# Inferring the semantics of temporal prepositions in Italian

**Tommaso Caselli** **Valeria Quochi**

ILC-CNR
Via Moruzzi, 1 56123, Pisa, Italy
Dip. Linguistica "T.Bolelli", Università degli Studi di Pisa
Via S.ta Maria, 36, 56100, Pisa, Italy
tommaso.caselli,valeria.quochi@ilc.cnr.it

## Abstract

In this work we report on the results of a preliminary corpus study of Italian on the semantics of temporal prepositions, which is part of a wider project on the automatic recognition of temporal relations. The corpus data collected supports our hypothesis that each temporal preposition can be associated with one prototypical temporal relation, and that deviations from the prototype can be explained as determined by the occurrence of different semantic patterns. The motivation behind this approach is to improve methods for temporal annotation of texts for content based access to information. The corpus study described in this paper led to the development of a preliminary set of heuristics for automatic annotation of temporal relations in text/discourse.

## 1 Introduction

In this work we report on the preliminary results of a corpus study, of contemporary Italian, on temporal relations that hold between a temporal adjunct and an event as a way to determine the semantics of temporal prepositions. We claim, following Schilder and Habel (2001), that the semantics of temporal prepositions is *rel (e, t)*, where *rel* is used to indicate the temporal relation associated with a certain preposition, *t* represents the meaning of the Temporal Expression (timex), and *e* the meaning of the event description involved.

Prepositions introducing a temporal adjunct are explicit signals of temporal relations. The ability to determine temporal relations between timexes introduced by prepositions and events is fundamental for several NLP tasks like Open-Domain Question-Answering systems (Hartrumpf et al. 2006, and Pustejovsky et al. 2002) and for Textual Entailment and Reasoning.

The corpus data collected seems to support our hypothesis that each temporal preposition can be associated with one prototypical temporal relation, and that deviations from the prototype can be explained as determined the occurrences of different semantic pattern.

The work described in this paper is part of a larger project we are conducting on temporal discourse processing in Italian, as proposed in Mani and Pustejovsky (2004).

## 2 Background

This section presents a brief overview of the TimeML specification language (Pustejovsky et al. 2005), which has been used as the starting point for this work, and some theoretical issues on Italian prepositions.

### 2.1 TimeML

The TimeML specification language (Pustejovsky et al. 2005) offers a guideline for annotation of timexes, events and their relations. Like other annotation schemes[1], TimeML keeps separated temporal expressions and events, tagged, respectively, with **TIMEX3** and **EVENT**. In addition, two other tags are used: **SIGNAL** and **LINK**.

The **EVENT** tag is used to annotate events, defined as something which occur or happen, and

[1] Filatova and Hovy (2001), Schilder and Habel (2001), Setzer (2001).



*Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 38–44,
Prague, Czech Republic, June 2007. 

states, defined as situations in which something holds true.

Temporal expressions, or timexes, like day times (*noon*, *the evening*, *1p.m*…), dates of different granularity (*yesterday, February 2 2007, last week, last spring, last centuries*…), durations (*five hours, in recent years*…) and sets (*twice a day*…), are annotated with the **TIMEX3** tag. This tag is based on specifications given by Ferro et al. (2001) and Setzer (2001). Each timex is assigned to one of the following types: DATE, for calendar times, TIME, for times of the day, even if indefinites (e.g. 'the evening'), DURATION, for timexes expressing a duration, and SET, for sets of times. Each timex is further assigned a value, according to the ISO 8601 specifications (for instance, *3 anni* '3 years' is normalized as "P3Y", i.e. a "period of 3 years").

Function words which explicitly signal a relation between two elements (timex and event, timex and timex, or event and event) are tagged with **SIGNAL**.

Finally, the **LINK** tag is used to specify the relation between two entities. It may indicate a temporal relation (**TLINK**), a subordinating relation (**SLINK**) or an aspectual relation (**ALINK**). The **TLINK** tag, which is pivotal for the present work, comprises 15 relations, only 13 of which are purely temporal. The 13 relations can be seen as derived from Allen's (1984) temporal logic, and 6 of them are binary relations - one being the inverse of the other. These relations (`simultaneous`, `includes`, `is_included`, `during`, `inv_during`, `begin`, `end`, `begun_by`, `ended_by`, `before`, `after`) make explicit the temporal relation holding between two elements.

