Automatic Extraction of Textual Information in Spanish[1]

CARLOS SUBIRATS RÜGGEBERG
Laboratorio de Lingüística Informática
Universidad Autónoma de Barcelona

1. Lexical syntax focuses on the study of the lineal projection of the dependency relation between predicates and arguments in those syntactic constructions, which convey meaning, as well as on the semantic and formal study of the derivational relations, which can be established within such constructions. Lexical syntax, therefore, specifies the hierarchy of classes in the lexicon of predicates and the relations between the sentences, which constitute the lineal projections of these classes.

By the same token, lexical semantics focuses firstly on the study of the relations of semantic dependency within the lexicon, such as hyponymy and meronymy and secondly, on the lineal semantic relations, such as antonymy. In the case of the morpho-phonological forms, which correspond to more than one entry and therefore are polysemic, lexical semantics assigns to each entry a non-ambiguous paraphrase expressed in a lexical and syntactic sublanguage. These paraphrases should be extensions of the hierarchical and lineal semantic relations of the corresponding entries.

This paper describes an application based on lexical syntax and semantics, which enables the automatic extraction of information in Spanish texts. This application extracts the sentential information from the relational characterization of the syntactic constructions, which serve as carriers of meaning. To achieve this, it detects the relations between predicates and arguments in order to unambiguously determine the information transmitted by these relations. Actually, this is composed of three integrated modules, which cyclically carry out the following processes:

(1) TAGGING OF LEXICAL ITEMS**:** This not only means the tagging of simple forms (chains of characters between two blank spaces), but also of idioms (lexical items composed of two or more simple forms);

(2) RESOLUTION OF AMBIGUITIES: This process totally or partially eliminates the ambiguities introduced by the tagger;

(3) DETERMINATION OF THE HIERARCHY OF DEPENDENCIES BETWEEN PREDICATES AND ARGUMENTS: This hierarchy carries the sentential information (Harris 1991).

2. The tagger has access to an electronic dictionary composed of 600,000 lexical items: 550,000 simple forms and 50,000 idioms. This dictionary includes not only the invariable forms such as adverbs and conjunctions, but also the forms of those classes of words with morphological inflection, such as verbs, nouns and adjectives. The dictionary is built up by automatically expanding a lexicon of canonical forms, composed of 92,000 items, of which 66,000 are simple forms and 26,000 are idioms (Subirats 1992). The items of the dictionary are accompanied by the following information:

-the canonical form or forms to which the item is associated;

-the class of words to which the above mentioned canonical forms belong to;

-the inflectional morphological properties of verbal, nominal and adjectival items (in relation to a specific canonical form).

Since idioms (leaving aside verbal idioms) constitute more than one third of the lexicon of the Spanish language, the lexical analysis of a text necessarily requires a module to tag idioms. This module would have the following features:

> (1) In certain contexts, idioms may also be interpreted as a sequence of simple forms. Thus, idioms must be tagged both as idioms and as a sequence of simple forms. For instance, in a sentence like *A la ministra de defensa no le gusta que le griten* (The Secretary of Defense does not like to be shouted at), *ministra de defensa* (Secretary of Defense) must be interpreted as an idiom. In contrast, in the sentence *Los políticos le hablaron a la ministra de defensa* (The politicians talked to the Secretary of Defense /The politicians talked to the Secretary about defense), *ministra de defensa* may be construed as an idiom or as a sequence of simple forms which belong to two prepositional phrases, namely, *a la ministra* (to the Secretary) and *de defensa* (about defense).

> (2) Non-ambiguous idioms that do not admit a componential interpretation would only tagged as idioms. The non-ambiguous idioms which are easiest to recognize are those that include forms which appear only in idioms. For instance, *a troche y moche* (helter-skelter/ pell-mell) is easily identified as a non-ambiguous adverbial idiom, because the forms *troche* and *moche* have no meaning in themselves, and only appear in the above mentioned adverbial idiom. However, there are other idioms that are not ambiguous, and which do not include forms that only belong to idioms.

For example, the relative pronoun *lo que* (that which), which is formed by the clitic pronoun *lo* (it) and the relative pronoun *que* (that), is not ambiguous because it does not admit a componential interpretation in any context. The possibility or impossibility of interpreting an idiom as a sequence of simple forms in a particular context is an idiosyncratic property. As a result, it is specified in the dictionary.

