Eliciting specialized frames from corpora using argument-structure extraction techniques

Beatriz Sanchez Cardenas, Carlos Ramisch

Reference:

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

Authors: Beatriz Sánchez Cárdenas and Carlos Ramisch
Length: 8702 words (excluding references)

Beatriz Sánchez-Cárdenas
Research group LexiCon
Department of Translation and Interpreting
University of Granada
Calle Buensuceso, 11
18002 Granada (Spain)
(+34) 958244104
bsc@ugr.es
http://lexicon.ugr.es/sanchezcardenas

Carlos Ramisch
Aix Marseille Univ, Université de Toulon, CNRS, LIS
Parc Scientifique et Technologique de Luminy
163 Avenue de Luminy - Case 901
13288 Marseille Cedex 9 (France)
(+33) 4 86 09 06 72
Carlos.Ramisch@lis-lab.fr
http://pageperso.lis-lab.fr/~carlos.ramisch

Abstract

Frame Semantics provides a powerful cross-lingual model to describe the conceptual structure underlying specialized language. However, building specialized frames is challenging because of the complex nature of predicate-argument structures, and because of the domain-specific uses of general-language predicates. This article presents a semi-automatic method to elicit semantic frames from specialized corpora. Its goal is to discover lexical patterns that reveal the structure of specialized frames and to populate them with corpus-based data. Firstly, we automatically extracted verb-noun

triples from corpora using bootstrapping to identify noun-verb-noun phraseological patterns. Secondly, we annotated each noun-verb-noun triple with the lexical domain of the verbs and the semantic class and role of the noun filling each argument slot. We then used these annotations and patterns to classify similar triples. This allowed us to make generalizations and infer the structure as well as the types of lexical units that belong to these specialized frames. We evaluated our methodology using specialized corpora of environmental science texts in English and in Spanish.

Keywords

Frame semantics, frame-based terminology, corpora, corpus-based extraction, argument structure

Eliciting specialized frames from corpora using argument-structure extraction techniques

Abstract

Frame Semantics provides a powerful cross-lingual model to describe the conceptual structure underlying specialized language. However, building specialized frames is challenging because of the complex nature of predicate-argument structures, and because of the domain-specific uses of general-language predicates. This article presents a semi-automatic method to elicit semantic frames from specialized corpora. Its goal is to discover lexical patterns that reveal the structure of specialized frames and to populate them with corpus-based data. Firstly, we automatically extracted verb-noun triples from corpora using bootstrapping to identify noun-verb-noun phraseological patterns. Secondly, we annotated each noun-verb-noun triple with the lexical domain of the verbs and the semantic class and role of the noun filling each argument slot. We then used these annotations and patterns to classify similar triples. This allowed us to make generalizations and infer the structure as well as the types of lexical units that belong to these specialized frames. We evaluated our methodology using specialized corpora of environmental science texts in English and in Spanish.

Keywords

Frame semantics, frame-based terminology, corpora, corpus-based extraction, argument structure

1. Introduction

The study of phraseology in scientific texts tends to focus either on general scientific formulaic templates or on the study of terms for their inclusion in specialized dictionaries. However, the description of the language used in a given scientific or technical domain should go far beyond merely collecting an inventory of terms that are used to instantiate general-language constructs (L'Homme 2004, Hanks 2004, Williams 2005, Granger and Meunier 2008, Faber 2012). In fact, a significant part of specialized language is composed of structured lexico-grammatical constructs used to express complex concepts that are typical of a given domain. There is thus the need to develop specialized lexicons that provide this type of information.

This is particularly evident in translation. Translators dealing with specialized texts often have problems transposing the meaning of a sentence across languages because a superficial knowledge of the terms in a text is not sufficient. In addition to translating terms, it is necessary to translate actions and processes along with the entities that participate in them. For instance, a description of *earthquake* should include the entities that generally cause this event as well as its effect on other entities. This would afford translators a more in-depth knowledge of the concept and allow them to express it more idiomatically in the target language.

In our opinion, such a description should stem from the analysis of specialized corpora in the source and target languages. In this endeavor, domain-specific corpora are a rich

source of information. Given that verbs carry most of the semantic load of the sentence, they are essential to define the underlying conceptual structure of specialized texts (Fellbaum 1990; L'Homme 2012, 1998). Thus, the identification of noun-verb combinations in corpora is crucial to build structured descriptions.

The corpus-based construction of specialized lexical resources requires both linguistic and domain expertise, as well as suitable tools for performing corpus inquiries. Computational tools can support, enhance and facilitate corpus analysis to confirm and generalize linguistic introspection. Therefore, one often needs to run complex queries to model morphosyntactic and syntactic co-occurrence patterns, which in turn are proxies for predicate-argument structure.

Our research combined the principles of Frame-based Terminology (Faber 2012, 2015; Faber and León Araúz 2014) with computational tools for corpus searches, semantic annotation, and frame specification. For automatic corpus searches, we used the MWEtoolkit, a software application that extracts co-occurrence patterns from corpora using multi-level queries that support regular-expression operators (Ramisch 2015). This approach lies in the roots of a considerable amount of literature over the last 20 years on the identification of knowledge patterns in specialized texts (Faber et al. 2009, Feliu 2004, Condamines 2002, Condamines and Rebeyrolle, Meyer et al. 2001, Meyer et al. 1999, inter alia).

The output of the initial queries was analyzed and expanded, which helped to successively design new queries. This bootstrapping strategy eventually converged towards the description of complete variational patterns. The manual annotation of the semantics of these lexical co-occurrence patterns allowed us to cluster them, which ultimately led to the emergence of similar structures specifying the conceptual architecture of a concept in a specialized domain.

The first contribution of this article was the method used to identify phraseological patterns. More specifically, the terms designating specialized concepts in a knowledge base were used as seeds to create queries for the identification of specialized phraseological patterns. Multilevel corpus searches generated co-occurrence candidates that went beyond the immediate neighborhood of the words. Then, a bootstrapping strategy was defined for query expansion that uses the results of previous corpus searches to perform new ones, thus improving the lexical coverage of the results. The second contribution was the semantic annotation, which provided insights into the conceptual and linguistic behavior of terms in specialized corpora with a view to identifying a more abstract and language-independent representation.

The remainder of this article is structured as follows. Section 2 presents the theoretical background for our research. Section 3 explains the materials and the methodology used to query the corpora by means of lexical patterns. Section 4 describes the construction of specialized frames as well as the linguistic models that were used to semantically

annotate the data extracted from the corpora. Section 5 presents and discusses the results obtained and shows an example of the frame construction process. Section 6 presents conclusions derived from this research and our plans for future work.

2. Cognitive linguistics applied to specialized language

From the perspective of cognitive linguistics, language and cognition are a continuum stemming from our perception of the world (Langacker 1987). The assumption is that the linguistic system is not a set of rules independent of our experience. Rather, conceptual structure and language are embodied. As a consequence, language and thought are governed by physical experience from the world. One of the models that best suits these premises is Frame Semantics.

Frame Semantics (Fillmore 2006) is a model of knowledge representation that describes concepts according to their location in the whole conceptual system to which they belong. Its main underlying idea is that, since language reflects our cognition, it is possible to describe any language in terms of cognitive structures, known as *frames*. A frame is a schematic representation of a situation. In order to define a frame, it is first necessary to identify the main participants, or frame elements, of each schematized situation. Frame elements that are essential to specify the meaning of the frame are called *core frame elements*. FrameNet is resource that describes English according to Frame Semantics (Fillmore et al. 2003, Ruppenhofer et al. 2016).[1]

1 https://framenet.icsi.berkeley.edu

In L'Homme's model of specialized frames (L'Homme et al. 2014, L'Homme's and Pimentel 2012, L'Homme 2012), terminological resources are mainly based on the FrameNet methodology and are inspired by the information contained in this resource. Nevertheless, her specialized frames differ somewhat from those in FrameNet at both the lexical and conceptual levels (L'Homme et al. 2016). For example, at the lexical level, words that become terms behave differently (e.g. *mouse, introduce*) and thus must be redefined. At the conceptual level, specialized domains include new frames. This means that it is necessary either to adapt existing frames or to create new ones. In order to create a specialized frame under this model, concordances associated with a concept are extracted from the corpus. Then, the participants (or frame elements) of verbs describing a given action are manually annotated with linguistic information. This annotation includes the participants (or frame elements) of the action and their nature, their thematic roles, as well as the syntactic function and syntactic group of the participant. Nonetheless, this process has the considerable drawback of being a highly time-consuming task. There are various terminological resources based on L'Homme's proposal of specialized semantic frames (e.g. DicoEnviro and Juridico).[2] A review of other frame-based resources can be found in San Martin (2016).

