## Semantic Relations Established by Specialized Processes Expressed by Nouns and Verbs: Identification in a Corpus by means of Syntactico-semantic Annotation

## Nava Maroto García; Marie-Claude L'Homme; Amparo Alcina

CES Felipe II, Universidad Complutense de Madrid; Observatoire de linguistique Sens-Texte (OLST), Université de Montréal; TecnoLeTTra Team, Universitat Jaume I

Traducción e Interpretación, Casa del Gobernador, C/ Capitán s/n, E-28300, Aranjuez, Spain

E-mail: marianava.maroto@ajz.ucm.es, mc.lhomme@umontreal.ca, alcina@trad.uji.es

#### Abstract

This article presents the methodology and results of the analysis of terms referring to processes expressed by verbs or nouns in a corpus of specialized texts dealing with ceramics. Both noun and verb terms are explored in context in order to identify and represent the semantic roles held by their participants (arguments and circumstants), and therefore explore some of the relations established by these terms. We present a methodology for the identification of related terms that take part in the development of specialized processes and the annotation of the semantic roles expressed in these contexts. The analysis has allowed us to identify participants in the process, some of which were already present in our previous work, but also some new ones. This method is useful in the distinction of different meanings of the same verb. Contexts in which processes are expressed by verbs have proved to be very informative, even if they are less frequent in the corpus. This work is viewed as a first step in the implementation – in ontologies – of conceptual relations in which activities are involved.

Keywords: argumental and circumstantial relations, activity concepts, semantic role annotation

#### 1. Introduction

This contribution presents a methodology for the analysis of terms referring to specialized processes in industrial ceramics (namely, the process of laying tiles on a surface). Processes are key concepts in many specialized fields and ceramics is no exception. From a conceptual point of view, they belong to the class "activity", obtained by the abstraction of processes, operations or events performed by or with entities. A process consists of a designed set of operations carried out in a particular order, using particular tools and that produce some change in the properties of entities as an outcome (Nuopponen, 2007). Processes can be expressed in discourse by means of both nouns (e.g. *tile-laying*) and verbs (e.g. *to lay*) and these establish different kinds of relations that can be identified in corpora.

This work aims to identify these relationships using a method based on the annotation of terms that denote processes and their participants (arguments and circumstants)<sup>1</sup>. Our annotation – based on the method developed within the FrameNet project (Ruppenhofer et al, 2010) – specifies the semantic and syntactic properties of participants (their semantic role, syntactic function and syntactic group).

The article is organized as follows. First (Section 2), we outline the objectives of this research. Then we offer a brief overview of previous work dealing with annotation of verbs both in general language and terminology (Section 3). We focus on work that is related to the specific analysis we set ourselves to carry out. Thirdly, the methodology followed in this research is described (Section 4). In section 5 we offer the results obtained

after the annotation of contexts for verbs and nouns, together with some examples of annotated contexts. Finally, conclusions and further research envisaged are outlined.

## 2. Objectives

In Maroto and Alcina (2009a) we explored the relations established by the process of tile firing expressed in Spanish by the noun *cocción*. We focused on a set of argumental and circumstantial relations in which one the concepts is the process of firing and the other could be one of the following: agent *(kiln operator)*, patient *(ceramic paste)*, product *(biscuit)*, state *(vitreous state)*, instrument *(kiln)*, method *(vitrification)* or effect *(contraction)*.

The analysis of nouns proved to be problematic when studying processes, because sometimes it is difficult to differentiate between separate meanings when the same noun can refer to both the action and its final result. For example, *alicatado* (the noun for wall tiling) can be used to refer both to the process of installing tiles on a wall and to the result (a group of tiles installed on a wall). Contextual information can help disambiguate the meanings in some but not all cases.

That is why in the present study we include both nouns and verbs that are likely to express processes, assuming that verbs will help us overcome some of these difficulties.

Our analysis relies on a fine-grained syntactico-semantic annotation that allows us to discover and represent the properties shared by predicate units and their participants. We hypothesize that contexts in which processes are expressed by verbs are likely to contain a formal expression of argumental and circumstantial relationships, and therefore could complement the analysis of contexts where processes are expressed by nouns.

