An empirical approach to a preliminary successful identification and resolution of temporal expressions in Spanish news corpora

María Teresa Vicente-Díez¹, Doaa Samy², Paloma Martínez¹
¹Universidad Carlos III de Madrid
Avda. de la Universidad 30, Leganés 28911, Madrid, Spain
{tvicente, pmf}@inf.uc3m.es

²Cairo University
Main Campus, Giza 12613, Egypt
dsamy@cu.edu.eg

Abstract
Dating of contents is relevant to multiple advanced Natural Language Processing (NLP) applications, such as Information Retrieval or Question Answering. These could be improved by using techniques that consider a temporal dimension in their processes. To achieve it, an accurate detection of temporal expressions in data sources must be firstly done, dealing with them in an appropriated standard format that captures the time value of the expressions once resolved, and allows reasoning without ambiguity, in order to increase the range of search and the quality of the results to be returned. These tasks are completely necessary for NLP applications if an efficient temporal reasoning is afterwards expected. This work presents a typology of time expressions based on an empirical inductive approach, both from a structural perspective and from the point of view of their resolution. Furthermore, a method for the automatic recognition and resolution of temporal expressions in Spanish contents is provided, obtaining promising results when it is tested by means of an evaluation corpus.

1. Introduction
Automatic management of temporal information has been subject of a growing interest. A wide range of Natural Language Processing (NLP) applications, such as Information Retrieval (IR) or Question Answering (QA) can highly benefit from temporal reasoning techniques that start with the identification and resolution of temporal expressions.

According to (Ahn et al., 2005), temporal expressions (henceforth, timexes) are natural language phrases that refer directly to time points or intervals. They not only convey temporal information on their own but also serve as anchors for locating events referred to in text. Thus, temporal reasoning demands to detect the time when the events occur. A successful detection requires an accurate identification of time expressions (recognition) in first instance and their resolution and return in an appropriated standard format (normalization) in second instance.

Recognition and normalization imply a number of challenges, some of which are related to the nature of the time expressions. For example, the majority of temporal expressions are deictic or relative (“el próximo mes” (“next month”)). That means that a date of reference is needed to solve and capture the underlying semantics of these expressions.

Other challenges are related to the lack of resources. Previous studies have faced the problem of the treatment of temporal information. However, the majority of the available resources are in English (Mani and Wilson, 2000; MITRE, 2007; Pustejovsky et al., 2003) and not all can be directly applicable to Spanish language. Few studies have addressed the temporal information in Spanish (Martínez-Barco et al., 2002, Saquete et al., 2004, Saquete et al., 2006, Saquete et al., 2006b). These studies mainly adopt a deductive approach which parts from the knowledge to the data. In this way, the novelty of our proposal lies in its empirical inductive approach applied to Spanish, as well as in the suggested time expressions typology.

This work presents an empirical method for the recognition and normalization of temporal expressions in Spanish news corpora. Through the analysis of the different types of timexes in the corpora, first, we present a proposal for a generic framework describing the typology of time expressions. Second, we describe how this typology is used in the management of temporal expressions both for their identification and resolution. The typology, together with the patterns that define these expressions, form up the knowledge base considered as a starting point for a successful automatic identification and resolution of temporal expressions in Spanish. A study of the training corpus and the analysis of the frequencies of the temporal expressions included are presented. Finally, results obtained of applying this method on an evaluation corpus are shown and discussed, just as those pending aspects suggested for future works.

2. Typology Proposal
Two perspectives are considered in this typology: the first is a structural perspective concerning structural form of the timexes; the second is concerned with the resolution of the timexes in relation to their reference point in time.

2.1 Structural Perspective Classification
According to this classification and from a structural perspective, two elements are considered as basic
constituents: the *time-unit* and the *time-modifier*. This is justified from a linguistic point of view, since time expressions are usually considered as phrases where the time-unit acts as the head of the phrase and the time-modifier acts as the modifier. Though the proposed framework is generic, the scope of the present study is limited to Spanish language.

