# Multilingual Extension of Temporal Expression Recognition using Parallel Corpora

**M. Puchol-Blasco E. Saquete P. Martínez-Barco**
Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
{marcel,stela,patricio}@dlsi.ua.es

## Abstract

This paper presents the automatic extension of TERSEO to other languages, a knowledge-based system for the recognition and normalization of temporal expressions, originally developed for Spanish.

TERSEO was extended to English and Italian through the automatic translation of the temporal expressions, and it was presented in previous works (see Saquete et al. (2004a)), but a new methodology has been designed with the purpose of obtaining better results in this issue.

This new methodology is based on the use of parallel corpora for extending the TERSEO temporal model to other languages. In this case, two different methods have been tested: (1) automatic translation of TERSEO patterns to other languages and (2) automatic corpora annotation in the target side of parallel corpora. The main idea is focused on annotating the Spanish side of a parallel corpora, projecting the analysis to the second language, and then obtaining new TERSEO patterns (1) and new annotated corpus (2).

The set of new patterns will be used to improve the current TERSEO language independent modules. Whereas the new annotated corpus will be used to train a ML system. This system will annotate new temporal expressions in the new language.

## 1 Introduction

Recently, the Natural Language Processing community has focused its interest on developing language independent systems, and what has been demonstrated by the growing number of international conferences and initiatives placing systems with multilingual/cross-language capabilities among the hottest research topics, such as the European Cross-Language Evaluation Forum[1] (CLEF).

On the other hand, regarding temporal reasoning, also systems featuring multilingual capabilities have been proposed. Among others, Moia (2001) emphasized the potentialities of such applications for different information retrieval related tasks.

As many other NLP areas, research in automated temporal reasoning has recently seen the emergence of machine learning approaches trying to overcome the difficulties of extending a language model to other languages (Carpenter, 2004; Ittycheriah et al., 2003). In this direction, the outcomes of the first Time Expression Recognition and Normalization Workshop (TERN 2004[2]) provide a clear indication of the state of the field. In spite of the good results obtained in the *recognition* task, *normalization* by means of machine learning techniques still shows relatively poor results with respect to rule-based approaches, and it remains an unresolved problem.

The porting process of systems to new languages (or domains) in the case of rule-based approaches is a very costly and time-consuming task due to the requirement of rewriting a large number of rules from scratch, or adapting them to each new language (Schilder and Habel, 2001; Filatova and Hovy, 2001). On the other hand, machine learning approaches (Katz and Arosio, 2001) can be extended with little human intervention through the use of language corpora. However, the large

[1] *http://www.clef-campaign.org/* last visited on 10/02/07
[2] *http://timex2.mitre.org/tern.html* last visited on 10/02/07

annotated corpora necessary to obtain high performance are not always available.

In this paper we describe a new procedure to build temporal models for new languages, starting from previously defined ones and using parallel corpora as a resource. While still adhering to the rule-based paradigm, its main contribution is the proposal of a methodology to automate the porting of a system from one language to another. In this procedure, we take advantage of the architecture of an existing system developed for Spanish (TERSEO, see (Saquete et al., 2005)), where the recognition model is language-dependent but the normalizing procedure is completely independent. In this way, the approach, using parallel corpora and alignment information, is capable of automatically learning the recognition model for a new language by adjusting the set of normalization patterns. Moreover, our approach is capable of automatically obtaining annotated corpus with temporal expression information that will be used to train Machine Learning approaches.

## 2 Extending TERSEO: from Spanish to other languages

TERSEO has been developed in order to automatically recognize temporal expressions (TEs) appearing in a Spanish written text, and normalize or resolve them according to the temporal model proposed in Saquete (2005), which is compatible with the ACE annotation standards for temporal expressions (Ferro et al., 2005).

For more information about TERSEO system architecture see Saquete (2005) and Saquete et al. (2004a).

The main purpose of this paper is to describe a new procedure to automatically build temporal models for new languages, starting from this previously defined model for Spanish.

In previous works, this issue was treated: an English model was automatically obtained from the Spanish one through the automatic translation of the Spanish temporal expressions to English. The resulting system for the recognition and normalization of English TEs obtained good results both in terms of precision (P) and recall (R) (Saquete et al., 2004b). Later, another system extension was made for obtaining an Italian temporal model from Spanish model (see Saquete et al. (2004a)). In this case, worse results were obtained compared with the English model, due to the lack of resources. With this new approach, our main aim is to obtain better results than previous approaches.