### 2.2 Temporal PPs in Italian

Italian prepositions can be divided into two main groups: monosyllabic like *a*, *da*, *in*, *per*, *tra*, -and polysyllabic ones like *fino a* 'up to', *dopo* 'after',, *prima* 'before'…This difference at a surface level reflects a difference also at a semantic level: monosyllabic prepositions are either semantically empty elements (i.e. when they are particles preselected by the VP), or they bear a very abstract relational meaning, which gets specialized on the basis of the co-text; polysyllabic prepositions, on the other hand, have a more specific meaning of their own. For instance, the preposition *dopo* 'after' always means "subsequently, afterwards", disregarding its co-text; which makes the identification of the relation between the elements involved an easier task. In addition to this, most prepositions, both polysyllabic and monosyllabic, belong to different semantic fields, e.g. spatial, temporal, manner or other.

For the purpose of this work, any preposition followed by a timex, as defined in TimeML (Section 2.1), is considered a temporal preposition. Consequently, we will speak of Temporal PP for any sequence of the form "preposition + timex".

In Italian, as in many other languages, the form that Temporal PPs, or temporal adjuncts, may take is influenced by the aspect and actionality of the VP. In traditional grammars, for instance, it is claimed that they can be introduced by *in* if the lexical aspect denotes a telic event (e.g. (1)) and by *per* if the lexical aspect denotes a process or a particular subclass of telic events, i.e. achievements (e.g. (2)). Moreover, these kinds of Temporal PPs necessarily refer to the conclusion of the process denoted by the events and thus are incompatible with the progressive aspect:

1) *a.* Maria ha pulito la stanza *in mezz'ora*.
   [Maria cleaned the room in half an hour]
   *b.* La pizza arriva *in cinque minuti*.
   [The pizza will arrive in five minutes]
2) *a.* Marco ha lavorato *per due ore*.
   [Marco has worked for two hours]
   *b.* Marco mi prestò il libro *per due giorni*.
   [Marco lend me his book for two days]

The influence of the aspect and actionality of the VP has an impact also in the identification of their meaning. In particular, in example 1) *a.* the preposition signals that the event of cleaning the room lasted for half an hour, while in the example 1) *b.* the event of arriving takes place after five minutes from the utterance time. In example 1), thus, the same Temporal PP, i.e. IN + timex, has two different meanings, signalled by the relations `includes` and `after`. The different temporal relations are determined by two different semantic patterns: [DURATIVE_Verb] + *in* + [TIMEX type: DURATION] for 1) *a*, and [TELIC_Verb] + *in* + [TIMEX type: DURATION], for 1) *b*.

## 3 The corpus study

In order to verify our hypothesis that the most frequent temporal relations represents the prototypical meaning of a temporal preposition[2], a corpus study has been conducted. It is important to note that we do not refer to frequency *tout court*, but is frequency with respect to a certain semantic pattern. Since we want to develop a system for automatic annotation of temporal relations, a 5 million word syntactically shallow parsed corpus of contemporary Italian, drawn from the PAROLE corpus, has been used[3].

All occurrences of a prepositional chunk with their left contexts has then been automatically extracted and imported into a database structure using a dedicated *chunkanalyser* tool[4]. This automatically generated DB was then augmented with ontological information from the SIMPLE Ontology, by associating the head noun of each prepositional chunk to its ontological type, and has been queried in order to extract all instances of Temporal PPs, by restricting the nouns headed by prepositions to the type "TIME", which is defined in SIMPLE as "all nouns referring to temporal expressions" (SIMPLE Deliverable 2.1: 245).

To identify the meaning of temporal prepositions, therefore, we considered sequences of the form:

*Fin Vb Chunk + Prep Chunk: semtype= TIME*

where *Fin Vb Chunk* is a shallow syntactic constituent headed by a finite verb and corresponds to the "anchoring" event, and *Prep Chunk* is the prepositional phrase that represents an instance of a timex. To get a more complete picture of the distribution of Temporal PPs in text, we extracted sequences from zero up to a maximum of two intervening chunks, obtaining a set of about 14,000 such sequences.

A first observation is about the distribution of the Temporal PPs. As illustrated in Table 1 (below) Temporal PPs tend to occur immediately after the event they are linked to.

[2] We assume and extend Haspelmath's (forth.) proposal on the explanatory and predictive power of frequency of use.

[3] The corpus was parsed with the CHUNK-IT shallow parser (Lenci et al. 2003).

[4] By courtesy of Ing. E. Chiavaccini.

| Sequence | Distance | # Occurrences |
|---|---|---|
| Fin_Vb + PP (Time) | 0 | 5859 |
| Fin_Vb + PP (Time) | 1 | 4592 |
| Fin_Vb + PP (Time) | 2 | 3677 |

Table 1. Occurrences of Temporal PPs with respect to the distance from the event.