<sup>&</sup>lt;sup>1</sup> Arguments are fundamental for the definition of the term (for example, the agent or the patient), whereas circumstants express the circumstances in which the process takes place, but

In example [1] we find a context for the verb *colocar* (to lay). Through syntactico-semantic annotation we can identify the patient of the process (in this case *azulejos*, wall tiles), the means used to lay the wall tiles (elastic organic adhesives, *adhesivos orgánicos elásticos*), the technique used (thin-layer, *en capa* fina) and the mode in which the process takes place (using an open joint system, *utilizando el sistema de juntas abiertas*)<sup>2</sup>.

[1a] Colocar los azulejos mediante adhesivos orgánicos elásticos y en capa fina, utilizando el sistema de juntas abiertas.

[1b] Lay the wall tiles by means of elastic organic adhesives and using the thin-layer technique, applying the open-joint system.

The patient, the means, the technique and the mode are all related to the process expressed by the verb *to lay* through an argumental or circumstantial relation.

Both nouns and verbs are explored in context in order to identify and represent the semantic roles held by their participants (arguments and circumstants).

### 3. Previous work

In general language, several projects consider the annotation of verbs in order to describe their connections to their syntactic dependents, such as WordNet (Fellbaum, 1998), FrameNet (Ruppenhofer et al, 2010) or Verbnet (Kipper et al, 2004).

Due to their predicative nature, verbs are often described taking into account their arguments. Arguments are listed and can be labeled simply with variables (X, Y and Z, as in Mel'čuk et al, 1995). Arguments can also be further specified with general semantic labels (such as "something" or "someone"), as in WordNet (Fellbaum, 1998), or more specific ones (Agent, Patient, Instrument) as in VerbNet (Kipper et al, 2004) or with Frame Elements as in FrameNet (Ruppenhoffer et al, 2010).

In some of the resources mentioned in the paragraph above, annotation of sentences or contexts is viewed as a convenient method to discover and display information about predicate-argument relationships.

In contrast, terminology has traditionally focused on nouns, often neglecting the relevance of verbs. However, verbs play a key role in the development of processes, and should therefore be considered as a central part of the analysis. This is increasingly recognized in terminology and more and more projects include verbs. Some examples are L'Homme (2002), Nuopponen (2007), Lépinette and Olmo Cazevieille (2007) and De Vecchi and Estachy (2008).

Some of these researchers describe the argument structure (L'Homme, 1998) with the aim to take into account their syntactic and semantic properties. Others, such as Marshman et al (2009), use verbs to discover and distinguish various types of cause-effect relationships.

De Vecchi and Estachy (2008) adopt a so-called

*pragmaterminological approach*, according to which actions – and therefore the verbs that express them – are assigned a preferential role in the conceptual structuring of a domain.

We believe that if specialized processes are to be described for terminological purposes, verbs need to be considered in order to identify the participants and the relations they establish with the process.

## 4. Methodology

Our methodology was organized in the following steps. First, we selected a set of ten processes related to tile-laying that can be expressed by both nouns and verbs. Then, we selected and annotated contexts in which those nouns and verbs expressed the processes. Finally, we observed the participants that appeared in the contexts and compared them with our previous research.

Our analysis is based on the work carried out in the DicoInfo project (L'Homme, 2008). DicoInfo is a specialized online dictionary that provides detailed information on the lexico-semantic properties of terms. Pimentel and L'Homme (2011) describe a methodology to annotate contexts with labels that specify the arguments and circumstants related to a predicate, using semantic roles. This method is in turn based on the methodology devised within the FrameNet project (Ruppenhofer et al, 2010), where linguistic realizations of lexical units are annotated together with their participants (both arguments and circumstants, called *core frame elements* and *non-core frame elements* in Frame Semantics).

We focus on the process of tile laying in the ceramic industry, which includes aspects such as the selection of the right tile for each surface, the correct choice and application of a laying method and bonding materials, as well as the use of different instruments. Several entities (instruments, bonding materials, tiles) intervene in the process and should be taken into account in terminological resources.

