Surprise: Spanish FrameNet![1]

Carlos Subirats and Miriam R.L. Petruck
International Computer Science Institute, Berkeley CA
{subirats,miriamp}@icsi.berkeley.edu

Abstract

In this paper, we characterize a set of emotion frames and examine predicates that involve describing events of surprise in order to introduce the Spanish FrameNet project. We compare the Spanish LUs to those in English in order to characterize the similarities and differences between the lexicalization patterns in the two languages. Finally, we suggest that the existing descriptions of English language based frames can be used to describe the semantic and syntactic valences of the Spanish predicates. The larger project serves as a test case for the development of lexical resources based on the structure and content of the original FrameNet project.

1. Background to Spanish FrameNet

Spanish FrameNet (http://gemini.uab.es/SFN), henceforth SFN, is developing a corpus-based lexicon for a significant portion of the vocabulary of present-day Spanish in terms of Frame Semantics (Fillmore 1982, 1985). SFN will provide a body of semantically and syntactically annotated sentences from which reliable information will be reported on the semantic and syntactic valences of each item targeted for analysis. The resulting database is being structured along lines similar to those of the original FrameNet project (http://www.icsi.berkeley.edu/~framenet).

The basic assumption of Frame Semantics is that each word evokes a particular frame and possibly profiles some element or aspect of that frame. Semantic frames are schematic representations of situations involving various participants, props, and other conceptual roles, each of which is called a frame element (FE). The semantic arguments of a predicating word correspond to the frame elements of the frame (or frames) associated with that word. A frame semantic description of a lexical unit identifies the frames which underlie a given meaning and specifies the ways in which frame elements are realized in structures headed by the word (See Johnson, et al. 2002, Fillmore, et al. 2002, and Fillmore, et al. in press).

For example, consider the `Communication_response` frame which deals with communicating a reply or response to some prior communication or action, and whose frame elements minimally include `Speaker`, `Addressee`, `Trigger` and

[1] Spanish FrameNet is funded by the Ministry of Science and Technology of Spain through the grant TIC2002-01338 "Tratamiento automático de la información textual en español: procesamiento léxico, sintáctico y semántico" December 2002 - December 2005. This paper has been supported in part by an ICSI fellowship to Carlos Subirats during some of 2002 and 2003. Spanish FrameNet is developed in cooperation with the FrameNet Project at the International Computer Science Institute in Berkeley, California. The authors would like to thank their colleagues, Beau Cronin, Collin F. Baker, Michael Ellsworth, Charles J. Fillmore, Marc Ortega, Sira Palazuelos, Josef Ruppenhofer, Petra Steiner, and Abby C. Wright for their assistance.

Message. The sentence in (1), below, is a canonical example of a verb in the Communication_response frame.

(1) Sara le **respondió** a Max que ella no iría.
Sara him **responded** to Max that she not would-go
Sara **responded** to Max that she would not go.

Here, *Sara* fills the role of Speaker; *Max* is the Addressee; and *que ella no iría* is the Message. Note that the Trigger is not expressed in this sentence, but may be realized in other types of sentences, as shown in (2), with the noun phrase *la pregunta de Max.*

(2) Sara **respondió** la pregunta de Max.
Sara **responded** the question of Max
Sara **responded** to Max's question.

Each frame element tag is part of a set of three tags, consisting of the frame element (i.e. the semantic tag), the grammatical function, and the phrase type of the annotated constituent. The example in (3) shows these triples for sentence (2), above.

(3) Sara **respondió** la pregunta de Max.
Speaker Trigger
Ext DirObj
NP NP

Here, notice that *Max,* the Addressee, is part of the larger NP that instantiates the FE Trigger, information which is included in the database by tagging just the phrase *de Max* with the FE label Addressee on a secondary FE layer.

Note that we use the term *External* (*Ext*) for subjects of target verbs, as well as for any constituent that controls the subject of a target verb. As shown, the verb *respondió* allows for sentences of the type given in (1), with Speaker, Addressee and Message expressed, as well as that given in (2), with Speaker, Addressee, and Trigger expressed. The mappings between the semantic and syntactic information given in the triples of annotation for the set of sentence types in which a given lexical unit occurs constitutes its *valence*. SFN's goal is to annotate corpus citations and to discover the valence patterns for a large number of words showing how those valence patterns are instantiated in actual sentences.

The next section describes the corpus and software used in SFN.

