

# Linguistic knowledge for specialized text production

Miriam Buendía-Castro, Beatriz Sánchez-Cárdenas

University of Granada

Department of Translation and Interpreting, Buensuceso 11, 18002 Granada (Spain)

E-mail: mbuendia@ugr.es, bsc@ugr.es

## Abstract

This paper outlines a proposal for encoding and describing verb phrase constructions in the knowledge base on the environment EcoLexicon, with the objective of helping translators in specialized text production. In order to be able to propose our own template, the characteristics and limitations of the most representative terminographic resources that include phraseological information were analyzed, along with the theoretical background that underlies the verb meaning argument structure in EcoLexicon. Our description provides evidence of the fact that this kind of entry structure can be easily encoded in other languages.

**Keywords:** phraseological information, verb meaning, translation

## 1. Introduction

In recent years, the methodology used in the documentation and phases of terminological extraction during the translation process have changed considerably because of the new ways of organizing and obtaining information from the Internet. If in the past, translators relied exclusively on lexicographic and terminographic resources, such as paper dictionaries, today online resources have become the principal source of documentation.

Translators work primarily with situation-based specialised texts and are under a permanent time constraint. Therefore, in order to be able to work quickly and efficiently, they need auxiliary materials that contain accurate up-to-date information, and which provide access to the network of conceptual relations between terms. Apart from source language and target language conceptual and translation equivalences, such auxiliary materials should also include explanations of meaning, typical contexts, and phraseological information.

Most specialized dictionaries only partially meet translator needs because they very often fail to specify how a term really behaves in specialised discourse (Kerremans, 2010). This is also true of a great number of terminological knowledge bases that are available online. Very frequently, these term bases are not consistent in the treatment of specialized knowledge units. The information that they provide is limited to a series of entries whose design is not systematic, and which do not include sufficient conceptual information. Furthermore, despite the fact that most terminographers agree that phraseological information in terminographic products are useful, few specialized resources actually contain word combinations (L'Homme, 2009: 260).

In this paper we focus on the description of the phraseological information currently being implemented in EcoLexicon (<http://ecolexicon.ugr.es>), a visual online thesaurus of environmental science. Phraseological information becomes especially necessary for the production of the target language text in the translation process. In this phase, the translator needs information regarding collocations as well as the semantic and syntactic structures in which terms participate. In this sense, the more collocations a translation dictionary contains, the better it can fulfil its function (Bergenholtz & Tarp, 2010: 33).

In order to evaluate EcoLexicon for translation, a questionnaire was designed (López-Rodríguez et al., forthcoming). It was completed online by 3rd year students of the Faculty of Translation and Interpreting of the University of Granada (Spain). The analyses of the results proved that phraseology along with contexts of use, were judged by students to be some of the most relevant information that they needed for text production in the translation process.

## 2. EcoLexicon

EcoLexicon (<http://ecolexicon.ugr.es>) is a visual online thesaurus of environmental science, which currently contains more than 3,000 concepts and over 14,500 terms in English, Spanish, German, Modern Greek, Russian, and French. It is based on a theoretical approach known as Frame-based Terminology (Faber, 2009, 2011, 2012). One of the things that makes EcoLexicon different from other online resources is that for each concept, a dynamic network is displayed that links the search concept to all related concepts in terms of a closed inventory of semantic relations (see Figure 1). Additionally, it also includes the following: (i) the definition of the concept; (ii) information concerning its attributes; (iii) graphical resources that illustrate the concept (Faber et al., 2007; Prieto-Velasco, 2009); (iv) contexts of use (Reimerink et

