyecto, recursos, software, supervisión, validación y visualización: Beatriz Sánchez Cárdenas