> (3) The tagger should indicate the ambiguities of idioms which include other idioms. The most common cases of inclusion are the following:

SIMPLE INCLUSION: The adverbial idiom *hoy por hoy* (nowadays) includes the adverbial idiom *por hoy* (for today). The square brackets with numbers in subscript mark the extension of each, and illustrate the inclusion of one within the other:

$${}_{1}[\textit{hoy}\ \ {}_{2}[\textit{por hoy}]_{2}\ ]_{1}$$

DOUBLE INCLUSION: The adverbial idiom *medio en broma medio en serio* (half joking, half seriously) includes two adverbial idioms, *en broma* (jokingly) and *en serio* (seriously):

$${}_{1}[\textit{medio}\ {}_{2}[\textit{en broma}]_{2}\ \textit{medio}\ {}_{3}[\textit{en serio}]_{3}\ ]_{1}$$

NESTED INCLUSION: The adverbial idiom *de una vez por todas* (once and for all) includes *de una vez* (right now) which in turn, includes *una vez* (once):

$${}_{1}[\ {}_{2}[\textit{de}\ {}_{3}[\textit{una vez}]_{3}\ ]_{2}\ \ \textit{por todas}]_{1}.$$

COMBINATION OF INCLUSIONS SUCH AS THE DOUBLE AND THE NESTED INCLUSIONS: The adverbial idiom *de una vez para siempre* (once and for all) includes two idioms, *de una vez* (right now) and *para siempre* (forever), but at the same time, *de una vez* includes the idiom *una vez*, as pointed out above:

$${}_{1}[\ {}_{2}[\ \textit{de}\ \ {}_{4}[\ \textit{una vez}\ ]_{4}\ ]_{2}\ \ {}_{3}[\textit{para siempre}]_{3}\ ]_{1}$$

This typology includes the most common ambiguities that appear in idioms, when they are considered as lexical items. However, this typology does not include other ambiguities that appear in certain noun or prepositional phrases, such as intersecting idioms. For instance, in the noun phrase, *agua de riego por aspersión* (water irrigation by sprinkling), there is an ambiguity caused by the intersection of the noun phrases, *agua de riego* (water for irrigation) and *riego por aspersión* (irrigation by sprinkling). Likewise, the prepositional phrase *a la fuerza aérea* (to the air force) shows another ambiguity which is caused by the intersection of *a la fuerza* (by force), and *fuerza aérea* (air force).

3. The tagger in our application searches each text for the occurrence of simple and compound lexical items, and looks them up in the dictionary. It substitutes the lexical items for the information associated with them in the dictionary, and formalizes the lexical information (including ambiguities) in an automaton (Fig. 1). In general, taggers display their output, in other words, the lexical items of the text, and the lexical, categorial, and morphological information assigned by the tagger, in columns or in two-line blocks. This format, though user-friendly, does not allow for the formalization of ambiguities related to simple or compound lexical items, and thus is an obstacle to further reprocessing.

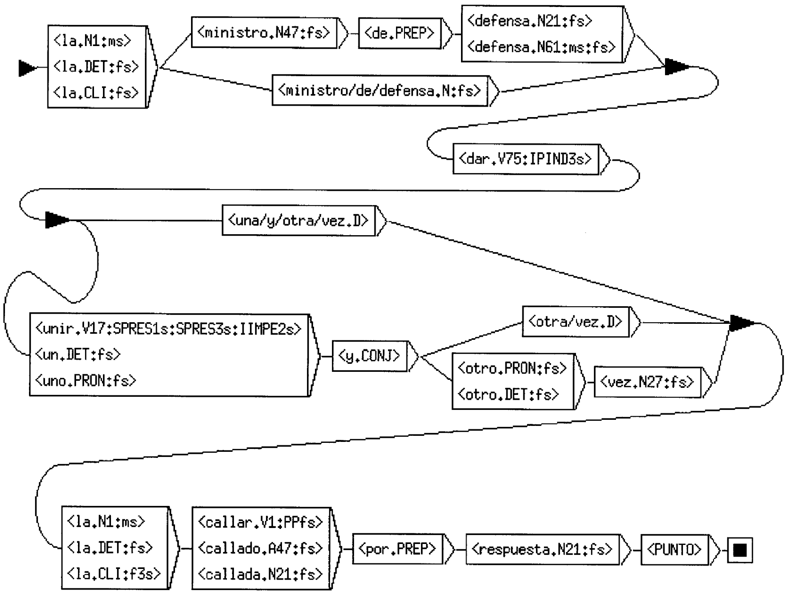

Fig. 1. Automaton-like representation of the tagged sentence *La ministra de defensa dio una y otra vez la callada por respuesta* (The Secretary of Defense gave over again and over again her silence as a response[3]./The Secretary of Defense kept silent and refused to reply). Values of tags are specified at the Appendix.