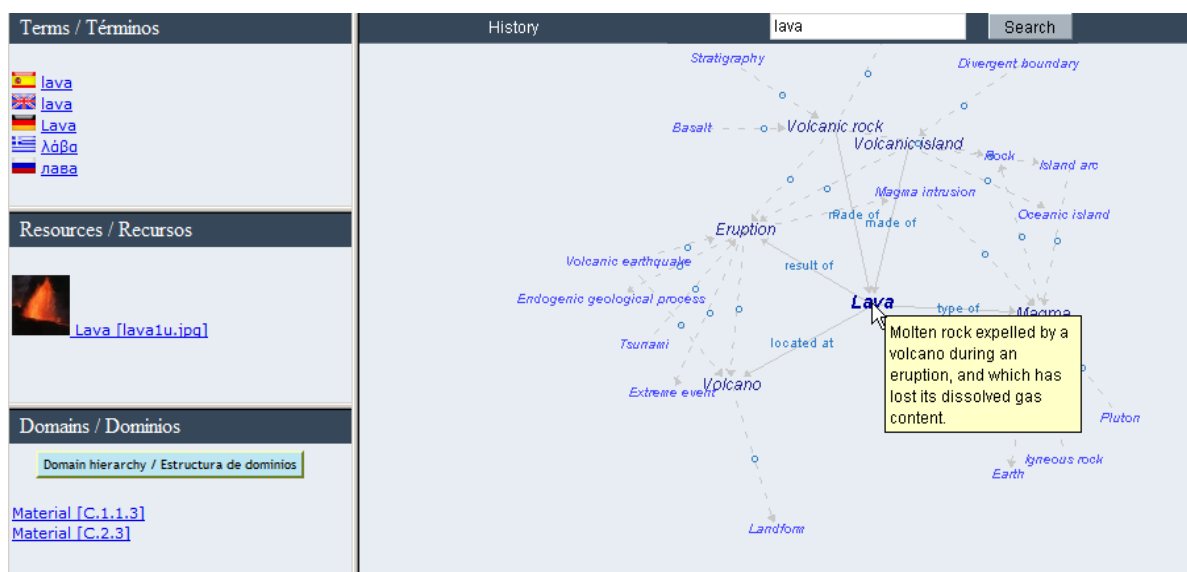


Figure 1: Interface of EcoLexicon. LAVA concept display

al., 2010); (v) the terminological units that designate the concept in different languages; (vi) phraseological information concerning each terminological unit (currently being implemented) (Montero & Buendía, 2010, forthcoming).

### 3. Terminographic resources containing phraseological information

As part of our study, we analyzed the characteristics and limitations of the most representative terminographic products that list collocations (DiCoInfo<sup>1</sup>, DiCoEnviro<sup>2</sup>, DAFA<sup>3</sup>, DICOFE (Verlinde et al., 1995-2003), the Accounting Dictionaries<sup>4</sup>, and Termium Plus<sup>5</sup>). From this preliminary analysis, we established the following requirements for an ideal terminological resource:

1. The resource should be available online and be based on the latest research trends in specialized communication and cognition.
2. The resource should be bilingual or multilingual with correspondences between the phraseological units in different languages. In the verbal lexicon, verb meaning is linked to argument structure, and there is rarely isomorphism between languages. Thus, a useful resource should establish equivalences between lexical domains rather than between individual verbs.
3. The resource should guarantee accessibility in order to allow search modes adapted to different user needs and situation profiles (Bergenholtz & Tarp, 2004, 2010).
4. Entries in the resource should be designed so that

they are easy for users to understand (without complicated metalanguage). Moreover, they should also not include too much information, i.e. an excessive number of collocations.

Based on these premises, we describe the design of an ideal entry for verbs (see Tables 1 and 2), using the verbs *cracher* (French) and *spew* (English) when they are activated by the entity VOLCANO in an AGENT role. Such an entry is focused on text production and, more specifically, on the context of the translation process.

### 4. Representation of verb meaning and argument structure in EcoLexicon

From a theoretical perspective, our representation of verb meaning and argument structure is based on the following convergent models: Frame-Based Terminology (Faber, 2009, 2011, 2012), Role and Reference Grammar (Van Valin & LaPolla, 1997; Van Valin, 2005) and the Lexical Grammar Model (Faber & Mairal, 1999; Ruiz de Mendoza & Mairal, 2008). These models were chosen because they are all lexically-based and able to specify combinatorial parameters, which facilitate text production in the target language. To this end, verb arguments are characterized from a semantic and actantial perspective. In order for the users to be able to select the type of information that best fits their needs, in EcoLexicon the entry takes into account its lexical domain, semantic frame, conceptual class, semantic roles, syntactic function, argument examples, and contexts.

