Jornada de Seguimiento de Proyectos, 2004
Programa Nacional de Tecnologías Informáticas

Automatic processing of textual information in Spanish (Tratamiento automático de la información textual en español: procesamiento léxico, sintáctico y semántico) TIC2002-01338

Carlos Subirats-Rüggeberg*
Autonomous University of Barcelona

Abstract

The goal of this project is to create an online lexical resource for Spanish with the collaboration of the Berkeley FrameNet Project[1], based on frame semantics and supported by corpus evidence. The "starter lexicon" will be available to the public by January 2006, and will contain at least 1000 lexical items -verbs, predicative nouns, and adjectives- representative of a wide range of semantic domains. The aim is to document the range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses, through:

- **human approved and automatic annotated example sentences, and**
- **automatic capture and organization of the annotation results.**

The resulting database will be in a platform-independent format, and it will be able to be displayed and queried via the web and other interfaces.
Sentences for annotation will be automatically extracted from a 350 million-word corpus. This corpus will be processed with tools which allow automatic processing of lexical and syntactic information in Spanish by means of:

- **electronic dictionaries,**
- **and syntactic transducers, and**
- **automata intersection algorithms.**

The project deliverables will include:

- **a corpus database of 20,000 semantic and syntactically annotated sentences, and**
- **semantic frames and frame elements definitions.**

Keywords: lexical semantics, Spanish, corpus linguistics, frame semantics, semantic role, annotation, natural language processing, tagging, parsing

---

* Email: Carlos.Subirats@uab.es
[1] http://www.icsi.berkeley.edu/framenet

1 Goals of the project

1.1. Basic concepts

A **semantic frame** is a script-like structure of inferences, linked by linguistic convention to the meanings of linguistic units -in our case, lexical units. Each frame identifies a set of **frame elements** (FEs) -participants and props in the frame.
A **frame semantic description of a lexical unit** identifies the frames which underlie a given meaning and specifies the ways in which FEs, and constellations of FEs, are realized in constructions headed by the word. **Valence** descriptions provide, for each word sense, information about the sets of combinations of:

- FEs,
- grammatical functions and
- phrase types attested in the corpus.

The annotated sentences are the building blocks of the database. These are marked up in **XML** and form the basis of the lexical entries. This format supports searching by lemma, frame, frame element, and combinations of these.

1.2. Dictionary and thesaurus

The resulting database, that will be called Spanish FrameNet, will act both as a dictionary and a thesaurus.

- The **dictionary** features include:
  - definitions,
  - tables showing how frame elements are syntactically expressed in sentences containing each word,
  - annotated examples from the corpus:
    - human approved and
    - automatically annotated, and
  - an alphabetical index.
- Like a **thesaurus**, words are linked to the semantic frames in which they participate, and frames, in turn, are linked to wordlists and to related frames.

The Spanish FrameNet project is based on the evidence offered by a 350 million-word corpus[2] which includes both New World and European Spanish. The semantic and syntactic annotation is carried out by using the system developed by the Berkeley FrameNet Project, whose input are files that have been extracted from the corpus, POS tagged, lemmatized[3], and chunked[4]. Each Spanish FrameNet entry will provide links to other lexical resources, including Spanish EuroWordNet synsets and syntactic subcategorization frames. The project's deliverables will consist of the Spanish FrameNet database itself:

- lexical entries for individual word senses,
- frame descriptions, and
- annotated subcorpora.

---

[2] http://gemini.uab.es/SFN/SFN_Corpus.html
[3] http://gemini.uab.es/SFN/SFN_taggers_chunkers.html#Taggers
[4] http://gemini.uab.es/SFN/SFN_taggers_chunkers.html#Chunkers

1.3. Classes of frame elements

1.3.1 External FEs

External FEs are realized outside of the maximal phrase headed by the target lexeme. Externals satisfy an FE requirement of a target word in the following syntactic contexts:

- subjects of finite target verbs, *A Juan le encanta [la paella]*$_{\textbf{External}}$ (John loves paella) or target nouns and adjectives, by virtue of their grammatical relation to a support verb, such as the subjects of *[El presidente]*$_{\textbf{External}}$ *les dio un ultimátum a los terroristas* (The president gave the terrorists an ultimatum), *[Venezuela]*$_{\textbf{External}}$ *es rica en tradiciones* (lit. Venezuela is rich in traditions), etc.
- subjects (or objects) of controlling structures, *[Los políticos]*$_{\textbf{External}}$ *decidieron bajar los impuestos* (Politicians decided to lower taxes), *[Le]*$_{\textbf{External}}$ *obligaron a firmar el contrato* (They forced him to sign the contract)
- "extracted" constituents, etc.

1.3.2 Implicit FEs

Some FEs are conceptually "understood", but are not expressed in relevant positions in the sentence. In order for example sentences representing the same valence to be grouped together automatically, we introduced tags to bear the annotation for the missing FEs. We distinguish three types of Implicit FEs, Existential, Anaphoric, and Constructionally Licensed:

- Existential implicit FEs include the missing objects of *¿Ya has comido?* (Have you eaten yet?) and *Bebes demasiado* (You drink too much).
- Anaphoric implicit FEs include the missing objects of *Ellos decidirán* (They'll decide) and *¿Comprendes?* (Do you understand?).
- Constructionally licensed omissions, e.g., subject delition in Spanish, the implicit subjects of imperatives, etc., are not lexicographically relevant.

1.3.3 Conflated FEs

Many lexemes that can express several FEs as separate constituents can also express them as single constituents in which information about two FEs is conflated. In *Le han nombrado director general* (They have named him general director) the person and the office show up as separate constituents; but in *Han nombrado al director general* (They named the general director) we find a single constituent identifying both.

1.3.4 Incorporated FEs

We sometimes find FEs which are typically expressed as separate constituents in the valence patterns of one lexeme, but are typically incorporated in other lexemes in the same frame. For example, in the Shoot projectiles frame, the Firearm is a separate constituent in *Les dispararon con una ametralladora* (They shot at them with a machine gun), but incorporated in *Les ametrallaron* (They machine gunned them).

1.3.5. Complex Frames

The frames underlying some lexical units are best understood as comprising more than one frame, through simple **frame inheritance**, **multiple frame inheritance** with specifications of FE, **binding**, and **frame composition:**

- **Frame Inheritance:** The combinatorial properties of, say, *empujar* (push) are not only those determined by the unique meaning of the verb and the immediate frame in which it participates (CAUSE_TO_MOVE), but also by the fact that it is an action verb (involving agent, patient, optional instrument, etc.), and that it is an event verb (allowing specification of temporal and locational parameters). Thus properties of more general frames are inherited by more specific ones.

- **Multiple Frame Inheritance:** Consider the sense of *discutir* (argue), which inherits from both the frames DISPUTE and CONVERSATION. While *discutir*, like all CONVERSATION words, involves "reciprocal talk", some of its properties are inherited from the grammar of *disputing* or *fighting*. In this, it differs from other conversation verbs, like *charlar* (chat), *comentar* (comment), etc. Both DISPUTE and CONVERSATION frames, in turn, inherit the RECIPROCITY frame, which allows variable syntactic realization of the participants: either joint, as in *Ellos discutieron* (They argued), or disjoint, as in *Él discutió con ella* (He argued with her). CONVERSATION is also heir to the SPEAKING frame, while DISPUTE inherits the ATTACKING frame.
  Compare *despreciar* (scorn), *admirar* (admire), *criticar* (criticize), and *adular* (flatter). All of these verbs belong to a class of JUDGMENT verbs, involving one person passing judgment on the behavior of another. Of these, *criticar* and *adular* are also speaking verbs, and in these cases the Judge in the JUDGMENT frame is also the Speaker in the SPEAKING frame. In addition, while *criticar* allows non-identity of the Evaluee of the JUDGMENT frame and the Addressee of the SPEAKING frame (as in *Él me criticó en la prensa* (He critized me in the newspapers), *adular* requires a single participant for both FEs.