In the next two subsections we describe the term selection process (4.1) and the selection of contexts and annotation method (4.2).

# 4.1 Term selection

First, we looked at the terms belonging to the thematic subarea "Procesos de colocación" (Tile-laying processes) contained in the termbase *Ceramica*, described in Civera and Alcina (2000) and Civera (2002). This database consists of 4,617 terminological records organized in seven broad thematic areas (i.e. products, manufacturing, raw materials and additives, laying processes, quality, units of measure and, finally, organizations and institutions). It contains mainly nouns such as *colocación* (tile laying), which derive from verbs, but few verbs per se are included.

We also carried out the automatic extraction of verbs in the corpus TxtCeram using the extractor TermoStat (Drouin, 2003). The TxtCeram corpus is a monolingual specialized corpus in Spanish about ceramic industry

 $<sup>^2</sup>$  In section 4.2 we show how the syntactico-semantic annotation of this context has been carried out using the Oxygen XML editor.

containing 34 specialized texts (2,444,791 words). TermoStat is a term extraction tool that compares specialized and non-specialized corpora in order to extract candidate terms. It was particularly interesting for this project since the tool can be parameterized to identify specific term patterns of parts of speech.

We used the extractor to locate candidate verbs from the corpus and compared the obtained list with the noun terms in the *Ceramica* database. In this way, we identified the verbs present in the corpus corresponding to the nouns that are recorded in the terminological database. For example, in *Ceramica* we found the noun *ajuste* (fitting), but the verb *ajustar* (to fit) had no terminological record although it appeared in the list of candidate terms produced by TermoStat.

Both the exploration of the terminological database *Ceramica* and the corpus yielded a number of verbs and their derived nouns. We selected the ten most frequent verbs and nouns related to the tile-laying process (which we considered to be the most relevant processes in the field).

Table 1 shows the list of ten verbs and nouns selected for the analysis, together with the frequency of occurrence of each noun and verb in the corpus<sup>3</sup>. We offer an approximate translation of the terms into English.

Verbs	Freq.	Noun	Freq.
colocar	476	colocación	1659
(to lay)		(tile-laying)	
ajustar	148	ajuste	208
(to fit)		(fitting)	
solar	102	solado	503
(to floor)		(flooring)	
revestir	97	revestimiento	2788
(to cover a wall)		(wall-covering)	
anclar	60	anclaje	85
(to anchor)		(anchor)	
pavimentar	54	pavimentación	102
(to pave)		(paving)	
rejuntar	20	rejuntado	521
(to grout)		(grouting)	
asentar	20	asentamiento	49
(to settle)		(settling)	
alicatar	18	alicatado	458
(to tile a wall)		(wall-tiling)	
encolar	7	encolado	73
(to glue)		(gluing)	

Table 1: Terms selected for the analysis.

### 4.2 Context selection and annotation

After selecting the verbs, we observed the contexts in which each verb and its corresponding noun appeared and carefully selected contexts for each of them. Contexts were selected on the basis of their "richness", i.e. ideally, they had to contain many arguments and circumstants and thus yield valuable information on the relationships observed between the predicative unit and its participants. For those verbs that appeared frequently in the corpus we chose a minimum of seven and a maximum of 20 contexts in which at least two participants were present.

Once the contexts were selected for annotation, the XML-Editor oXygen was used in order to create terminological records. A detailed explanation of the structure of the terminological records and of the methodology to develop each record exceeds the scope of this paper, but a comprehensive description can be found in DesGroseillers (2011). Context annotation with oXygen allows us to indicate the following in the sentence *Colocar los azulejos mediante adhesivos orgánicos elásticos y en capa fina, utilizando el sistema de juntas abiertas* (Lay the wall tiles by means of elastic organic adhesives and using the thin-layer technique, applying the open-joint system.):

- The annotated lexical unit: e.g. colocar (to lay)
- The type of participant (arguments or circumstants<sup>4</sup>): e. g. *azulejo* (wall tile)
- The semantic role of the participant: e.g. *paciente* (patient)<sup>5</sup>
- The syntactic function of the participant: e.g. *objeto* (object)
- The syntactic group of the participant: e.g. SN (noun phrase)

The annotation process allowed us to identify and explicitly show the relations established between the annotated terms and the participants in the process.