Frame-based Terminology (Faber 2012, 2015) offers a slightly different perspective on specialized frames. This model applies the premises of Frame Semantics to the study of

2 Number of lexical units in DicoEnviro (November 2017): 973 in English, 1,277 in French, 172 in Spanish, 34 in Portuguese. Number of lexical units in DicoInfo (November 2017): 852 in English, 1,105 in French, 100 in Spanish.

the conceptual organization that underlies specialized domains. It shares many assumptions with Sociocognitive terminology (Temmerman 1997, 2000), Socioterminology (Gaudin 2003) and with the Communicative Theory of Terminology (Cabré 2003). Frame-based Terminology conceives specialized frames as schematic knowledge representations of the cognitive architecture of the expert, making specialized communication possible. Thus, predicates and their arguments correspond to a generic cognitive structure. Given that frames can provide a way to organize concepts and their relations in a specialized domain, this model has been applied to the study of specialized language. Since frames are not universal, they are only valid for a specific culture (Faber and Vidal Claramonte 2017). However, frames can be generalized for a cluster of cultures that share a certain number of features (e.g. Western culture). The same is true for specialized languages, where specific semantic frames are shared by all the experts of a domain as has been shown by neuroimaging fMRI experiments (Faber et al. 2017). An example of a specialized semantic frame in Environmental Science (Faber 2012, 2015) is the Environmental Event. One instantiation of this frame is evoked by the entry of *atmospheric event*, which includes the nouns and the verbs that participate in this concept:

ATMOSPHERIC EVENT:

- **Source-of**: atmospheric conditions (e.g. *low pressure*) can *form/originate/evolve into* an atmospheric event (e.g. *hurricane*, *cyclone*).
- **Movement_of**: atmospheric events can *rotate/spin/move* in a direction

- **Effect_of**: atmospheric events can (i) *impact/strike/make contact* with a landform (e.g. *coast, shoreline, area*); (ii) *trigger/produce/cause* water events (e.g. *flooding, waves*) or geological events (e.g. *landslide, debris flow*).

These elements participate in the action. Linguistically, they are generally nouns or noun phrases acting as the arguments of a verb.[3] According to Faber and León-Araúz (2016: 159) context codifies the pragmatic information that should be included in term entries and in the frame in which the concept is embedded. Specialized semantic frames can be used for many purposes, such as the elaboration of specialized dictionaries, the development of tools for writing specialized texts and their semi-automatic syntactic-semantic analysis. Moreover, they provide a way to cluster semantically related lexical units so as to account for the variability and language-independent dimensions of language production. Since specialized semantic frames are language independent, the instantiation of frames for more than one language is the basis for sophisticated multilingual resources, which can even be used to develop translation tools. Specialized frames allow users to understand concepts on the cognitive level, and to be able to use them on the lexical level.

This model is adopted and implemented in EcoLexicon, an environmental knowledge base developed by the LexiCon research group at the University of Granada (Spain).[4] This resource currently contains 4,385 concepts and 23,252 terms in four languages

3 Whereas some nouns such as *erosion* may also represent actions and events, we do not account for them explicitly in our study.

4 http://ecolexicon.ugr.es/en/index.htm

(English, Spanish, German, and Modern Greek) with three other languages under development (French, Russian and Dutch).[5] This article adopts the principles of Frame-based Terminology for the structure of specialized lexical resources and proposes a methodology for the creation of specialized frames.

3. Extraction methodology

As previously mentioned, our objective was to develop a systematic corpus-based method of extracting knowledge to support the frame creation process. This section describes how verb arguments were extracted from our corpora. For this purpose, we decided to use MWEtoolkit, a computational tool for making corpus queries and filtering their results. However, it was first necessary to do the following: (a) collect and clean the corpora; (b) pre-process them with automatic syntactic analyzers; (c) convert them into a suitable format.

3.1. Corpus description

The corpora for this research were collected within the context of a larger research project for the creation of a lexicon on environmental sciences. From these larger corpora, we extracted two sub-corpora of Spanish and English texts on volcanic activity. The corpora contain both academic texts and scientific outreach texts on topics such as volcanoes, magma, eruptions, tectonic plates, etc. The corpora include manually

5 Number of lexical units (November 2017): 5,257 in English, 4,864 in Spanish, 4,069 in German, 4,927 in Greek, 824 in French, 661 in Russian, 52 in Dutch.

selected documents as well as web-crawled texts, collected with WebBootCat with the help of seed keywords (Baroni et al. 2006).

Our corpora were initially a set of documents in text format. The first step involved the manual deduplication of the selected documents. All characters were then automatically converted to UTF-8, removing sentences that contained spurious characters that had not been correctly converted (e.g. some Spanish diacritics). A simple sentence-splitting program based on punctuation heuristics was also applied.[6] After removing single-word sentences, the whole corpora were run through UDPipe, a suite for natural language processing (Straka et al. 2016).[7] UDPipe read the corpus sentence by sentence and performed the following tasks:

1. Each sentence was tokenized into minimal units (roughly corresponding to words).
2. Each word was tagged with its part of speech (POS) using the Universal Dependencies tagset (Nivre et al. 2016).
3. Each word was associated with its canonical form or lemma.
4. A syntactic dependency tree was built with the relations between words, also using the Universal Dependencies tagset.

Although the output of the last step was in the end not used in our queries, we plan to explore syntactic dependencies in future work. The tokenization performed by UDPipe split Spanish contractions such as *del* (*de+el*, '*of+the*') and *al* (*a+el*, '*to+the*'), and

6 We used the sentence splitter included in Europarl: http://www.statmt.org/europarl

7 The models used by UDPipe were trained on UD v1.4: http://ufal.mff.cuni.cz/udpipe

replaced them with the individual tokens that compose them. After automatic parsing, the resulting corpus was in CONLL-U format, a tab-separated text file with one token per line and one piece of information (surface form, POS, lemma, etc.) per column. In order to run queries on large text collections, we used the MWEtoolkit to build an index for the corpus, including POS and lemmas, which permitted fast searches. The resulting specialized corpora on volcanic activity consisted of 33,837 sentences and 609,116 words in English, and 49,664 sentences and 1,222,944 words in Spanish.

3.2. Query and filtering tools

We focused on the volcanic activity concept. Corpus queries were performed with the MWEtoolkit, a computational tool for the discovery of multiword units in corpora (Ramisch 2015).[8] Although it was initially conceived as a tool for the construction of lexicons for fixed multiword expressions (Linardaki et al. 2010), it can also be used for any kind of corpus work that involves extracting co-occurrence patterns. This was our case, except that the co-occurring words were not directly used as phraseological lexical units, but rather to build profiles of selectional preferences for specialized frames. With a view to finding lexical elements that instantiate relations between concepts, we used morphosyntactic information and regular expression operators to define queries whose results were subsequently filtered. This was implemented as shell scripts which, in turn, used the MWEtoolkit scripts to extract candidates, count them, calculate association measures, and sort them. This tool works in two steps.

8 http://mwetoolkit.sourceforge.net

Step 1: Querying the corpora. Corpus queries take the form of multi-level regular expressions. Textual patterns were used to match verbs that combine with the target terms. Once we tuned the queries, we encapsulated them into easy-to-use scripts, which eliminated the tangle of the regular expressions and increased their readability.

```
[lemma="volcano" POS="NOUN"]
[]{repeat={0,3} ignore=true}
[pos="VERB"]
[]{repeat={0,3} ignore=true}
[lemma="lava" pos="NOUN"]
```

Figure 1. Query search for the relation between volcano *and* lava

Figure 1 shows an example of the MWEtoolkit query for verbs lexicalizing the relation between *volcano* and *lava*.[9] In this query, the strings between square brackets correspond to a token. For instance, the first token is a word whose lemma is *volcano,* whereas the last token is a word whose lemma is *lava*. Both have a constraint on the POS tag, which must be *NOUN*. Thanks to Universal Dependencies, the constraints on POS are identical for all languages. The middle token has no constraint on its lemma, meaning it can have any realization. However, there is a constraint on its POS tag, which must be *VERB*. The second and fourth elements are placeholders for any sequence of up to three words that can appear between nouns and verbs, and between verbs and nouns. These intervening words are never retrieved as part of the match (we set *ignore=true*) and are discarded from the output, since they will often correspond to

9 Each element of the query is shown on a different line for increased readability.

adverbs, determiners, and prepositions that do not carry much useful information for this task. When the query was made, MWEtoolkit then searched for matches in the indexed corpus. The output was a set of triples formed by a first noun (n_1), a verb (v), and a second noun (n_2). It was assumed that these triples captured co-occurrence patterns that reflected the verb's argument structure. However, a certain number of triples were spurious and had to be discarded (for instance, in the sentence *the volcano that you see contains no lava*, the query yielded the triple *volcano-see-lava*).

Step 2: Filtering and sorting the results. One advantage of the MWEtoolkit is that it includes many filters that help to clean query results. In a large-scale experiment, this speeds up lexicographic work by looking only at the most relevant output. We used a standard association measure, pointwise mutual information (PMI), to sort the query results described above in descending order.