Since the tagger assigns all possible lexical, morphological and/or categorial information to the lexical items, it also includes ambiguities. Therefore, a disambiguation process is required. Within our system, disambiguation is carried out by an automata intersection algorithm[2] (Ortega

1997). This algorithm intersects minimized deterministic automata (i.e. the output of the tagger) with deterministic subsequential transducers, specifying contextual lexical constraints (Silberztein 1993; Subirats 1998). The result of intersecting an input automaton with a transducer may consist in adding or removing one or more transitions of the output automaton.

The automata and transducer transitions can be labeled with one symbol or with a chain of symbols of their alphabets. Thus, for each transition:

The automata intersection algorithm can compare independent symbols or chains of symbols. Furthermore, the transducers of our application can use variables, which take their value from the automata with which they intersect. Variables may take the value of an entire transition. In the case of transitions labeled with chains of symbols, variables can take the value from one or more symbols of the chain.

The aim of using an intersection algorithm is to transduce input sentences displayed in the form of automata, in other automata, which are canonical syntactic representations of the input sentences. The latter allow the non-ambiguous identification of the relations between predicates and arguments, which convey meaning (Harris 1991).

For example, the automaton of Fig. 1, which results from tagging the sentence *La ministra de defensa dio una y otra vez la callada por respuesta* (= The Secretary of Defense gave over again and over again her silence as a response), presents the following ambiguities:

(1) *la* (the/it/her) can be interpreted as a determiner, noun, or a clitic pronoun;

(2) *ministra de defensa* (Secretary of Defense), as mentioned above, can be interpreted as an idiom or as a concatenation of simple forms. In the latter case, *defensa* "the act or process of defending" can be interpreted as a feminine noun or alternatuively, as *defensa* (a back/fullback on a football team), a noun that can be either a feminine or masculine;

(3) *una y otra vez* (over and over again) can be interpreted as:

(3.1) an adverbial idiom;
(3.2) the coordination of *una* (a, one) and the idiom *otra vez* (once again);
(3.3) a concatenation of simple forms.

By the same token, in (3.2) and in (3.3), *una* also admits an interpretation as the first or third person singular of the present subjunctive or as the second person singular of the imperative of the verb *unir* (join). Similarly, in (3.3), *otro* (another) can be either a pronoun or a determiner.

(4) *callada* (silence) can be interpreted as a past participle, as an adjective or as a noun.

The ambiguities described above in (1-4) can be eliminated by intersecting the automaton-like representation of the tagged sentence where these ambiguities appear with transducers, which formalize certain restrictions of Spanish lexical syntax. As a matter of fact, the ambiguity referred to in (1) can be removed by intersecting the automaton in Fig.1 with the transducer in Fig. 2. The latter specifies that if the singular feminine determiner *la* (the) is followed by a singular feminine noun, it must only be tagged as a singular feminine determiner and the following noun as a singular feminine noun. Likewise, the ambiguities referred to in (3) and (4) above can be simultaneously eliminated, since they appear in relation with the non-ambiguous verbal idiom *dar la callada por respuesta* (give silence has a response). Actually, as it can be seen in Fig. 4, the intersection of the automaton of Fig. 1 with the transducer of Fig. 3 allows the following transductions:

(1) The adverbial idiom *una y otra vez* (over and over again) is extracted and placed after the verbal idiom. In the input automaton, *una y otra vez* is placed inside the verbal idiom *dar la callada por respuesta,* between its verbal nucleus *dar* and the fixed string, *la callada por respuesta*. Actually, given its place in the input automaton, it admits only an interpretation as an adverbial idiom

(2) *Dar* and *la callada por respuesta* are taken, and the remaining ambiguities associated to its non-idiomatic interpretation resolved. For example, *callada* is given its correct interpretation.

(3) The verbal idiom *dar la callada por respuesta* is assigned the temporal variable *(VAR-2,* cf. Fig. 3*)*, which in the input automaton, is a tag of its verbal nucleus *dar.*

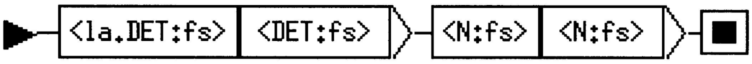

Fig. 2. Transducer which enables the disambiguation of *la* (the), when it precedes a feminine noun. The left label of each transition corresponds to its input and the right label, to its output .