This section presents the procedure we followed to extend our system to other languages, starting from the Spanish model already available, and parallel corpora that is sentence aligned. The Spanish model was manually obtained and evaluated showing high scores for precision (88%).

The method presented here changes completely the view of extending TERSEO system, compared to the ones used until this moment, since it is based on the use of aligned parallel corpora from Spanish to other languages. This methodology has been previously used in other NLP research areas, such as WSD (see Gliozzo et al. (2005)), or Magnini and Strapparava (2000)), although it had not been used before in the multilingual extension of systems that recognize and normalize temporal expressions.

Two different methods are used with parallel corpora: (1) automatic translation of TERSEO patterns for the target side language and (2) automatic corpora annotation in the target corpus. The main idea is to annotate the Spanish side of parallel corpora, projecting the analysis to the second language, and then obtaining new TERSEO patterns (1) and new annotated corpus (2).

Obtaining a temporal model in the target language consists of using the initial temporal model used in TERSEO (a set of patterns and normalization rules that are necessary for temporal expression recognition and normalization) and translating it to other languages using a parallel corpora and alignment systems.

Obtaining a tagged corpus in the target language is not an innovating task. Lately, the NLP area has been interested in the treatment of parallel corpora for the automatic tagging of corpus in a target language. The paper of Rebecca Hwa et al. (2005) can be taken as an example. These authors deal with the problem of syntactic annotation of corpus in a language different to English. In order to solve it, they used a good English parser to annotate the source language of an aligned parallel corpora. After this, they projected the English annotations with alignment information and obtained the target side of the parallel corpora annotated with syntactic information.

Analogous to this concept, we wanted to extrapolate this problem to the recognition of the temporal expressions. Nowadays, there are a lot of lan-

guages lacking corpus tagged with temporal information. A possibility of overcoming this problem represents the method proposed here.

As shown in Figure 1, for performing the tasks of target corpus generation and patterns translation, some input information is required: 1) TERSEO information, that consists of processing the Spanish side of the parallel corpora with a POSTagger, the Recognition module, the Argument Detection and the Parameters Validation. 2) Alignment Information that is obtained from the Token Alignment and 3) the POStagging of the non-Spanish side of the parallel corpora. Besides, for patterns translation, another input is required: TERSEO Spanish patterns.

TERSEO modules used in this process are the same as the ones used by the initial system. The other modules will be explained thoroughly in following sections.

### 2.1 Token Aligner

Token alignment is necessary to obtain the alignment information used on Patterns Translation and Target Corpus Generation modules for projecting a temporal expression from the Spanish side of the parallel corpus to the non-Spanish side.

The alignment of parallel corpora at token level is an important subject in current research. Authors as Franz Josef (Och and Ney, 2003) work in this area of research, contributing to the scientific community with tools like GIZA++, which will be used in our future evaluation of this method.

### 2.2 Patterns Translation

The patterns translation process consists of obtaining each temporal expression in the source parallel corpora language (Spanish) with TERSEO, and tagging them with a category and some argument information. Using alignment information, a projection between that information can be made from the Spanish side of parallel corpora to the non-Spanish side. Due to the fact that patterns were used for tagging the Spanish side, the projection gives us information to translate those patterns to the non-Spanish side language.

TERSEO has a set of patterns that can be applied to the different modules. These patterns have four different types of elements: (1) single words (e.g. 'el' 'día' 'siguiente'), (2) elements with lexical and morphological information (e.g. 'días' ADV-T), (3) elements described by regular expressions (e.g. NUM→ [0-9]+), and (4)