In practice, each triple was counted in the corpus from which it was extracted. Occurrence counts were thus obtained for the triples, denoted as $c(n_1,v,n_2)$, as well as for the individual words composing it, denoted as $c(n_1)$, $c(v)$ and $c(n_2)$. These were then combined into a single relevance score, PMI, which estimated the extent to which the co-occurrence of these items was unexpected with respect to random co-occurrence in a corpus of N words (Church and Hanks, 1990):

$$PMI(n_1, v, n_2) = \log_2 \frac{c(n_1, v, n_2) \times N^2}{c(n_1) \times c(v) \times c(n_2)}$$

High PMI values indicate that the association of the triple n_1, v, n_2 is relevant, whereas low PMI values indicate that this is just random co-occurrence. Therefore, the triples extracted from the corpus were ranked in descending order, based on their PMI. The most relevant information thus appeared at the beginning of the resulting list. The list of extracted and ranked triples was then stored for further manual analysis in the frame creation process (see Section 4).

Since MWEtoolkit commands and query language are complex and cannot be easily memorized, we encapsulated them into a command-line tool **search-triples.sh** that has three parameters: **<noun1> <verb> <noun2>**. This script can query the corpus for sequences of words with lemmas noun1, verb, and noun2, allowing 0 to 3 intervening words to appear in between the verb and each noun. Each of the three elements can be underspecified by using the special keyword *ANY*, which means that the query will return any nominal or verbal lemma in that position. For example, the query *ANY ANY ANY* extracts all noun-verb-noun pairs where there are no more than three intervening words between the verb and each noun. The three parameters can also specify a set of lemmas, such as *(lava|magma|rock)*, which means that the query will return any triple containing the lemmas *lava*, *magma*, or *rock*. Afterwards, the tool counts the triples and individual words and calculates the PMI score, sorting the output in descending order of relevance. Results are stored in a TSV file (tab-separated values), editable in spreadsheet editors.

3.3. Search result bootstrapping

A bootstrapping method was then used to identify which verbs were associated with each term in the corpora. This was done incrementally by starting from a set of seed nouns. The previously described query and filtering strategy involved the specification of the lemmas of (some) elements for which co-occurrence triples (noun-verb-noun) were subsequently extracted. In our work, these lemmas were obtained by a bootstrapping procedure in which the results of queries were used to build new ones, gradually expanding the representativity of the results until a large portion of the phraseological spectrum was covered.

The initial set of seed terms was obtained from a terminological inventory of the domain, in our case, volcanic activity. First, a seed pair of nouns was used in the noun1 and noun2 positions in the queries. This search returned a set of verbs that connected the concepts designated by these nouns. These verbs were then reused in conjunction with one of the two initial nouns to extract other nouns that might appear in the noun1 or noun2 positions. Every time a query was run, one of the three elements was underspecified with *ANY* while the two others were specified as a set of possible words, according to the current query results. In the end, we obtained a set of triples that covered many variation patterns for the construct of the conceptual frame.

For instance, according to our base lexicon, the concepts ERUPTION and LAVA are semantically related to the concept VOLCANO (VOLCANO causes ERUPTION; LAVA is

located at VOLCANO). Thus, we formulated the queries "volcano ANY eruption" and "volcano ANY lava". These queries retrieved verbs such as *eject*, *spew* and *cause*.

These verbs were used to look for other nouns in the second argument with the search "volcano (eject|spew|cause) ANY", which retrieved results such as *steam*, *continent*, or *explosion*. This procedure was performed until no new relevant nouns or verbs were found, thus covering most of the combinatorial patterns of the concept VOLCANO in the corpus. The inventory of triples was used as raw material for the manual annotation and elicitation of specialized frames.

4. Frame construction based on argument structure generalization

The methodology described led to the extraction of noun-verb-noun triples associated with a concept and which play a significant role in specialized frame construction. For instance, the volcanic eruption frame includes verbs such as *emit*, *spew* and *expel* in the form of triples (*volcano-expel-ash*). This linguistic information extracted from corpora presumably corresponds to the knowledge shared by the domain experts.

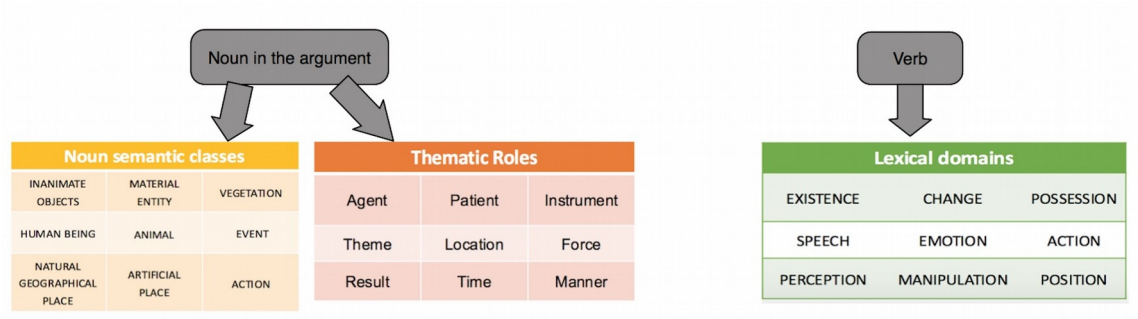

Figure 2. Description of nouns using semantic classes and thematic roles, and verbs using lexical domains

The automatically extracted triples were then annotated with linguistic information, which in turn led to further generalizations regarding argument structure and underlying conceptual meaning. To this end, verbs were assigned to a lexical domain, whereas the noun phrases in the verb argument slots were attributed a thematic role and a semantic class (see Figure 2).

Once the annotation of lexical domains, semantic classes and thematic roles was completed for a significant number of triples, they were clustered based on their shared argument structure. This highlighted the underlying conceptual structure of the concept. This section describes the linguistic models used to describe and annotate the linguistic information extracted from the corpora in order to infer similar argument structures.

4.1. Characterizing the nuclear meaning of verbs

We operated on the premise that the meaning of a verb constrains the number and meaning of its arguments. Verbs whose definitions shared the same generic term were grouped together to obtain similar argument structures. We based our annotation on the Lexical Grammar Model (Faber and Mairal, 1999, 2017; Mairal and Faber 2002), stemming from Martín Mingorance's Functional-Lexematic Model, integrating Coseriu's Lexematics (1977) and Dik's Stepwise Lexical Decomposition (1978). After an analysis of 12,000 English verbs, Faber and Mairal (1999) obtained eleven macro-classes of verbs, each defined by a nuclear term or genus: EXISTENCE (*to be, to*

happen), CHANGE (*to become*), POSSESSION (*to have*), SPEECH (*to say*), EMOTION (*to feel*), ACTION (*to do, to make*), COGNITION (*to know, to think*), MOVEMENT (*to move, to go, to come*), PHYSICAL PERCEPTION (*to see, to hear, to taste, to smell, to touch*), MANIPULATION (*to use*) and POSITION (*to put, to be*).

For example, in *the volcano spews ashes*, and *the volcano emits lava*, both *spew* and *emit*, belong to the lexical domain of MOVEMENT, since their nuclear meaning is *to cause something to go out.* The generic term in definitions is obtained through semantic factorization (Dik 1978, Faber and Mairal 1999, Sánchez Cárdenas 2011), based on the information in lexicographic resources. When a verb can no longer be decomposed in terms of a more general one, the genus representing its lexical domain has been reached. Figure 3 shows this process for the verb *to spew* whose genus is *to go*, which belongs to the MOVEMENT lexical domain (Sánchez Cárdenas and Faber 2014):

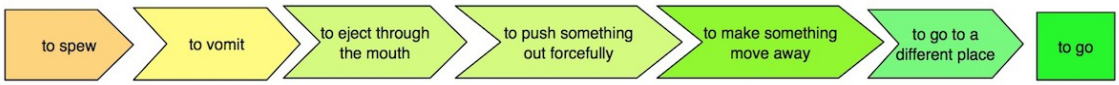

Figure 3. Lexical decomposition of the MOVEMENT verb to spew

Moreover, each subdomain has an internal hierarchical structure. Sánchez and Faber (2014) studied French verbs related to volcanic activity in the lexical domain of MOVEMENT (*dégager, émettre, laisser échapper, exhaler, rejeter, cracher, éjecter*). These verbs were found to maintain dependency links, based on the following semantic

properties: MANNER, INTENTION, FIGURE, GROUND, CAUSE, SOURCE, GOAL, PATH, and DIRECTION. The hierarchical entailment is shown in Figure 4.