2. Spanish FrameNet Corpus and Software

SFN uses a 300 million-word corpus which includes both New World and European Spanish texts. The corpus includes a variety of Spanish texts from different genres, primarily newspapers, newswire texts, book reviews, and humanities essays. The project uses the Corpus Workbench software from the Institut für Maschinelle Sprachverarbeitung of the University of Stuttgart for searching the Spanish corpus and creating subcorpora of sentences for annotation.[2] The SFN corpus is tagged with an in-house tool which uses an electronic dictionary of 600,000 forms. This dictionary is expanded automatically from a dictionary that contains 93,000 lemmas. The output of the tagger is a set of deterministic automata, one per corpus sentence, whose transitions are tagged with the lexical and morphological information of the word form electronic dictionary (Subirats and Ortega 2000).

Automatic processes select example sentences and create subcorpora of different syntactic constructions in which a given lexical unit may occur. The sentence extraction is carried out by an automata intersection algorithm (Ortega 2002) that finds the intersection of the output of the tagger with transducers that specify the syntactic forms. The extraction and subcorpora creation processes provide annotators with examples of each possible syntactic configuration in which a given lexical item can occur. Annotators then select sentences for annotation that illustrate the ways in which frame elements are realized syntactically.

Figure 1 shows an actual annotated sentence from the database.

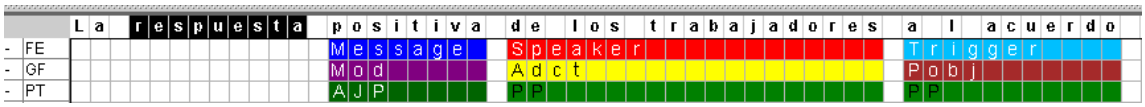

Figure 1. Annotation of a sentence in the `Communication_response` frame

The figure presents the text of the sentence on each of the three levels below it, frame element, grammatical function, and phrase type, as shown by the abbreviations FE, GF, and PT in the leftmost column. The target word *respuesta* is highlighted with a black background, and its dependents are annotated with appropriate frame element tags. In addition, each constituent tagged with a frame element, also receives a grammatical function tag and a phrase type tag.

SFN uses the same annotation software and database structure as that of the Berkeley project. Figure 2 shows part of the Framenet DeskTop software graphical user interface (GUI) for annotation in SFN. The FrameNet DeskTop is divided into a navigation frame on the left and a content space on the right. The navigation frame holds a tree that provides lexicographers with direct access to the main objects in the database, including frames, frame elements and lexical units. Any object in the list may be expanded further, the final one being the example sentence. On the left side of Figure 2, the `Communication_response` frame has been selected, under which is a list of the FEs of the frame. Below the frame elements, there is a list of lexical units in the frame, each of which expands to a set of subcorpora, each of which in turn expands to a list of

[2] http://www.ims.uni-stuttgart.de/

sentences. In Figure 2, there are two lexical units that have been expanded to show the names of their subcorpora: the verb *desmentir* - 'deny' and the noun *respuesta* – 'response'. The content space on the right of the DeskTop is divided into three sections. The top section is for viewing a subcorpus and selecting a sentence.[3] The middle section is for annotating the selected sentence, and the bottom section provides lists of labels available for each of the different annotation layers. Figure 2 shows the annotation of the selected sentence, with the bottom section of the DeskTop's content space open to the frame element layer.[4]

