



REPORT ABOUT CELCT ACTIVITY FOR THE LODE PROJECT

Valentina Bartalesi Lenzi and Rachele Sprugnoli
bartalesi@celct.it – sprugnoli@celct.it

30 June 2011

Table of Contents

1.	Introduction.....	2
2.	TimeML.....	2
3.	Details on the Annotation	2
4.	Statistical Data.....	2
	4.1 General Statistics	2
	4.2 Distribution of TIMEX3s by TYPE.....	3
	4.3 Distribution of TIMEX3s by VAL	3
	4.4 Distribution of Events by CLASS.....	3
	4.5 Distribution of Events by TENSE.....	3
	4.6 Distribution of Events by ASPECT	4
	4.7 Distribution of Events by POS.....	4
	4.8 Distribution of Events by VFORM.....	4
	4.9 Distribution of Events by MOOD.....	4
	4.10 Distribution of Events by POLARITY	4
	4.11 Distribution of Events by MODALITY	4
	4.12 Distribution of TLINKs by RELTYPE.....	5
	4.13 Distribution of SLINKs by RELTYPE	5
	4.14 Distribution of ALINKs by RELTYPE	5
5.	General Remarks.....	5
	References	6

1. Introduction

This document reports on the activities developed by CELCT within the LODE (Logic-based web tool for DEaf children) project. The goal of our work was to manually annotate a set of Italian tales following the TimeML specifications adapted to the Italian language (Caselli et al., 2011).

2. TimeML

TimeML (Pustejovsky et al., 2003) is a markup language for the annotation of Events (i.e. actions, states, and processes - <EVENT> tag), Temporal Expressions (i.e. durations, calendar dates, times of day and sets of time - <TIMEX3> tag), Signals (e.g. temporal prepositions and subordinators - <SIGNAL> tag) and various kind of dependencies between Events and/or Temporal Expressions (i.e. temporal, aspectual and subordination relations - <TLINK>, <ALINK> and <SLINK> tags respectively).

3. Details on the Annotation

Our annotation follows the It-TimeML guidelines, version 1.3.1 (Caselli, 2010), developed within a collaboration between ILC-CNR and CELCT.

The first prototype of *CAT (CELCT Annotation Tool)*, a stand-alone application, has been used to perform the annotation and to export the files in an XML in-line format.

Six tales have been annotated:

- Il paese di Ventolà
- Picnic
- L'aquila e il sasso
- La lepre e la tartaruga
- Francesco e la dieta
- Il gallo e il ragno

For what concern the annotation effort, the work involved 2 annotators and required a total of 10 person/days.

4. Statistical Data

This section shows statistical data about the performed annotation.

4.1 General Statistics

Number of files	6
Number of tokens (without punctuation)	1,853
Number of TIMEX3	42
Number of Event	389
Number of Signal	52
Number of SLINK	65
Number of ALINK	9
Number of TLINK	222

4.2 Distribution of TIMEX3s by TYPE

	Total
DATE	22 (52.38 %)
TIME	11 (26.19 %)
DURATION	8 (19.05 %)
SET	1 (2.38 %)
<i>TOTAL</i>	42

4.3 Distribution of TIMEX3s by VAL

	Total
XXXX-XX-XX	19 (45.24 %)
PIW	2 (4.76 %)
TMO	5 (11.91 %)
T04:00	1 (2.38 %)
P1D	1 (2.38 %)
T14:00	1 (2.38 %)
TEV	2 (4.76 %)
TNI	1 (2.38 %)
TMI	1 (2.38 %)
PTXH	3 (7.15 %)
PXD	1 (2.38 %)
TXX:XX	1 (2.38 %)
PRESENT_REF	2 (4.76 %)
PTXM	1 (2.38 %)
XXXX-WXX-07	1 (2.38 %)
<i>TOTAL</i>	42

4.4 Distribution of Events by CLASS

	Total
ASPECTUAL	10 (2.57%)
I_ACTION	22 (5.66%)
I_STATE	25 (6.43%)
OCCURRENCE	227 (58.35%)
PERCEPTION	22 (5.66%)
REPORTING	24 (6.17%)
STATE	59 (15.16%)
<i>TOTAL</i>	389