Fig. 1 shows an example of the annotation of a context of the verb *colocar* (to lay). As can be seen, the annotated context includes four participants, namely:

- PATIENT of the action of laying: azulejos (wall tiles),

MEANS used as bonding material during the operation: adhesivos orgánicos elásticos (elastic organic adhesives)
TECHNIQUE employed in the laying process: en capa fina (thin layer method)

- MODE in which the process takes place: *utilizando el sistema de juntas abiertas* (using the open joint system)

<sup>&</sup>lt;sup>3</sup> The frequency of the verbs has been calculated both looking at the results from the TermoStat extractor (which lemmatizes verbs) and complementing the analysis with the results of text analysis tools.

<sup>&</sup>lt;sup>4</sup> Throughout the article we have used the term "argument" to refer to those elements which are fundamental for the definition of the term. However, in the annotation of contexts we have used the label "Act" (abbreviation for "actant") because this is how it has been implemented in previous projects in oXygen, such as the DicoInfo. "Actant" is another term to refer to these fundamental elements.

<sup>&</sup>lt;sup>5</sup> The labels used for semantic roles differ from the labels used to indicate Frame Elements and are reminiscent of the roles mentioned in Fillmore (1968).

## 5. Results

The annotation of 385 contexts allowed for the identification of the following participants (an example is offered to illustrate each one):

- AGENT: colocador (tiler)
- PATIENT: baldosa (tile)
- INSTRUMENT: *llana* (trowel)
- METHOD: *mediante la utilización de morteros* (through the use of mortar)
- TECHNIQUE: en capa fina (thin-layer method)
- DESTINATION: suelo (floor)
- MEANS: material de agarre (bonding material)
- MANNER: anclaje mecánico (mechanical anchor)
- MODE: *prescindiendo de las juntas de colocación* (without laying joints)



Fig. 1: Annotated context of the verb colocar (to lay).

The first four participants were already present in Maroto and Alcina (2009a), but the syntactico-semantic annotation allowed us to discover new ones. For example, a new participant stems from the distinction between MEANS, introduced to refer to the bonding material used in the laying process (for example, mortar) and INSTRUMENT, used for the entities employed for carrying out the process itself (for example, trowel).

The role TECHNIQUE was also added in order to

annotate the way in which ceramic products are laid, often expressed through an adverbial phrase such as *en capa fina* (thin-layer). This new role makes clearer the distinction from MODE, METHOD and MANNER and it can be said to be domain specific.

The study has also shown that in the particular case of tile-laying processes, there are relations that were present in our previous study (Maroto and Alcina, 2009a) but are hardly or not at all represented for this kind of processes. This is the case of the cause-effect and the process-result relations.

The cause-effect relation has been identified in the corpus, but it could not be annotated directly using semantic roles because there is no direct syntactic relation, as can be seen in the following example extracted from the corpus:

[2a] Así, tenemos que recordar las causas más importantes provocadoras de lesiones generalizadas: Asentamiento de las estructuras (CAUSE), con los subsiguientes fenómenos de deformaciones diferenciales (EFFECT) sobre los elementos constructivos interrelacionados con ellas.

[2b] In this way, let us remind the most important causes that provoke generalized damage: **structure settling** (CAUSE), with the subsequent phenomena of **differential distortions** (EFFECT) in the constructive elements related to them.

We have not found examples of the process-product relation either, but this is due to the fact that the laying process is not a manufacturing process, and therefore, no new objects are produced as an outcome, as was the case with the firing process.

Nevertheless, we decided not to eliminate these relations from the classification, as they have proved useful in previous analysis of processes of different nature (such as the firing process).