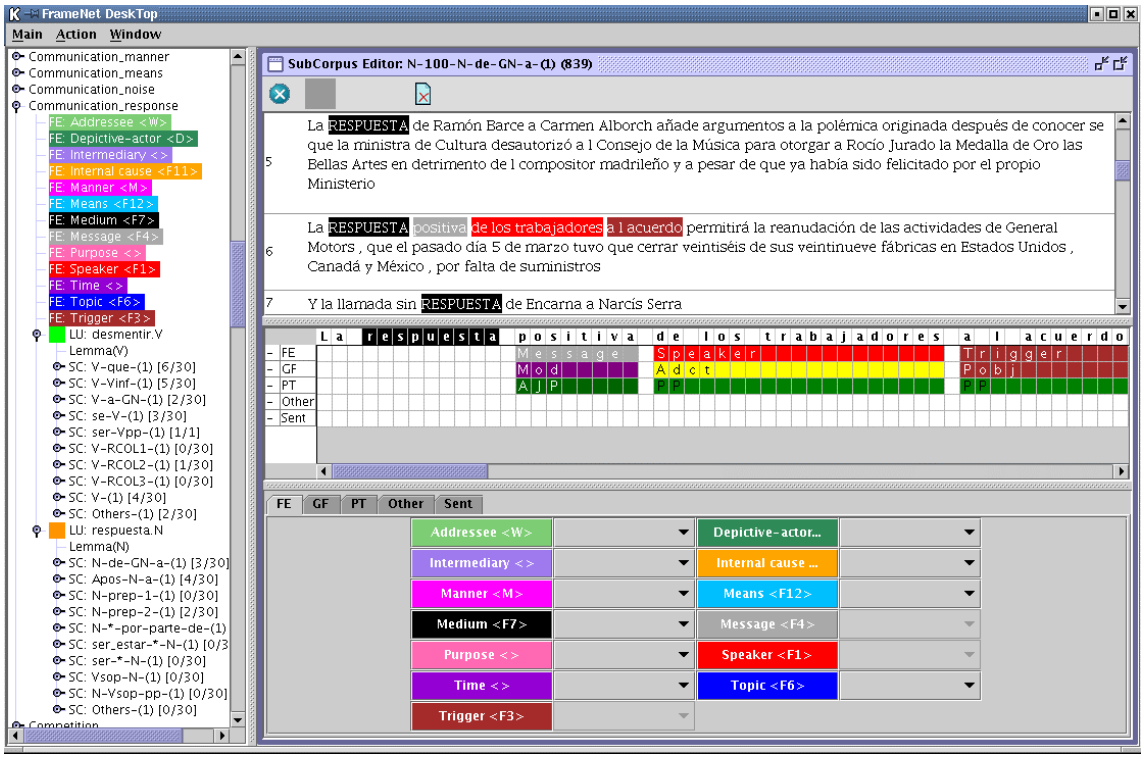

Figure 2. Annotation of a sentence in the `Communication_response` frame

The FrameNet annotation tools allow the user to mark selected constituents in the extracted data according to the frame elements that they exemplify. The result of the annotation process is a set of annotated sentences exemplifying how each frame element in a semantic frame is realized syntactically in respect to a given target word. This can be seen in Figure 2 for sentence 6 in the upper right frame.

SFN uses a MySQL database, and consists of a lexicon with entries for argument taking nouns, verbs and adjectives. Each entry represents a lexical unit, i.e. a pairing of a lemma with a semantic frame. To illustrate, the lemma *calcular* would be paired with

[3] Here the subcorpus is shown in regular mode, though KWIC viewing mode is also possible.
[4] This paragraph has been adapted from Fillmore, Petruck, Ruppenhofer, and Wright (in press).

two different semantic frames (at least). In its mathematical sense, *calcular* - 'calculate' (i.e. do the math) would belong in a calculation frame, while in its other sense it would belong to a cognition frame, quite close in connotation to the colloquial English *guestimate* (i.e. *guess* + *estimate*).

SFN is studying areas of the lexicon that parallel existing English FrameNet descriptions. Our experience tells us that most of the frames defined so far are valid cross-linguistically, because frames are meant to characterize conceptual structure at a basic level of description. It has yet to be determined at what level of description the parallels cease.

The following frames have been defined in the Spanish FrameNet database and LUs have been annotated in each.

Communication Frames	**Emotion Frames**
Conversation Communication_response Questioning Request Statement	Cause_emotion Experiencer_object Experiencer_subject Stimulus_subject

Table 1: Spanish FN Communication and Emotion Frames

In the next section, we provide a brief description of the emotion frames, along with definitions and examples for the relevant frame elements.

3. Emotion Predicates in Spanish

Words in theses frames concern emotions brought about in an `Experiencer`, either as a result of an `Agent`'s action or an outside `Stimulus`. The words are categorized in terms of the four different frames, defined in part by the valence patterns of the frame elements. To illustrate, in the `Experiencer_subject` frame, the `Experiencer` is the subject of the target verb as in (4a) where *Max* is the `Experiencer`, while in the `Experiencer_object` frame, the `Experiencer` is the object of the target verb, as in (4b) where *Sara* is the `Experiencer`.

(4)

a. Max se alarmó de que el motor se hubiera incendiado.
   Max panics of that the engine RFL had caught fire
   Max panics that the engine caught fire.

b. A Sara le fastidian las interrupciones.
   To Sara her bother the interruptions
   Interruptions bother Sara.