According to the Lexical Grammar Model (Faber & Mairal, 1999; Ruiz de Mendoza & Mairal, 2008) verbs in any language can be classified into one of ten general domains: EXISTENCE, CHANGE, POSSESSION, SPEECH, EMOTION, ACTION, COGNITION, MOVEMENT, PHYSICAL PERCEPTION and MANIPULATION. Each of these lexical domains is characterized in terms of a general verb or genus that can

<sup>1</sup> <<http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/>>

<sup>2</sup> <<http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/>>

<sup>3</sup> <<http://www.projetdafa.net/>>

<sup>4</sup> <<http://www.asb.dk/en/research/researchcentresandteams/researchcentres/centreforlexicography/centreprofile/>>

<sup>5</sup> <<http://www.termiumplus.gc.ca/site/termium.php?lang=eng&cont=001>>

be compared to Wierzbicka's (1995) semantic primitives. The genus of a given verb is obtained by decomposing it into more general meanings following Dik's (1978) Stepwise Lexical Decomposition.

Once the lexical domain to which the verb belongs is specified, its semantic frame is provided. As is well known, a frame can be defined as a prototypical situation of an event, a relation, or an entity (Fillmore et al., 2003) that is characterized by the configuration of its semantic roles. For example, in the sentence *The hurricane hit Atlanta*, the verb *hit* belongs to the IMPACT frame described by two main frame elements: an IMPACTOR that makes sudden, forcible contact with the IMPACTEE. Since frames refer to general situations, they include the verbs that can be used to depict each specific type of context. For example, the verbs *crash*, *collide*, *impact*, *smash* and *strike* belong to the IMPACT frame. Thus, they share the same actantial configuration which is described by using core frame elements (IMPACTOR, IMPACTEE) and non-core elements (CAUSE, FORCE, MANNER, PLACE, RESULT, etc). Nevertheless, these tags, though very intuitive, have certain disadvantages. On the one hand, the list of frame elements is open-ended and unconstrained. In other words, each time a new frame is described, new frame elements are created. On the other hand, the tagging of frame elements relies primarily on the intuition of the analyzer. For this reason, we decided to combine semantic frames with the thematic relations in Role and Reference Grammar (Van Valin and LaPolla, 1997; Van Valin, 2005) to characterize the semantic behavior of arguments.

Conceptual classes are derived from the incipient ontology currently being implemented in EcoLexicon (León Araúz & Magaña, 2010). Top-level concepts can be divided into objects, events, attributes, and relations. Concepts can be concrete or abstract. In EcoLexicon, abstract concepts include theories, equations, and units for measuring physical entities, which are generally used by experts to describe, evaluate, and simulate reality. In contrast, concrete concepts are those occupying space and involving time. They can be applied to natural entities, geographic accidents, water bodies, constructions, and the natural and artificial process events in which they can take part. For example, a VOLCANO is conceptualized both as a GEOLOGICAL PHENOMENON (which includes entities such as OCEANIC PLATE, EARTHQUAKE, ERUPTION or CLIFF), and as a NATURAL DISASTER, composed of concepts such as EARTHQUAKE, HURRICANE, TORNADO, LANDSLIDE or FLOOD.

Regarding semantic roles, both thematic roles and macroroles as described in Role and Reference Grammar (Van Valin & LaPolla, 1997; Van Valin, 2005) are taken into account. Thematic relations (AGENT, THEME, GOAL) describe the semantic behavior of verb arguments. They are generalizations across frame element roles (giver, thinker, runner, etc.) described by FrameNet (Fillmore et al., 2003) and are useful when describing the verb lexicon because they allow the user to understand and predict relations between verbs and nominal forms. According to Van Valin (2005), among others, the lexical representation of a verb can be derived from its logical structure. To this end, verbs are classified into *Aktionsart* categories (state, activity, achievement, semelfactive, accomplishment and active accomplishment), based on inclusion tests (Sánchez Cárdenas, 2010). Since each of these classes has a specific configuration, verb argument structure can be derived from the aspectual nature of the verb. Each argument is then linked with a thematic relation based on the premises of Role and Reference Grammar (Van Valin & LaPolla, 1997; Van Valin, 2005). Finally, argument structure is represented in terms of the macroroles, ACTOR and UNDERGOER. As generalizations of thematic relations, macroroles are part of the semantic and syntactic interface, and help to distinguish between core arguments from non-core arguments. Macroroles are then linked to syntactic functions.