4.5 Distribution of Events by TENSE

	Total
PAST	183 (47.04%)
PRESENT	78 (20.05%)
FUTURE	2 (0.52%)
NONE	126 (32.39%)
<i>TOTAL</i>	389

4.6 Distribution of Events by ASPECT

	Total
PERFECTIVE	153 (39.33%)
IMPERFECTIVE	124 (31.88%)
PROGRESSIVE	0
NONE	112 (28.79%)
<i>TOTAL</i>	389

4.7 Distribution of Events by POS

	Total
NOUN	43 (11.05%)
VERB	334 (85.86%)
ADJECTIVE	12 (3.09%)
PREPOSITION	0
OTHER	0
<i>TOTAL</i>	389

4.8 Distribution of Events by VFORM

	Total
INFINITIVE	55 (14.14%)
GERUND	3 (0.77%)
PARTICIPLE	8 (2.06%)
NONE	323 (83.03%)
<i>TOTAL</i>	389

4.9 Distribution of Events by MOOD

	Total
CONDITIONAL	1 (0.26%)
IMPERATIVE	4 (1.03%)
SUBJUNCTIVE	5 (1.28%)
NONE	379 (97.43%)
<i>TOTAL</i>	389

4.10 Distribution of Events by POLARITY

	Total
POS	376 (96.66%)
NEG	13 (3.34%)
<i>TOTAL</i>	389

4.11 Distribution of Events by MODALITY

	Total
VOLERE	1
POTERE	2
DOVERE	1
<i>TOTAL</i>	4 / 389 (1.03%)

4.12 Distribution of TLINKs by RELTYPE

	Total
BEFORE	94 (42.34%)
AFTER	9 (4.06%)
INCLUDES	17 (7.66%)
IS_INCLUDED	40 (18.02%)
MEASURE	4 (1.80%)
SIMULTANEOUS	29 (13.06%)
IAFTER	0
IBEFORE	5 (2.25%)
IDENTITY	20 (9.01%)
BEGINS	3 (1.35%)
ENDS	0
BEGUN_BY	0
ENDED_BY	1 (0.45%)
<i>TOTAL</i>	222

4.13 Distribution of SLINKs by RELTYPE

	Total
MODAL	28 (43.08%)
FACTIVE	13 (20.00%)
EVIDENTIAL	22 (33.85%)
COUNTER_FACTIVE	2 (3.07%)
CONDITIONAL	0
NEG_EVIDENTIAL	0
<i>TOTAL</i>	65

4.14 Distribution of ALINKs by RELTYPE

	Total
CULMINATES	0
TERMINATES	1 (11.11%)
CONTINUES	3 (33.33%)
INITIATES	5 (55.56%)
REINITIATES	0
<i>TOTAL</i>	9

5. General Remarks

On the basis of the analysis of quantitative data presented in the previous section, some observations can be drawn:

- 1) TIMEX3s are underspecified in many cases because there are frequent temporal expressions like *un giorno/una sera* which are not anchored to a precise date.
- 2) The low percentage of I-ACTION and I-STATE events (5.66% and 6.43% respectively) indicates a reduced usage of subordinated clauses, in accordance with the simple syntactic structure of tales.
- 3) The clear prevalence of past tense (47.04%) characterize verbs in the corpus. This is a proper feature of tales with respect to other types of texts: for example, corpora of news stories have the majority of verbs at the present tense.
- 4) Events are mainly expressed by verbs (85.86%) and not by nouns (11.05%).

- 5) Tales contain temporal sequences of events: in particular, events in consecutive clauses are linked with temporal relations of type BEFORE (42.34%).

References

- J. Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. *IWCS-5, Fifth International Workshop on Computational Semantics*.
- T. Caselli. 2010. It-TimeML: TimeML Annotation Scheme for Italian, Version 1.3.1. *Technical Report*, September 23 2010
- T. Caselli, V. Bartalesi Lenzi, R. Sprugnoli, E. Pianta, I. Prodanof. 2011. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In *Proceedings of LAW V*, Portland, Oregon, USA, June 19-24 2011