UNIVERSITÀ DEGLI STUDI DI PADOVA
CORSO DI LAUREA IN INGEGNERIA INFORMATICA
TESI DI LAUREA

INDIVIDUAZIONE AUTOMATICA DI ESPRESSIONI TEMPORALI IN TESTI ITALIANI

RELATORE: Ch.mo Prof. *Carlo Minnaja*
Dip. *di Matematica Pura e Applicata*

LAUREANDO: *Laura Pattaro*

Padova, 11 Aprile 2006

a chi coltiva la speranza
e a chi l'ha solo smarrita per la via...

alla mia famiglia

Indice

Sommario

L'informazione temporale riveste un ruolo di grande importanza in vari campi di ricerca, in particolare nel campo dell'elaborazione del linguaggio naturale. Lo sviluppo di strumenti software efficaci per il riconoscimento e la gestione di informazioni temporali conduce, fondamentalmente, al problema dell'interpretazione della sintassi e della semantica testuale, analogamente a quanto accade per ogni problema di estrazione di informazione da un testo.

Questa tesi introduce il concetto di *espressione temporale*, illustrando le caratteristiche e le problematiche relative alla classificazione, al riconoscimento e alla rappresentazione di tale tipologia di espressioni.

Verranno descritti un insieme di strumenti per la *tokenizzazione* e l'etichettatura di testi creato da un gruppo di ricerca dell'Università di Edimburgo e un programma, sviluppato da uno studente dell'Università di Padova, che fa uso di tale tool per l'individuazione automatica di espressioni quantitative in documenti scritti in lingua italiana.

Verrà proposta una nuova versione di tale programma per l'italiano, orientata alla sola estrazione di informazione temporale. Il progetto si basa sull'analisi morfologica e sintattica della lingua e l'implementazione tiene conto dei risultati della fase analitica. La verifica delle prestazioni e un confronto con il programma originale mostrano che, con tale approccio, sono stati raggiunti buoni risultati.

Introduzione

L'informazione temporale è oggetto di grande interesse in molti ambiti di ricerca, poiché fornisce un parametro naturale per l'organizzazione di altri tipi di informazione. In particolare, il problema dell'interpretazione semantica di informazioni di tempo si pone come sfida nel campo dell'elaborazione del linguaggio naturale e contribuisce allo sviluppo della ricerca.
La presente tesi tratta l'individuazione automatica di espressioni temporali in testi italiani.

Nel primo capitolo verrà illustrato il concetto di *tempo* e verrà fornita una chiara definizione di *espressione temporale*, corredata di caratterizzazioni e problematiche legate alla classificazione. Si vedrà come l'informazione di tempo viene espressa tramite il linguaggio naturale e come possa essere formalmente rappresentata per la trattazione automatica.

Il secondo capitolo illustra una serie di strumenti (LT TTT, Text Tokenization Tool) realizzati dal Language Technology Group dell'Università di Edimburgo per l'elaborazione automatica di file di testo mirata al riconoscimento e all'etichettatura di specifici elementi sintattici. In seguito, sarà descritto un programma, implementato da uno studente dell'Università di Padova (v. [16]), che utilizza il sistema LT TTT e che è finalizzato al riconoscimento e all'annotazione di espressioni quantitative (numeriche, percentuali, temporali) in testi italiani.

Nel terzo capitolo viene proposta una nuova versione del programma scritto

per l'italiano. Questa versione limitata al riconoscimento delle sole espressioni temporali e finalizzata a ottenere una maggiore efficienza rispetto al programma originale. Verranno riportate la fasi di progetto e le caratteristiche salienti dell'implementazione. All'analisi delle prestazioni per la verifica del raggiungimento degli obiettivi, seguiranno un confronto col programma originale ed alcune osservazioni sulle scelte di progetto effettuate e sulle problematiche rimaste aperte. Risulterà evidente come il nuovo approccio presentato, basato sulle peculiarità lessicali e grammaticali della lingua di riferimento, possa fornire un metodo di individuazione efficace e più esaustiva.

Per facilitare la comprensione dei passaggi coinvolti, in Appendice sono presentati utili richiami relativi alla grammatica italiana.

Capitolo 1

L'informazione temporale

Time is a pervasive dimension of reality
as everything evolves as time elapses.
(F. Grandi)

1.1 Panoramica

Il tempo, come lo spazio, fornisce una coordinata fondamentale della nostra realtà. Ogni avvenimento, ogni entità naturale o artificiale, ogni evento climatico o astronomico, è associato necessariamente a una collocazione spazio-temporale precisa alla quale non sempre è possibile risalire, ma che, qualora fornita, risulta di grande utilità (basti pensare, ad esempio, alla ricostruzione della "storia passata"). La possibilità di cogliere e gestire contenuti di informazione temporale è, pertanto, di fondamentale importanza. A tale scopo, si rivela di grande utilità la disponibilità di sistemi computazionali in grado di estrarre ed archiviare informazioni temporali contenute in documenti digitali nonchè di acquisire e immagazzinare conoscenze relative agli aspetti temporali delle realtà da essi stessi modellate.

Così come viene espressa in linguaggio naturale, l'informazione temporale può anche essere ricavata da testi in linguaggio naturale. Come osservano Pustejovsky e Mani [28], viviamo in un mondo dinamico dove ogni evento è correlato ad altri e ad ogni azione ne consegue un'altra, un mondo quindi dove i fatti e le proprietà

associati alle entità cambiano col trascorrere del tempo. Se non fossimo in grado di identificare gli eventi all'interno di insiemi di dati espressi in linguaggio naturale e di collocare tali eventi temporalmente, rischieremmo di perdere gran parte delle informazioni stesse. Attualmente, l'interpretazione del contenuto dell'informazione costituisce ancora un problema (e.g. si pensi alla dipendenza dal contesto, questione che interessa gran parte delle informazioni fornite nei notiziari: il contesto può essere interno o esterno alla frase e il tempo a cui si fa riferimento può essere determinato soltanto mediante la conoscenza di ulteriori elementi); pertanto, la necessità di interpretare e trattare l'informazione temporale costituisce una delle maggiori sfide nel campo dell'elaborazione del linguaggio naturale e apporta un notevole contributo al suo sviluppo.

L'elaborazione del linguaggio naturale negli ultimi anni ha fatto molti progressi, questo grazie agli studi avanzati in diversi ambiti: l'analisi morfologica, il *part-of-speech tagging*, l'estrazione di tipo *Named-Entity*, il *parsing*. Si è visto, inoltre, che l'avanzamento della ricerca può essere accelerato mediante l'uso di metodi di studio basati sui *corpora*[1].

Gli ambiti e le applicazioni che possono beneficiare dell'uso dei corpora e dello sviluppo nel trattamento delle informazioni di tempo sono diverse. Si riporta qualche esempio:

- la ricerca nel campo del *Question Answering* (rivolta alla ricostruzione di risposte a domande del tipo: "Quando... ?"), in cui sono di fondamentale importanza il riconoscimento e la normalizzazione[2] delle espressioni temporali;

- l'*estrazione di informazione*[3] (che prevede, ad esempio, la normalizzazione

[1]Il termine *corpus* -plurale: corpora- indica un vasto insieme di documenti di testo raccolti da diverse fonti; i corpora vengono utilizzati per attività di elaborazione del linguaggio naturale. Gli approcci moderni per l'estrazione di informazione fanno ampio uso di insiemi di testi etichettati; richiedono la progettazione e il controllo di schemi di annotazione e uno sforzo combinato di attività di etichettatura manuale e automatica, il tutto al fine di ottenere i corpora da utilizzare, a loro volta, per la prova e la valutazione di classificatori di informazione.

[2]Col termine *normalizzazione* si intende la rappresentazione dell'espressione in qualche formato standard; v. anche 1.3.3

[3]La ricerca di informazioni all'interno di documenti in formato digitale è condotta mediante

di riferimenti temporali per l'inserimento nelle basi di dati);

- l'analisi cronologica di eventi rivolta alla ricostruzione di semplici avvenimenti o di periodi storici, o all'elaborazione di testi per lo sviluppo di riassunti automatici (e.g. per la composizione automatica di biografie), ambiti in cui risulta fondamentale la fase di ordinamento temporale dell'informazione;

- la ricerca nel campo delle traduzioni automatiche (che necessita di traduzione e normalizzazione dei riferimenti temporali);

- l'elaborazione di domande temporali per le basi di dati; Androutsopoulos (citato in [32]) fornisce una rappresentazione semantica che consente di mappare domande temporali (espresse in linguaggio naturale) in estensioni temporali del linguaggio d'interrogazione SQL;

- l'estensione della dimensione temporale ai documenti Web[4] (e.g. per l'individuazione di versioni diverse dello stesso documento; v. [4]).

Nel resto del capitolo e nei capitoli seguenti si tratteranno le espressioni temporali, dopo averne fornito una definizione precisa. Si potrebbe fare una distinzione tra il concetto di *informazione* temporale finora inteso e quello di *espressione* di tempo (intrinsecamente dotata di una maggiore specificità), ma si preferisce, nel seguito, considerare equivalenti questi due concetti (i casi in cui assumeranno significati distinti saranno chiaramente deducibili dal contesto).

sistemi di Information Retrieval (IR) e di Information Extraction (IE); i due sistemi danno due differenti tipi di supporto (v. [19]): le applicazioni di IR consentono semplicemente di individuare testi che possono contenere materiale pertinente a quanto cercato, lasciando all'utente l'onere di estrarre l'informazione dal testo attraverso la lettura; l'IE è un processo che analizza direttamente i testi, mediante applicazioni configurate in modo appropriato, fornendo in uscita all'utente solo l'informazione specifica desiderata, sotto forma di dati non ambigui in formato fisso.

[4]Può essere utile, per crearsi un'idea del contesto, la consultazione della bibliografia sugli aspetti temporali ed evolutivi del World Wide Web stesa da F. Grandi: *An Annotated Bibliography on Temporal and Evolution Aspects in the World Wide Web*, A TIMECENTER Technical Report, Sept 2003.

1.2 Rappresentazione del tempo

Per trattare l'informazione temporale, è utile introdurre, innanzitutto, un modello standard per la rappresentazione del *tempo*. Nel lavoro di Dyreson e Snodgrass [12] viene suggerita una semplice metafora geometrica in cui il tempo è rappresentato come una linea dotata di estremi (i.e. un segmento), assumendo l'universo limitato dal punto di vista temporale. Di conseguenza, un *istante* è identificato come un *punto* sulla linea del tempo; un *periodo* è il tempo compreso tra due istanti e un *intervallo* è inteso come una lunghezza o un segmento di cui non sia specificata la collocazione lungo la linea temporale. All'intervallo è dato, in altri lavori, lo stesso significato qui assegnato al periodo (i.e. corrisponde, sostanzialmente, a un intervallo con collocazione nota); qui verrà assunta questa seconda definizione.

Ad ogni possibile implementazione di tale struttura viene associata una quantità temporale, chiamata *chronon*, che consiste nella più piccola unità di tempo rappresentabile; la linea del tempo risulta così suddivisa in un numero finito di chronons.

In letteratura, le varie entità temporali (intese come diverse "quantità" di tempo: secondo, minuti, giorni, etc) individuabili concretamente in un determinato contesto sono generalmente indicate col nome di *granularità* e saranno specificate nel paragrafo successivo.