1. dégager: libérer une substance
Les volcans dégagent du souffre.
\+ SOURCE (initial location)
1.1. émettre: dégager quelque chose (de soi).
Le volcan de l'île de Vulcano émet des fumerolles.
\+ FIGURE (object that undergoes the event)
1.1.1. laisser échapper: émettre ce qui était retenu.
Le volcan laisse échapper des gaz.
1.1.2. exhaler: émettre des substances gazeuses.
Le volcan a exhalé 20 millions de tonnes de CO2 .
\+ GOAL (final location)
1.1.2.1. rejeter: exhaler hors de soi.
Le volcan de l'Eyjafjallajokull rejette des cendres.
\+ SOURCE (initial location)
1.1.2.1.1. cracher: rejeter hors de la bouche.
Le volcan a craché un panache de cendres.
\+ MANNER (how the event develops)
1.1.2.1.2. éjecter: rejeter avec force.
Le volcan a éjecté d´énormes masses de laves et de cendres.

Figure 4. Verbs hierarchical entailment in the field of volcanology (Source: Sánchez and Faber 2014)

Finally, it should be noted that lexical domains are a useful way to differentiate senses of polysemous verbs occurring in sentences with the same configuration:

1. [*The storm* Theme] *reached* [*category 1 hurricane intensity* Magnitude] → EXISTENCE
2. [*The storm* Theme] *reached* [*the coast* Location] → MOVEMENT

4.2. Characterizing the ontological nature of nouns

In order to describe the lexico-grammatical behavior of verbs, it is also necessary to characterize their arguments not only from the general perspective of their thematic roles (Section 4.3), but also from the perspective of the ontological identity of the head nouns in phrases. For instance, the verb *destroy* in the sentence *The lava flow destroyed the tropical forest* has two arguments: an Agent/Force whose noun phrase belongs to the semantic category of "Geological entity" and a Patient whose noun phrase could be classified as "Natural place".

Describing the ontological nature of a noun is useful to describe the selectional preferences[10] of the verbs with which the noun co-occurs (Hatier et al. 2016). This type of knowledge is paramount in text production. Semantic classes specify the ontological nature of nouns in predicative expressions, and thus can help to predict interlinguistic argument structure equivalences (Sánchez and Buendía 2012; Buendía and Sánchez 2016). Many attempts have been made by philosophers, linguists, and computational linguists to classify reality into a system of hierarchical conceptual classes. Although there have been some successful initiatives for general language (Huyghe 2015; François et al. 2007; Fellbaum 1998, Flaux and Van de Velde 2000; Dubois and Dubois-Charlier 1997), there is no classification as yet for Environmental Science.

10 According to a well-grounded linguistic tradition (Firth 1961; Sinclair 1991; Hanks 2004; Halliday et al. 2014) selectional restrictions are crucial in order to describe the linguistic behavior of words (Hanks 2012). However, these restrictions are not clear-cut. Rather, they behave sequentially in a range that goes from highly prototypical (thus very frequent) to highly improbable (or very infrequent). For this reason, *selectional preferences* is a preferable term over *selectional restrictions* (Hanks 2012).

Such a classification would presumably be based on a well defined set of criteria as well as their granularity and the procedures followed to coerce nouns into a specific semantic class. Since an exhaustive answer to these questions would exceed the scope of this article, we can only provide a brief outline of our methodology. From a linguistic perspective, it is commonly accepted that noun typologies can be either referential or non-referential. Referential categorizations of nouns are ontological or taxonomical since they take into account the properties of the nouns' referents. Some examples of referential noun classes, which Gross (1994, 2008) calls "object classes", are Natural Objects, Human Beings, and Emotions. This classification can also be based on universal binary properties such as concrete / abstract or human / non-human.

Furthermore, non-referential classifications (Huyghe 2015) use linguistic criteria, which lead to classes such as relational nouns (*neighbor, father, victim*), partitive nouns (*head, piece, handlebar*), collective nouns (*herd, ensemble, crowd*), referentially autonomous nouns (*table, school, tree*) or referentially non-autonomous nouns (*quantity, circle*). We have taken into account some of these linguistic properties to make decisions regarding the inclusion of certain nouns in a class. For instance, the property *countable / uncountable* can differentiate events and actions, since the first ones are countable (*death, match, explosion*) whereas the latter are uncountable (*erosion, absorption, deforestation*). According to classic linguistic parameters, our classes are organized in terms of four binary distinctions: [concrete / abstract], [animate / inanimate], [human /

non-human], [natural / artificial].[11] The combination of these properties can be confined into six coarse-grained categories:

1. +[concrete, animate, human]
2. +[concrete, animate, non-human]
3. +[concrete, inanimate, natural]
4. +[concrete, inanimate, artificial]
5. +[abstract, inanimate, natural]
6. +[abstract, inanimate, artificial]

These classes pre-exist the study of corpora, since they are based on general language-independent notions. Nonetheless, one drawback of such a typology is that these classes have a very general semantic spectrum. As a solution, each class was organized into more fine-grained corpus-specific semantic categories. Each class was defined with a battery of distributional tests (Gross 1994, 2008; Flaux and Van Velde 2000), based on the semantico-syntactic properties of each noun class. These tests are based on the selectional preferences of nouns in their lexical environment and on the preferences that other elements, mostly verbs, impose on these nouns. We illustrate our categorization procedure with the following semantic classes: natural entities, events and actions.

4.2.1. Natural entities

11 Since nouns that are relevant for terminological purposes have an autonomous existence, the distinction between categorematic and syncategorematic was not considered, i.e., nouns with autonomous existence (*wind*, *tree*, *eruption*) versus nouns ontologically dependent on another noun (*quantity*, *volume*, *meter*).

The main feature of natural entities is that their referents have spatial properties, such as distance or volume, and sensorial properties, such as appearance. They have not been created by humans, and they can be perceived by the senses. This category can be divided into inanimate objects, material entities, and natural geographic places. Their distributional properties are described in the following sub-sections.

4.2.1.a. Inanimate objects: *mineral, rock, branch*

This category includes countable nouns referring to entities which are concrete, inanimate and natural, such as *mineral*, *rock*, and *branch*. Although the referent of *branch* belongs to a living entity (*a plant*), *branch* and *plant* behave differently in terms of selectional preferences. Natural entities cannot be the theme of creation predicates with a human creator, though they can be created by natural processes:

a) ?This rock / branch / mineral has been created by Peter.

Given the fact that they are material and autonomous entities, they can be weighed and measured:

b) This rock / branch / mineral weighs …

c) This rock / branch / mineral measures...

4.2.1.b. Material entities: *gas, smoke, ash, lava*

The referents of material entities also refer to concrete entities, often describing substances (*gas, smoke*) and elements involved in natural processes (*ash, lava*). In short,

the referents of the category of material entity have the same attributes as inanimate objects, but they are uncountable:

*d) *one gas / smoke / ash / lava*

e) one rock / branch / mineral

4.2.1.c. Natural geographical places: *forest, riverbed, shore*

Natural geographical places refer to natural entities that can assume the role of locations for other entities and events, such as *forest, riverbed, shore*. These nouns share the properties of inanimate objects and material entities. Consequently, they can be perceived by the senses (concrete) and they cannot be created by humans (natural). Unlike inanimate objects, however, they can be measured but not weighed:

f) This forest / riverbed / shore measures…

g) ?This forest / riverbed / shore weighs…

Another characteristic of natural geographical places is that, unlike inanimate objects, they can have the thematic role of Location (Flaux and Van de Velde 2000: 48) and, accordingly, their internal extension can be perceived and explored:

h) To be in / to get to / to leave from the forest / riverbed / shore.

i) ?To be in the rock / branch / mineral.

j) To visit / take a walk through the forest / riverbed / shore.

k) ?To visit / take a walk through the rock / branch / mineral.

4.2.2. Actions and events

The referents of the categories action and event are abstract situations that take place:

l) The explosion / eruption / treatment has occurred yesterday.

m) The explosion / eruption / treatment took place yesterday.

n) There has been an explosion / eruption / treatment.

The class of action and event nouns has a high degree of complexity, given their spatial and temporal properties. Unfortunately, even though various studies have been carried out in this domain, an exhaustive typology of this kind of nouns has not as yet been established (Huyghe 2015: 11).

There are many possible ways of classifying actions and events, according to several non-exclusive parameters such as agentivity or referential autonomy (Flaux and Van Valde 2000). Generally speaking, actions have a homogenous internal development whereas events are heterogeneous. In this regard, if we eliminate the temporal part of them, actions remain the same, but not events. For example, an evaporation (for instance of a lake) that lasts 5 minutes less is still an evaporation, but an explosion that lasts for 5 minutes less might not be an explosion any more. In other words, events do not have internal limits (e.g. *swimming*) as compared to actions, which have a beginning and an end that constitutes their goal (e.g. *voyage*).

In contrast to events, actions are not susceptible to individualization and so they cannot be modified by number adjectives meaning "more than one":

o) ?Two stabilizations / contaminations

p) Two eruptions / explosions / transformations

To indicate the temporal extension of events, it is necessary to add their duration.

q) The stabilization / treatment / contamination lasted for 20 minutes.

Following this process, it was possible to establish a list of the semantic classes for the nouns in the corpora. Table 1 summarizes the noun typology used in our annotation.

