TimeEL: Recognition of Temporal Expressions in Greek texts

Prokopis Prokopidis[1], Elina Desipri[1],
Haris Papageorgiou[1] & George Markopoulos[2]
[1]Institute for Language and Speech Processing & [2]University of Athens

1 Introduction

The unprecedented abundance of available information in our digital epoch has foreshadowed the fast growth of technology dealing with information access to unstructured data. Large collections of text need to be analyzed, annotated and organized in order to be searchable and retrievable from machines. Applications trying to capture a document's topic or "what a document is about" have to extract events or facts specifying "Who did/will do what to whom and how, when and where". Event and fact extraction is often employed in the framework of various natural language processing applications such as information extraction, e-government and Business Intelligence, recommendation systems, automatic summarization and question-answering. Events and facts are usually studied in terms of either their semantic structure (participants or arguments) or of the spatiotemporal dimension (Filatova & Hovy 2001). Thus, recognition and normalization of temporal expressions (TIMEXes) is considered a challenging prerequisite stage of analysis for several applications and has attracted a growing interest in the Natural Language Processing community during recent years. With this motivation, in this paper, we present work for the development of TimeEL, a rule-based software module that performs recognition of TIMEXes in Greek texts.

The remainder of this paper is structured as follows. In Section 2, we provide a short overview of the state of the art in developing relevant resources and tools. Section 3 discusses our efforts in constructing a Greek resource annotated for temporal expressions. In Section 4, we present the methodology in developing the TimeEL recognizer and we provide first evaluation results and error analysis. Finally, in Section 5, we conclude with discussion of ongoing work on normalization of recognized expressions and manual annotation of events.

2 Related Work

Initial efforts for the creation of corpora annotated for temporal expressions were conducted in the framework of the Message Understanding Conferences. One of the tasks of MUC-6 (MUC 1995) was the recognition of absolute temporal references to dates and times (*September 1, 1986; 10 P.M.*) in English texts. Annotations of indexical (*tomorrow; next week*) and relational (*ten days after the earthquake*) expressions were integrated in the resources of MUC-7 (MUC 1998). The TIDES scheme (Ferro et al. 2005) introduced the interpretation of temporal

expressions via the TIMEX2 XML element, which included an ISO standard compatible normalization of each expression (*2009-10-29-T18:30* for *29 October, 6:30 p.m*). In the same scheme, durations and underspecified expressions were marked and linked to reference times. An English corpus, which was annotated with the TIDES scheme and consisted of 306K words, was used in the Time Expression Recognition and Normalization (TERN) Evaluation task[1] organized in 2004 by the Automatic Content Extraction (ACE) program.

Expanding on the TIDES scheme, the TimeML mark-up language (Pustejovsky et al. 2005) included a specification and guidelines for time-stamping events (i.e. anchoring event predicates to temporal expressions) and for the annotation of temporal, subordination and aspectual links between events. The TimeBank 1.2 Corpus (Boguraev et al., 2007) is a resource that contains 183 English documents from the news domain that have been manually annotated according to the TimeML 1.2.1 specification. The TimeBank corpus has been used as a reference standard in the framework of the Temporal Relation Identification task of the SemEval-2007 event (Verhagen et al. 2007). It will also be used in the context of TempEval-2[2], where a subtask for automatic systems will be the determination of the type (*time, date, duration, set*) and the normalized value of reference temporal annotations.

Manual temporal annotation is a very time consuming and it has been reported (Verhagen & Moszkowicz 2009) that temporal and event annotation projects have not managed so far to produce resources similar in volume and level of consistency to other efforts focusing on morphological and/or syntactic annotation. Moreover, there are only a few projects involving annotation in languages other from English. This is perhaps the reason why researchers have even proposed the automatic import of the temporal annotations from English to another language by having the original texts translated by human translators (Forăscu 2008).

In parallel to these annotation efforts, several research groups have presented rule-based, statistical or hybrid systems for automatic recognition and normalization of temporal expressions. Mani & Wilson (2000) present a system that uses manual and automatically learnt rules and report an 83.2 F_1 score in recognizing and normalizing TIMEXes. One of the systems in TERN 2004 (Negri & Marseglia 2004) used approximately 1000 hand crafted rules and achieved a 92.6 F_1 score in recognizing TIMEX2 spans in English texts. In the ACE 2007 TERN[3], evaluation data consisted of 2028 time expressions and the task included recognizing TIMEX spans together with specific attributes about the expressions (*Value, Modifier, Anchor value, Anchor directionality, Set*). The best system achieved a 61.6 score, which was calculated based on weighted matches of the attributes and the extent of TIMEX2 spans. For Greek, Lucarelli et al. (2007) use

[1] http://timex2.mitre.org/tern.html
[2] http://www.timeml.org/tempeval2/
[3] http://www.itl.nist.gov/iad/mig//tests/ace/2007/

patterns extracted semi-automatically from training data to recognize temporal expressions annotated according to the MUC-7 guidelines. They report an F_1 score of 96.46 on a corpus of financial documents consisting of 205 K tokens and 1244 temporal expressions.