1.3 Le espressioni di tempo

Per trattare l'individuazione automatica di espressioni temporali all'interno di testi italiani, ossia il riconoscimento e l'etichettatura delle stesse mediante strumenti informatici è utile, innanzitutto, fare chiarezza sul contesto.

1.3.1 Definizione

In accordo con Ahn, Adafre e de Rijke [18], si definiscono *espressioni temporali* le "(parti di) frasi in linguaggio naturale che si riferiscono direttamente a punti o a intervalli nel tempo"; esse non portano semplicemente informazione temporale di per sé stesse, ma servono anche per localizzare eventi a cui si fa riferimento in un determinato contesto.

In senso lato, l'informazione di tempo può essere concepita come l'insieme degli aspetti temporali di entità, eventi, documenti, notizie, etc.

1.3.2 Caratterizzazione

Espressioni di tempo comuni in linguaggio naturale comprendono: nomi (*lunedì*, *gennaio*), aggettivi (*corrente*, *presente*), proposizioni subordinate (*Quando sarai tornato...*); esse sono, cioè, costituite da singole parole o insiemi di parole tra loro combinate in sintagmi o proposizioni [5]. Il carattere temporale di una frase, tuttavia, talvolta non è espresso da specifici termini ma è determinato da altre associazioni implicite tra le parti che costituiscono la frase stessa, e.g. dai tempi verbali[6]. Il tempo del verbo e l'aspetto dell'azione [7] che esso esprime spesso sono in grado di specificare la collocazione temporale degli eventi; in alcuni casi, invece, possono solo indicare la dimensione cronologica (i.e. passato, presente o futuro) in cui l'azione si è svolta, ma non il momento preciso. Katz e Arosio [34] propongono un metodo indipendente dalla lingua e indipendente dalla teoria per etichettare relazioni di tempo interne alle frasi (anche implicite).

In questa tesi non saranno prese in considerazione l'analisi verbale e le relazioni cronologiche implicite; il lavoro sarà concentrato sui riferimenti di tempo espliciti.

[5] Un'ampia panoramica della lessico associato al contesto temporale e alcuni richiami di grammatica sono riportati in Appendice.

[6] Attraverso di essi si è in grado di esprimere il rapporto tra due momenti diversi, consentendo così di stabilire la relazione cronologica tra due eventi.

[7] L'*aspetto* è il modo in cui si presentano le azioni rappresentate dal verbo relativamente alla propria *durata*, alla propria *compiutezza* e al proprio *svolgimento*; e.g. un'azione può essere momentanea o avere una durata, può presentarsi nel nascere, nel proprio sviluppo o alla propria conclusione. Tempo e aspetto sono espressi attraverso la struttura e le desinenze del verbo.

Han e Lavie [29] propongono una lista di espressioni rappresentative della complessità del contesto temporale; il relativo elenco viene riportato di seguito poiché fornisce, attraverso svariati esempi, una buona caratterizzazione delle espressioni temporali:

- *11 aprile 2006*: le espressioni più semplici sono precisamente collocabili lungo la linea del tempo (v. "Granularità", 1.3.2);
- *di mercoledì*: molte espressioni sono *sotto-specificate*, ossia non hanno una precisa collocazione nel tempo e non si riferiscono né a un'unità temporale né a un intervallo;
- *sabato o domenica*: le disgiunzioni logiche fanno uso di congiunzioni come "o";
- *alle 4*: è un caso di ambiguità linguistica, potrebbe riferirsi alle 4 della mattina o alle 4 del pomeriggio;
- *oggi*: espressioni di questo tipo si collocano lungo la linea del tempo tramite un riferimento implicito a un momento temporale determinato;
- *la settimana scorsa*: in questo caso l'informazione comporta una traslazione nel tempo; nello stesso tempo, è fornisce la propria granularità (nell'esempio, ci si riferisce a un tempo non più esteso di una settimana);
- *giorni fa*: la traslazione nel tempo non può essere direttamente valutata, poiché mancano sia la possibilità della collocazione temporale, sia la possibiità di determinare la granularità dell'espressione;
- *il primo venerdì del mese*: è una tipica "espressione ordinale", specificata da un numero ordinale e un intervallo di tempo;
- *sabato e domenica*: è un esempio di enumerazione di entità temporali;
- *da oggi al 2007*: è un intervallo, specificato da un punto d'inizio e un punto finale (implicitamente, si assegna la stessa granularità (v. 1.3.2) ad entrambi i punti: in questo caso, l'unità è il giorno);

- *ogni giorno di maggio* : esprime una ricorrenza all'interno di un intervallo temporale ed ha un significato diverso da *ogni giorno*, in cui l'intervallo non è specificato;

- *due volte al giorno*: esprime una frequenza all'interno di un intervallo, ma non specifica il momento dell'azione.

Granularità

Il concetto di granularità risale ai tempi antichi (v. [23]), quando gli eventi naturali periodici, come la rotazione terrestre, l'alternanza delle fasi lunari, la rivoluzione della terra attorno al sole, vennero identificati come unità di tempo (*granularità*) e vennero introdotti i calendari come mezzo per esprimere relazioni tra le diverse unità.

Nella scienza computazionale, la capacità di fornire e mettere in relazione tra di loro le rappresentazioni di tempo riguardanti una stessa realtà ma specificate a diversi *livelli di granularità* costituisce un importante campo di ricerca ed è anche uno dei requisiti di base per lo sviluppo di molte applicazioni (e.g. la progettazione di database temporali e la loro interoperabilità, la conversione di dati temporali, il data mining, la specificazione e la verifica di sistemi a tempo reale, il *reasoning* orientato al tempo, l'elaborazione del linguaggio naturale).

Nella maggioranza delle applicazioni pratiche, il dominio del tempo viene suddiviso secondo un determinato livello di granularità e ciascuna delle parti in cui lo stesso viene partizionato è percepita come un'unità indivisibile (*granulo*). La descrizione di un evento, pertanto, può utilizzare le varie unità per fornire una qualificazione temporale dell'evento stesso ad un livello di astrazione appropriato.

L'introduzione della granularità del tempo in una rappresentazione formale, tuttavia, comporta alcune problematiche. Ad esempio, ci si può trovare a dover risolvere il problema di assegnare un significato specifico all'associazione di frasi caratterizzate da domini temporali diversi, inoltre, può essere necessario passare da un dato dominio temporale a un'altro che può presentare una granularità differente (più fine o più grossolana).

Generalmente, l'unità di base assunta per l'informazione temporale e per le basi di dati temporali è il giorno.

Indeterminatezza

Il modello di tempo descritto nel paragrafo 1.2 introduce anche il concetto di determinatezza e indeterminatezza; si parla di istante

- *determinato*, quando è noto il chronon in cui l'istante è localizzato;
- *indeterminato*, quando è noto l'insieme di chronon in cui è potenzialmente localizzato l'istante, ma non si è in grado di individuare il chronon preciso.

In generale, si parla di *indeterminatezza* di un'espressione temporale quando non è noto il tempo esatto al quale essa si riferisce.

Dyreson e Snodgrass, nel campo delle basi di dati temporali [17], parlano di informazione di tipo "don't know exactly when" e spiegano da dove può avere origine tale tipo di informazione; ad esempio:

- *granularità*: la granularità della base di dati può essere diversa da quella che caratterizza l'evento considerato (e.g., sappiamo che un fatto è accaduto durante un determinato anno ma la base di dati conserva informazioni con granularità di un giorno);
- *tecniche di assegnazione di date*: molte tecniche di questo tipo sono intrinsecamente imprecise (si pensi alla datazione col metodo del Carbonio-14);
- *pianificazioni*: spesso non si è in grado di specificare esattamente la data prevista della fine di un progetto;
- *eventi accaduti in tempi imprecisati o sconosciuti*: non sempre è noto il tempo in cui un evento è accaduto (si pensi, ad esempio, alla necessità di conservare nella base di dati di un ufficio anagrafico una data di nascita non nota o imprecisa).

Grandi e Mandreoli (v. [2], [3]), impegnati nello sviluppo di tecnologie XML e nella codifica di semantica temporale per applicazioni orientate alla conservazione dell'eredità culturale e per la realizzazione di versioni digitalizzate di documenti storici, propongono una classificazione delle espressioni temporali indeterminate basata su una distribuzione di probabilità a gradini. Le espressioni vengono suddivise in quattro categorie:

- viene usata un'espressione a granularità di più alto livello per denotare un'espressione indeterminata con granularità di livello inferiore (e.g. "L'abbazia fu consacrata nel 1276"), si parla di granularity mismatch;
- si fa riferimento alla parte iniziale di un intervallo temporale (e.g. "All'inizio del XV secolo");
- si fa riferimento alla parte finale di un intervallo temporale (e.g. "Alla fine del XV secolo");
- il riferimento è relativo a un *intorno* temporale (e.g. "Intorno al 1850").

1.3.3 Riconoscimento e normalizzazione

Nell'ambito dell'individuazione automatica di informazione temporale si presenta la necessità di:

- specificare quali tipi di informazione sia opportuno etichettare come espressioni temporali e quali no (riconoscimento);
- interpretare le espressioni individuate secondo uno schema di classificazione e rappresentarle in un formato standard (normalizzazione), assegnando ad esse un opportuno valore.

Una proposta attuale e interessante di uno schema per le marcature di espressioni temporali è costituita dallo standard per le annotazioni steso dalla società MITRE[8] [30]. Tale standard fornisce delle linee-guida rivolte a chi si occupa di

[8]La MITRE Corporation è una organizzazione senza fini di lucro costituita per scopi pubblici. Si occupa di ingegneria dei sistemi, ingegneria dell'informazione etc. Gestisce tre centri di ricerca

annotazione manuale (e.g. per la costruzione di corpora etichettati) e a chi sviluppa strumenti software per l'estrazione di informazione temporale da documenti. Il processo di annotazione proposto è decomposto in due passi corrispondenti ai due obiettivi sopra esposti: il primo prevede l'individuazione di un'espressione temporale in un documento, il secondo consiste nell'identificazione del valore di tempo che tale espressione indica, o a cui l'autore intende riferirsi; questo approccio risulta applicabile a una grande varietà di generi di documenti.

Lo schema di annotazione citato è stato progettato principalmente secondo questi criteri: *semplicità*, per poter essere utilizzato facilmente; *precisione*, per poter essere applicato ad attività di elaborazione del linguaggio naturale; *naturalezza*, ovvero dovrebbe distinguere le espressioni come lo farebbe una persona nell'annotazione manuale; *espressività*, cioè dovrebbe essere in grado di specificare al meglio i valori di tempo, anche facendo uso di parametri e di indicazioni di granularità; *riproducibilità*, ovvero ogni linea-guida è definita sulla base di esempi, in modo tale da assicurare coerenza tra i risultati di etichettatori diversi.

Il documento precisa i tipi di espressioni per i quali è prevista la marcatura, fornisce i dettagli per codificare formalmente e coerentemente il significato delle espressioni e presenta indicazioni per la determinazione dell'estensione precisa delle espressioni (in termini di parole in esse contenute). La rappresentazione semantica utilizzata è altamente indipendente dalla lingua di riferimento ed è prevista, per il futuro, l'estensione dello schema per l'utilizzo con corpora multilingue.