A time-unit can be simple, if it is formed up from one type of time-units, or complex, if it is formed up from more than one unit (e.g. “el mes de agosto” (“the month of August”)).

- **Time-units** include:
  - *Time measurement units* (e.g. “hora” (“hour”), “minuto” (“minute”), “semana” (“week”).
  - *Deictic units* (e.g. “hoy” (“today”), “ayer” (“yesterday”), etc.
  - *Named units*: non-numeric (e.g. weekdays: “lunes” (“Monday”), months: “enero” (“January”), seasons: “invierno” (“Winter”) or numeric (e.g. “1998”, “12/10/2007”).

- **Modifiers**, according to their position in the expression, can be classified into *pre-modifiers* (e.g. “último” (“last’)) and *post-modifiers* (e.g. “después” (“after’)).

However, modifiers can also be classified according to their semantic content. For example: *ordinal-modifiers* (e.g. “primer” (“first’)), *frequency-modifiers* (e.g. “cada” (“each’), etc.

### 2.2 Resolution Perspective Classification

On the other hand and adopting another perspective which considers the resolution of the expression, temporal expressions can be classified into six types, according to the way they are defined:

- **Absolute temporal expressions**: are those that are completely defined by themselves. They don’t need another point in time to be a reference that allows their resolution, e.g. 25/10/2007’.

- **Relative temporal expressions**: they make reference to another point of time that is needed to know in order to be completely determined. In this case, resolution is not possible immediately, but certain previous calculus are required, e.g. “ayer” (“yesterday”). The reference date needed should be taken form the analyzed document: it can be obtained from the context (Reference Time), or it can be considered as the date of creation of the document (Creation Time).

- **Intervals**: time sets with two boundaries: date of init and date of end. In this way, intervals can be also considered absolute or relative, according to its boundaries, e.g. “desde mayo a junio” (“from May to June”). In our proposal, intervals will be considered as two temporal units joined by a connector.

- **Sets**: expressions referring to repeated events, they denote how often something happens, e.g. “cada día” (“every day”), “los lunes” (“Mondays”).

- **Durations**: expressions that indicate a period of time meaning how long something lasts, e.g. “durante dos meses” (“during two months”).

- **Named Dates**: expressions directly translatable, that correspond to a festivity, a holiday, etc., e.g. “el día de Navidad” (“Christmas Day”) = “25/12”.

### 3. Timexes Identification and Resolution Method

Some timexes occur with a higher frequency in the domain. They correspond with syntactic patterns whose generalization constitutes a guaranteed success percentage for the subsequent detection and normalization of all the expressions that accomplish them. Table 1 presents the most frequent temporal expressions in the training corpus.

| EXPRESSION | % occur. | freq. | #occur. |
|------------|----------|-------|---------|
| YYYY (e.g. “2007”) | 11.14% | 132 |
| YYYYMMDD (e.g. “20070527”) | 11.21% | 129 |
| “hoy” (“today”) | 7.65% | 88 |

Table 1 Examples of highest occurrence frequencies in training corpus

Once determined the most relevant patterns in training corpus, a grammar for automating the recognition task has been developed, as well as a proposal for the resolution and normalization of the maximum number of the temporal expressions detected.

Table 2 presents the description of a number of the most frequent patterns that accomplish the predicates of the recognition grammar whereas Table 3 presents some examples of resolution rules implemented, together with an example of their operation mode.

In Table 4 the components of the patterns are shown in detail.