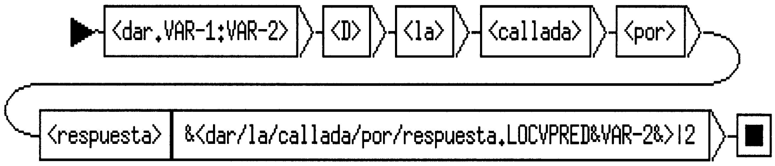

Fig. 3. '&' means that the transduction is carried out with the following string: '|' marks the creation of a new transition in the output automaton; '2' stands for a variable referring to the number in the corresponding transition of the input automaton and depending upon the value of this transition. *LOCVPRED* is a code that is introduced in the transduction, and which means *predicative verbal idiom*.

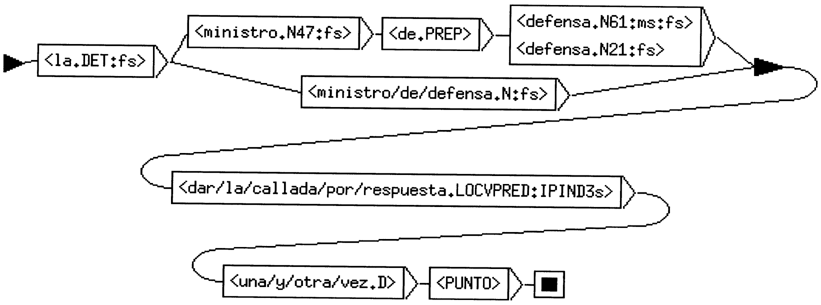

Fig. 4. Automaton resulting from the cyclical intersection of Fig. 1 automaton with Fig. 2 and 3 transducers.

Dar la callada por respuesta is a predicative verbal idiom, whose argument requirement includes only one nonsentential argument: it is a P_n and its lineal projection is N_1 *dar la callada por respuesta*. Thus, in Fig.4, its argument structure can be parsed with the transducer in Fig. 5. The latter parses noun phrases which are optionally preceded by a determiner followed by any number (including zero) of prepositional phrases introduced by *de* (of). Since *ministra de defensa* is the only possible argument of the verbal idiom *dar la callada por respuesta*, the intersection of the automaton in Fig. 4 with the transducer in Fig. 5 allows the parsing of the argument structure of the sentence, and also the categorical rejection of the non-idiomatic interpretation of *ministra de economía*. Consequently, the automaton in Fig.6 determines the relations of the canonical syntactic construction, which permits the identification in a non-ambiguous way of the meaning conveyed.

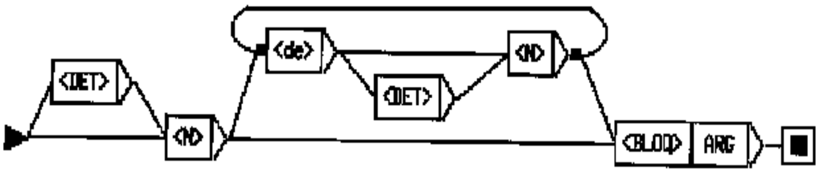

Fig. 5. Transducer which parses noun phrases; *BLOQ* indicates that the transduction must introduce two new transitions, namely '{' and '}*ARG*', which mark the beginning and the end of the noun phrase respectively. *ARG* stands for *argument*.

4. These methods of tagging, disambiguating, and finally detecting the relation between predicates and arguments unify processes, which have until now been treated separately. At the same time, they extract information from Spanish texts with a single automata intersection algorithm. Since the unified morphological and syntactic processes formalized in transducers constitute extensions of Spanish lexical syntax, the proposal of treating these processes globally is also a way of verifying its empirical scope as a simulation model of the speaker's linguistic knowledge.

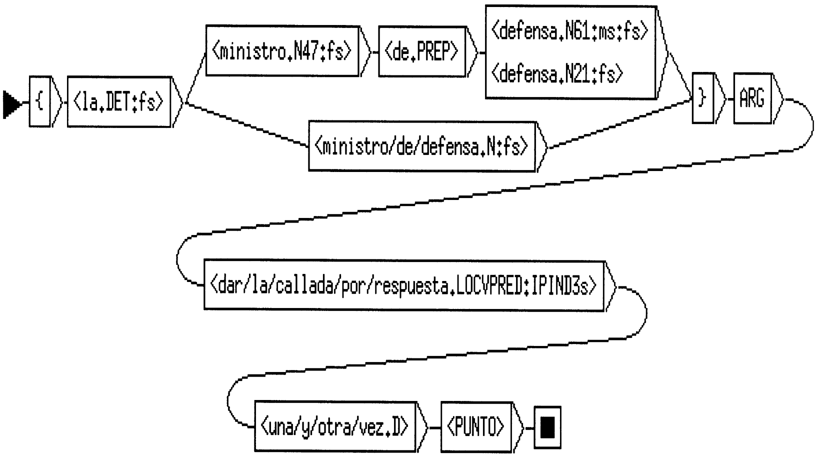

Fig. 6. Automaton resulting from the intersection of the automaton of Fig. 4 with the transducer of Fig. 5. The braces state the beginning and the end of the parsed argument.