Formal features			**Name of the category**	**Examples**
concrete	**animate**	**+human**	**HUMAN BEING**	*Paul, people, victim*
		-human	**ANIMAL**	*sheep, dog, cow*
			VEGETATION	*tree, plant, flower*
	inanimate	**natural**	**INANIMATE OBJECT**	*mineral, rock, ash*
			WATER BODY	*lake, river*
			ATMOSPHERIC ENTITY	*atmosphere, sky*
			MATERIAL ENTITY	*gas, smoke, ash, lava*
			NATURAL GEO. PLACE	*forest, riverbed, shore*
			LANDFORM	*island, coast, continent*
abstract			**ACTION**	*earthquake, deflagration*
			EVENT	*flooding, erosion*
			WATER EVENT	*wave, tide*
		artificial	**MAGNITUDE**	*depth, temperature*
			ARTIFICIAL PLACE	*house, building*

Table 1. Semantic noun classes in the volcanology domain.

This preliminary typology of noun semantic classes will be completed when the whole environmental science domain has been studied in depth. Our plan is to eventually combine manual classification with automatic distributional tests that measure semantic

similarity between words. In a preliminary study that will be developed in future work, we conducted a pilot experiment of lexical patterns using distributional vectors. In this experiment we grouped together all the clusters sharing the same distribution (i.e., sharing the first noun, the verb and the second noun), such as: *volcano* [*erupts, ejects, emits*] [*lava, ash, gas, steam*].

4.3. Characterizing the thematic role of noun-verb pairs

Frame Semantics describes the participants of an action in terms of frame elements, which are semantic characterizations of the semantic behavior of noun phrases. For instance, the *commercial transaction* frame includes elements such as a buyer, a seller, goods, and money (Petruck 1996), whose behavior is lexicalized by verbs such as *to buy, to sell* or *to pay*. All of them codify our cognitive perception of the same event, each one focusing on a different perspective (the buyer, the seller, the payment).

However, one of the drawbacks of Frame Semantics is that the list of all the possible frame elements that describe the participants is never ending. As Frame Semantics linguists themselves claim, this inventory will only be completed once the whole language is described. Even if it was finished, the repertoire would be too large to allow generalizations in specialized language. The disadvantage of using an open inventory of thematic roles to describe argument structure is that it is difficult to use them to generalize linguistic behavior of lexical units sharing similar argument structures. A closed set of thematic roles is thus preferable.

This drawback is addressed in Construction Grammar (Goldberg 1995), a model that views grammar and meaning as a continuum. More specifically, the meaning of the words is completed by the meaning of their syntactic constructions. In order to describe argument structure, Goldberg (1995) uses *argument roles*, which are generalizations of frame elements, to characterize verbs and clauses. For instance, since giver, sender and lover are semantically coherent, they can be fused into the argument role Agent. Thus, the model states that each type of verb construction has a different configuration of argument roles. There are five classes of argument structure constructions: ditransitive, caused motion, resultative, intransitive motion and conative. For instance, the ditransitive construction has the structure *X causes Y to receive Z*, where *X* is the Agent, *Y* is the Recipient and *Z*, the Patient.

In specialized language, Buendía-Castro (2013: 377) proposes the following set of thematic roles to describe environmental sciences: Agent, Natural Force, Destination, Experiencer, Frequency, Geographical Location, Manner, Path, Patient, Situation, Origin, Theme, Time, and Result. This role set has the drawback of including some ontological noun properties in the characterization of the roles (e.g. Geographical Location vs. Destination). Based on previous proposals, our closed inventory of thematic roles for the arguments typical in Environment Science are the following:

Agent: A volitional participant that initiates an action or event that affects a Theme or Patient, and which can have a Result.

[*Humans* Agent] *have created* [*a climate catastrophe* Result].

[*People* Agent] *have caused* [*great damage* Result] [*to the environment* Patient].

Force: A non-volitional force, process, or event that produces a new entity or transforms a Patient, affects a Theme or produces a Result.

[*The volcanic eruptions* Force] *create* [*new islands* Patient].

[*The volcano* Force] *spews* [*lava* Theme].

[*The storm* Force] *ravages* [*the coast* Patient].

[*The hurricane* Force] *hits* [*the State of Florida* Theme].

Theme: A participant affected by an event that changes its possessor or location but not its internal structure.

[*The hurricane* Theme] *moves* [*towards the coast* Location].

[*The storm* Theme] *reached* [*the coast* Location].

Patient: An entity that undergoes a transformation as a consequence of an external action, namely by an Agent or a Force.

[*The storm* Force] *ravages* [*the coast* Patient].

[*The river* Force] *erodes* [*the landscape* Patient].

[*The volcanic eruptions* Force] *create* [*new islands* Patient].

Result: An event whose existence is produced as a consequence of an external Agent

or Force

[*Volcanic eruptions* Force] *can cause* [*continent shifts* Result].

Instrument: An instrument used by an Agent or to perform an event that affects a Patient.

[*The sludge* Patient] *is treated* [*by the wastewater treatment plant* Instrument].

Location: A place where an event occurs. (The roles of Path, Source, and Goal are considered as Location).

[*The hurricane* Theme] *formed* [*over the Atlantic Ocean* Location].

Manner: This role identifies the way in which the action takes place such as its frequency, intensity or mode.

[*The storm* Theme] *grew* [*in intensity* Manner].

The extracted triples were then manually annotated with these thematic roles. Since this is a time-consuming task, future research will explore the automatic annotation of corpora with thematic roles, as suggested by Hadouche et al. (2011).

4.4. Grouping similar argument structures

The last step before structuring and populating the specialized semantic frames was the automatic grouping of similar structures. For this purpose, all triples with the same structure were grouped together. For example, since the triples *volcano-eject-lava* and *volcano-emit-gas* were both tagged as Force (Landform) –MOVEMENT– Theme (Material Entity), they were considered to have the same structure. Similar frames were

regarded as having identical or similar semantic structures. Groupings of annotated triples were then shown to the lexicographers who were in charge of building the semantic frames.

5. Results and analyses

This section describes how the triples with the semantic participants related to volcanic activity were organized into semantic frames. As previously mentioned, some of the triples that had been extracted were excluded since they were not relevant (e.g. *volcano-be-big*). Those containing any inaccuracy (e.g. *volcano-eject-could*) were marked as errors. A total of 114 triples in English and 107 in Spanish were obtained. The triples marked as irrelevant/errors are analyzed in Section 5.1.

During the annotation process, each verb was tagged with a lexical domain. Nouns were also tagged with their semantic class, such as landform (e.g. *continent, island, volcano*), material entity (e.g. *gas, smoke, ash*), action (e.g. *eruption, explosion*) or human being (e.g. *people, victims, citizens*), as described in Section 4.2. Intransitive verbs were also included (e.g. *volcano-erupt*) and tagged as having a single argument. Then, the noun arguments were assigned a thematic role, such as Agent (e.g. *volcano, explosion, eruption*), Theme (e.g. *gas, smoke, ash*) or Result (e.g. *island, land, death*). Figure 5 shows an example of the English annotation.

n1	TR1	SC1	verb	Domain	n2	TR2	SC2
volcano	force	landform	belch	MOVEMENT	ash	Theme	material entity
volcano	force	landform	eject	MOVEMENT	ash	Theme	material entity
volcano	force	landform	expel	MOVEMENT	ash	Theme	material Entity
volcano	force	landform	spew	MOVEMENT	ash	Theme	material entity
volcano	force	landform	shift	MOVEMENT	continent	Theme	landform
volcano	force	landform	create	EXISTENCE	continent	Patient	landform
volcano	force	landform	create	EXISTENCE	island	Patient	landform
volcano	force	Landform	blow	MOVEMENT	lava	Theme	material entity
volcano	force	Landform	produce	EXISTENCE	lava	Patient	material entity
volcano	force	landform	spew	MOVEMENT	lava	Theme	material entity
volcano	force	landform	spew	MOVEMENT	magma	theme	material Entity
volcano	theme	landform	be	POSITION	island	location	landform
volcano	force	landform	eject	MOVEMENT	material	Theme	material Entity
volcano	force	landform	erupt	MOVEMENT	rock	Theme	material entity
volcano	force	landform	create	EXISTENCE	collapse	Result	event
volcano	force	landform	cause	EXISTENCE	destruction	Result	event

Figure 5. Example of the English corpus annotation

First, each triple (n_1-verb-n_2) was annotated with five different tags: one lexical domain for each verb (domain), a thematic relation (TR) and a semantic class (SC) for each of the two nouns (n_1 and n_2). Then, we automatically grouped all the results according to the five tags of each line. As a result, we were able to infer and describe various lexical structures in English and Spanish, as depicted in Figures 6 and 7.