To define the normalized output value the international standard ISO 8601 (2004) for representation of dates and times is used. It is based on the Gregorian calendar and on the 24-hour timekeeping system for representation of times. Both in dates and times representations the extended format is used. When a complete representation of a calendar date is needed, the extended format is YYYY-MM-DD, where YYYY stands for the year number, [MM] represents the calendar month and [DD] means a calendar day. When dealing with an expression of time the extended format to represent it is hh:mm:ss, where [hh] represents hours, [mm] minutes and [ss] seconds.
| PATTERN ID | INPUT FORMAT | RESOLUTION RULE | EXAMPLE |
|------------|--------------|----------------|---------|
| P01 BASIC_DATE | YYYY[-|/|][MM[-|/|][DD] | Day =DD Month=MM Year=YYYY | 20051231 | NA | 2005-12-31 |
| P02 BASIC_DATE_INV | DD[-|/|][MM[-|/|][YY][YY] | Day =toDD (DAY) Month=toMM(MONTH_NAME) Year=YYYY | [el] 31 de diciembre de 2005 | NA | 2005-12-31 |
| P03 REL_DEICTIC_UNIT | [ART|PREP]? | Day =getDD(Creation_Time) + 1 Month=getMM(Creation_Time) Year=getYYYY(Creation_Time) | mañana (tomorrow) | mañana (tomorrow) | 2005-12-31 | 2006-01-01 |
| P04 REL_POST_MODIF_DAY_ | N | Day =getDD(Reference_Time) - N Month=getMM(Reference_Time) Year=getYYYY(Reference_Time) | tres días antes (three days before) | 2004-10-15 | 2004-10-12 |
| P05 REL_TIME | [ART]? HOUR[;|:][H|h] Minute[;|:][M|m] [GMT]? | Day=UNDEFINED Month=UNDEFINED Year=UNDEFINED Hour=HOUR Minute=MINUTE | [las] 22:00 GMT | 2005-12-31 | XXXX-XX-XX 22:00 |
### 4. Experimentation and Results

#### 4.1 Corpora

The training corpus used for the present study is composed of a set of news in Spanish language (Newswire), containing several temporal expressions in each document. Its size is approximately 67000 words from 3 different sources.

For evaluation, the corpus used has roughly 54000 words, also taken from the same 3 sources. These are: AFP - Agence France-Presse, APW - Associated Press Worldstream, and XIN – Xinhua.

These corpora were originally developed for the TERN (Temporal Expression Recognition and Normalization) task for Spanish of the ACE 2007 evaluation, proposed by NIST (ACE, 2007), in which we took part (NIST, 2007; Vicente-Díez et al., 2007). All details of the corpora are shown in Table 5.

#### 4.2 Results

A sample of the results obtained is presented and discussed. In this stage, we have focused on expressions categories with high frequency of occurrence. Table 6 shows temporal expressions recognition and normalization results following the proposed method.

---

**Table 4 Pattern-components of recognition grammar**

| PATTERN COMPONENT | DESCRIPTION |
|-------------------|-------------|
| NUMERABLE         | WEEKDAY_NAME|MONTH_NAME|SEASON_NAME|TIME_MEASUREMENT_UNIT |
| WEEKDAY_NAME      | lunes|martes|…|domingo (Monday|Tuesday|…|Sunday) |
| MONTH_NAME        | enero|febrero|…|diciembre (January|February|…|December) |
| MONTH_NAME_SHORT  | ene|feb|…|dic (jan|feb|…|dec) |
| SEASON_NAME       | primavera|verano|otoño|invierno (Spring|Summer|Autumn|Winter) |
| TIME_MEASUREMENT_UNIT | año|mes|dia|hora|noche|siglo|centuria|minuto|segundo|decada|mañana|tarde| (year|month|day|hour|night|century|century|minute|second|decade|morning|evening|…) |
| DEICTIC_UNIT      | hoy|ahora|ayer|anoche|mañana|anteayer|anteanoche|pasado mañana (today|now|yesterday|last night|the day before yesterday|the night before last|the day after tomorrow) |
| TIME_MODIF        | PRE_MODIF|PRE_MODIF_ORDINAL|PRE_MODIF_FREQUENCY|POST_MODIF |
| PRE_MODIF         | pasado|ultimo|anterior|presente|proximo|posterior|siguiente|hace|hacia|dentro de |past|last|previous|present|next|later|following|ago|ago|in |
| PRE_MODIF_ORDINAL | primero|segundo|…|décimo (first|first|second|…|tenth) |
| PRE_MODIF_FREQUENCY | cada (each) |
| POST_MODIF        | pasado|ultimo|anterior|presente|proximo|posterior|siguiente|venidero|que |viene|antes|después |past|last|previous|present|next|later|following|future|in the future|before|after |
| CARDINAL          | CARDINAL_ALPH|CARDINAL_NUM |
| CARDINAL_ALPH     | uno|dos|tres|cuatro|… (one|two|three|four|…) |
| CARDINAL_NUM      | [1-9][0-9]? |
| ART               | ART_DEF|ART_INDEF |
| ART_DEF           | el|la|los|las (the) |
| ART_INDEF         | un|una|unos|unas (a|an) |
| DEMOS             | este|esta|astos|estas (this) |
| PREP              | a|ál|de|en |at the|of the|in |
| YYYY              | [0-9]? |
| MM                | [0-9]? |
| DD                | [0-9]? |
| MONTH             | [1-9] |
| DAY               | [1-9]|[1-2]|5-9|3|0-1 |
| TIME              | HOUR[:|H|h]MINUTE |
| HOUR              | [0-1][0-9]|2|0-3 |
| MINUTE            | [0-5][0-9] |