Thanks to RRG thematic relations and macroroles, syntactic structures can be separated from semantic ones. This facilitates the description of thematic relations that appear in more than one syntactic position. For example in the sentences *The volcano ejects ashes* and *Ashes are ejected from the volcano*, the NP *ashes* appears in the syntactic positions of Direct Object and Subject but in both sentences, it has the thematic relation of THEME and the macrorole of UNDERGOER.

Argument examples illustrate how noun phrases are encoded in each language. Even if the specialized texts express the same ideas (e.g. a volcanic eruption), each language can designate this process entity in different ways. For example, in French, the THEME of the verb *cracher* is usually *un panache de cendres*, whereas in English, the THEME of the equivalent verb, *spew*, is frequently *a cloud of ash*.

<b>CRACHER (FRENCH)</b>		Lexical domain: <b>MOVEMENT</b> Frame: <b>Cause Motion</b>			
<b>conceptual class</b>	<b>macrorole</b>	<b>thematic relation</b>	<b>syntactic function</b>	<b>argument example</b>	<b>context</b>
GEOLOGICAL PHENOMENON	ACTOR	Agent + Source	S	- le volcan	<b>Le volcan</b> a craché des cendres.
MATERIAL: RESULT OF A NATURAL PROCESS	UNDERGOER	Theme	COD	- de la lave - panache de cendres - des fumerolles - des gaz	Le volcan crache <b>des cendres</b> .

Table 1. Template for the verb *cracher* in the terminological entry for *volcan* in EcoLexicon

<i>SPEW</i> (ENGLISH)		Lexical domain: <b>MOVEMENT</b> Frame: <b>Cause Motion</b>			
conceptual class	macrorole	thematic relation	syntactic function	argument example	context
GEOLOGICAL PHENOMENON	ACTOR	Agent + Source	S	- the volcano	<b>The volcano</b> in Iceland continues to spew a huge ash cloud.
MATERIAL: RESULT OF A NATURAL PROCESS	UNDERGOER	Theme	COD	- cloud of ash - lava - plumes of ash and steam	The volcano erupted under Eyjafjöll glacier and spew <b>ashes</b> into the air.

Table 2. Template for the verb spew in the terminological entry for volcano in EcoLexicon

Finally, contexts refer to frequency-related information extracted from concordances which can help users to produce text in the same way as a native speaker.

These templates will be available in EcoLexicon through the terminological unit *volcan* (French) (Table 1), and its correspondence in English, the verb *spew*, available through the term *volcano* (Table 2). As shown, both verbs belong to the lexical domain of MOVEMENT and activate the frame of CAUSE MOTION.

## 5. Conclusion

In this study, we present a proposal for encoding and describing phraseological information in EcoLexicon, a knowledge base on the environment. This proposal is designed with the objective of helping translators in specialized text production. Since translators are constantly in contact with at least two languages, they inevitably run the risk of directly mapping source language structures onto the target language text. Since the terminological entry proposed describes the combinatorial patterns of terms and verbs, it allows users to predict the linguistic behavior of a term in a given target language.

The purpose of this analysis was to study the different domains and verbs most frequently activated in each domain and frame, with a view to creating rules that can be computationally systemized to enhance and automate verb argument extraction. In other words, the fact that the verb *cracher* can be used in the lexical domain of MOVEMENT, within the frame of cause motion, when the AGENT is a natural phenomenon and the THEME is a material, implies that any verb which fulfils this rule could be used in the same context. This is the case of *vomir* or *spit*, which could also be used in this same context previously analyzed: *Le volcan a vomi des cendres, An Indonesian volcano spit lava and smoke.*

This kind of entry structure has the advantage of being easily encoded in other languages. Since in specialized languages, verbs have restricted uses and contexts, the

kinds of noun phrases that are likely to appear in each position are highly predictable from corpus analysis.

## 6. Acknowledgements

This research has been carried out within the framework of the project RECORD: Representación del Conocimiento en Redes Dinámicas [Knowledge Representation in Dynamic Networks, FFI2011-22397], funded by the Spanish Ministry for Science and Innovation.