volcano	FORCE	LANDFORM	***cause***	EXISTENCE	***death collapse, destruction, shift***	RESULT	EVENT
volcano	FORCE	LANDFORM	***form, be, make, create***	EXISTENCE	***landmass, continent, land, island***	PATIENT	LANDFORM
volcano	FORCE	LANDFORM	***form, produce***	EXISTENCE	***lava ash steam***	THEME	MATERIAL ENTITY
volcano	FORCE	LANDFORM	***emit, blow, erupt, spew, belch, shoot, eject, expel***	MOVEMENT	***steam, lava, ash, material, gas, magma, rock***	THEME	MATERIAL ENTITY
volcano	THEME	LANDFORM	***locate***	POSITION	***island***	LOCATION	LANDFORM

Figure 6. Grouped annotation of the English corpus

volcán	FORCE	LANDFORM	***destruir***	EXISTENCE	***ciudad***	PATIENT	ARTIFICIAL PLACE
volcán	FORCE	LANDFORM	***ocasionar, generar, provocar***	EXISTENCE	***víctima***	PATIENT	HUMAN BEING
volcán	FORCE	LANDFORM	***formar***	EXISTENCE	***isla, coordillera, estructura***	PATIENT	LANDFORM
volcán	FORCE	LANDFORM	***ocasionar, producir, esparcir, provocar, causar***	EXISTENCE	***explosión, muerte, actividad, catástrofe, contaminación, agrietamiento, daño, avalancha, depresión, destrucción, inundación, tsunami***	RESULT	EVENT
explosión, efecto, enfriar, cráter, volcán, cono, presión	FORCE	LANDFORM	***registrar, producir***	EXISTENCE	***lava, ceniza, tierra, vapor, lluvia***	THEME	MATERIAL ENTITY
erupción, volcán	FORCE	LANDFORM	***emitir, expulsar, formar***	MOVEMENT	***agua, ceniza, lava, material geológico, roca, tierra, vapor, hidrógeno, ácido, gas, ceniza***	THEME	MATERIAL ENTITY
volcán	LOCATION	LANDFORM	***registrar, incluir, presentar***	POSITION	***vapor***	THEME	MATERIAL ENTITY
volcán	LOCATION	LANDFORM	***registrar***	POSITION	***episodio, actividad, erupción, exhalación***	THEME	EVENT

Figure 7. Grouped annotation of the Spanish corpora

5.1. Error analysis

As previously mentioned, some of the triples could not be annotated because of the semantic or syntactic nature of the nouns and verbs automatically retrieved from the

corpora. Of the 114 triples from the English corpus, 65 were relevant (57.02%). Of the 107 triples from the Spanish corpus, 88 were relevant (82.24%). Although the results were acceptable in the Spanish corpora, the percentage of relevant results was somewhat less satisfactory for the English one. The error analysis reflects the lower percentage of relevant findings, which will lead to the enhancement of this protocol in future research. Whereas we only show examples in English, the same problems happened in Spanish, though in a lower proportion.

The main problems in relation to verb extraction stemmed from the following:

a) General-language verbs that were not relevant to the specialized field, either because of their specific use in scientific writing (***volcano*** *lava flows* ***suggest*** *that…*) or because the term (*volcano)* was extracted by MWEtoolkit as an argument of the verb and thus was sometimes erroneously considered as the subject of the verb (*The government ordered the evacuation of people living near the* ***volcano***, ***stating*** *that…).*
b) Constructions such as the *going-to* future where it was difficult to distinguish the main verb from auxiliary elements (e.g. *a* ***volcano*** *is* ***going*** *to erupt).*
c) Verbal periphrasis and idioms (***take*** *form*) where the second element was missing.
d) Phrasal verbs (***develop*** *into*).

As for the nouns present in the triples, the main obstacles encountered during the annotation process were the following:

a) The presence of syncategorematic nouns that required a second noun to be accurately understood, such as *column, amount, flow* (e.g. ***volcanoes*** *expel* ***columns*** *of ashes*).

b) Multi-word terms such as *eruptive pulse* or *tectonic plate shift*, where only the second noun was extracted.

c) Triples such as *volcano-erupt-explosion*, *volcano-erupt-plate*, in which the second noun did not correspond to an argument but rather to an adjunct or non-obligatory complement (e.g. *for a* ***volcano*** *to* ***erupt****, the* ***plates*** *of the earth crust collide and a volcano erupts; A revived Japanese* ***volcano*** *has* ***erupted*** *with its biggest* ***explosion***).

d) Subordinate relative clauses that led to spurious triples. For example, *lava is molten rock that a* ***volcano expels*** *during an* ***eruption*** produced *volcano-expel-eruption* instead of *volcano-expel-molten rock.*

However, these problems could be solved in future research by contrasting those triples with a classic manual concordance search in the same corpora, using a tool such as Sketch Engine. Nevertheless, in this experiment we decided not to compensate for the mistakes encountered, and thus ignored those triples that were not self-explanatory, in order to test the reliability and accuracy of this protocol. However, as reflected in the results obtained, the annotation of the excluded triples would not necessarily lead to the

creation of new frames apart from the ones described below. Future research would need to corroborate this hypothesis.

Recent advances in computational text processing allow the representation of words as vectors in a semantic space (i.e. word embeddings). One of their most interesting properties is that they allow us to compare words based on vector operations (e.g. cosine similarity). In the near future, we plan to study the use of word embeddings to represent nouns and verbs in our triples. Hence, our software could be enhanced to cluster similar triples, suggest unattested ones, and filter out spurious ones. For instance, the triples (*volcano, spew, lava*) and (*crater, expel, magma*) could be automatically grouped before manual annotation, since their components are close in the word embeddings space. This means that the similarity is high between *volcano* and *crater*, *spew* and *expel* and *lava* and *magma*. Moreover, if these two triples are observed in the corpus, we could suggest new ones, which have never been attested, such as (*volcano, eject, rock*), since the verb *to eject* is similar *to spew* and *rock* is similar to *lava*. Finally, vector representations could help detect spurious triples such as (*volcano, eject, column*) since the word *column* would be considered dissimilar to all other nouns occupying the third position in triples involving *volcano* and *eject*.

Future research should explore more complex events and face new challenges such as the identification of causal relations, such as the triple *tectonic plate-cause-earthquake*. For instance, the description of *tectonic plate* would presumably retrieve from the

corpora verbs such as *erode, collide, and shift*, as well as nouns like *earthquake*, *eruption*, and *continent* in the form of triples (*tectonic plate-cause-earthquake*). Nevertheless, the real Agent of the action is the noun *movement of tectonic plate* (Barrière 2001).

5.2. Semantic frames of volcanic activity

As has been explained, each triple was annotated with five different tags (see Figure 2). All of this information was then automatically re-sorted and similar argument structures were grouped together. The results showed different lexico-grammatical structures associated with a concept. The process was performed in parallel in English (Figure 6) and Spanish (Figure 7). Figures 6 and 7 show the shared elements, which highlight the underlying conceptual architecture. The clusters of the triples shown in Figures 6 and 7 are meaningful since they reveal the lexical patterns of the volcanic activity in both languages.

These lexical templates point to the cognitive structure of the volcanic event. Figure 8 shows its graphical representation. As can be seen, each semantic class, represented with different colors, has a prototypical conceptual relation with another semantic class through a different thematic role. For instance, the semantic class LANDFORM, can act as a Location in the lexical domain of POSITION, where the second argument is a MATERIAL ENTITY with the thematic role of Theme. LANDFORM can also take the role of Force in the domain of EXISTENCE. In this case, the second argument can

either be an EVENT with the thematic role of a Result or Patient, thus making reference either to a destructive EVENT (Result) or to the creation of a new LANDFORM (Patient). The lexicalization of verbs expressing these relations varies in English and Spanish.

From a cognitive perspective, *volcano* participates in three lexical domains: EXISTENCE, MOVEMENT and POSITION. Besides containing one or various argument structures, each domain accounts for a different dimension of volcanic activity since they represent the different scenarios in which *volcano* participates.