The annotation process has also permitted in some cases the distinction between different meanings of the same verb, as can be seen in the two following examples. In example [3] the patient of the process of *alicatar* (to tile) may be the tiles, whereas in [4] the patient seems to be the surface on which tiles are placed. In this case, tiles have a different semantic role, namely, they could be considered as the means, and the surface would be the patient. We therefore have two different senses, 'to place tiles' and 'to cover a surface using tiles':

[3a] *Cuando se alicatan* (ANNOTATED VERB) *azulejos* (PATIENT) *de cierta capacidad de absorción de agua* [...], las baldosas deben remojarse antes de su colocación.

[3b] When **wall tiles** (PATIENT) of a certain water absorption capacity **are tiled** (ANNOTATED VERB) [...], these tiles should be soaked before being laid.

[4a] Aplique las placas de yeso sobre el rodapié ya existente, para preparar la superficie (PATIENT) que se alicatará (ANNOTATED VERB) posteriormente con las nuevas baldosas (MEANS).

[4b] Apply the plaster plaques on the existing baseboard in order to prepare the **surface** (PATIENT) that **will later be tiled** (ANNOTATED VERB) **with the new tiles** (MEANS).

An outstanding fact we noticed is that, although verbs are less frequent in the corpus, their contexts are often very informative because it is easier to identify the participants without ambiguity. For example, the AGENT is identified as the subject of the verb, as in the following example, where the agent *colocador* (tiler) is the subject (indirectly expressed) of the verb *colocar* (to lay):

[5a] Para cubrir tres metros cuadrados de adhesivo, aplicado con la **llana** (INSTRUMENT), con estas baldosas en 15 minutos, un **colocador** (AGENT) tendría que **colocar** (ANNOTATED VERB) una **baldosa** (PATIENT) cada seis o siete segundos.

[5b] In order to cover three square meters of adhesive –applied using a **trowel** (INSTRUMENT)– with these tiles in 15 minutes, a **tiler** (AGENT) should **lay** (ANNOTATED VERB) one **floor tile** (PATIENT) every six or seven seconds.

In fact, the role AGENT was only found in contexts where the process was expressed by a verb.

## 6. Conclusion

The annotation of semantic roles in context allowed us to identify terms related to processes, whether expressed by a verb or a noun. We have discovered new semantic roles and thus refined our previous classification of relations. Although most semantic roles could be said to be common to most processes, some field-specific roles have also been discovered.

Semantic role annotation has proved a useful way to distinguish between different meanings of certain terms. The annotation of verbs, however, seems to be more informative and allows us to overcome the problem of ambiguity posed by nouns that may refer to the final result or to the actual process of tile-laying.

The methodology presented in this contribution could be reused in other specialized fields where processes are relevant in order to identify terms related to processes.

One of our future perspectives is the development of an ontology of industrial ceramics in which concepts (represented as classes) are formally and explicitly related to other concepts through attributes and conceptual relations (Maroto and Alcina, 2009b). Some the participants identified during of the syntactico-semantic annotation can be later implemented as classes in our ontology. The conceptual relations that link the concepts expressed in the context can also be explicitly and formally represented in the ontology and used in automatic tasks such as information retrieval. In this way, for example, a user may be able to retrieve a concept by asking the system about a possible relation (Maroto and Alcina, 2009b).

A further practical application of the annotation

methodology is the extraction of collocations that could be included in a dictionary based on an ontology.

## 7. Acknowledgements

The present research has been possible thanks to the grant programme "José Castillejo" for young post-doctoral researchers funded by the Spanish Ministry of Education.

This research is inscribed in the project ONTODIC-II. Methodologies and techniques for the elaboration of dictionaries of collocations based on ontologies. Terminographic resources for e-Translation, funded by the Spanish Ministry of Science and Innovation (ref. TIN2009-07690).