3 Annotated Resource

In this section, we provide details on the corpus we have used for development and evaluation purposes of our recognizer. This corpus, which amounts to 26.5K tokens and 1.7K sentences, comprises financial web documents, sport-related articles and transcribed documentaries about political affairs. An adaptation of the TIDES scheme for Greek was compiled by three students of linguistics (Roditi 2008), who worked using the Callisto annotation tool[4] to mark the extent of 601 TIMEX2 expressions, including 244 dates, 224 durations, 39 times and 28 set-denoting expressions.

All markables have as their linguistic head a proper lexical trigger relevant to the concept of time, which must be able to be oriented on a timeline. Table 1 contains indicative examples of markables together with some time-related items that were not considered triggers.

	Lexical Triggers	Non-Triggers
Noun	περίοδος, μισάωρο, διήμερο, (ε)βδομάδα, έτος, χιλιετία, πρωινό	χρονοδιάγραμμα, χρονοχρέωση, αρχή
Proper Name	Δευτέρα, Σαββατοκύριακο. Ιανουάριος, Φεβρουάριος	
Specialized time pattern	8:00 π.μ., 7/11/1981, 7-11-1981, 10ος μ.Χ.	
Adjective	καθημερινός, μεταμεσονύχτιος, περσινός, μονοετής	επόμενος, μακρινός βραχυπρόθεσμος,
Adverb	καθημερινά, ημερήσια, εβδομαδιαία, μηνιαία, ετήσια, πολυετώς	μακροπρόθεσμα, συχνά, πάλι, ξανά, επιτέλους
Time noun/adverb	χτες, αντιμεθαύριο, πρόπερσι	
Number	3 (έφτασε στις 3), πέντε (στις πέντε του μηνός) ΄80, ΄90, εξήντα.	

Table 1 Sample Lexical Triggers and Non-Triggers for Greek.

Non-triggers include words that are marked as parts of TIMEXes (“προηγούμενη” in “προηγούμενη χρονιά”) but not triggers themselves; subordinating conjunctions introducing temporal clauses (“καθώς”, “αφού”,

[4] http://callisto.mitre.org/

“αφότου”); frequency adverbs; and proper nouns with a non-temporal reference (“το θωρηκτό Οκτώβρης”).

Each markable is annotated as a TIMEX2 XML element accompanied by the following attributes: VAL contains the normalized, ISO-8601 compatible version of the expression. MOD is used for annotation of temporal modifiers. ANCHOR_VAL contains a normalized version of an expression that anchors the expression under examination. ANCHOR_DIR captures the directionality between VAL and ANCHOR_VAL. SET is a boolean attribute for expressions denoting sets of times. NON_SPECIFIC is used for non-referential expressions. Table 2 contains expressions that exemplify these attributes.

Text	Annotation
Την Πέμπτη το απόγευμα.	<TIMEX2 VAL="2009-02-20TEV">Την Πέμπτη το απόγευμα</TIMEX2>
Θα λείπω την άλλη βδομάδα.	<TIMEX2 VAL="2008-W12">την άλλη βδομάδα</TIMEX2>
Το Σ/Κ θα γράψω το άρθρο	<TIMEX2 VAL="2002-W38-WE">το Σ/Κ</TIMEX2>
Έζησε τον 19ο αιώνα	<TIMEX2 VAL="19">τον 19ο αιώνα</TIMEX2>
Η Μάχη του Μαραθώνα το 490π.Χ.	<TIMEX2 VAL="BC0490"> το 490 π.Χ. </TIMEX2>
όχι περισσότερες από 4 ημέρες	<TIMEX2 VAL="P4D" MOD="EQUAL_OR_LESS">όχι περισσότερες από 4 μέρες</TIMEX2>
Η συμμετοχή τους κατ’ έτος	<TIMEX2 VAL=”P1Y” NON_SPECIFIC=”YES”>κατ’ έτος</TIMEX2>
Τα επόμενα τρία χρόνια θα είναι δύσκολα.	<TIMEX2 VAL="P3Y" ANCHOR_DIR="STARTING" ANCHOR_VAL="2007-05-06"> Τα επόμενα 3 χρόνια</TIMEX2>

Table 2 Temporal expressions and their TIMEX2 annotations.