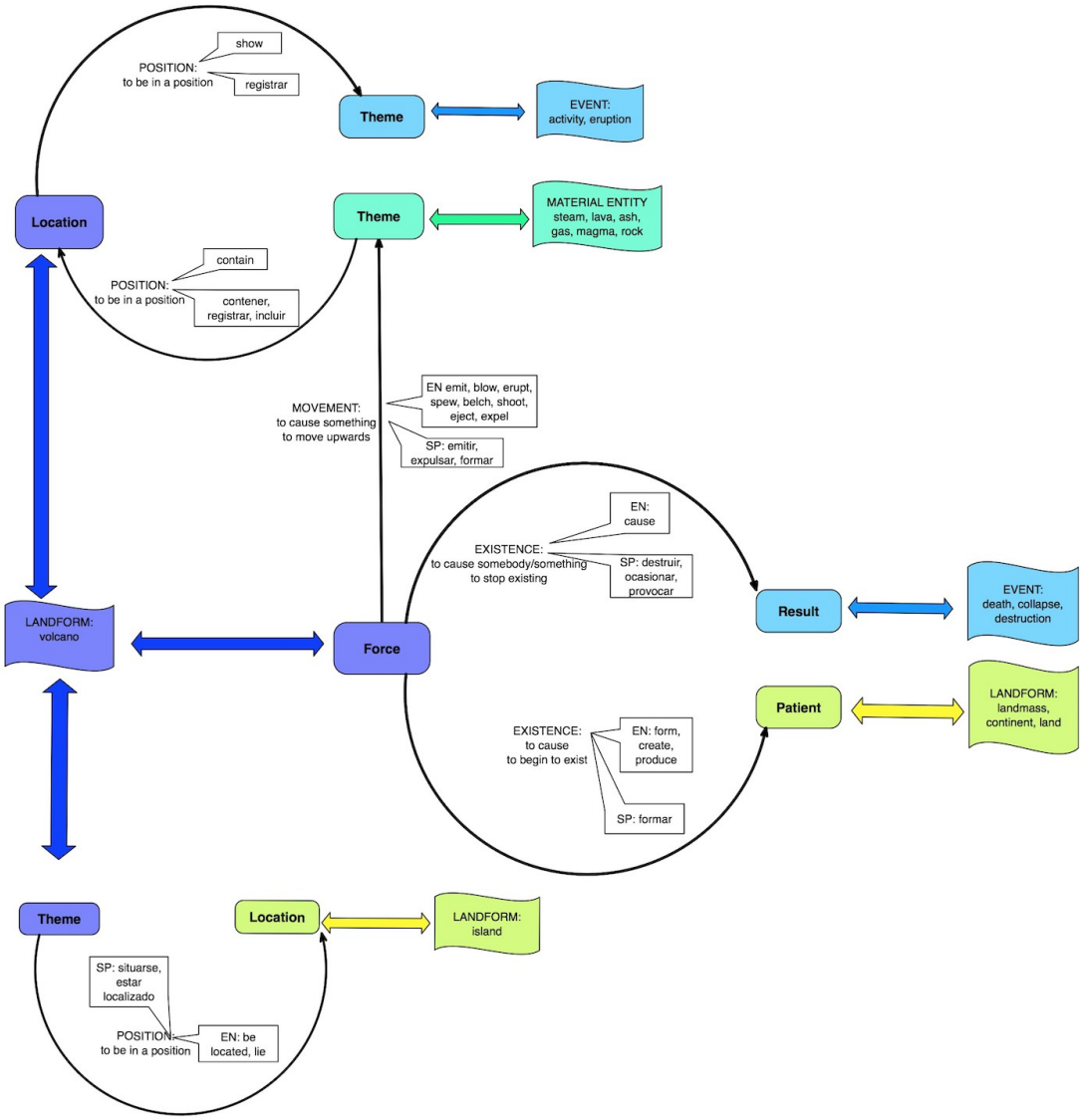

Figure 8. Semantic frame of "volcanic activity"

In the domain of EXISTENCE, volcanoes are natural forces that either participate in the creation of new landforms (*continents, islands*), or in the destruction of places (*cities*). They can produce casualties (*death, victim*) as well. Volcanoes can also trigger natural events such as continent shifts or tsunamis. In the lexical domain of MOVEMENT, a

material (*ash, lava, rock*) is expelled from the volcano. Lastly the lexical domain of POSITION is the source of structures that account for a geographical location.

In e-lexicography, the representation of concepts according to their cognitive behavior is not new. Nevertheless, an innovative aspect of our research is its frame elicitation protocol, which is more objective and less dependent on introspection. Secondly, we have annotated information extracted from the corpora, instead of working on the concordances themselves. From our point of view, this has a greater generalization potential, since we have analyzed all the extractions obtained instead of only a sample. Thirdly, our extraction, annotation, and analysis have been done in two corpora in parallel. This made it possible to obtain a representation of terms based on the particular features of each language.

This representation is informative since it shows the most prototypical lexical patterns in each language. Nevertheless, given the fact that this is a cognitive representation, it does not reflect exact word combinations. It is evidently necessary to know which verb-noun combinations are the most frequent. For example, although the verbs *form* and *cause* express the creation of an event, the *form* tends to be used with the nouns *continent* and *shift* whereas *cause* is used more frequently with *death, collapse* and *destruction*.

In conclusion, the results obtained indicated the existence of three kinds of cognitive structures in English and Spanish. Volcanoes are seen as destructive entities causing damage, as natural forces expelling geological materials or as natural forces creating new landforms. This means that the same entity can be conceptualized from different perspectives, and its argument structure varies accordingly.

Finally, this representation of specialized semantic frames can be used as a definitional template, which is a schematic representation of the most prototypical relations established by the concepts that are members of the same semantic frame (San Martín and León-Araúz 2013: 3). As such, it facilitates definition writing for terminographers since it has been proved that semantic frames are very useful for defining specialized concepts (Durán-Muñoz 2017).

6. Conclusions and future work

As shown in our study, a term can activate different frames in different lexical domains, depending on the semantic categories and thematic relations of its arguments. The advantages of multilingual semantic frames for terminological and translation purposes are numerous. In fact, such a representation is a proxy that allows the inference of the cognitive structures underlying scientific texts. Although results in the examples might sound obvious because everyone is more or less familiar with volcanoes, this kind of conceptual representation is less accessible when it is a question of more specialized concepts, such as *aquifer depletion* or *schistosity*. Becoming

familiar with the frame structures of these concepts is a good way for non-expert users (such as translators) to understand them.

On the other hand, describing terms in relation to the semantic frames associated with their concepts is a different approach to describing equivalence for verbs. Indeed, finding the equivalent verb associated with a term is a difficult task, since verbs are not generally described in terminological resources, at least not at this level of abstraction. General-purpose bilingual dictionaries cannot solve this problem either, since they do not provide this type of information or include specialized senses and uses. The multilingual frame structures proposed might help to solve this problem since they show verb equivalence is based on argument structure.

In order to be able to include semantic frames in terminological knowledge bases, it is first necessary to design the structure of these templates and secondly, to populate them with corpus-based information. The methodology designed in this paper serves this purpose. Since these templates are only intended for internal research use, future work will also focus on the design of a user friendly interface for their visualization. New techniques for the semi-automatic extraction of frame structures will be further explored. In this sense, we plan to annotate the corpus syntactically. Syntactic dependencies could mitigate some of the errors obtained in the triples, thus leading to more accurate results.

Acknowledgements

This research was carried out within the framework of project FFI2017-89127-P, Translation-oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. Partial funding was also received from project PARSEME-FR (ANR-14-CERA-0001), and the PARSEME Cost Action (IC1207).

References

Baroni, Marco, Adam Kilgarriff, Jan Pomikálek and Pavel Rychlý. 2006. WebBootCaT: instant domain-specific corpora to support human translators. In *Proceedings of EAMT. 11th Annual Conference of the European Association for* Machine Translation. Oslo, Norway: 247–252.

Barrière, Caroline. 2001. Investigating the causal relation in informative texts. Terminology. *International Journal of Theoretical and Applied Issues in Specialized Communication*, 7(2): 135-154.

Buendía-Castro, Miriam and Beatriz Sánchez-Cárdenas. 2016. Using Argument Structure to Disambiguate Verb Meaning. In *Proceedings of the XVII EURALEX international congress*, ed. by T. Margalitadze, G. Meladze,. Tbilisi: Ivane Javakhishvili Tbilisi University Press: 482-490

Cabré Castellví, María Teresa. 2003. "Theories of terminology: Their description, prescription and explanation". *Terminology*, 9(2): 163-199.

Condamines, Anne and Josette Rebeyrolle. 2001. Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB). *Recent advances in computational terminology*: 127-148.

Condamines, Anne. 2002. Corpus analysis and conceptual relation patterns. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 8.1: 141-162.

Coseriu, Eugenio. 1977. *Principios de semántica estructural*, Madrid, Gredos.

Church, Kenneth Ward and Patrick Hanks. 1990. "Word association norms, mutual information, and lexicography." *Computational linguistics*, 16(1): 22-29.

Dik, Simon. 1978. *Functional Grammar*, Dordrecht: Foris Publications.

Dubois, Jean and Françoise Dubois-Charlier. 1997 Synonymie syntaxique et classification des verbes français. *Langages*: 51-71.

Durán-Muñoz, Isabel. 2017. "Producing frame-based definitions". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 22(2): 223-249.

Faber, Pamela and León-Araúz, Pilar. 2014. "Specialized knowledge dynamics". In *Dynamics and Terminology: An interdisciplinary perspective on monolingual and multilingual culture-bound communication*, ed. by R. Temmerman and M.V. Campenhoudt, 16, 135, Amsterdam/Philadelphia: John Benhamins Publishing Company.

Faber, Pamela and M. C. África Vidal Claramonte, 2017. "Food terminology as a system of cultural communication". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 23(1): 155-179.

Faber, Pamela and Pilar León-Araúz 2016. "Specialized knowledge representation and the parameterization of context". *Frontiers in psychology*, 7: 196.