L'intento delle linee-guida è fornire gli strumenti per l'interpretazione delle espressioni di tempo indicanti: *quando* è accaduto un dato evento, *quanto a lungo* dura un dato evento, oppure *quanto spesso* si verifica un dato evento.
Vi sono rappresentati, perciò, tre diversi tipi di valori di tempo [7]:

punti nel tempo (rispondono alla domanda: *quando?*): possono essere date di calendario, momenti del giorno o combinazioni delle due; sono rappresentate tramite valori in formato ISO [9].

e sviluppo finanziati dal governo federale degli Stati Uniti e gestisce anche un proprio programma di ricerca e sviluppo che esplora l'uso di nuove tecnologie.

[9] ISO-8601, ftp://ftp.qsl.net/pub/gismd/8061v03.pdf, http://www.so.org/iso/en

durate : rappresentano un periodo di tempo in formato ISO;

frequenze : fanno riferimento a insiemi di punti nel tempo, anziché a punti isolati.

Vengono, inoltre, affrontati alcuni dei problemi semantici caratteristici delle informazioni temporali, come l'indeterminatezza (v.1.3) e la questione delle espressioni di tempo sotto-specificate (v.1.3); in questi casi, lo standard non propone direttamente l'assegnazione di un valore ai parametri della rappresentazione formale, ma consente, all'utenza che ne fa uso, di specificare a piacimento il valore stesso.

L'identificazione di espressioni temporali è limitata a quelle che contengono parole di tempo riservate usate in un senso temporale, chiamate "lexical trigger" (e.g. giorno, settimana, ora, lunedì, future,etc.) e specificate nel dettaglio all'interno del documento.

1.4 Etichettatura di espressioni temporali

Una metodologia di etichettatura proposta in quest'ambito è quella sopra citata, definita dalla società MITRE, alla quale è stato assegnato il nome "TIMEX2". Sono state, successivamente definite estensioni di TIMEX2, denominate "TimeML" e "TIMEX3". Mani descrive, in [32], tali metodi di marcatura, dei quali si riporta una breve descrizione.

1.4.1 TIMEX2

TIMEX2 è caratterizzato da regole abbastanza complesse. Come illustra Mani [32], è stato sviluppato originariamente dal programma TIDES del DARPA e adottato dal governo degli Stati Uniti nel task RDC (Scoperta e Caratterizzazione di Relazioni) all'interno del programma ACE (Estrazione Automatica dei Concetti) e in due workshop estivi di TimeML dell'ARDA. Costituisce lo schema di annotazione di riferimento per la marcatura dell'estensione di espressioni temporali inglesi (con tags TIMEX2) e la normalizzazione dei loro valori secondo il formato ISO-8601 (1997), con qualche ampliamento. Lo schema TIMEX2 rappresenta formalmente

il significato delle espressioni indicanti punti collocati nella linea temporale, durate di eventi, tempi *fuzzy* (i.e. confusi, non precisi). Lo stesso distingue tra usi specifici e non-specifici delle espressioni temporali (indefiniti, abituali,...). La rappresentaizone degli insiemi di tempo, inoltre, è caratterizzata da qualche estensione (periodicità, granularità). In realtà, gli annotatori che utilizzano i tag TIMEX2 si scostano dalle linee guida indicate producendo errori sistematici.

Per l'utilizzo dei tag TIMEX2 nella classificazione delle espressioni temporali, sono possibili diversi approcci (v. [32])

1.4.2 TimeML

TimeML (Pustejovsky et al. 2004) è un *metadato* standard per la marcatura di eventi e il loro ancoraggo temporale all'interno di documenti. E' stato applicato soprattutto ad articoli di notiziari in lingua inglese. Lo schema di annotazione integra due altri schemi: TIMEX2 del TIDES (v. sopra), STAG di Sheffield (Setzer & Gaizauskas, 2000) e altri lavori emergenti (Katz & Arosio, 2000); TIMEX2 è storicamente il segmento più vecchio di TimeML e ne costituisce la parte più robusta. TimeML individua una varietà di espressioni di eventi, inclusi aggettivi che indicano staticità temporale, infiniti sostantivati e, inoltre, gestisce i tempi verbali. La rappresentazioni di eventualità, in TimeML, hanno vari attributi, tra cui: il tipo di evento, il tempo verbale, l'aspetto etc. Gli avverbi temporali includono preposizioni temporali e congiunzioni (prima, dopo, mentre). E' possibile rappresentare anche espressioni di tempo, aggiungendo modifiche a TIMEX2, pervenendo così ad uno schema di annotazione chiamato TIMEX3 (v. sottoparagrafo seguente). Il nodo principale, però, sta nel collegare eventualità e tempi: e.g. ancorando un evento a un tempo, oppure ordinando eventi e/o tempi; ciò viene realizzato mediante un nuovo strumento chiamato TLINK. Tale collegamento prende in considerazione anche eventi reali rispetto ad eventi ipotetici, invece, attraverso SLINKS. TimeML annota anche verbi che non caratterizzano un evento distinto, ma indicano una fase particolare di un altro evento; ciò si ottiene tramite ALINK che lega il verbo in questione all'evento. Un lavoro recente di Hobbs & Pustejovsky (2004) mappa la struttura di TimeML ad una teoria formale del tempo (si parla della "piccola

Ontologia del Tempo", DAML), che consente di porre interrogazioni formali a un sistema di reasoning.

1.4.3 TIMEX3

La specificazione di TIMEX2 è problematica per gli etichettatori. Un'estensione che ha portato vantaggi è il sistema di etichettatura TIMEX3, il quale presenta alcune semplificazioni ottenute con l'aggiunta di due attributi oltre al campo relativo al valore (*value*): la quantificazione sull'insieme (*quant*) e la specificazione della frequenza all'interno dell'insieme (*freq*). TIMEX3 permette, inotre, di attribuire dei tag anche a espressioni dipendenti dal contesto (cosa già possibile con TimeML); inoltre presenta uno stile funzionale per la codifica di traslazioni specificate nelle espressioni temporali (i.e., la prossima settimana). Non è ancora verificata, tuttavia, la piena realizzabilità di queste estensioni di TIMEX3.

1.5 Le basi di dati temporali

Le basi di dati costituiscono uno strumento fondamentale per la conservazione di grandi quantitativi di dati, in quanto ne consentono l'immagazzinamento e permettono di accedervi efficientemente mediante strumenti di interrogazione; negli ultimi anni,la quantita' crescente di informazioni accessibili in rete ha proposto nuove sfide alla ricerca in quest'ambito, sia in campo accademico sia in ambito industriale.

In questo contesto le informazioni possono essere strutturate (e.g. nel caso provengano da basi di dati relazionali o orientate agli oggetti), parzialmente strutturate o completamente disorganizzate (e.g. quando siano costituite da semplici raccolte di documenti o immagini). Le basi di dati, intese in maniera classica, classificano gli aspetti statici delle informazioni; di conseguenza, le informazioni in esse contenute corrispondono ad entita' valide, per il mondo reale, al momento dell'inserimento. In molte applicazioni, in particolare per la trattazione delle informazioni semistrutturate, riveste interesse non solo l'informazione temporale statica contenuta nei dati, ma anche l' informazione sull'evoluzione cronologica dei

dati e sui loro aspetti dinamici (si consideri, ad esempio, l'accesso a versioni successive di un determinato documento Web; oppure collezioni di dati dinamici di varia natura: finanziaria, medica, relativa alla gestione del personale). L'informazione di questo tipo necessita di una base di dati *temporale*, la quale preveda:

- l'inserimento degli aspetti dinamici dei dati,

- linguaggi di interrogazione ampliati per l'accesso alle coordinate temporali dei dati stessi.

Oliboni, Quintarelli e Tanca [33] introducono per la gestione dei dati semistrutturati un modello basato su grafi etichettati, nel quale conservare gli aspetti dinamici e statici dei dati limitando la quantità di etichette necessarie alla rappresentazione dell'informazione; propongono, inoltre, un linguaggio di interrogazione simile all'SQL e tecniche di controllo del modello per consentire interrogazioni relative agli aspetti statici e dinamici dei dati.

1.5.1 Tempo di validità e tempo di transazione

Alle basi di dati temporali sono associati due concetti di base: il tempo di transazione e il tempo di validità.

Per *tempo di transazione* si intende, in generale, il periodo di tempo in cui un fatto è *valido* in una base di dati, da quando è inserito nel sistema a quando viene cancellato; riguarda dunque l'evoluzione dei dati rispetto al sistema in cui sono memorizzati. Il *tempo di validità* riguarda, invece, l'evoluzione dei dati rispetto alla realtà applicativa che essi descrivono; corrisponde al tempo in cui un fatto è valido nel mondo reale.

In anni recenti, il gruppo di ricerca nel campo delle basi di dati del CSITE-CNR si è dedicato all'introduzione di aspetti temporali sul Web, adattando ed estendendo concetti e tecniche derivati dalla ricerca sulle basi di dati temporali. L'inizio della ricerca risale al 1997 ed esplora l'applicabilità al Web delle nozioni di tempo di transazione e di tempo di validità. In questo contesto, tali nozioni assumono sfumature proprie: il tempo di transazione riguarda la disponibilità e

l'attribuzione di una versione a risorse del Web (spostandosi lungo un'ideale linea che rappresenti il tempo di transazione, si avrebbe accesso a diverse versioni della stessa pagina Web); il tempo di validità si riferisce, invece, alla validità temporale reale dell'informazione contenuta delle risorse stesse (i.e. documenti Web). Grazie al concetto di tempo di validità, risulterebbe possibile:

- realizzare documenti dotati di un'intrinseca caratterizzazione temporale (si potrebbe parlare, cioè, di documenti storici),

- consentire una visualizzazione *temporalmente selettiva* delle risorse Web.

Questo si rivela utile nell'ottica della creazione di applicazioni dedicate all'eredità culturale. In quest'ambito, lo stesso gruppo ha proposto [4] un'infrastruttura di tipo XML/XSL ("The Valid Web", [5] e [1]) per la gestione di dati variabili nel tempo e dei quali si voglia mantenere eventuali versioni progressive prodotte da modifiche degli stessi. L'infrastruttura è fornita di: modelli per la rappresentazione dei dati, linguaggi di interrogazione, strutture di indicizzazioni, etc. La sua finalità d'uso è la gestione dell'informazione storica contenuta in documenti Web multimediali e prevede un'ulteriore estensione per la codifica di altre caratteristiche temporali avanzate, come l'indeterminatezza, le granularità e i diversi calendari. Tutto ciò permette un'elaborazione efficiente dei dati temporali attraverso un semplice ambiente Web.

Come detto in 1.3, il concetto di indeterminatezza dei dati temporali si estende anche nel campo delle basi di dati. Dyreson e Snodgrass, a questo proposito, ritengono (v. [17]) che si debba distinguere tra l'indeterminatezza contenuta nei dati e quella contenuta nelle interrogazioni: i sistemi di gestione delle basi di dati temporali dovrebbero essere sviluppati in modo tale da poter fronteggiare tale problema, consentendo all'utente di controllare (tramite interrogazioni) l'indeterminatezza delle informazioni derivate dai dati inseriti e fornendo, nelle risposte alle interrogazioni, anche le informazioni relative all'indeterminatezza introdotta dalle interrogazioni stesse. Inoltre, i risultati delle interrogazioni dovrebbero garantire un certo grado di precisione anche in presenza di indeterminatezza.