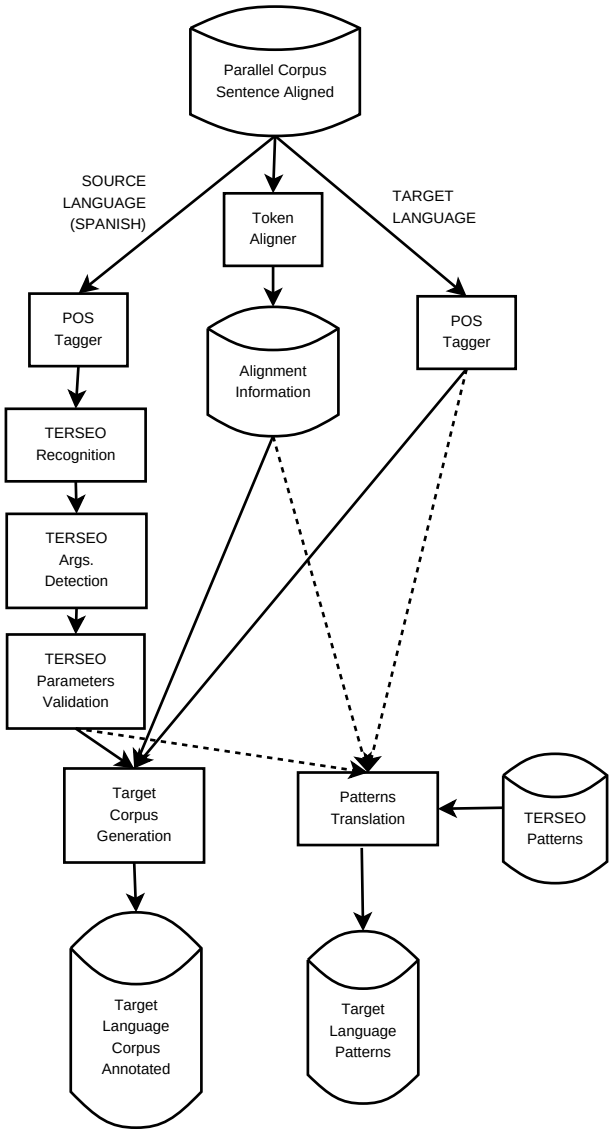


Figure 1: Multilingual extension procedure.

dictionary entries (e.g. DIA_SEMANA→'lunes' 'martes' 'miércoles'...).

The translation of all these elements will be performed as follows:

1. **Single words:** patterns that contain single words in temporal expressions will be translated into words of the aligned target language.

2. **Elements with lexical and morphological information:** lexical and morphological information of the source language will be translated into the one obtained in the target language.

3. **Regular expressions:** as regular expressions correspond to invariable language elements

(e.g. numbers), the regular expressions are not translated.

4. **Dictionary entries:** the dictionary entries will be translated element by element, as if they were single words. As these pattern types are indispensable for TERSEO recognition, when an element of the dictionary entry is not contained in the source language corpus, bilingual dictionaries will be used. Furthermore, these entries usually are not ambiguous, and good performance can be obtained with bilingual dictionaries.

It is important to consider the variations of the positions of elements that exist between different languages. Positions will be considered in the target pattern, guided by the alignment information. Also, it is possible that some element of the initial patterns does not correspond with any element of the target language. In such case, this element will be removed from the target pattern.

The target patterns will correspond directly to the source patterns, therefore, if a categorization pattern has the ID X, the target pattern will have the same identifier. We base this conclusion on the Direct Categorization Correspondence, that will be explained next. Using the same idea, we estimate that there is a direct relation between the patterns of source argument detection with those of the target.

**Direct Categorization Correspondence (DCC):** given two temporal expressions X and Y, and having the premise that TE normalization is language independent, it can be assumed that $Category_Y = Category_X$, due to the fact that categorization is intimately related with the resolution of the temporal expression.

Figure 2 can be seen as an example. In a first moment, there are two Spanish patterns: a dictionary entry and another one that contains single words, and we want to translate these patterns to English. In a first step, TERSEO returns that 'el lunes siguiente' (sentence contained in the source language corpus) is a temporal expression and it is categorized with the pattern category ID 1. Besides, TERSEO obtains that 'lunes' is an argument with the value 'DIA_SEM'. Once the Token Aligner returns the Alignment Information, it will be obtained that 'el lunes siguiente' is equivalent to 'next Monday' in the way that the Figure shows. Following these steps, the translated patterns will be finally achieved.