While all emotion predicates require an Experiencer (the person having the emotion) and a Stimulus (the source or cause of the emotion), as exemplified in (4), above, some require that the Experiencer be the External (e.g. *alarmase*, as in 4a), while others require that the Stimulus be the external (e.g. *fastidiar*, as in 4b).

In the Stimulus_subject frame, either a Stimulus brings about a particular emotion or experience in the Experiencer or saliently fails to bring about a particular experience, an example of which is given in (5).

(5) Esta historia es **asombrosa** (para nosotros).
This story is amazing (for us)
This story is amazing (for us).

For many LUs in this frame, it is not necessary for the Experiencer to be expressed, although it can be.

The Cause_emotion frame covers those words used for scenarios in which an Agent seeks to bring about an internal mental or emotional state in the Experiencer. For instance, the verb *tranquilizar* - 'calm' as in (6), below, provides an illustration of a canonical example of words in this frame, where the presence of the gerundive form *contando* in the dependent clause makes clear that Carlos acted with the intention of calming Maria.

(6) Carlos **tranquilizó** a María contándole la verdad.
Carlos calmed to Maria telling-her the truth
Carlos calmed Mary by telling her the truth

In contrast, absent further contextual and pragmatic information about the intentionality of the Agent, (7) is ambiguous, even though the so-called "default" interpretation is that Carlos did something with the intention of calming Maria. While human agents have intentionality, they don't necessarily have control over the effect of their actions. This brings about the ambiguity in (7).

(7) Carlos **tranquilizó** a María al contarle la verdad.
Carlos calmed to Maria to tell-her the truth
Carlos calmed Maria by telling her the truth.

Table 2 lists the emotion frames in Spanish FN and English FN. Of the four emotion frames defined in SFN, Cause_emotion is the only one that requires an Agent, whether or not expressed in the sentence.[5] As shown in Table 2, the Cause_emotion frame in SFN corresponds to Cause_to_experience in FN.

[5] Spanish allows subject deletion, hence Agent in subject position need not be expressed in a sentence.

SpanishFN Emotion Frames	English FN Emotion Frames[6]
Cause_emotion	Cause_to_experience
Experiencer_object	Experiencer_object
Experiencer_subject	Experiencer_subject
Stimulus_subject	Subject_stimulus

Table 2: Emotion Frames in SFN and FN

The most significant difference between Spanish and English emotion predicates is that with Spanish `Experiencer_object` predicates, the `Experiencer` is an indirect object, while in the analogous English sentence it is the external argument, as illustrated in (9) with the verb *gustar* -'to like'.

(9) Me **gusta** este libro.
Me please this book
I like this book.

Thus, while Spanish *gustar* is an `Experiencer_object` verb, English *like* is an `Experiencer_subject` verb.

4. Motivating the Lexical Units and Determining Frame Membership

Part of the work of SFN is to determine what forms constitute independent lexical units, and to which frame each belongs. Consider the examples in (10).

(10)
a. Juan **sorprendió** a María al contarle la verdad.
Juan surprised to Maria on explaining-her the truth
Juan surprised Maria by telling her the truth.

b. María **se** **sorprendió** de que Juan cantase.
Maria REFL surprised of that Juan sang
Maria got surprised when Juan sang.

c. María está **sorprendida** de que Juan cante.
Maria is surprised of that Juan sang
Maria is surprised that Juan sang.

Sentence (10a) characterizes a complex scene, which includes the onset of an event, characterized by (10b), and the ongoing state, characterized by (10c). Thus, it is

[6] Emotion frames in FN are currently under review, and there are likely to be some changes regarding which LUs are in the Experiencer_object frame.

noteworthy that *sorprendió* (3rd-person singular of *sorprender*) in (10a), the morphologically simplest form, is used to express a complex event. Moreover, the morphologically more complex forms *se sorprendió* (3rd-person singular of *sorprenderse*) in (10b) and *sorprendido* (*sorprender* + past participle suffix) in (10c) characterize the simpler parts of the complex event. In addition, they are formed by adding linguistic material to the simpler form: the reflexive clitic pronoun *se* is added to *sorprender* to form *sorprenderse*; and the past participle suffix -(*i*)*da* is added to *sorprender* to form the past participle used in construction with *estar* - 'to be'.

We will now argue that there are three separate lexical units in (10a)-(10c). To begin with, *se sorprendió* (in 10b) appears to be an instance of the middle-*se* construction. However, unlike other predicates that occur in the middle-*se* construction[7], the verb *sorprenderse* does not allow a *by*-REFLEXIVE phrase, as shown in (11).