Capitolo 2

Strumenti per l'individuazione delle espressioni temporali

Il programma che sarà presentato nel capitolo 3 consiste in un riadattamento di un lavoro per l'estrazione di informazioni quantitative da testi italiani, sviluppato qualche anno fa, per la Tesi di Laurea, da F. Faccioni (studente di Ingegneria Informatica presso l'Università di Padova).

Lo strumento principale su cui si basano i programmi appena citati è un sistema di tokenizzazione di testi (*Text Tokenization Toolkit*) realizzato all'Università di Edimburgo per l'estrazione di informazione da testi in inglese.

- Nella prima sezione del Capitolo sarà brevemente descritto il sistema di tokenizzazione.
- Nella seconda sezione sarà illustrato il programma implementato da Faccioni.

2.1 Il sistema TTT

Text Tokenization Toolkit (TTT, [15]) è un sistema software sviluppato dal Language Technology Group (LTG) dell'Università di Edimburgo; LTG è un gruppo dedicato alla ricerca e allo sviluppo di software nell'ambito dell'elaborazione del linguaggio naturale.

Il sistema, disponibile all'indirizzo Web *http://www.ltg.ed.ac.uk/software/ttt/* in formato binario per Solaris, include una serie di strumenti concepiti per la tokenizzazione e il mark-up di testi in lingua inglese. Fornisce componenti per segmentare il testo in moduli di varie estensioni predefinite (come paragrafi, frasi, parole e altri tipi di token) e, in particolare, consente all'utente di definire personalmente le regole per la produzione di mark-up dedicato ad applicazioni particolari; tali regole devono essere create all'internodi opprtuni file grammaticali. Gli strumenti forniti elaborano testi in formato XML ed sono in grado di convertire *mark-up non-XML* in *mark-up XML*.

I tool disponibili non sono usati singolarmente, ma vengono combinati tra loro in pipeline: ciascuno di essi può aggiungere, modificare o rimuovere parte di mark-up relativo ai dati in ingresso. Diverse combinazioni degli stessi tool possono, così, produrre differenti risultati anche se applicate agli stessi testi. Una pipeline, normalmente, comprende un parte di elaborazione *a livello carattere* (dedicata, ad esempio, la suddivisione dei paragrafi in parole) e una parte cosiddetta *a livello XML* (che, ad esempio, effettua il raggruppamento di particolari tipi di parole in insiemi più ampi).

Il sistema TTT comprende, in particolare:

- una directory denominata *bin*, contenente i seguenti file binari eseguibili:

  fsgmatch (Fast SGml MATCH), nucleo del sistema TTT: *transducer* general purpose che elabora un flusso di dati in ingresso e lo riscrive sulla base di regole fornite in un file grammaticale. Può essere utilizzato per cambiare l'input in vari modi; tutte le grammatiche fornite con il sistema TTT, tuttavia, limitano la sua azione all'aggiunta di informazioni

di mark-up. *Fsgmatch* ha due modalità operative, secondo le quali l'input è considerato come flusso di caratteri (*fsgmatch* a livello carattere) o come flusso di elementi SGML/XML (*fsgmatch* a livello SGML).

lttok è un programma che identifica singole parole e le etichetta come elementi XML; può essere usato in alternativa a *fsgmatch*; per default, il suo avvio causa anche l'attivazione del programma *ltstop*. *lttok* è guidato da regole che consentono, tramite la specificazione di espressioni regolari da individuare nel testo in ingresso, il riconoscimento di stringhe come potenziali parole. In generale, il programma classifica come parole le stringhe riconosciute; talvolta, tuttavia, possono essere specificati altri tipi di entità da associare al riconoscimento (e.g. simboli di punteggiatura).

ltstop è un programma che identifica ed etichetta punti finali di frasi (tecnicamente, è un disambiguatore di confini di frase): per ogni punto, tenta di determinare se è un indicatore di fine-frase oppure se è parte di un'abbreviazione, o se coincide con entrambi. La distinzione tra i casi è possibile grazie alla disponibilità di accesso a una lista di abbreviazioni conosciute e a una lista di non-abbreviazioni conosciute (se *ltstop* non riconosce un'abbreviazione e la confonde con un punto, significa che l'abbreviazione non è contenuta nella lista ed è possibile aggiungerla).

ltpos assegna etichette di tipo POS (Part-Of-Speech) alle parole di testi in lingua inglese, basandosi su modelli probabilistici; può gestire sia testo in formato ASCII sia file in formato XML.

sgdelmarkup è un programma general-purpose a cui è assegnato il compito di rimuovere parte del mark-up posto da elaborazioni precedenti; gestisce file in formato XML.

sgmltrans traduce file XML in un altro formato (HTML, LaTeX o altro tipo). Il file che governa il funzionamento di *sgmltrans* contiene una lista ordinata di regole; ciascuna di esse descrive gli elementi ai quali dev'essere applicata e specifica le stringhe da fornire in uscita. La docu-

mentazione precisa che questo programma è sperimentale, non è molto efficiente ed è dotato di funzionalità limitate.

sgmlperl è una versione di *sgmltrans* molto potente, in quanto consente che i corpi delle regole (contenute in file XML) siano costituiti da programmi Perl standard.

sggrep lavora come il programma *grep*, cercando espressioni sotto forma di stringhe di formato regolare. L'output è in formato XML e può essere usato come ingresso per una successiva chiamata di *sggrep*.

- una directory denominata DOC, contenente la documentazione relativa al sistema (disponibile al sito *http://www.ltg.ed.ac.uk/software/ttt/tttdoc.html*).
- una directory denominata GRAM, contenente le grammatiche che agiscono *a livello carattere* (directory *char*) e *a livello SGML* (directory *sgml*).
- una directory denominata LEX, contenente i file **.lex* con le liste di vocaboli a cui fanno riferimento le grammatiche (lessico esterno).
- una directory denominata OUTPUT, verso cui viene reindirizzato l'output del programma, contenente.
- una directory denominata RES, contenente le regole che specificano i programmi precedentemente elencati.
- una directory denominata SCRIPTS, contenente programmi Perl per conversioni di caratteri e stringhe.

Una descrizione del sistema TTT è fornita in [15].

L'home page del Language Technology Group di Edimburgo si trova all'indirizzo *http://www.ltg.ed.ac.uk/index.html*

2.2 Un programma per l'italiano

In passato il sistema TTT è stato utilizzato da Faccioni per la realizzazione di un programma finalizzato all'individuazione di espressioni quantitative in lingua italiana.

Il programma utilizza tre file grammaticali per le fasi di riconoscimento delle espressioni quantitative:

- *numbers2.gr*, in cui sono definite le regole che permettono di individuare ed etichettare numeri e quantità rappresentati in modo testuale;

- *numex2.gr*, in cui sono definite le regole per l'etichettatura di quantità monetarie e numeri percentuali;

- *timex2.gr*, in cui sono definite le regole per l'individuazione di espressioi;

L'interesse, in questa tesi, è rivolto al problema dell'identificazione delle espressioni temporali; si focalizza perciò l'attenzione sulla gestione di tale finalità.

Il programma di Faccioni, a questo riguardo, fa uso di due file specifici: il file lessicale *timex2.lex* e il file grammaticale *timex2.gr*.

timex2.lex contiene una lista di termini caratteristici del contesto temporale e una piccola quantità di aggettivi generici che posono essere affiancati a termini legati al tempo.

timex2.gr guida l'azione di *fsgmatch* a livello SGML: contiene, infatti, una serie di regole che consentono, tramite la specificazione di espressioni regolari da individuare nel testo in ingresso, il riconoscimento di parole o gruppi di parole come potenziali espressioni temporali; si parla di "match" avvenuto (ossia di riconoscimento avvenuto), quando l'espressione regolare definita da una regola (detta RULE) si trova in esatta corrispondenza *a livello sintattico* con una serie di parole contenute nel file elaborato; ciò determina l'assegnazione, a tale serie di parole, dell'etichettatura definita per le espressioni temporali. Le RULEs definite in questo file si riferiscono al riconoscimento

- delle tipologie di termini appartenenti al lessico;
- di orari del giorno;
- di date riferite a un calendario;
- di molte altre espressioni più o meno complesse, ottenute dalla combinazione delle classi precedenti tra di loro e/o con articoli e preposizioni semplici e articolate.

Le *etichette* identificano le espressioni temporali come elementi di tipo *TIMEX*, i quali sono dotati di un attributo *type* che può assumere tre valori distinti: *DATE*, *TIME*, *DURATION* (v. $TTT/RES/general.dtd,xml).

Ognuno di questi elementi fornisce un potenziale tipo di classificazione per l'espressione temporale; tramite il file $TTT/UTPUT/SCRIPTS/generaltrans.dtd,xml è possibile specificare, per ciascuno di essi, i parametri che definiscono l'aspetto che esse devono assumere nel file in uscita (e.g. possono essere evidenziate mediante uno sfondo colorato lungo la propria estensione, tramite caratteri in grassetto e/o corsivo, etc).

La documentazione dettagliata si trova in [16].

Capitolo 3

Nuova stesura del programma per l'italiano

In questo Capitolo viene approntato il progetto di riadattamento del programma sviluppato da Faccioni. Poiché l'interesse di questa tesi è limitato alla sola estrazione di espressioni temporali, verranno tralasciate le parti del programma dedicate all'individuazione di entità numeriche non legate al tempo.

Il programma originale riconosce numerose espressioni temporali, dalle più semplici alle più complesse; da un'analisi si evince, tuttavia, che opportune modifiche consentirebbero di individuarne una gamma più vasta.

Il modo più immediato di procedere potrebbe consistere, in prima analisi, nell´estendere il file relativo al lessico e nell'introdurre nel file grammaticale ulteriori RULEs derivate; tuttavia, Faccioni osserva, nel suo lavoro [16], che la definizione di nuove RULEs permetterebbe di "creare una casistica più esaustiva" per il problema, ma in tal modo si finirebbe per rientrare "nelle competenze di un linguista". Un'analisi approfondita richiederebbe, certamente, considerevoli competenze di linguistica e lo sviluppo di un programma il più esaustivo possibile che dovrebbe, preferibilmente, essere progettato in collaborazione con un esperto linguista (fermo restando che l'esaustività rimarebbe un'utopia). E', tuttavia, possibile mostrare che, pur limitandosi a uno studio più superficiale, i risultati possono essere notevolmente migliorati senza aumentare di molto il numero delle

regole grammaticali.

La lettura del programma mette in luce anche un altro aspetto: nonostante Faccioni si fosse prefisso come scopo quello di "sfruttare il software messo a disposizione e le sue idee base, non quello di adattarlo alla lingua", il programma per l'italiano sembra ricalcare, per certi aspetti, la struttura di quello originale scritto per la lingua inglese e le prestazioni, probabilmente, risentono di ciò.

Sulla base di queste considerazioni, si propone di affrontare il problema dalle radici, ossia dalla lingua italiana e dalle sue peculiarità morfologiche e grammaticali.

Le modifiche apportate al programma originale, illustrate nei primi due paragrafi, interessano sia il file del lessico esterno *timex2.lex* sia il file *timex2.gr* che contiene le regole grammaticali; seguono: un confronto con il programma originale, la verifica degli obiettivi e alcune osservazioni.
Si conclude il Capitolo con un esempio di elaborazione di un testo.