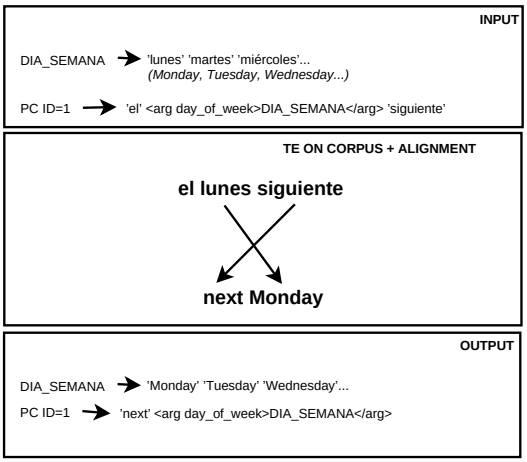


Figure 2: Patterns translation example.

Once the target patterns have been obtained, TERSEO can perform using them and therefore, identifying and normalizing temporal expression in target language.

### 2.3 Target Corpus Generation

Once we have found and categorized TEs and the arguments inside them in the Spanish side of the parallel corpora, alignment information is taken for projecting TE information into the non-Spanish side of the parallel corpora.

The tagging of the non-Spanish side is based on four features: the two features previously obtained (noun and POS) and two new features extrapolated from the Spanish side parallel corpora. The first new feature is information about the time expression (including TEs boundaries and categorization). The second new feature is information about arguments contained into time expressions (including the boundaries).

The output corpus is tagged following this set of rules:

1. **Format:** each line on the target corpus corresponds to a token and its features (one feature per column). A blank line will be the sentence separator.

2. **Time Expression:** when a TE is found in the source language, it is projected to the target language corpus as follows: alignment information about the first element position on TE will be used for tagging the TE beginning in target corpus. The tag associated to TE beginnings is '(TE TE_ID*' (TE_ID corresponds to a temporal expression category).

Alignment information about the last element position on TE will be used for tagging the TE ends in target corpus. The tag associated to TE ends is '*)'. It is possible to have a temporal expression on just one token. In this case, the tag for TE column will be '(TE TE_ID*)'. Other elements in TE column will be tagged with '*'. As can be seen, temporal expression in target language will be contained between '(TE*' and '*)' on TE column.

3. **Arguments:** when an argument is found in the source language, it is projected to the target language corpus as follows: alignment information about the first element position on the argument will be used for tagging the argument beginning in the target corpus. The tag associated to the argument beginnings is '(ARG ARG_ID*' (ARG_ID corresponds to ID argument assigned in the Arguments Detection module). Alignment information about the last element position on the argument will be used for tagging the argument ends in the target corpus. If beginning and end argument position are the same, the tag assigned on the argument column will be '(ARG ARG_ID*)'. Other elements in argument column will be tagged with '*'. As in Time Expression, an argument will be contained between '(ARG*' and '*)'.

Once all TE in the Spanish side of the parallel corpora have been translated to the non-Spanish side, the corpus tagged with TE can been used for training Machine Learning systems. Once the training is performed, the new TE recognition system based on Machine Learning is ready to obtain categorization and arguments for unseen new sentences in target language. After this, it is only necessary to apply TERSEO TE Resolution module for temporal expression Normalization.

An example for this process is shown in Figure 3. In this example the temporal expressions in the English side of the parallel corpora wants to be obtained. The first step is recognizing temporal expressions on the Spanish corpus with TERSEO. In this example, the sentence 'El lunes siguiente iremos a la ciudad' ('*We will go to the city next Monday*' in the English side of the parallel corpora) has been found. Somultaneously, alignment information is obtained. As we only need alignment information about elements contained in temporal expressions, only this information type is shown in Figure 3. Therefore, only alignment information about 'el lunes siguiente' must be known. Once all previous rules have been applied, the output corpus corresponds to the elements at the bottom of Figure 3.

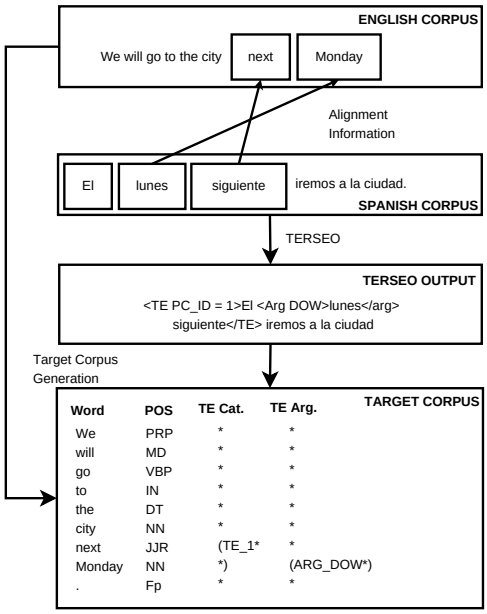


Figure 3: Target corpus generation example.