(11) *Juan **se** **sorprendió** por sí solo de que María cantase
Juan REFL surprised by him alone of that María sang

Thus, *sorprenderse* is not comparable to the middle-*se* construction and hence cannot be analyzed as a construction. Furthermore, if there were just one lexical unit for the type of predicate in (10a)-(10c), we would expect that the morphologically complex forms would have to be derived from the simpler form. However, this is not the case. First, there is no necessary relationship between the existence of the participle forms and that of the reflexive forms. To wit, there are participle forms without corresponding reflexives, as in (12a) and (12b); and there are reflexive forms without corresponding participles, as in (13a) and (13b). Thus, the existence of a participle form does not entail the existence of a reflexive; and the existence of a reflexive form does not entail the existence of a participle. This provides evidence against the hypothesis that the forms in (10a)-(10c) are one lexical unit.

(12)
a. Max está **encantado** de que Eva vea sus tesoros.
Max is delighted of that Eva sees his treasures
Max is delighted that Eva sees his treasures.

b. *Max se **encanta** de que Eva vea sus tesoros**.**
Max REFL delights of that Eva sees his treasures

(13)
a. Max se **alegra** de que haya llegado**.**
Max REFL becomes-glad of that has arrived
Max becomes glad that he/she has arrived.

[7] For example, *El niño se **durmió** por sí solo* (the child REFL slept by himself alone) – 'The child fell asleep by himself' illustrates the middle-*se* construction with *durmió* –'sleep'. Sentences of this sort can only be related to a transitive: *El padre **durmió** al niño* (The father "slept" the child) – 'The father put the child to sleep'.

b. *Max está **alegrado** de que haya llegado.
Max is glad of that has arrived

In addition, there are reflexive forms without corresponding non-reflexive forms, as shown in (14), which further supports the argument that the reflexive form is lexical, and not constructional.

(14)

a. Jorge **se** **enorgullece** de los logros de su partido.
Jorge REFL takes-pride of the successes of his party
Jorge takes pride in his party's successes.

b. *Jorge **enorgullece** a su partido.
Jorge takes-pride to his party

c. Juan **se** **extrañó** de que no hubiera llamado.
Juan REFL was-weirded-out of that not had called
Juan was weirded out that he/she hadn't called.

d. *Juan extrañó a Eva de que no hubiera llamado.
Juan weirded-out to Eva of that not had called

Having described the forms and provided evidence to support the claim that there are three lexical units (*sorprender*, *sorprenderse*, and *sorprendido*), we can now consider the meanings conveyed by them. Of the three lexical units under consideration here, the verb *sorprender* has the most complex meaning, despite being morphologically the simplest. It is a causative, and belongs to the `Cause_emotion` frame, in which an Agent seeks to bring about an emotion in an `Experiencer`. As shown in (10a), Juan, the subject of the verb *sorprendió*, is the `Agent` who seeks to surpise Maria, the `Experiencer`. Each of the morphologically more complex forms has a less complex meaning: *sorprenderse* is an inchoative, as it refers to just the beginning of an event; and *sorprendido* is a stative, as it refers to the ongoing state of being surprised. Both of these belong in the `Experiencer_subject` frame, because the `Experiencer` (of the emotion) is realized as the subject of the verb, as seen in (10b) and (10c).

5. Summary Reports

Automatic processes generate reports that show the results of the annotation. For instance, the Lexical Entry Report summarizes the syntactic realizations of the frame elements and the valence patterns of the lexical unit in two tables. The two parts of this report are illustrated in Figures 3 and 4 respectively, for one of the three lexical units discussed here, *sorprender* - 'surprise' in the `Cause_emotion` frame.[8]

[8] Although not discussed here, the verb *sorprender* also occurs in the `Experiencer_object` frame, as

Frame Elements and Their Syntactic Realizations

The Frame elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realizations(s)
Agent	3 exx	NP.Ext 3 exx
Cause	1 exx	VPndo.AObj 1 exx
Experiencer	1 exx	---- 1 exx

Figure 3

As seen in Figure 3[9], the frame element `Agent` is realized as a Noun Phrase that is an External argument; `Cause` is realized as a Verb Phrase with a gerundive verb that is an AObj (Adverbial Object); and `Experiencer` is null instantiated.[10]

Valence Patterns:

These frame elements occur in the following syntactic patterns:

Number Annotated	Patterns	
1 exx TOTAL	Agent	Experiencer
1 exx	NP Ext	--- ---
2 exx TOTAL	Agent	
2 exx	NP Ext	
1 exx TOTAL	Cause	
1 exx	VPndo AObj	

Figure 4

Figure 4 shows the valence patterns, that is, the syntactic and semantic combinations in which the frame elements can occur for the Cause_emotion verb *sorprender*. For instance, in the first example of the table, the `Agent` is a Noun Phrase and External Argument and the `Experiencer` is null instantiated.