The data in Table 1 show that Temporal PPs have a behavior similar to modifiers, like adjectives anchoring on the time axis of the event they refer to.

### 3.1 Annotating Temporal Relations

To identify the semantics of temporal prepositions, a subcorpus of 1057 sequences of *Fin Vb Chunk + Prep Chunks (Time)* was manually annotated by one investigator with temporal relations in a bottom-up approach.

The tags used for the temporal relation annotation were taken from the TimeML **TLINK** values (see Section 2.1). This will restrict the set of possible relations to a finite set. To ease the task, we excluded the inverse relations for `includes`, `during`, `begin`, and `end`. In order to understand the role of the co-text, we also marked the types of timexes according to the TimeML **TIMEX3** tag (ibid.). In this annotation experiment we did not consider information from the VP because it will be relevant to explain the deviations from the prototype.

. To facilitate the assignment of the right temporal relation, we have used paraphrase tests. All the paraphrases used have the same scheme, based on the formula *rel (e, t)*, illustrated in the 3):

3) *The event/state of X is R timex.*

where X stands for the event identified by the *Fin Vb Chunk*, R is the set of temporal relations and timex is the temporal expression of the Temporal PP. This means that the sequence in 4):

4) [[$_{\mathrm{Vfin}}$[Sono stato sposato] [[ $_{\mathrm{PP}}$[*per* quatto anni]]
'I have been married for four years'

can be paraphrased as 5):

5) The state of "being married" happened *during* four years.

The only temporal relation that is not paraphrased in this way is `simultaneous`, which corresponds to 6):

6) *The event/state X HAPPENS(-ED) AT timex.*

## 4 Results

Among the 1057 sequences in our sub-corpus, we found that only 37.46% (for a total of 449 excerpts) where real of instances of Temporal PPs, the others being either false positives or complex timexes, i.e. timexes realized by a sequence of a NP followed by a PP introduced by "*di*" (of), as in the following example:

7) [$_{NP}$[la notte]] [$_{PP}$[di Natale]
'the Christmas night'

In Table 2 (below) we report the temporal prepositions identified in the corpus:

| Temporal Preposition | # occurrences |
|---|---|
| In 'in' | 91 |
| A 'at/on' | 64 |
| Da 'from/since/for' | 37 |
| Dopo 'after' | 1 |
| Attraverso 'through' | 1 |
| Di 'of' | 43 |
| Durante 'during' | 5 |
| Entro 'by' | 9 |
| Fino a 'up to' | 6 |
| Fino da 'since' | 3 |
| Oltre 'beyond' | 1 |
| Per 'for' | 50 |
| Tra 'in' | 3 |
| Verso 'towards' | 1 |

Table 2. Instances of temporal prepositions in the corpus.

The relative low number of real Temporal PPs can negatively influence the analysis and the identification of the semantics of the temporal prepositions. In order to verify whether the data collected could represent a solid and consistent baseline for further analysis, we analysed all instances of false positive timexes. With the exception of a few cases, which could have been easily recognized by means of a Timex Grammar, we found out that 482/608 instances are represented by nouns which have some sort of temporal value but whose assignment to the semantic type "Time" in the Ontology do not correspond to the given definition (Section 3), e.g: *colazione* 'breakfast', *scuola* 'school', *presidenza* 'presidency', and many others.

Therefore, we performed a new extraction of sequences excluding all instances of false positives. The new results are very different since more than 56.03% of all prepositional chunks are Temporal PPs. This provides support to the fact that the sequences extracted from the sub-corpus, though small in number, can be considered as a consistent starting point for identifying the semantics of temporal prepositions. In particular, the prepositions presented in Table 2 correspond to the most frequent prepositions which give rise to temporal relations between timexes and events. Though small, the 449 sequences prove to be reliable: we have identified a total of 320 temporal relations, as illustrated in Table 3:

| Temporal Relation | # occurrences |
|---|---|
| Includes | 87 |
| During | 72 |
| Before | 11 |
| After | 11 |
| Imm_before | 1 |
| Imm_after | 2 |
| Simultaneous | 5 |
| Beginning | 52 |
| Ending | 10 |
| No Temporal Link | 60 |
| No Assigned | 9 |

Table 3. Kinds of Temporal Relation Identified.