3.1 Il file lessicale *timex2.lex*

Il file relativo al lessico esterno ($TTT\LEX\timex2.lex) è stato esteso mediante l'aggiunta di termini caratteristici del contesto temporale e di aggettivi ad essi correlati. Nello stesso risultano, così, presenti le seguenti tipologie di termini:

- i nomi dei giorni della settimana (tag DY e DYY);

- i nomi dei mesi (tag MS);

- i nomi delle stagioni e i relativi aggettivi (tag SEA, ASEA, ASEAP);

- nomi che rappresentano intervalli più o meno ampi di tempo (istante, secondo, anno, millennio, etc.), raggruppati secondo la similarità d'uso (e.g. si dice "l'anno scorso" e "il mese scorso", ma difficilmente si sentirà dire "il secondo scorso"; dunque, *anno* e *mese* apparterranno allo stesso gruppo nel file timex2.lex, *secondo* starà in un altro gruppo) e associati ai tag TU, TUP, DP, DPP, DS, DSP, DU, DUP;

- termini indicati particolari orari del giorno (tag TN);

- termini indicanti momenti del giorno legati al moto terrestre e, per questo, definibili mediante un intervallo temporale (tag MOM);

- termini relativi ai momenti dedicati ai pasti (poiché, in linea di massima, hanno una collocazione temporale definibile con buona approssimazione mediante un intervallo di tempo);

- termini collocabili, rispetto al presente, in maniera determinata (stasera, oggi, domani);

- aggettivi dotati una **connotazione** temporale (BA, BAP, AGG) o privi di essa (UP, UPP, QT, QTP);

- preposizioni che possono introdurre sintagmi dotati di significato temporale;

- nomi di giornate o periodi festivi;

- avverbi;

- locuzioni avverbiali;

Viene, innanzitutto, riportato il contenuto del file, evidenziando in grassetto i vocaboli e le modifiche introdotti (la lista dei termini lessicali è riportata ordinata per righe); successivamente, verranno esposte le motivazioni delle principali variazioni introdotte ed eventuali osservazioni.

domenica	DY	lunedì	DY **DYP**
martedì	DY **DYP**	mercoledì	DY **DYP**
giovedì	DY **DYP**	venerdì	DY **DYP**
lunedi	DY **DYP**	**martedi**	DY **DYP**
mercoledi	DY **DYP**	**giovedi**	DY **DYP**
venerdi	DY **DYP**	sabato	DY
sabati	**DYP**	domeniche	**DYP**
gennaio	**MS**	febbraio	**MS**
marzo	**MS**	aprile	**MS**
maggio	**MS**	giugno	**MS**
luglio	**MS**	agosto	**MS**
settembre	**MS**	ottobre	**MS**
novembre	**MS**	dicembre	**MS**
minuto	TU	**secondo**	**TU**
istante	**TU**	**attimo**	**TU**
momento	**TU**	ora	TU
mezzora	TU	mezz'ora	:: TU
secondi	TUP	minuti	TUP
ore	TUP	**istanti**	**TUP**
attimi	**TUP**	mezzogiorno	TN
mezzodì	**TN**	mezzanotte	TN
colazione	EP	pranzo	EP
merenda	EP	aperitivo	EP
cena	EP	oggi	**GG**
stasera	**GG**	**stanotte**	**GG**
stamattina	**GG**	**stamane**	**GG**
stamani	**GG**	ieri	**GG**
domani	**GG**	**domattina**	**GG**
dopodomani	**GG**	l' altro ieri	:: **GG**
ieri l' altro	:: **GG**	primavera	SEA

estate	SEA	autunno	SEA
inverno	SEA	**primaverile**	**ASEA**
estivo	**ASEA**	**estiva**	**ASEA**
autunnale	**ASEA**	**invernale**	**ASEA**
primaverili	**ASEAP**	estivi	**ASEAP**
estive	**ASEAP**	autunnali	**ASEAP**
invernali	**ASEAP**	mattina	DP
mattino	DP	pomeriggio	DP
dì	**DP**	sera	DP
notte	DP	mattinata	DP
serata	DP	nottata	DP
giornata	DP	giorno	DP
volta	**DP**	mattine	DPP
pomeriggi	DPP	sere	DPP
notti	DPP	mattinate	DPP
serate	DPP	nottate	DPP
giornate	DPP	giorni	DPP
volte	**DPP**	**momento**	**DS**
tempo	**DS**	**periodo**	**DS**
momenti	**DSP**	**tempi**	**DSP**
periodi	**DSP**	ultimo	UP
ultima	UP	primo	UP
prima	UP	**stesso**	**UP**
stessa	**UP**	**medesimo**	**UP**
medesima	**UP**	inizio	PT
fine	PT	metà	PT
principio	**PT**	**termine**	**PT**
arco	**ACM**	**corso**	**ACM**
mezzo	**ACM**	ultimi	UPP
ultime	UPP	primi	UPP
prime	UPP	**mezze**	**UPP**
stessi	**UPP**	**stesse**	**UPP**

medesimi	**UPP**	**medesime**	**UPP**
attuale	**BA**	precedente	BA
seguente	BA	**successivo**	**BA**
successiva	**BA**	**antecedente**	**BA**
scorso	BA	scorsa	BA
prossimo	BA	prossima	BA
passato	**BA**	**passata**	**BA**
venturo	**BA**	**ventura**	**BA**
futuro	BA	**futura**	**BA**
tardo	BA	tarda	BA
nostro	**BA**	**nostra**	**BA**
breve	**BA**	**a venire**	**:: BA**
attuali	**BAP**	precedenti	BAP
seguenti	BAP	**successivi**	**BAP**
successive	**BAP**	**antecedenti**	**BAP**
scorsi	BAP	scorse	BAP
prossimi	BAP	prossime	BAP
passati	BAP	passate	BAP
venturi	**BAP**	**venture**	**BAP**
futuri	BAP	**future**	**BAP**
nostri	**BAP**	**nostre**	**BAP**
lunghi	**BAP**	**lunghe**	**BAP**
brevi	**BAP**	**a venire**	**:: BAP**
giorno	DU	bimestre	DU
trimestre	DU	**quadrimestre**	**DU**
semestre	DU	stagione	DU
anno	DU	weekend	DU
settimana	DU	mese	DU
lustro	**DU**	decennio	DU
secolo	DU	millennio	DU
epoca	**DU**	**week end**	**:: DU**
fine settimana	:: DU	bimestri	DUP

trimestri	DUP	**quadrimestri**	**DUP**
semestri	**DUP**	stagioni	DUP
decenni	DUP	secoli	DUP
millenni	DUP	**epoche**	**DUP**
anni	DUP	weekends	DUP
settimane	DUP	mesi	DUP
lustri	**DUP**	fine settimana	:: DUP
alba	**MOM**	**tramonto**	**MOM**
imbrunire	**MOM**	**vespero**	**MOM**
entro	EO	**oltre**	**EO**
non oltre	**:: EO**	**entro e non oltre**	**:: EO**
capodanno	**HOL**	natale	HOL
epifania	HOL	**quaresima**	**HOL**
pasqua	HOL	pentecoste	HOL
avvento	HOL	**ognissanti**	**HOL**
mercoledì delle ceneri	**:: HOL**	**vacanze**	**HOL**
ferie	**HOL**	tutti i santi	:: HOL
martedi grasso	:: HOL	festa del papa	:: HOL
festa della mamma	:: HOL	anno nuovo	:: HOL
tutti i santi	:: HOL	**poco**	**QT**
tanto	**QT**	**molto**	**QT**
parecchio	**QT**	**pochi**	**QTP**
poche	**QTP**	**alcuni**	**QTP**
alcune	**QTP**	**tanti**	**QTP**
tante	*QTP*	**diversi**	**QTP**
diverse	**QTP**	**molti**	**QTP**
molte	**QTP**	**parecchi**	**QTP**
parecchie	**QTP**	**sporadici**	**QTP**
sporadiche	**QTP**	**svariati**	**QTP**
svariate	**QTP**	**adesso**	**AVV**
allora	**AVV**	**allorché**	**AVV**
allorquando	**AVV**	**ancora**	**AVV**

appena	**AVV**	**attualmente**	**AVV**
contemporaneamente	**AVV**	**dacché**	**AVV**
dapprima	**AVV**	**dianzi**	**AVV**
dopo	AVV	durante	AVV
finalmente	**AVV**	**finché**	**AVV**
finora	**AVV**	**frattanto**	**AVV**
frequentemente	**AVV**	già	AVV
ieri	AVV	**immediatamente**	**AVV**
infine	**AVV**	**intanto**	**AVV**
istantaneamente	**AVV**	**lungamente**	**AVV**
mai	**AVV**	**meno presto**	**AVV**
mentre	**AVV**	**momentaneamente**	**AVV**
oggi	AVV	**ogniqualvolta**	**AVV**
ognitanto	**AVV**	**ora**	**AVV**
oramai	**AVV**	**ormai**	**AVV**
poi	**AVV**	**poscia**	**AVV**
precedentemente	**AVV**	presto	AVV
prima	AVV	**prestissimo**	**AVV**
quando	**AVV**	**quandunque**	**AVV**
raramente	**AVV**	**recentemente**	**AVV**
sempre	**AVV**	**spessissimo**	**AVV**
spesso	**AVV**	**sovente**	**AVV**
subito	**AVV**	**subitissimo**	**AVV**
successivamente	**AVV**	**talora**	**AVV**
talvolta	**AVV**	tardi	AVV
tardissimo	**AVV**	**temporaneamente**	**AVV**
testè	**AVV**	**tosto**	**AVV**
tuttora	**AVV**	**a breve**	**:: LOC**
alle origini	**:: LOC**	**all'epoca di**	**:: LOC**
all'improvviso	**:: LOC**	**al momento giusto**	**:: LOC**
al momento opportuno	**:: LOC**	**al tempo in cui**	**:: LOC**
a lungo	**:: LOC**	**a mano a mano che**	**:: LOC**

a quel punto	**:: LOC**	**a questo punto**	**:: LOC**
a volte	**:: LOC**	**ben presto**	**:: LOC**
da allora	**:: LOC**	**da ora in poi**	**:: LOC**
da quando	**:: LOC**	**da sempre**	**:: LOC**
da subito	**:: LOC**	**dopo che**	**:: LOC**
dopoché	**:: LOC**	**dopo di**	**:: LOC**
d'ora in poi	**:: LOC**	**d'ora in avanti**	**:: LOC**
di buon'ora	**:: LOC**	**di punto in bianco**	**:: LOC**
fin da allora	**:: LOC**	**fino a che**	**:: LOC**
fino ad ora	**:: LOC**	**fin quando**	**:: LOC**
fintanto che	**:: LOC**	**fino a quando**	**:: LOC**
fra poco	**:: LOC**	**in avvenire**	**:: LOC**
in contemporanea	**:: LOC**	**in futuro**	**:: LOC**
in origine	**:: LOC**	**in passato**	**:: LOC**
in precedenza	**:: LOC**	**in seguito**	**:: LOC**
l'avvenire	**:: LOC**	**man mano che**	**:: LOC**
meno presto	**:: LOC**	**meno tardi**	**:: LOC**
meno raramente	**:: LOC**	**meno spesso**	**:: LOC**
nel contempo	**:: LOC**	**nel frattempo**	**:: LOC**
nel momento che	**:: LOC**	**non appena**	**:: LOC**
ogni qual volta	**:: LOC**	**ogni tanto**	**:: LOC**
ogni volta che	**:: LOC**	**ogni volta**	**:: LOC**
or ora	**:: LOC**	**per il futuro**	**:: LOC**
per l'addietro	**:: LOC**	**per l'avvenire**	**:: LOC**
per ora	**:: LOC**	**per ore**	**:: LOC**
per sempre	**:: LOC**	**più presto**	**:: LOC**
più raramente	**:: LOC**	**più tardi**	**:: LOC**
poco fa	**:: LOC**	**prima o poi**	**:: LOC**
tutte le volte che	**:: LOC**	**tutt'ora**	**:: LOC**
una volta	**:: LOC**	**un tempo**	**:: LOC**
per quando	**:: LOC**	**prima di**	**:: LOC**
prima che	**:: LOC**	**una volta che**	**:: LOC**

via via che :: LOC
(i tag preceduti dalla coppia di punti "::" caratterizzano espressioni composte da più parole: i punti marcano il punto di separazione tra il termine lessicale e il tag)