---

**Table 5 Corpora features description**

| Training corpus | Evaluation corpus |
|-----------------|-------------------|
| **reference**   | ACE 2007 Training V1.0 | ACE 2007 Evaluation Source Data V2.0 |
| **authors**     | (Walker, C. et al., 2006) | Linguistic Data Consortium (LDC), Philadelphia |
| **authors**     | (Walker, C. et al., 2007) | Linguistic Data Consortium (LDC), Philadelphia |
| **# files**     | 225 | 168 |
| **corpus size** | 484 KB | 395 KB |
| **# words**     | 67 K | 54 K |
| **dates of news** | January-April 2005 | June 2005 |
With a few patterns the number of timexes recognized exceeds an 82% of total in the evaluation corpus, while an 81% of detected timexes are correctly resolved and normalized applying the resolution rules previously described. Figures for false alarms (FA), errors (ERR) and missing (MISS) objects represent a low percentage of total detections.

| ID   | IDENT PATTERN                  | IDENT RESULTS | NORM RESULTS |
|------|--------------------------------|---------------|--------------|
|      | #DETEC                        | #NORM         |             |
|      | OK                            | OK            |             |
|      | %OVER                         | %OVER         |             |
|      | TOTAL CORPUS                  | TOTAL DETECT  |             |
| P01  | BASIC_DATE                    | 153           | 12,53        |
| P02  | BASIC_DATE_INV                | 11            | 0,90         |
| P03  | BASIC_DATE_TIME               | 14            | 1,15         |
| P04  | DAY_MONTH_NAME_SHORT          | 61            | 4,50         |
| P05  | MONTH_NAME_SHORT_DAY          | 15            | 1,23         |
| P07  | YEAR                          | 124           | 10,16        |
| P09  | DAY_MONTH                     | 35            | 2,87         |
| P15  | REL_DEICTIC__UNIT             | 85            | 7,78         |
| P16  | PRE_MODIF__NUMERABLE          | 60            | 6,55         |
| P17  | TIME_MEASUREMENT__UNIT_POST__ | 8             | 0,66         |
| P19  | PREP_DAY                      | 0             | 0            |
| P20  | PREP_MONTH_NAME               | 0             | 0            |
| P21  | PREP_YEAR                     | 0             | 0            |
| P22  | DIRECT__TRANSLATION           | 1             | 0,16         |

| TOTAL | TOTAL_CORPUS/                  | 1333          | 1212         |
|       | CORPUS/                          | 100           | 100          |
| FA    | PARTIALLY IDENTIFIED            | 110           | 8,25         |
| ERROR | MISPRINTS/INCORRECTLY NORMALIZED| 17            | 1,28         |
| MISS  | NOT IDENTIFIED/NORMALIZED       | 104           | 7,80         |
| TOTAL | CORRECTLY IDENTIFIED/           | 1102          | 81,19        |
| OK    | NORM                            | 82,7          | 984          |