## 5 Inferring Preposition Semantics

The analysis we propose for each single preposition provides information on its semantics. Such information is obtained on the basis of the frequency[5] with which a given temporal relation is associated or coded by that preposition. We claim, as already stated, that temporal relations coded by prepositions are signals of a certain semantic pattern. Different temporal relations coded by the same preposition signal different semantic pattern. According to the frequency with which a temporal relation, or a semantic pattern, occurs, it is considered either as the prototypical (i.e. most frequent) meaning or as a deviation from the norm, whose

[5] Note that what counts is relative frequencies, and not absolute frequencies.

explanation relies in the analysis of the semantic pattern in which it occurs. It is for this reason that a major role in this analysis is played by the types of timexes which follow the preposition. Keeping track of their types, according to the TimeML classification (Section 2.1), is very useful mainly for cases where the same temporal preposition codes different temporal relations depending on the type of the timex by which it is followed. In other words, it is a way to assess the semantic pattern which has been used to code that meaning. In the following sections we will focus on the semantics of the most frequent temporal prepositions, that is *in* 'in', *a* 'at, on', *per* 'for'[6], *da* 'for, since, from'. Cases of low frequency temporal relations are not analyzed here because they would require both more data and a separate investigation.

### 5.1 Prepositions *per* and *da*

These two prepositions, although they encode different temporal relations, are presented in a unique subsection due to their extremely similar coherent distribution across temporal relations. In particular, the 80% (40/50) of *per* identifies a DURING temporal relation, and 83.78% (31/37) of *da* identifies a BEGIN temporal relation.

From these data, we can represent the semantics of *per* as follows:

8) **λ(e, λ(t, DURING (e, t))**

and that of *da* as:

**9) λ(e, λ(t, BEGIN (e,t))**

### 5.2 The Preposition *in*

The preposition *in* is by far the most used temporal preposition. In our corpus there are 91 occurrences of this preposition, distributed as follows:

INCLUDES (57/91: 62.63%)
DURING (19/91: 20.87%)
AFTER (6/91: 6.59%)
BEGIN (3/91: 3.29%)
SIMULTANEOUS (2/91: 2.19%)
No LINK (2/91: 2.19%)
END (1/91: 1.09%)

Following our idea that the most frequent relation represents the prototypical meaning of the preposition; we claim that Temporal PPs introduced by *in* tend to code a relation of inclusion, semantically represented as:

10) **λ(e, λ(t, INCLUDES( e, t)).**

Since this preposition is not exclusively used with this meaning, the data forces us to provide an explanation for the other relations identified, in particular for DURING, AFTER and BEGIN.

Considering the DURING relation, we analyzed the types of timexes governed by the preposition but found that type distinctions did not help. Nevertheless, we observed a clearcut regularity analysing the normalized values of the timexes involved: we found that, whenever the timexes are definite quantified intervals of time (e.g. *2 days, 3 years, half an hour*) or temporally anchored instants, *in* encodes the temporal relation of DURING, thus deviating from the default interpretation represented in 10).

The relation AFTER shares with DURING the restriction on the normalized values of the timexes. However, for the AFTER relation there is a strong contribution from the VP, as claimed in traditional grammars. In such cases, it is the actionality of the VP that forces the interpretation of *in* to express the AFTER relation. In fact, this relation appears to occur only with achievement verbs, which inherently focus on the *telos* – or ending point (see example 1) *b* Section 1).

Finally, the BEGIN relation can be found only with aspectual verbs, e.g. *iniziare* 'begin' or *riprendere* 'resume'. In these cases the preposition does not really work as a temporal preposition, but more as a particle selected by the verb.

### 5.3 The Preposition *a*

The preposition *a* presents a non-trivial distribution, which makes it difficult to identify a prototypical value:

INCLUDES (20/64: 31.25%)
No LINK (19/64: 29.68%)
BEGINS (7/64: 10.93%)
ENDS (4/64: 6.25%)
SIMULTANEOUS (2/64: 3.12%)

[6]Note that the Italian preposition "*per*" corresponds only to a subset of uses of the English preposition "for" as in the example:

a) Suonò *per* un'ora [She played *for* an hour.]

However, with NoLINK relations the preposition *a* does not have a temporal value, rather it is used to express either quantities of time (and it usually corresponds to "how many times an event occurs or happens") or it can be considered as a particle selected by the VP. Therefore, if we exclude the NoLINK relations, we can consider that a Temporal PP introduced by *a* typically expresses a relation of inclusion. Further support to this observation can be observed in the possibility of substituting *a* with *in*, at least in the temporal domain. The semantics of the preposition is the following:

11) **λ(e, λ(t, INCLUDES(e, t)).**

As for the BEGINS and ENDS relations, the behaviour is the same as for the preposition *in*, i.e. they are activated by aspectual verbs.