Le principali modifiche apportate al file, oltre all'aggiunta di nuovi vocaboli, sono le seguenti: sono stati mantenuti i gruppi originali (in qualche caso modificando i tag) e si è scelto di conservare quasi tutti i termini presenti nel documento di partenza, salvo alcune preposizioni (poi inserite direttamente nella costruzione delle regole del file grammaticale) e qualche vocabolo (ritenuto essere collocato in maniera inappropriata); inoltre:

- sono stati inseriti nomi e tag relativi alla forma plurale dei giorni settimanali (l'inglese, diversamente, prevede solo la forma singolare);
- è stato creato il gruppo "MOM" contenente i momenti del giorno legati al sorgere e al calare del sole;
- sono stati inseriti, nei gruppi appropriati, ulteriori termini riguardanti intervalli temporali (e.g. quadrimestre) e festività;
- sono stati aggiunti, negli opportuni gruppi, ulteriori aggettivi dotati -o meno- di aspetto temporale;
- sono stati creati i gruppi "ASEA" e "ASEAP" degli aggettivi relativi alle stagioni, nelle forme singolare e plurale;
- sono stati popolati due nuovi gruppi contenenti gli avverbi e le locuzioni temporali, i quali hanno la particolarità di costituire di per sé stessi delle entità temporali e non necessitano di essere accompagnati da articoli o preposizioni.

3.2 Il file grammaticale *timex2.gr*

Il file relativo alla grammatica ($TTT\GRAM\sgml\timex2.gr) è stato modificato per ottenere *una maggiore leggibilità e una ristrutturazione delle RULEs basata sull'analisi sintattica (cercando di limitare, per quanto possibile, un eventuale aumento del numero delle RULEs stesse).*
Tali scelte sono state concretizzate nella maniera descritta di seguito.

Sono stati, innanzitutto, rivisti gli elementi di tipo ENTITY:

- sono stati eliminati quelli non utilizzati nella lingua italiana;

- sono stati aggiunti i seguenti (per semplificare le RULEs):

```
<!ENTITY IN
"W/#~^[Ii]n|[Nn]ell[aeo]|[Nn]ell'|[Nn]e[il]|[Nn]egli$">

<!ENTITY DA
"W/#~^[Dd]all['aeo]|[Dd]ai|[Dd]al|[Dd]a|[Dd]agli$">

<!ENTITY QUESTO
"W/~^[Qq]uest['aeio]|[Qq]uell['aeio]|[Qq]uel|[Qq]uei|
[Qq]uegli$">
```

- sono stati modificati alcuni nomi (per coerenza linguistica).

In secondo luogo, sono state aggiornate le RULEs relative alle voci contenute nel file timex2.lex: quasi tutti i gruppi lessicali sono stati considerati etichettabili, ad eccezione dei seguenti:

- gli aggettivi che indicano quantità, contrassegnati con QT o QTP;

- gli aggettivi contrassegnati con i tag UP o UPP.

Questi aggettivi, infatti, non hanno un riferimento intrinseco al tempo; spesso, però, sono associati a nomi con i quali formano espressioni temporali (*qualche* giorno, *ultimo* anno). Come accennato in precedenza, gli altri tipi di aggettivi hanno, invece, di per sé stessi un significato legato al tempo (v. i tag BA, BAP: *antecedente*, *tardo*, *futuro*, ...); solo nelle RULEs di questi ultimi, perciò, compare l'indicazione di assegnazione del tag di *espressione temporale*:

```
<RULE name="mom" targ_sg="TIMEX[TYPE='DATE']"">.
```

Tutti i nomi delle RULEs sono stati cambiati, assegnando all'attributo *name* il valore (riportato in lettere minuscole) del tag associato al gruppo di termini riconosciuti dalla RULE stessa (e.g. nel primo esempio riportato sotto, alla RULE che riconosce le parole contrassegnate con il tag MOM sarà assegnato l'attributo *name="mom"*). Per le nuove parole inserite nel lessico sono state create le seguenti regole:

```
<!-- alba...tramonto - MOM -->
<RULE name="mom" targ_sg="TIMEX[TYPE='DATE']">
 <REL match="look-up" type="REF">
  <ARG bind='$1'>&WRD;</ARG>
  <ARG bind='$2'>MOM *</ARG>
 </REL>
</RULE>

<!-- entro...oltre - EO -->
<RULE name="eo" targ_sg="TIMEX[TYPE='DATE']">
 <REL match="look-up" type="REF">
  <ARG bind='$1'>&WRD;</ARG>
  <ARG bind='$2'>EO *</ARG>
 </REL>
</RULE>

<!-- avverbi - AVV -->
<RULE name="avv" targ_sg="TIMEX[TYPE='DATE']">
```

```
 <REL match="look-up" type="REF">
  <ARG bind='$1'>&WRD;</ARG>
  <ARG bind='$2'>AVV *</ARG>
 </REL>
</RULE>

<!-- locuzioni avverbiali - LOC -->
 <RULE name="loc" targ_sg="TIMEX[TYPE='DATE']">
  <REL match="look-up" type="REF">
   <ARG bind='$1'>&WRD;</ARG>
   <ARG bind='$2'>LOC *</ARG>
 </REL>
</RULE>

<!-- quantità singolare - QT -->
<RULE name="qt">
 <REL match="look-up" type="REF">
  <ARG bind='$1'>&WRD;</ARG>
  <ARG bind='$2'>QT *</ARG>
 </REL>
</RULE>

<!-- quantità plurali - QTP -->
<RULE name="qtp">
 <REL match="look-up" type="REF">
  <ARG bind='$1'>&WRD;</ARG>
  <ARG bind='$2'>QTP *</ARG>
 </REL>
</RULE>

<!-- aggettivi temporali - AGG -->
<RULE name="agg" targ_sg="TIMEX[TYPE='DATE']">
 <REL match="look-up" type="REF">
```

```
  <ARG bind='$1'>&WRD;</ARG>
  <ARG bind='$2'>AGG *</ARG>
 </REL>
</RULE>
```

Per quanto riguarda le RULEs rimanenti, finalizzate perlopiù al riconoscimento di espressioni più o meno complesse, è stato fatto riferimento, in modo specifico, alla grammatica italiana (in Appendice se ne trovano richiami essenziali), per coprire il più possibile la varietà delle espressioni temporali tipiche della nostra lingua, senza aumentare a dismisura numero di RULEs.

Sono stati individuati due metodi distinti.

1) Il primo metodo consiste nell'elencare, nel modo più esauriente possibile, i vari *sintagmi* (v. Appendice) componibili a partire dalla lista di parole contenute nel file timex2.lex.

Costruiamo, per ogni gruppo di sostantivi di tale file, i possibili gruppi *nominali* e *preposizionali* (v. Appendice), facendo uso di uno pseudo-linguaggio basato su regole sintattiche simili a quelle della programmazione in Perl.

Tale linguaggio utilizzerà:

- i tag (in carattere maiuscolo) che individuano i gruppi del lessico, per rappresentare gli elementi appartenenti a ciascun insieme;
- i nomi (in carattere maiuscolo) assegnati agli elementi di tipo ENTITY, per rappresentare il corrispondente insieme di articoli o preposizioni articolate;
- i tag (in carattere maiuscolo) relativi agli aggettivi del lessico, per rappresentare i diversi gruppi di aggettivi;
- la lettera N (in carattere maiuscolo) per indicare un numero cardinale di valore arbitrario;
- i nomi di RULEs già definite (in caratteri minuscoli);
- vocaboli (in carattere minuscolo) che rappresentino *sé stessi* e non rimandino ad altre entità.

Si adottano, inoltre, le seguenti convenzioni:

- le parentesi quadre saranno usate per indicare che l'entità tra di esse contenuta è opzionale;

- il simbolo di concatenazione semplice "|" tra più elementi significherà la presenza di uno solo di tali elementi, i.e. quello che avrà la più ampia corrispondenza con la parola del contesto; nelle RULEs espresse in XML, esso sarà associato al valore DISJ dell'attributo *match*;

- il simbolo di concatenazione doppia "||" significherà, invece, la presenza del primo elemento che troverà corrispondenza con la parola del contesto *match="DISJF"*).

Vediamo, ad esempio, come può essere costruito un sintagma contenente il nome di un giorno della settimana (l'espressione in pseudo-linguaggio evidenziata in grassetto è seguita da un esempio esplicativo riportato in corsivo):

prima	**DI**	**DY**	[**BA**]
prima	*di*	*sabato*	*[prossimo]*

Qui di seguito viene riportato un elenco di sintagmi codificati nel linguaggio specificato e relativi ai gruppi di sostantivi contenuti nel file timex2.lex.