4 The TimeEL recognizer

In this section, we describe the TimeEL software module for recognition of temporal expressions in Greek texts. At a pre-processing stage, the raw textual input to the module is enriched with annotations generated automatically by a pipeline of natural language processing tools for robust processing of Greek texts (Papageorgiou et al. 2002). During this stage, a tokenizer and a sentence splitter segment the input text. A transformation-based tagger trained on a manually annotated corpus assigns part of speech tags to all tokens in the input. A module that queries a morphological lexicon of 70K lemmas lemmatizes the tagged tokens. Finally, a pattern grammar compiled into finite state transducers is used by a chunker that recognizes non-recursive phrase and clause boundaries. All

annotations for each text are merged in a file that can be imported, examined and edited in the GATE environment for developing NLP applications (Cunningham et al. 2002).

Following this stage, the annotated files are processed by the core of the TimeEL module, a grammar that was developed and tested using GATE's JAPE framework for writing cascades of pattern-based rules over lexical items and annotations. The cascades of a JAPE grammar consist of phases (i.e. groups of rules), where the output of each phase is fed to the next phase. The left-hand side of each rule contains the conditions that have to be met for the rule to fire, while the right-hand side adds or deletes annotations. The conditions can refer to both lexical items and annotations (like tags and lemmas) generated by previous processors. This allows for development of compact rules that are easy to read, modify and extend.

In its initial phase, the grammar defines some macros concerning, among others, parts of days, duration adjectives and adverbial pre- and post-modifiers of temporal expressions. Macros may refer to previously defined macros and can be reused in later stages of the grammar. Thus, the macro SETEXPRESSIONS in Table 3 refers to the previously defined MONTHNAMES; the former is reused in a SetExpresssions rule that creates a TIMEX annotation that may be modified in later phases of the grammar. Other rules included in this initial phase mark, among other expressions, duration ("μονοετής", "μονοετής") and season adjectives ("φθινοπωρινός", "θερινός"), nouns denoting parts of the day ("πρωί", "μεσάνυχτα"), etc. Another set of rules combines numbers and numerals with time units ("μέρες", "βδομάδες") to mark expressions like "δυο μήνες" and "δεύτερο εξάμηνο".

Macro	Macro
Macro: MONTHNAMES ({Token.lemma=="Ιανουάριος"} | ... {Token.lemma=="Δεκέμβριος"})	Macro: SETEXPRESSIONS ({Token.lemma == "κάθε"} ((DAYNAMES)| (MONTHNAMES)))
Rule	
Rule: SetExpressions /* Name of the rule */ ((SETEXPRESSIONS)) /* Left-Hand Side: Condition with a macro */ :SetExpr -->:SetExpr.TIMEX /* RHS: A TIMEX annotation is created */	

Table 3 A rule for recognition of SET TIMEXes using a macro referring to another macro

In the next phase, the markables are extended to the boundaries of non-recursive phrases recognized by the chunker. This allows us to glue together adjacent expressions ("την Πέμπτη το πρωί", "τη νύχτα της 20ης Απριλίου 1967"). Moreover, during this phase pre- and post-modifiers ("στα τέλη",

"περίπου", "αργότερα") are included in the annotations to mark expressions like "στις αρχές του αιώνα" and "ένα μήνα αργότερα".

In the last, post-processing phase of the grammar, context is inspected for the removal of spurious expressions (e.g. "11-10" when it is preceded by the "με" preposition and refers to the score of a sport event). Finally, expressions representing time ranges ("το διάστημα 01-21 Αυγούστου") are split into two distinct annotations which denote the star and the end point of the range, in concordance with the scheme followed during the manual annotation of the resource.

Evaluation results obtained with the TimeEL recognizer are provided in Table 4. We present results against two test sets. The first contains all documents from the annotated resource, while the second concerns a small subset of the corpus containing only financial documents. We show the number of expressions in the reference test set and the number of items output by the system in the REF and SYS columns, respectively. We also report the number of correct (COR), partially overlapping but not identical (PCOR), missing (MISS) and spurious (SPUR) items.

Test Set	REF	SYS	COR	PCOR	MISS	SPUR	PREC	REC	F_1
Whole (A)	601	643	492	80	29	71	82.7	88.5	85.5
Finance (B)	77	81	72	4	1	5	91.3	96.1	93.6

Table 4 Evaluation results for the TimeEL recognizer on two annotated collections

The final three columns show the precision, recall and F_1 score, which are defined as follows:

(1) Precision = Correct + ½ Partial / Correct + Spurious + Partial
(2) Recall = Correct + ½ Partial) / Correct +Missing + Partial
(3) $F_1 = (1+1^2)$ Precision * Recall / (1^2 * Precision) + Recall)

As shown in (1) and (2) above, partially correct items generated by the system are assigned a half weight. This way we do not discard as missed items manually annotated relational expressions including dependent clauses ("*τον μήνα* που είδα τον πατέρα μου"), where the system recognizes only the head of the expression. The difference in the results against the two different test sets is partly explained by the fact that set B contains relatively less relational expressions compared to its superset, and thus less partial matches.