## 7. References

- Bergenholtz, H. & Tarp, S. (2004). The concept of dictionary usage. *Nordic Journal of English Studies*, 3(1), pp. 23-36.
- Bergenholtz, H. & Tarp, S. (2010). LSP Lexicography or Terminography? The Lexicographer's Point of View. In P.A. Fuertes-Olivera (Ed.), *Specialized Dictionaries for Learners*. Berlin/New York: Mouton de Gruyter, pp. 27-37.
- Dik, S. (1978). *Functional Grammar*. Dordrecht: Foris Publications.
- Faber, P. (2009). The cognitive shift in Terminology theory. In A. Vidal & A.J. Franco (Eds.), *Una visión autocrítica de los estudios de traducción*. MonTI, 1(1), pp. 107-134.
- Faber, P. (2011). The dynamics of specialized knowledge representation: Simulational reconstruction or the perception-action interface. *Terminology*, 17(1), pp. 9-29.
- Faber, P. (Ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/New York: Mouton de Gruyter.
- Faber P. & Mairal R. (1999). *Constructing a Lexicon of English Verbs*. Berlin/New York: Mouton de Gruyter.
- Faber, P., León Araúz, P., Prieto Velasco, J.A. & Reimerink, A. (2007). Linking images and words: the description of specialized concepts. *International Journal of Lexicography*, 20, pp. 39-65.
- Fillmore, C.J., Johnson C.R. & Petruck M.R.L. (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3), pp. 235-250.

- Kerremans, K. (2010). A comparative study of terminological variation in specialised translation. In C. Heine & J. Engberg (Eds.), *Reconceptualizing LSP: Online proceedings of the XVII European LSP Symposium 2009*. Aarhus: Aarhus School of Business, Aarhus University.
- L'Homme, M.C. & Leroyer, P. (2009). Combining the semantics of collocations with situation-driven search paths in specialized dictionaries. *Terminology*, 15(2), pp. 258-283.
- León Araúz, P. & Magaña, P.J. (2010). EcoLexicon: contextualizing an environmental ontology. In *Proceedings of the Terminology and Knowledge Engineering (TKE) Conference 2010*. Dublin: Dublin City University.  
<<http://lexicon.ugr.es/pub/leonmagana2010>> accessed 15 October 2011.
- López-Rodríguez, C.I., Buendía-Castro, M. & García-Aragón, A. (forthcoming). User needs to the test: evaluating a Terminological Knowledge Base on the environment by trainee translators. *Jostrans*.
- Montero Martínez, S. & Buendía Castro, M. (forthcoming). La sistematización en el tratamiento de las construcciones fraseológicas: el caso del Medio Ambiente. In *Actas del XXIX Congreso Internacional de AESLA: Empirismo y herramientas analíticas para la Lingüística Aplicada del siglo XXI*. Salamanca.
- Montero Martínez, S. & Buendía Castro, M. (2010). Las construcciones fraseológicas en la Terminología basada en Marcos. Conference presented at the *Seminario de Lingüística española. Los retos informáticos del español en lexicología, terminología y fraseología*, organized by the Fundación Duques de Soria and the University of Antwerpen, Antwerpen, 26-27 February 2010.
- Prieto Velasco, J.A. (2009). *Traducción e imagen: la información visual en textos especializados*. Granada: Tragacanto.
- Reimerink, A., García de Quesada, M. & Montero-Martínez, S. (2010). Contextual information in terminological knowledge bases: A multimodal approach. *Journal of Pragmatics*, 42(7), pp. 1928–1950.
- Ruiz de Mendoza Ibáñez, F.J. & Mairal Usón, R. (2008). Levels of description and constraining factors in meaning construction: an introduction to the Lexical Constructional Model. *Folia Linguistica*, 42(2), pp. 355–400.
- Sánchez Cárdenas, B. (2010). Paramètres linguistiques pour la constitution d'un dictionnaire électronique bilingue (français/espagnol) destiné à la traduction. Le cas des verbes de comptage. Doctoral thesis, University of Granada.
- Van Valin, R.D.Jr. (2005). *Exploring the syntax-semantics interface*. Cambridge: Cambridge University Press.
- Van Valin, R.D.Jr. & LaPolla, R. (1997). *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.
- Verlinde, S., Folon, J., Binon, J. & Van Dyck, J. (1995-2003). Dictionnaire contextuel du français économique, 4 volumes (L'entreprise + Exercisier, Le commerce + Exercisier, Les finances, L'emploi). Antwerp: Garant.
- Wierzbicka, A. (1995). Universal semantic primitives as a basis for lexical semantics. *Folia Lingüística*, 29(1-2), pp. 149-169.