- **Frame Composition:** In numerous cases a frame is complex because it contains another frame as one of its parts. Compare *Sacudió el mantel* (He shook the tablecloth) and *Sacudió las migas del mantel* (He shook the crumbs out of the tablecloth). The shaking (i.e. direct manipulation) is applied to the direct object in the first sentence, but is only a component of the full scene associated with the second sentence. We believe that Frame Semantics, combined with a feature-value representation of event structure, will provide new insights into much current work on this type of regular polysemy, which deals with patterns of valence variation.

1.4. Spanish Corpus

Our corpus, currently under construction, includes both New World and European Spanish. It is composed of texts of different genres, primarily newspapers, newswire texts, book reviews, and humanities essays.These texts of various origins and genres make a grand total of 350 million words. Our research project wishes to acknowledge the support of:

- Anthropos Editorial (Barcelona, Spain),

- Diario ABC (Madrid, Spain), and
- El Mundo (Madrid, Spain),

which made it possible for this research project to use excerpts of their texts and publications as the evidential basis for the inquiry into the behaviour of Spanish words. The SFN Corpus also includes the *Spanish Newswire Text,* Vol. 2, made available through the Linguistic Data Consortium[5]. The IMS Corpus Workbench of the Institut für Maschinelle Sprachverarbeitung[6] of the University of Stuttgart has been used to explore, extract, and sort example lines and sentences from the SFN Corpus.

1.5. Taggers and lemmatizers

The project Corpus is tagged with an application which uses an electronic dictionary of 600,000 forms[7], which are expanded from a dictionary of 93,000 lemmas:

- 67,000 single-word lexical units, like *marea* (tide), *inmoralidad* (immorality), *allí* (there), etc.;
- 26,000 multi-word lexical units, like *muerte cerebral* (brain death), *carga de profundidad* (depth charge), *ácido graso no saturado* (non saturated fat acid), etc.

Each form of the expanded dictionary is associated to a set of tags[8] that specify:

- the lemma to which it is associated,
- its POS, and
- the inflectional propreties of:
  - verbs (mood, tense, person, and number), and
  - nouns, adjectives, and past participles (gender and/or number).

The output of the tagger is a deterministic finite automata[9] or a 'flat' text. Disambiguation of the tagged sentences is carried out with contextual rules which apply to both formats.

1.6. Parsing Spanish Texts

Parsing is carried out with an intersection automata algorithm which uses subsequential transducers (Ortega 2002[10]) to detect:

- compound verbal forms, like *estamos estudiando* (are studying), *ha estado trabajando* (has been working), etc.
- idiomatic verbs, like *dar por sentado* (take for granted), *romper el hielo* (break the ice), etc.
- noun phrases and adjective phrases

The automata intersection algorithms intersect a text, previously converted into deterministic finite automata (DFA) whose transitions have been tagged using information from an electronic dictionary, with *p*-subsequential transducers. These transducers specify lexical properties of multiword lexical units as well as formal properties of syntactic constructions in Spanish. The output of the intersection of the DFA (text) and the transducers (lexical and syntactic constructions) is transduced DFA where:

---

[5] http://www.ldc.upenn.edu
[6] http://www.ims.uni-stuttgart.de
[7] http://gemini.uab.es/carlos/INFO-DICO/formas.html
[8] http://gemini.uab.es/carlos/INFO-DICO/etiquetario.html
[9] http://elies.rediris.es/elies10/3.htm#f5.2
[10] http://gemini.uab.es/SFN/Reports.html#Ortega

- single word as well as multi word lexical units are unambiguously POS tagged, lemmatized,
- inflectional morphological properties of verbs, nouns and adjectives are specified, and
- certain syntactic constructions, like, compound verb forms, noun phrases or prepositional phrases are parsed.

The transduced corpus is intersected with transducers which specify the syntactic constructions of sentences that have to be annotated. These sentences are then formatted in XML and downloaded into the FNDesktop, a tool developed by the Berkeley FrameNet project that we have adapted to Spanish. This tool is used by the linguists to select and annotate sentences.