Giorni della settimana - DY

```
di DY, entro [(il [BA])|QUESTO] DY [BA], dopo [(il [BA])|QUESTO] DY
[BA], per [(il [BA])|QUESTO] DY [BA], per tutt[oa] IL [BA] DY, fino
A [BA|UP] DY [BA], DA [BA|UP|QUESTO|qualche] DY [BA], da UN DY,
durante IL [DY|DYP], in QUESTO DY, ogni DY, IL [BA|UP] DY [BA], IL
DY prima|dopo, prima DI [BA|UP] DY, [per] tutt[ie] IL DYP
```

Mesi - MS

```
a MS, in MS, di MS, entro MS, entro [IL [BA|UP] mese di] MS [BA],
dopo [il mese di]|[IL BA] MS [BA], per [tutto] [IL [BA|UP]] MS,
```

per [tutto] il mese di MS, fino A [mese di] MS, DA [mese di] MS, DA [BA|UP] MS [BA], durante [tutto] MS, tutto MS, prima DI [BA|UP] MS [BA], metà MS, A PT DI [BA|UP] MS [BA], a fine MS

Parte della giornata - DP

A DP, IL DP [prima|dopo], di DP, entro [IL|qualche] DP, entro IL [BA|UP] DP [BA|UP|ASEA], dopo IL [BA|UP] DP [BA|UP|ASEA], per [tutto/a] IL [BA|UP] DP [BA|UP], fino A [BA|UP] DP [BA|UP|ASEA], DA [BA|UP] DP [BA|UP], IN DP, da qualche DP, durante IL DP, ogni DP, metà DP,A PT DI DP

Parti della giornata - DPP

IN [BAP|UPP] DPP [BAP|UPP], IL [BA|UP] DPP [BAP|UPP], IL DPP [prima|dopo], entro [QTP|N] DPP, per [QTP|N] DPP, da [QTP|NUM] DPP, dopo [QTP|NUM] DPP, durante IL [BAP|UPP] DPP [BAP|UPP|ASEAP], [TT] IL DPP

Periodo: giorno,... millennio - DU

IN [BA|UP] DU [BA|UP], IL [BA|UP] DU [BA|UP], IL DU [prima|dopo], entro [IL|qualche] DU, entro IL [BA|UP] DU [BA|UP], dopo IL [BA|UP] DU [BA|UP], per [TT] IL [BA|UP] DU [BA|UP], fino A [BA|UP] DU [BA|UP], DA [BA|UP] DU [BA|UP], da qualche DU, durante IL [BA|UP] DU [BA|UP], ogni DU, A PT DI [BA|UP] DU [BA|UP], metà DI [BA|UP] DU [BA|UP]

Periodi: giorni,... millenni - DU

IN [BAP|UPP] DUP [BAP|UPP], IL DUP [prima|dopo], entro [QTP|N] DUP, entro IL [BAP|UPP] DUP [BAP|UPP], dopo IL [BAP|UPP] DUP [BAP|UPP], per [TT] IL [BAP|UPP] DUP [BAP|UPP], [fino A [BAP|UPP] DU [adj]], DA [BAP|UPP] DUP [BAP|UPP], [QTP|N] DU, durante IL [BAP|UPP] DUP [BAP|UPP], [TT] IL DUP

Secondo... ora - TU

```
in UN|ogni TU, ogni TU, entro UN|qualche TU, da UN|qualche|QUESTO
TU, durante QUESTO TU, IL|UN TU prima|dopo
```

Secondi... ore - TU

```
entro [QTP|N] TUP, da [QTP|N] TUP, dopo [QTP|N] TUP, per [QTP|NUM]
TUP, IL TUP prima|dopo
```

Oggi...l'altro ieri - GG

```
GG, da GG, per GG, entro GG, fino a GG
```

Stagioni - SEA

```
a SEA, DI SEA, IL [BA|UP] SEA [BA|UP], IN [BA|UP] SEA [BA|UP],
durante IL [BA|UP] SEA [BA|UP], entro IL [BA|UP] SEA [BA|UP], per
IL [BA|UP] SEA [BA|UP], fino A [BA|UP] SEA [BA|UP], A PT DI [BA|UP]
SEA [BA|UP], metà SEA
```

Momento, periodo, tempo - DS

```
[in] ogni|QUESTO|UN|qualche DS, per un|qualche DS, fino a un|qualche
DS fa, IN DS [stesso] [in cui]
```

Momenti, periodi, tempi - DSP

```
[in] ogni|QUESTO|UN|qualche [BA] DS, per un|qualche DS, fino a
un|qualche DS fa, IN DS [stesso] [in cui], fino A [BA] DS
```

Alba... tramonto - MOM

```
A MOM, all'ora DI MOM, al momento DI MOM, durante IL MOM,
sull'imbrunire, verso IL MOM, dopo IL MOM, fino AL MOM
```

Momenti_pasti - EP

```
a EP, all'ora DI EP, nel momento DI EP, prima DI EP, dopo IL EP
```

Feste - HOL

```
a HOL, prima DI [BA|UP] HOL [BA|UP], dopo [IL] HOL, durante IL
HOL, per [IL] [BA|UP] HOL [BA|UP], entro [IL] [BA|UP] HOL [BA|UP],
A PT DI [BA|UP] HOL [BA|UP]
```

Mezzogiorno, mezzodì, mezzanotte - TN

```
a TN, dopo TN, prima di TN, per TN, entro TN, fino a TN
```

Una volta definiti i sintagmi, si procedere raggruppando tutti quelli che presentano la stessa struttura e creando, poi, le opportune RULEs per il loro riconoscimento (le RULEs possono, dapprima, essere espresse nello stesso pseudo linguaggio, per facilitarne la successiva scrittura nel formato XML).

2) Il secondo metodo consiste nell'elencare le domande alle quali tipicamente rispondono i *complementi di tempo* e le *subordinate temporali*; in una fase successiva vengono individuati i sintagmi che possono dare risposta a tali quesiti. Si può, anche in questo caso, utilizzare in un primo momento lo pseudo-linguaggio, per poi passare alla stesura delle RULEs in XML.

Per maggiore chiarezza, viene proposto un esempio nel quale l'espressione in pseudo-linguaggio è evidenziata in grassetto ed è seguita da tre esempi esplicativi riportati in corsivo:

UN|qualche|QT	**DP|DU|TU**	**prima**
una	*notte*	*prima*
qualche	*ora*	*prima*
un	*anno*	*prima*

Si riporta, dui seguito, un elenco esemplificativo di domande seguite da relative risposte codificate nello pseudo-linguaggio precedentemente descritto.

Per quanto tempo?

```
per [N|QTP] DPP|DUP|TUS, per UN|qualche|UC DP|DU|TU, per ore
e ore, per TT IL DU|DP|TU|DUP|TUP|DPP
```

In quanto tempo?

```
In N|QTP DPP|DUP|TUP
```

Quanto tempo dopo?

```
[dopo] N|QTP|IL|UN DPP|DUP|TUP, Num|QTP DPP|DUP|TUP [dopo],
[dopo] qualche|UN|QUESTO DU|DP|TU [dopo]
```

Quanto tempo prima?

```
N|QTP DPP|DUP|TUP prima, UN|qualche|QT DP|DU|TU prima
```

Fra quanto tempo?

```
fra N|QTP DPP|DUP|TUP, fra qualche|UN DP|DU|TU
```

Da quanto tempo?

```
da N|QTP DPP|DUP|TUP, da qualche|UN DP|DU|TU
```

Quanto tempo fa?

```
UN|qualche DP|DU|TU fa, N|QTP DPP|DUP|TUP fa
```

Entro quanto tempo?

```
entro qualche|UN|QT DP|DU|TU, entro N|QTP DPP|DUP|TUP
```

Per quando?

```
per [IL BA] DY [BA], per IL|QTP DP, per IL [BA] DY|SEA|MS|DU|DP,
per GG, per IL [UP] DP|SEA [BA], per [IL PT|BA]  DP|DU|SEA|MS|HOL
```

Fino a quando?

```
Fino A DP, fino A [BA] DY|SEA|DP|DU|MS, fino A UP DP DI MS,
DP DI SEA, fino A PT DI HOL|SEA|MS
```

Quando?

```
a|verso HOL, a|in|verso [IL] SEA, GG, dopo EP|TN|MS, dopo IL
SEA|MOM, IN DPP|DUP ASEA, IN [N] [BA], DY|SEA|DP|DU|MS,
IN SEA|DP|DU|MS[BA], A MOM|EP|TN|MT, DI SEA, DI DY|DP, prima
DI MOM|SEA|MS, durante IL DP|DU|EP, entro GG|MT|TN, MOM A EP,
IN ACM DI SEA|DU|DP|HOL|MS, UN TU
```

Applicando questi due metodi è possibile individuare un buon numero dei sintagmi che si presentano più frequentemente nei testi scritti. La revisione delle RULEs è stata condotta, perciò, sulla base dei risultati di questa fase analitica. Non è stata, tuttavia, effettuata una vera e propria scelta tra i metodi proposti: è stata piuttosto preferita una combinazione dei due procedimenti, soprattutto a causa del poco tempo a disposizione per l'implementazione.

L'impostazione dell'implementazione è la seguente: alcune RULEs contenute nel programma originale sono state conservate, altre sono state modificate e altre ancora completamente riscritte. Vediamo ora quali sono, nel concreto, i cambiamenti.

I nomi della maggior parte delle RULEs sono stati cambiati per rendere più immediata l'associazione tra ogni singola RULE, il lessico esterno e il risultato della RULE stessa.

Si è, poi, osservato che le RULEs create per definire date, indicazioni di secoli ed espressioni che rappresentano orari, essendo piuttosto semplici e lineari, non sono interessate da analisi sintattiche particolari; sono state, dunque, in gran parte conservate, salvo qualche eccezione.

Riportiamo un esempio delle modifiche, estratto dalle RULEs relative alle date. La sequenza prodotta dalla prima regola è del tipo "giorno 30" (come indicato nel commento che la precede), quella prodotta dalla seconda è del tipo "sabato 30":

```
<!-- giorno 30-->
<RULE name="namegiorno-data" targ_sg="TIMEX[TYPE='DATE']">
 <REL match="W/\#~^giorno$"></REL>
 <REL match="W/#~^([1-9]|1[0-9]|2[0-9]|3[01])$"></REL>
</RULE>
```

```
<!-- giovedi 30-->
<RULE name="giorno-data" targ_sg="TIMEX[TYPE='DATE']">
 <REL type="REF" match="nomi-giorni-lex"></REL>
 <REL match="W/#~^([1-9]|1[0-9]|2[0-9]|3[01])$"></REL>
</RULE>
```

Si propone di costruire una nuova regola per definire la *data* (i.e. 1-31); tale RULE verrà riutilizzata poi all'interno delle due regole precedenti:

```
<!-- 30 -->
<RULE name="data" targ_sg="TIMEX[TYPE='DATE']">
 <REL match="W/#~^([1-9]|1[0-9]|2[0-9]|3[01])$"></REL>
</RULE>

<!-- giorno 30-->
<RULE name="giorno_data" targ_sg="TIMEX[TYPE='DATE']">
 <REL match="W/#~^giorno$"></REL>
 <REL type="REF" match="data"></REL>
</RULE>

<!-- giovedi 30-->
<RULE name="dy_data" targ_sg="TIMEX[TYPE='DATE']">
<REL type="REF" match="dy"></REL>
<REL type="REF" match="data"></REL>
</RULE>
```

Con il metodo appena presentato è stato incrementato di una unità il numero delle RULEs, in apparenza senza alcun vantaggio; in realtà, si è guadagnato in *modularità* e *leggibilità*.