Faber, Pamela and Ricardo Mairal Usón. 2017. "The Functional Lexematic Model: past, present and future". In *Estudios de Filología Inglesa*, ed. by J.A.H.C Cutillas Espinosa, R. Manchón Ruiz and F. Mena Martínez, , Murcia: Editum, 315-340.

Faber, Pamela and Ricardo Mairal. 1999. *Constructing a Lexicon of English Verbs*, New York, Mouton de Gruyter.

Faber, Pamela, Juan Verdejo-Román, Pilar León-Araúz, Arianne Reimerink and Gloria Guzmán Pérez-Carrillo. 2017. Specialized knowledge processing in the brain: an fMRI study. In *Terminological Approaches in the European Context*, ed. by P. Faini Newcastle-upon-Tyne: Cambridge Scholars Publishing: 168-182.

Faber, Pamela, Pilar León Araúz and Jose Antonio Prieto Velasco. 2009. Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies*, 1(1): 1-23.

Faber, Pamela. 2012. *A cognitive linguistics view of terminology and specialized language*, 20. Berlin: Walter de Gruyter.

Faber, Pamela. 2015. "Frames as a framework for terminology". *Handbook of Terminology*, 1(14), ed. by Kockaert, H.J. and Steurs, F, 1:14-33. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Feliu, Judit. 2004. *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. PhD Thesis. Universitat Pompeu Fabra.

Fellbaum, Christiane J. 1990. English verbs as a semantic net, *International Journal of Lexicography*, 3/4: 278-301.

Fellbaum, Christiane. 1998. *WordNet: An electronic lexical database*. Blackwell Publishing Ltd.

Fillmore, Charles J. 2006. "Frame semantics". *Cognitive linguistics: Basic readings*, 34, 373-400.

Fillmore, Charles, Christopher Johnson and Miriam Petruck. 2003. "Background to FrameNet". *International Journal of Lexicography*, 16(3), 235-250.

Firth, John Ruppert. 1961. *Papers in Linguistics 1934-1951*. Oxford University Press.

Flaux, Nelly, Danièle Van de Velde. 2000. *Les noms en français : esquisse de classement*. Paris: Ophrys.

François, Jacques, Dennis Le Pesant and Danielle Leeman. 2007. "Présentation de la classification des Verbes français de Jean Dubois et Françoise Dubois-Charlier". Langue française, (1), 3-19.

Gaudin, François. 2003. *Socioterminologie. Une approche sociolinguistique de la terminologie*, ed. Duculot.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Granger, Sylviane and Fanny Meunier (eds). 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins Publishing.

Gross, Gaston. 1994. "Classes d'objets et description des verbes". *Langages*, 15-30.

Gross, Gaston. 2008. "Les classes d'objets". *Lalies*, (28), 111-165.

Hadouche, Fadila, Guy Lapalme, Marie-Claude L'Homme. 2011. "Attribution de rôles sémantiques à des actants", In *Actes de Traitement automatique des langues TALN*, Montpellier, France.

Halliday, Michael, Christian Mim Matthiessen and Christian Matthiessen. 2014. *An introduction to functional grammar*. London: Routledge.

Hanks, Peter. 2004. "Corpus pattern analysis". *Euralex Proceedings* (Vol. 1): 87-98.

Hanks, Peter. 2012. "How people use words to make meanings: Semantic types meet valencies". *Input, Process and Product: Developments in Teaching and Language Corpora*, 54-69.

Hatier, Sylviain, Magdalena Augustyn, Hoai Thi Thu Tran, Rui Yan, Anges Tutin and Marie-Paule Jacques. 2016. French Cross-disciplinary Scientific Lexicon: Extraction and Linguistic Analysis. In *Proceedings of the XVII EURALEX International Congress*. Tbilisi: Georgia, 355-366.

Huyghe, Richard. 2015. "Les typologies nominales: présentation". In Langue française, (1), 5-27.

L'Homme, Marie-Claude and Janine Pimentel. 2012. "Capturing syntactico-semantic regularities among terms: An application of the FrameNet methodology to terminology." In *LREC*: 262-268.

L'Homme, Marie-Claude, Robichaud Benoît and Carlos Subirats Rüggberg. Discovering frames in specialized domains. 2014. LREC: 1364-1371.

L'Homme, Marie-Claude, Subirats, Carlos, and Robichaud, Benoît. 2016. "A Proposal for combining general and specialized frames". In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon, CogALex-V*, 156-165.

L'Homme, Marie-Claude. 1998. Le statut du verbe en langue de spécialité et sa description lexicographique. Cahiers de lexicologie: 61-84.

L'Homme, Marie-Claude. 2012 Le verbe terminologique: un portrait de travaux récents. SHS Web of Conferences. Vol. 1. EDP Sciences: 93-107.

L'Homme, Marie-Claude. 2004. A Lexico-semantic Approach to the Structuring of Terminology. In Proceedings of Computerm: 7-14.L'Homme, Marie-Claude. 2012. "Adding syntactico-semantic information to specialized dictionaries: an application of the FrameNet methodology", *Lexicographica*, 28, 233-252.

Langacker, Ronald W. 1987. Foundations of cognitive grammar: Theoretical prerequisites (Vol. 1). Stanford university press.

Linardaki, Evita, Carlos Ramisch, Aline Villavicencio, Aggeliki Fotopoulou. 2010. "Towards the Construction of Language Resources for Greek Multiword Expressions: Extraction and Evaluation". In *Proceedings of the LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*. Valetta, Malta. ELRA: 31-40.

Mairal Usón, Ricardo and Pamela Faber. 2002. "Functional Grammar and lexical templates", *Functional Grammar Series*: 39-94.

Meyer, Ingrid, Kristen Mackintosh, Caroline Barrière and Tricia Morgan. 1999. Conceptual sampling for terminological corpus analysis. Proceedings of TKE'99, ed. by P. Sandrini: 256-267.

Meyer, Ingrid. 2001. Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In *Recent Advances in Computational Terminology*, ed. by D. Bourigault, C. Jacquemin, C. & M.C. L'Homme, (eds). Amsterdam/Philadelphia: John Benjamins, 279-302.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty and Daniel Zeman. 2016. "Universal Dependencies v1: A Multilingual Treebank Collection". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia. ELRA.

Petruck, Miriam R. 1996. Frame semantics. *Handbook of pragmatics*: 1-13.

Ramisch, Carlos. 2015. "Multiword Expressions Acquisition: A Generic and Open Framework". In *Theory and Applications of Natural Language Processing series, XIV*. Springer.

Ruppenhofer, Josef, Ellsworth, Michael, Petruck, Miriam R., Johnson, C. R., and Scheffczyk, Jan. 2016. *FrameNet II: Extended theory and practice*. Institut für Deutsche Sprache, Bibliothek.

Salomão, Maria Margarida Martins, Tiago Timponi Torrent and Thais Fernandes Sampaio. 2013. "A linguística cognitiva encontra a linguistica computacional:

notícias do projeto Framenet Brasil". Cadernos de Estudos Linguísticos 55 (1): 7-34.

San Martín, Antonio and Pilar León Araúz. 2013. "Flexible Terminological Definitions and Conceptual Frames." In *Proceedings of the International Workshop on Definitions in Ontologies* (DO2013), ed. by S. Seppälä and A. Ruttenberg, Montreal: Concordia University.

San Martín, Antonio. 2016. "La representación de la variación contextual mediante definiciones terminológicas flexibles". PhD Thesis. University of Granada.

Sánchez Cárdenas, Beatriz and Buendía Castro, Miriam. 2012. Inclusion of Verbal Syntagmatic Patterns in Specialized Dictionaries: The Case of EcoLexicon. In Proceedings of the 15th EURALEX International Congress. Ruth Vatvedt Fjeld and Julie Matilde Torjusen (eds): 554-562. Oslo: EURALEX.

Sánchez Cárdenas, Beatriz, and Pamela Faber. 2014. "A functional and constructional approach for specialized knowledge resources", ed. by N. Brian, C. Periñán Pascual, *Language Processing and Grammars: The Role of Functionally Oriented Computational Models*, Amsterdam: John Benjamins.

Sánchez Cárdenas, Beatriz. 2011. Structuration hiérarchique du lexique verbal à travers la propriété de troponymie. Revista de Lingüística y Lenguas Aplicadas, 6(1): 329-340.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford University Press.

Straka, Milan, Jan Hajič, Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis,

POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia. ELRA.

Temmerman, Rita. 1997. "Questioning the univocity ideal. The difference between socio-cognitive Terminology and traditional Terminology". *HERMES-Journal of Language and Communication in Business*, 10(18), 51-90.

Temmerman, Rita. 2000. Towards new ways of terminology description: The sociocognitive approach, 3, Amsterdam: John Benjamins Publishing.

Williams, Geoffrey. 2005. "English Collocation Studies: The OSTI report". *International Journal of Lexicography*, 18(3): 391-393.