Table 6 Results in evaluation corpus

5. Conclusions and Future Work

In this work, an empirical method of detecting and solving temporal expressions in Spanish Newswire is presented. Its evaluation shows promising results, with high figures obtained over the evaluation corpus.

Several aspects should be taken into account in future versions. First of all, the increasing of the number of the temporal expressions properly recognized through the completion of the recognition grammar specification, adding other patterns for expressions that are not currently considered.

Also resolution rules should be improved, adding treatment for repetitions (i.e: “cada día”, (“each day”)), vague expressions (i.e. “hace algunos días” (“days ago”)), etc.

In the same way, we consider a high-priority task the research about context extraction mechanisms that facilitate the resolution of relative temporal expressions.

Another aspect to be done is the implementation of dictionaries with a broader coverage of directly translatable temporal expressions, such as party days, festivities, etc. (i.e. “día de la Madre” (“Mother’s Day”), traditionally celebrated the 1st Sunday of May).

Finally, we propose the introduction of machine learning techniques in future versions, expecting that the performance of the identification of timexes (Ahn et al., 2005), working with different and heterogeneous source documents, was increased.

6. Acknowledgements

This work has been partially supported by the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267), and by the Spanish Ministry of Education under the project BRAVO (TIN2007-67407-C03-01).

7. References

ACE (2007). The ACE 2007 (ACE07) Evaluation Plan. National Institute of Standards and Technology, Information Technology Laboratory – Information Access Division (IAD).

Ahn, D., Fissaha, S. and de Rijke, M. (2005). Extracting Temporal Information from Open Domain Text: A Comparative Exploration. J. Digital Information Management, 3(1), pp. 14–20.

ISO 8601 (2004). Data elements and interchange formats –Information interchange - Representation of dates and times.

Mani, I. and Wilson, G. (2000). Robust Temporal Processing of News. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, pp. 69–76.

Martínez-Barco, P., Saquete, E., and Muñoz, R. (2002). A Grammar-Based System to Solve Temporal Expressions in Spanish Texts. In ForTAL’02:
MITRE Corporation (2007). TimeBank.
http://www.cs.brandeis.edu/~jamesp/arda/time/timebank.html

NIST (2007). National Institute of Standards and Technology 2007 Automatic Content Extraction Evaluation Official Results (ACE07) v.2.
http://www.nist.gov/speech/tests/ace/ace07/doc/ace07_eval_official_results_20070402.htm

Pustejovsky, J., Castaño, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A. and Katz, G. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. In Proceedings of the IWCS-5 Fifth International Workshop on Computational Semantics.

Saquete, E., Martinez-Barco, P., Muñoz, R., and Vicedo, J.L. (2004). Splitting Complex Temporal Questions for Question Answering Systems. In ACL’2004: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Morristown, NJ, USA.

Saquete, E., Martinez-Barco, P., Muñoz, R., Negri, M., Speranza, M. and Sprugnoli, R. (2006). Multilingual Extension of a Temporal Expression Normalizer using annotated corpora. In Proceedings of the Workshop Cross-language Knowledge Induction at EACL 2006.

Saquete, E., Muñoz, R. and Martinez-Barco, P. (2006b). Event ordering using TERSEO system. Data & Knowledge Engineering, vol. 58 (1), pp. 70-89.

Vicente-Diez, M.T., de Pablo-Sánchez, C. and Martínez, P. (2007). Evaluación de un Sistema de Reconocimiento y Normalización de Expresiones Temporales en Español. Procesamiento de Lenguaje Natural, vol. 39, pp. 113-120.