## 8. References

- Civera, P. (2002). Traducción científico-técnica y terminología en el sector de la industria cerámica. In Alcina, A. and S. Gamero (2002). La traducción científico-técnica en la sociedad de la información. Castellón: Publicaciones de la Universitat Jaume I, pp. 167--176.
- Civera, P., Alcina, A. (2000). Elaboración de una herramienta de comunicación plurilingüe para el sector cerámico: un diccionario terminológico multilingüe en soporte informático. In XL Congreso de la Sociedad Española de Cerámica y Vidrio. Onda (Castellón).
- DesGroseilliers, S. (2011). Oxygen. Guide de l'utilisateur. Une guide pour la création de fiches terminologiques dans le DicoInfo et DicoEnviro. Observatoire linguistique Sens-Texte (OLST), Université de Montréal. Internal document.
- De Vecchi, D., Estachy, L. (2008). Pragmaterminologie: les verbes et les actions dans les métiers. In *II TOTh Conference (Terminologie et Ontologie: théories et applications). Annecy, France, June 2008*: pp. 35--52.
- Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1), pp. 99--117.
- Fellbaum, C. (ed.). (1998). *WordNet: An Electronic Lexical Database. Language, Speech and Communication.* Massachussets: MIT.
- Fillmore, C. J. (1968). The case for case. In Bach, E., Harms, R.T. (eds.). *Universals in linguistic Theory*. New York: Holt, Rinehard & Winston, pp. 1--88.
- Kipper, K., Snyder, B., Palmer M. (2004). Extending a verb-lexicon using a semantically annotated corpus. In Fourth International Conference on Language Resources and Evaluation (LREC 2004). Lisbon, Portugal, May, 2004.
- Léppinette, B., Olmo Cazevieille, F. (2007). Un dictionnaire syntaxico-sémantique bilingue de la zootechnie. *Cahiers de linguistique* 33(1), pp. 83--102.
- L'Homme, M.C. (1998). Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie*, 73(2), pp. 61--84.
- L'Homme, Marie-Claude. 2002. What can verbs and adjectives tell us about terms? In *Proceedings of*

*Terminology and Knowledge Engineering (TKE). Nancy, France, August, 2002.* 

- L'Homme, M.C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire* 217, pp. 78--103.
- Maroto, N., Alcina, A. (2009a). Conceptual relations established by processes: the case of "cocción". In *Proceedings of the Eighth International Conference Terminology and Artificial Intelligence (TIA)*, *Toulouse, France, November, 2009.* On line proceedings available at: http://www.irit.fr/TIA09/thekey/articles/maroto\_alcina concrel process.pdf. Last access: March 2012.
- Maroto, N., Alcina, A. (2009b). Formal description of conceptual relationships with a view to implementing them in the ontology editor 'Protégé'. *Terminology* 15(2), pp. 232-257.
- Marshman, E., L'Homme, M.C., Surtees, V. (2009). Marqueurs de la relation cause-effet : stabilité et variation dans des corpus de nature différente. In *Eighth International Conference Terminology and Artificial Intelligence (TIA), Toulouse, France, November, 2009.* On line proceedings available at: http://www.irit.fr/TIA09/thekey/articles/lhomme-mars hman-surtees.pdf. Last access : March 2012.
- Mel'čuk, I., Clas, A., Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvaine-la-Neuve: Duculot.
- Nuopponen, A. (2007). Terminological modelling of processes. An experiment. In Antia, B. (ed.). Indeterminacy in Terminology and LSP. Studies in honour of Heribert Picht. John Benjamins, Amsterdam/Philadelphia: pp. 199--216.
- Pimentel, J., L'Homme, M.C. (2011). Annotation syntaxico-sémantique de contextes spécialisées: application à la terminologie bilingue. In Van Campenhoudt, M., Lino, T., Costa, R. (dirs.). (2011). Passeurs de mots, passeurs d'espoir. Lexicologie, terminologie et traduction face au défi de la diversité. Actes des 8e Journées scientifiques du Réseau de chercheurs Lexicologie, terminologie, traduction. Lisbonne, 15-17 Octobre 2009. Paris: Éditions des archives contemporaines, pp. 651--670.
- Ruppenhoffer, J., Ellsworth, M., Petruck, M., Johnson, C. R., Scheffczyk, J. (2010). *FrameNet II: Extended Theory and Practice*. On-line book available at: http://gemini.uab.es:9080/SFNsite/papers/framenet-bo ok. Last access: March 2012.