## 6 Conclusion and Future Work

In this preliminary study we showed that prepositions heading a Temporal PP can be associated with one default temporal relation and that deviations from the norm are due to co-textual influences. The prototypical semantics of temporal prepositions can be represented as in 8)-11).

We also showed that the normalized values of timexes play a major role in the identification of temporal preposition semantics, more than the bare identification of their types. Instances of deviations from the prototypical meaning which could not be explained by differences in the timexes forced us to analyse the VPs, thus providing useful information for the definition of the heuristics.

An important result of this work is the definition of a preliminary set of heuristics for automatic annotation of temporal relations in text/discourse. Our study also suggests a possible refinement of the SIMPLE Ontology aimed at its usability for temporal relation identification; and it can be seen as a starting point for the development of a Timex Grammar.

In the next future we intend to implement this set of heuristics with a machine learning algorithm to evaluate their reliability. All wrongly annotated relations could be used for the identification of the relevant information to determine the contribution of the VP.

Some issues are still open and need further research, in particular it will be necessary to investigate the role of some 'complex' Temporal PPs (e.g. *in questo momento* 'in this moment', which can be paraphrased as 'now'), and how to extract the meaning of Temporal PPs as suggested in Schilder (2004).

## References

Allen F. James. 1984. Towards a General Theory of Action and Time. *Artificial Intelligence*, (23):123-54

Ferro Lisa, Mani Inderjeet, Sundheim Beth and Wilson George. 2001. *TIDES Temporal Annotation Guidelines: Version 1.0.2*. MITRE Technical Report, MTR 01W0000041

Filatova, Elena and Hovy, Eduard. 2001. Assigning Time-Stamps To Event –Clauses. *Proceedings of the ACL Workshop on Temporal and Spatial Information, Toulouse, France, 6-1 July*, pages 88-95

Haspelmath, Martin. 2007 (forthcoming). Frequency vs. iconicity in explaining grammatical asymmetries (ms).

Lassen Tine. 2006. An Ontology-Based View on Prepositional Senses. *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions* pages 45-50.

Lenci Alessandro, Montemagni Simonetta and Vito Pirrelli. 2003. CHUNK-IT. An Italian Shallow Parser for Robust Syntactic Annotation, in *Linguistica Computazionale* (16-17).

Mani Inderjeet and James Pustejovsky. 2004. Temporal Discourse Models for Narrative Structure. *ACL Workspoh on Discourse Annotation*

Hartrumpf Sven, Helbig Hermann and Rainer Osswald. 2006. Semantic Interpretation of Prepositions for NLP Applications. *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, pages 29-36.

Pustejovky James, Belanger Louis, Castaño José, Gaizauskas Robert, Hanks Paul, Ingria Bob, Katz Graham, Radev Dragomir, Rumshisky Anna, Sanfilippo Antonio, Sauri Roser, Setzer Andrea, Sundheim Beth and Marc Verhagen, 2002. *NRRC Summer Workshop on Temporal and Event Recognition for QA Systems*.

Pustejovsky James, Ingria Robert, Saurí Roser, Castaño José, Littman Jessica, Gaizauskas Robert, Setzer Andrea, Katz Graham and Inderjeet Mani. 2005. The Specification Language TimeML. *The Language of Time: A Reader*, Mani Inderjeet, Pustejovsky james and Robert Gaizauskas (eds), OUP.

Ruimy N., et al. 1998. The European LE-PAROLE Project: The Italian Syntactic Lexicon. *Proceedings of the LREC1998*, Granada, Spain.

Saint-Dizier Patrick.2006. *Syntax and Semantics of Prepositions,* (ed.), Springer, Dordrecht, The Netherlands.

Schilder Frank and Habel Christopher. 2001. Semantic Tagging Of News Messages. *Processing of the ACL Workshop on Temporal and Spatial Information, Toulouse, France, 6-1 July*, pages 65-72

Schilder Frank. 2004 Extracting meaning fron Temporal Nouns and Temporal Prepositions. *ACM Transactions on Asian Language Information Processing*, (3) 1:33-50

Setzer Andrea. 2001. *Temporal Information in Newswire Article: an Annotation Scheme and Corpus Study*, Ph.D. Thesis, University of Sheffield.

SIMPLE Work Package D2.1, available at < http://www.ub.es/gilcub/SIMPLE/simple.html>.

Van Eynde Frank. 2006. On the prepositions which introduce an adjunct of duration. *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions* pages 73-80.