## 3 Conclusions and Further Work

In this paper we have presented an automatic extension of a rule-based approach to TEs recognition and normalization. The procedure is based on building temporal models for new languages starting from previously defined ones. This procedure is able to fill the gap left by machine learning systems that, up to date, are still far from providing acceptable performance on the normalization task.

Two methods for extending a temporal model to other languages have been presented: one based on pattern translation and another based on automatic corpus annotation that will be used later in ML systems. Both methods are based on using parallel corpora from Spanish to other languages.

For further work, our first task will be evaluating the proposed system, comparing the knowledge-based extension vs. the parallel corpora-based extension. Several token aligner tools will be used, comparing results between them. If good results are obtained, other languages will be treated.

## 4 Acknowledgments

This work has been suported by the Generalitat Valenciana throught the research grant BFPI06/182 and the project GV06/028 (Tratamiento bilingüe Valenciano-Castellano de preguntas temporales complejas en los sistemas de búsqueda de respuestas), the Spanish Ministery of Science and Technology (project TIN2006-15265-C06-01: TEXT-MESS - Knowledge discovery and Representation in Human Language Technology) and the European Union (project FP6-IST-2005-33860: QALL-ME - Question answering learning technologies in a multilingual and multimodal environment).

## References

B. Carpenter. 2004. Phrasal Queries with LingPipe and Lucene. In *13th Text REtrieval Conference*, NIST Special Publication. National Institute of Standards and Technology.

L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. 2005. Tides.2005 standard for the annotation of temporal expressions. Technical report, MITRE.

E. Filatova and E. Hovy. 2001. Assigning time-stamps to event-clauses. In ACL, editor, *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, pages 88–95, Toulouse,France.

A. Massimiliano Gliozzo, M. Ranieri, and C. Strapparava. 2005. Crossing parallel corpora and multilingual lexical databases for wsd. In *CICLing*, pages 242–245.

R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3):311–325.

A. Ittycheriah, L.V. Lita, N. Kambhatla, N. Nicolov, S. Roukos, and M. Stys. 2003. Identifying and Tracking Entity Mentions in a Maximum Entropy Framework. In ACL, editor, *Proceedings of the NorthAmerican Chapter Association for Computational Linguistic (NAACL) Workshop WordNet and Other Lexical Resources: Applications, and Customizations*.

G. Katz and F. Arosio. 2001. The annotation of temporal information in natural language sentences. In ACL, editor, *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, pages 104–111, Toulouse,France.

B. Magnini and C. Strapparava. 2000. Experiments in word domain disambiguation for parallel texts.

T. Moia. 2001. Telling apart temporal locating adverbials and time-denoting expressions. In ACL, editor, *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, Toulouse,France.

F. Josef Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

E. Saquete, P. Martínez-Barco, and R. Mu noz. 2004a. Automatic multilinguality for time expression resolution. In R. Monroy, G. Arroyo-Figueroa, L. Enrique Sucar, and J. Humberto Sossa Azuela, editors, *MICAI*, volume 2972 of *Lecture Notes in Artificial Intelligence*, pages 458–467. Springer.

E. Saquete, P. Martínez-Barco, and R. Mu noz. 2004b. Evaluation of the automatic multilinguality for time expression resolution. In *DEXA Workshops*, pages 25–30. IEEE Computer Society.

E. Saquete, R. Mu noz, and P. Martínez-Barco. 2005. Event ordering using terseo system. *Data and Knowledge Engineering Journal*, page (To be published).

E. Saquete. 2005. *Temporal information Resolution and its application to Temporal Question Answering*. Phd, Departamento de Lenguages y Sistemas Informáticos. Universidad de Alicante, June.

F. Schilder and C. Habel. 2001. From temporal expressions to temporal information: Semantic tagging of news messages. In ACL, editor, *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, pages 65–72, Toulouse,France.