APPENDIX

Category codes

A	*adjective*
AINT	*interrogative adjective*
APOS	*possessive adjective*
AREL	*relative adjective*
CLI	*clitic pronoun*

CONJ	*conjunction*
D	*adverb*
DET	*determiner*
INTE	*interjection*
LOC	*form belonging to an idiom*
N	*noun*
PINT	*interrogative pronoun*
PN	*predicative prepositional phrase*
PREL	*relative pronoun*
PREP	*preposition*
PRON	*pronoun*
V	*verb*

MORPHOLOGICAL CODES

1	*first person*
2	*second person*
3	*third person*
f	*feminine*
GER	*gerund*
ICOND	*conditional indicative*
IFUTU	*future indicative*
IIMPE	*imperative*
INF	*infinitive*
IPIMP	*imperfect past indicative*
IPIND	*simple past indicative*
IPRES	*present indicative*
m	*masculine*
n	*nonpersonal gender*
p	*plural*
PP	*past participle*
s	*singular*
SPIMA	*imperfect past subjunctive A*
SPIMB	*imperfect past subjunctive B*
SPRES	*present subjunctive*

Notes

[1] I would like to thank Marc Ortega for his valuable help in the development of this project. I would also like to thank Quim Borges and Sebastián Galera for their observations and suggestions in relation to the development of automata intersection algorithms. Last but not least, I would like to thank Mercedes García for her help in the translation into English of the Spanish original text. This project is financed by the Spanish *Ministerio de Educación* (TIC96-0804).

[2] The code of this algorithm has been created by Marc Ortega.

[3] Literal translation.

References

Díez Orzas, P.L. 1997. *La relación de meronimia en los sustantivos del léxico español: contribución a la semántica computacional.* Ph.D. dissertation, Universidad Autónoma de Madrid.

García-Page, M. 1996. Sobre las variantes fraseológicas en español. *Revista Canadiense de Estudios Hispánicos* 20.3:477-490.

Harris, Zellig S. 1991. *Language and Information. A Mathematical Approach.* Oxford: Clarendon Press.

Mohri, Mehryar. 1997. Finite-state transducers in language and speech processing. *Computational Linguistics* 23.2:269-311.

Ortega, Marc. 1997. *Extensión y composición de autómatas y transductores*. Laboratorio de Lingüística Informática, Universidad Autónoma de Barcelona.

Palacios, Roser. 1996. *Operadores de primer nivel con complemento de régimen. Un estudio transformacional en el léxico*. Laboratorio de Lingüística Informática, Universidad Autónoma de Barcelona.

Roche, Emmanuel. 1996. Finite-state transducers: parsing free and frozen sentences. In A. Kornai, ed. *Proceedings of the ECAI 96 Workshop. Extended Finite State Models of Language,* pp.52-57. (Also in http://www.cs.rice.edu/~andras/confirmed.html)

Roche, Emmanuel and Shabes, Ives. 1995. Deterministic part-of-speech tagging with finite-state transducers. *Computational Linguistics* 21.2:227-253.

Sánchez León, F. 1997. *Análisis morfosintáctico y desambiguación en castellano*. Ph.D. dissertation, Universidad Autónoma de Madrid.

Silberztein, Max. 1993. *Dictionnaires électroniques et analyse automatique de textes*. Paris: Masson.

Subirats Rüggeberg, C. 1998. Bases de conocimiento lingüístico y análisis automático del discurso. *La Coronica Spanish Medieval Language and Litterature,* forthcoming.

Subirats Rüggeberg, C. 1992. Verbal, nominal and adjectival inflection in the Electronic Dictionary of Simple Forms of Spanish. *Lingvisticae Investigationes* 16.2:345-371.

Subirats Rüggeberg, C. 1987. *Sentential Complementation in Spanish.* Amsterdam: John Benjamins.

Voutilainen, Atro. 1997. *EngCG tagger, Version 2.* In T. Brondsted and I. Lytje, eds. *Sprog og Multimedier.* Aalborg: Aalborg Universitetsforlag. (Also in http://www.ling.helsinki.fi:80/~avoutila/cg/doc/aalborg/aalborg.ps