Another reason for worse results against set A is the fact that the spurious expressions generated by the system are relatively more in some documents not included in the financial collection. These false positives are in most cases metaphorical and non-temporal uses of temporal expressions ("το μέλλον" in "το μέλλον της Ελλάδας", "χρόνο" in "μάχης με τον χρόνο", "τώρα" in "έλα τώρα"). Missing items are relatively few in both sets and include expressions referring to

particular historically or culturally defined expressions ("της Φραγκοκρατίας", "αρχαιότητα"), which, according to the schema, should be annotated without a VAL attribute.

5 Conclusions and Future Work

We have presented a rule-based recognizer of temporal expressions that we have evaluated against a collection of Greek texts, manually annotated for TIMEX2 elements. As can be observed from the evaluation results, the overall performance of the recognizer is satisfactory. Nevertheless, some issues have clearly not been addressed yet. One of them is normalization of the recognized expressions. Towards this goal, we have started expanding the grammar with rules recognizing the type of each expression, distinguishing between dates, times, sets and durations. In initial experiments, we have observed a 94.8 accuracy in detecting the correct type for all manually annotated expressions in test set B. Detecting the correct value for the VAL attribute is a more challenging task, especially for indexical and relational expressions. We are currently experimenting in using the Document Creation Time and other explicit expressions in the documents, in order to select correct reference times for other, vague expressions.

In another line of the same research effort, we have started augmenting our corpus with annotations about events, following the TimeML scheme. We assign each event a type (*Occurrence, Perception, Reporting, Aspectual*, etc.), while instances of events are accompanied by attributes relevant to tense, aspect and polarity. Temporal links will eventually be added to the resource to represent the temporal relationship holding between events and temporal expressions.

References

Boguraev B., J. Pustejovsky, R. Ando, M. Verhagen. 2007. TimeBank evolution as a community resource for TimeML parsing. In *Language Resources and Evaluation*. 41(1).91–115.

Cunningham H., D. Maynard, K. Bontcheva, V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In the *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*. Philadelphia.

Ferro, L., L. Gerber, I. Mani, B. Sundheim and G. Wilson. 2005. TIDES 2005 Standard for the Annotation of Temporal Expressions. http://fofoca.mitre.org/annotation_guidelines/ 2005_timex2_standard_v1.1.pdf

Filatova E. and E.H. Hovy. 2001. Assigning Time-Stamps to Event-Clauses. In the *Proceedings of ACL Workshop on Temporal and Spatial Reasoning*, Toulouse, France.

Forăscu, C. 2008. GMT to +2 or How Can TimeML Be Used in Romanian. In *Proceedings of LREC-2008*. Marrakech, Morroco.

Lucarelli, G., X. Vasilakos and I. Androutsopoulos. 2007. Named Entity Recognition in Greek Texts with an Ensemble of SVMs and Active Learning. In *International Journal on Artificial Intelligence Tools*. 16.1015–1045.

Mani, I. and G. Wilson. 2000. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting on Association For Computational Linguistics*. Hong Kong.

MUC 1995. Proceedings of the 6th conference on Message Understanding Conference. California: Morgan Kaufmann.

MUC 1998. Proceedings of the 7th conference on Message Understanding Conference. California: Morgan Kaufmann.

Negri M. & L. Marseglia, Recognition and normalization of time expressions: ITC-irst at TERN 2004. Technical report. ITC-irst. Trento.

Papageorgiou H., P. Prokopidis, I. Demiros, V. Giouli, A. Konstantinidis and S. Piperidis. Multi-level XML-based Corpus Annotation. In Proceedings of the 3nd Language Resources and Evaluation Conference. Las Palmas.

Pustejovsky J., B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani. 2005. The Specification Language TimeML. In *The Language of Time: A Reader*. Oxford, : Oxford University Press.

Roditi I. 2008. An annotation scheme for Greek Temporal Experssions. MA Thesis. Athens: Institute for Language and Speech Processing.

Verhagen, M., R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Prague, Czech Republic.

Verhagen M. & J. L. Moszkowicz. 2009. Temporal Annotation and Representation. *Language and Linguistics Compass*. 3(2).517–536.