---

exemplified in *A Juan le sorprendió que María cantase* (to Juan him surprised that Mary sang) – 'It surprised John that Mary sang.'

[9] SFN is an ongoing project, and the number of sentences annotated is much larger than shown here.

[10] See Johnson, et al. (2002:11-13) for an explanation of null instantiation.

6. What's it Good For?

The Spanish FrameNet database will include a wealth of information useful for linguistic research by providing valence descriptions for a considerable amount of the vocabulary of Spanish. Such information can be used to study crosslinguistic differences in lexicalization patterns, as shown in Table 3.

	Stative being in a state	**Inchoative** entering into a state	**Causative** putting into a state
Spanish	**Experiencer_subject**		**Cause_emotion**
	estar V-PP	**V REFL**	**V**
	estar sorprendido	sorprenderse	sorprender
English	**Experiencer_subject**		**Cause_to_experience**
	be V-PP	**get V-PP**	V
	be surprised	get surprised	surprise

Table 3. Lexicalization Patterns of Spanish and English Emotion Predicates

Table 3 summarizes the differences in the lexicalization patterns of these predicates in Spanish and English. While both languages lexicalize the causative meaning with a verb (*sorprender* and *surprise*) and the stative meaning with an adjective (*estar sorprendido* and *to be surprised*), Spanish lexicalizes the inchoative meaning in the reflexive verb *sorprenderse* - 'to get surprised', while English uses a construction with *get* and the adjectival past participle *surprised*. In addition, while English has just one lexical unit *surprised* in the `Experiencer_subject` frame, Spanish has two: *sorprendido* used in conjunction with *estar* as a stative; and *sorprenderse* which is inchoative.

In addition to its being a resource for theoretical work in linguistics, the Spanish FrameNet database has more practical applications. For instance, Boas (2002) proposes to link German and English FrameNet to create a bilingual FrameNet dictionary. Similarly, it would also be possible to link Spanish and English FrameNet to create a Spanish-English bilingual dictionary. In principle, Spanish FrameNet would also be useful for designing machine translation systems, much the way the frames defined for English are being used to develop frames for representing Japanese texts with a view to using Frame Semantics in a machine translation system.[11]

References

Boas, H.C. 2002. Bilingual FrameNet Dictionaries for Machine Translation. In M. González-Rodríguez and C.P. Suárez-Araujo, eds. *Proceedings of the Third International Conference on Language Resources and Evaluation.* Las Palmas.

[11] This work is being conducted by J. Marc Gawron at San Diego State University.

Fillmore, C.J. 1982. Frame Semantics. In The Linguistic Society of Korea, ed. *Linguistics in the Morning Calm*. Seoul: Hanshin.

Fillmore, C.J. 1985. Frames and the Semantics of Understanding. In *Quaderni di Semantica*, Vol 6, No. 2, pp. 222-254.

Johnson, C.R., Fillmore, C.J., Petruck, M.R.L., Baker, C.F, Ellsworth, M., Ruppenhofer, J., and Wood, E.J. 2002. *FrameNet: Theory and Practice*. International Computer Science Institute, Technical Report-02009. Berkeley, CA.

Fillmore, C.J., Baker, C.F., and Sato, H. 2002. The FrameNet Database and Software Tools. In M. González-Rodríguez and C.P. Suárez-Araujo, Eds. *Proceedings of the Third International Conference on Language Resources and Evaluation.* Las Palmas, Spain.

Fillmore, C.J., Petruck, M.R.L., Ruppenhofer, J., and Wright, A.C. (in press) FrameNet in Action: the Case of Attaching. *International Journal of Lexicography* 16.3.

Ortega, M. 2002. *Intersección de autómatas y transductores en el análisis sintáctico de un texto.* MA Thesis, Polytechnic University of Catalonia, Spain.

Subirats, C., and Ortega, M. 2000. *Tratamiento automático de la información textual en español mediante bases de información lingüística y transductores. Estudios de Lingüística del Español* 10.