2 Accomplishments

The results of the project can be queried via web with the Report System and FrameSQL:

- Report System: http://oasis.uab.es:8080/farinaweb/Index
- FrameSQL[11]: http://sato.fm.senshuu.ac.jp/sfn20/notes/index2.html

3 Results and deliverables

The project's linguistic deliverables comprise lexical and frame databases:

- sample lexicons of Spanish nouns, verbs, and adjectives together with databases of frame descriptions characterizing the conceptual structures which support the word meanings;
- detailed exhibitions, through sorted collections of semantically and syntactically annotated sentences, of the ways in which phrases built around these words instantiate elements of the host frame.

The computational deliverables consist of a suite of tools adapted from FrameNet[12] developed and shaped for carrying out the various steps in this research:

- for manipulating the Spanish corpora which are POS-tagged, lemmatized and partially parsed;
- for parameterizing search and sort procedures for organized sampling of uses of specific lemmas;
- for designing tag sets to be used in annotation;
- for annotating sentence constituents by frame element;
- for parsing the portions of sentences in which a particular target word has been identified and the phrases which constitute its arguments have been blocked off;
- for generating summaries of semantic and syntactic combinatorial properties as discovered through the empirical work and for constructing lexical entries.

Our project which incorporates both linguistic-semantic and natural language processing techniques will also provide crucial information for many language engineering tasks such as:

[11] FrameSQL has been developed by Prof. Hiroaki Sato, Senshu University in Japan
[12] http://www.icsi.berkeley.edu/framenet/

- automatic word-sense disambiguation;
- automatic labelling of semantic roles:
- advanced semantically-driven language models for speech recognition

4 References

[1] Baker, C.F., C.J. Fillmore and B. Cronin. (2003). The Structure of the FrameNet Database, International Journal of Lexicography, Volume 16(3), (pp. 281-296).

[2] Baker, C.F. and J. Ruppenhofer (2002) FrameNet's Frames vs. Levin's Verb Classes. In J. Larson and M. Paster (Eds.) In Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society. (pp.27-38).

[3] Fillmore, C.J. (1985). Frames and the semantics of understanding. In Quaderni di Semantica 6(2), (pp. 222-254).

[4] Fillmore, C.J. (1982). Frame Semantics. In Linguistics in the Morning Calm (pp.111-137). Seoul: Hanshin Publishing Co.

[5] Fillmore, C.J., C.R. Johnson and M.R.L. Petruck. (2003). Background to FrameNet. International Journal of Lexicography 16(3), (pp. 235-250).

[6] Fillmore, C.J. and H. Sato. (2002). Transparency and Building Lexical Dependency Graphs. In J. Larson and M. Paster (eds.) Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society (pp. 87-99).

[7] Johnson, C.R., M.R.L. Petruck, C.F. Baker, M. Ellsworth, J. Ruppenhofer, and C.J. Fillmore. (2002). FrameNet: Theory and Practice. Berkeley, CA: International Computer Science Institute, Technical Report-02009.

[8] Ortega, M. 2002. *Transductores en el análisis léxico y sintáctico de un texto*. Proyecto tesis de licenciatura, Universidad Politécnica de Cataluña.

[9] Sato, H. (2003). FrameSQL: A Software Tool for FrameNet. In *ASIALEX '03 Tokyo Proceedings* (pp. 251-258), Asian Association of Lexicography.

[10] Subirats, C. and M. Ortega. (2000). Tratamiento automático de la información textual en español mediante bases de información lingüística y transductores. *Estudios de Lingüística del Español* 10: http://elies.rediris.es/elies10/

[11] Subirats, C. and M. R. L. Petruck. (2003). Surprise: Spanish FrameNet! *International Congress of Linguists. Workshop on Frame Semantics (July 29, 2003)*. Prague

TIC2002-01338

[12] Subirats, Carlos; Sato, Hiroaki. 2004. Spanish FrameNet and FrameSQL. *4th International Conference on Language Resources and Evaluation. Workshop on Building Lexical Resources from Semantically Annotated Corpora.* Lisbon.