Le RULEs create per definire i gruppi nominali e gruppi preposizionali basati sul lessico sono state quasi tutte scritte *ex-novo*, sulla base dell'analisi effettuata in precedenza; le stesse vengono riportate di seguito, seguite dalla corrispondente espressione in pseudo-linguaggio e dai relativi commenti.

```
<RULE name="avv_loc" type="DISJ">
<REL type="REF"  match="loc"></REL>
<REL type="REF"  match="avv"></REL>
</RULE>
```

loc|avv (individua le locuzioni e gli avverbi presenti in timex2.lex).

```
<RULE name="dp_tu_du" targ_sg="TIMEX[TYPE='DATE']">
 <REL type="GROUP" match="DISJ" m_mod="QUEST">
<REL match="&UN;" m_mod="QUEST"></REL>
<REL match="&IL;" m_mod="QUEST"></REL>
<REL match="W/#~^qualche$" m_mod="QUEST"></REL>
<REL type="REF"  match="qt" m_mod="QUEST"></REL>
<REL match="&QUESTO;" m_mod="QUEST"></REL>
 </REL>
 <REL type="GROUP" match="DISJF">
<REL type="REF"  match="dp"></REL>
<REL type="REF"  match="tu"></REL>
<REL type="REF"  match="du"></REL>
<REL match="W/#~^tempo$"></REL>
 </REL>
</RULE>
```

[UN|IL|qualche|qt|QUESTO] dp|||tu|du||tempo

```
<RULE name="dpp_tup_dup" targ_sg="TIMEX[TYPE='DATE']">
  <REL type="GROUP" match="DISJ" m_mod="QUEST">
<REL match="IL;"></REL>
<REL match="&CARD;"></REL>
<REL match="&QUANT;"></REL>
<REL type="REF" match="qtp"></REL>
```

```
<REL match="&QUESTO;"></REL>
</REL>
  <REL type="GROUP" match="DISJF">
<REL type="REF"  match="dpp"></REL>
<REL type="REF"  match="tup"></REL>
<REL type="REF"  match="dup"></REL>
   </REL>
</RULE>
```

[IL|CARD|QUANT|qtp|QUESTO] dpp||tup||dup

```
 <RULE name="tempo" targ_sg="TIMEX[TYPE='DATE']">
        <REL match="&QUESTO;" m_mod="QUEST"></REL>
<REL type="GROUP" match="DISJ" m_mod="QUEST">
<REL match="&UN;"></REL>
<REL match="&IL;"></REL>
<REL match="W/#~^qualche$"></REL>
<REL type="REF"  match="qt"></REL>
<REL match="&QUESTO;"></REL>
<REL type="REF"  match="up"></REL>
<REL type="REF"  match="ba"></REL>
        </REL>
   <REL type="GROUP" match="DISJ">
<REL type="REF"  match="dy"></REL>
<REL type="REF"  match="gg"></REL>
<REL type="REF"  match="ms"></REL>
<REL type="REF"  match="ds"></REL>
<REL type="REF"  match="hol"></REL>
<REL type="REF"  match="sea"></REL>
<REL type="REF"  match="dp_tu_du"></REL>
</REL>
   <REL type="GROUP" match="DISJF" m_mod="QUEST">
```

```
<REL type="REF"  match="ba"></REL>
<REL type="REF"  match="up"></REL>
<REL type="REF" match="asea"></REL>
         </REL>
 </RULE>
```

[UN|IL|qualche|qt|QUESTO|up|ba] dy|gg|ms|ds|hol|sea|dp_tu_du [ba||up||asea]

```
  <RULE name="tempi" targ_sg="TIMEX[TYPE='DATE']">
<REL match="&QUESTO;" m_mod="QUEST"></REL>
   <REL type="GROUP" match="DISJ" m_mod="QUEST">
<REL match="IL;"></REL>
<REL type="REF" match="qtp"></REL>
<REL type="REF" match="nums"></REL>
<REL match="&QUESTO;"></REL>
<REL type="REF" match="upp"></REL>
<REL type="REF" match="bap"></REL>
        </REL>
   <REL type="GROUP" match="DISJ">
<REL type="REF"  match="dsp"></REL>
<REL type="REF"  match="dyp"></REL>
<REL type="REF"  match="dpp_tup_dup"></REL>
        </REL>
   <REL type="GROUP" match="DISJF" m_mod="QUEST">
<REL type="REF" match="bap"></REL>
<REL type="REF" match="upp"></REL>
<REL type="REF" match="aseap"></REL>
        </REL>
 </RULE>
```

[IL |qtp|nums|QUESTO|up|ba] dsp|dyp|dpp_tup_dup [bap||upp||aseap]

```
 <RULE name="vari_tempi" type="DISJ">
<REL type="REF" match="tempo"></REL>
<REL type="REF" match="tempi"></REL>
<REL type="REF" match="parte_giorno"></REL>
<REL type="REF"  match="date"></REL>
<REL type="REF"  match="decadi"></REL>
<REL type="REF"  match="date_secoli"></REL>
  </RULE>
```

tempo|tempi|parte_giorno|date|decadi|date_secoli

```
  <RULE name="fra_tempo" targ_sg="TIMEX[TYPE='DATE']">
   <REL type="GROUP" match="DISJF">
  <REL match="W/#~^tra$"></REL>
  <REL match="W/#~^fra$"></REL>
        </REL>
  <REL match="W/#~^circa$" m_mod="QUEST"></REL>
<REL match="&UN;" m_mod="QUEST"></REL>
<REL type="GROUP" match="DISJF">
<REL type="REF"  match="vari_tempi"></REL>
        </REL>
  <REL match="W/#~^circa$" m_mod="QUEST"></REL>
 </RULE>
```

tra||fra [circa][UN] vari_tempi [circa]

```
  <RULE name="da_tempo" targ_sg="TIMEX[TYPE='DATE']">
  <REL match="&DA;"></REL>
<REL type="GROUP" match="DISJF" m_mod="QUEST">
       <REL match="W/#~^circa$" m_mod="QUEST"></REL>
       <REL match="W/#~^quasi$" m_mod="QUEST"></REL>
       <REL match="W/#~^ormai$" m_mod="QUEST"></REL>
```

```
        </REL>
   <REL type="GROUP" match="DISJF">
<REL type="REF"  match="vari_tempi"></REL>
<REL type="REF" match="ore"></REL>
<REL type="REF" match="date"></REL>
        </REL>
  <REL match="W/#~^circa$" m_mod="QUEST"></REL>
  <REL match="W/#~^ormai$" m_mod="QUEST"></REL>
 </RULE>
```

DA [circa||quasi||ormai] vari_tempi||ore||date [circa][ormai]

```
  <RULE name="entro_tempo" targ_sg="TIMEX[TYPE='DATE']">
<REL type="REF"  match="eo"></REL>
  <REL match="W/#~^circa$" m_mod="QUEST"></REL>
  <REL match="&IL;" m_mod="QUEST"></REL>
   <REL type="GROUP" match="DISJ">
<REL type="REF"  match="vari_tempi"></REL>
<REL type="REF" match="ore"></REL>
<REL type="REF" match="mom"></REL>
<REL type="REF" match="ep"></REL>
<REL type="REF" match="tn"></REL>
        </REL>
  <REL match="W/#~^fa$" m_mod="QUEST"></REL>
  <REL match="W/#~^circa$" m_mod="QUEST"></REL>
 </RULE>
```

eo [circa][IL] vari_tempi|ore|mom|ep|tn [fa][circa]

```
  <RULE name="tempo_fa" targ_sg="TIMEX[TYPE='DATE']">
<REL type="REF"  match="vari_tempi"></REL>
  <REL match="W/#~^fa$"></REL>
```

```
 </RULE>
```

vari_tempi fa

```
  <RULE name="verso_tempo" targ_sg="TIMEX[TYPE='DATE']">
  <REL match="W/#~^verso$"></REL>
<REL match="&IL;" m_mod="QUEST"></REL>
   <REL type="GROUP" match="DISJ">
<REL type="REF"  match="vari_tempi"></REL>
<REL type="REF" match="ore"></REL>
<REL type="REF" match="mom"></REL>
<REL type="REF" match="ep"></REL>
<REL type="REF" match="tn"></REL>
        </REL>
 </RULE>
```

verso [IL] vari_tempi|ore|mom|ep|tn

```
  <RULE name="in_tempo" targ_sg="TIMEX[TYPE='DATE']">
  <REL match="&IN;" m_mod="QUEST"></REL>
  <REL match="W/#~^circa$" m_mod="QUEST"></REL>
<REL type="REF"  match="vari_tempi"></REL>
  <REL match="W/#~^circa$" m_mod="QUEST"></REL>
 </RULE>
```

[IN][circa] vari_tempi [circa]

```
  <RULE name="prima_dopo" targ_sg="TIMEX[TYPE='DATE']">
<REL type="REF"  match="vari_tempi"></REL>
   <REL type="GROUP" match="DISJF">
<REL match="W/#~^dopo$"></REL>
<REL match="W/#~^prima$"></REL>
```

```
        </REL>
 </RULE>
```

vari_tempi dopo||prima

```
 <RULE name="dopo_tempo" targ_sg="TIMEX[TYPE='DATE']">
  <REL match="W/#~^non$" m_mod="QUEST"></REL>
   <REL type="GROUP" match="DISJF">
  <REL match="W/#~^più tardi$"></REL>
  <REL match="W/#~^prima$"></REL>
  <REL match="W/#~^dopo$"></REL>
  <REL match="W/#~^non più tardi$"></REL>
  <REL match="W/#~^non prima$"></REL>
  <REL match="W/#~^non dopo$"></REL>
</REL>
  <REL match="&DI;" m_mod="QUEST"></REL>
  <REL match="&IL;" m_mod="QUEST"></REL>
  <REL match="&UN;" m_mod="QUEST"></REL>
   <REL type="GROUP" match="DISJF">
<REL type="REF"  match="vari_tempi"></REL>
<REL type="REF" match="ore"></REL>
<REL type="REF" match="mom"></REL>
<REL type="REF" match="ep"></REL>
<REL type="REF" match="tn"></REL>
        </REL>
 </RULE>
```

[non] (più tardi)||prima||dopo [DI][IL][UN] vari_tempi||ore||mom||ep||tn

```
 <RULE name="fino_a" targ_sg="TIMEX[TYPE='DATE']">
   <REL type="GROUP" match="DISJF" m_mod="QUEST">
  <REL match="W/#~^fino$"></REL>
```

```
  <REL match="W/#~^intorno$"></REL>
        </REL>
  <REL match="&A;" m_mod="QUEST"></REL>
  <REL match="W/#~^circa$" m_mod="QUEST"></REL>
   <REL type="GROUP">
   <REL type="GROUP" m_mod="QUEST">
<REL match="W/#~^ora$"></REL>
<REL match="&DI;"></REL>
        </REL>
   <REL type="GROUP" match="DISJ">
<REL type="REF" match="vari_tempi"></REL>
<REL type="REF" match="mom"></REL>
<REL type="REF" match="ore"></REL>
<REL type="REF" match="ep"></REL>
<REL type="REF" match="tn"></REL>
<REL type="REF" match="dopo_tempo"></REL>
        </REL>
        </REL>
  <REL match="W/#~^fa$" m_mod="QUEST"></REL>
  <REL match="W/#~^circa$" m_mod="QUEST"></REL>
 </RULE>
```

[fino||intorno][A][circa] ([(ora DI)] vari_tempi|ore|mom|ep|tn|dopo_tempo) [fa][circa]

```
  <RULE name="di_tempo" targ_sg="TIMEX[TYPE='DATE']">
<REL match="&DI;" m_mod="QUEST"></REL>
   <REL type="GROUP" match="DISJF">
<REL type="REF" match="dy"></REL>
<REL type="REF" match="dp"></REL>
<REL type="REF" match="sea"></REL>
<REL type="REF" match="ms"></REL>
<REL type="REF" match="du"></REL>
```

```
<REL type="REF" match="gg"></REL>
<REL type="REF" match="vari_tempi"></REL>
    </REL>
 </RULE>
```

[DI] dy||dp||sea||ms||du||gg||vari_tempi

```
 <RULE name="ogni_tempo" targ_sg="TIMEX[TYPE='DATE']">
<REL match="W/#~^ogni$"></REL>
<REL type="REF" match="vari_tempi"></REL>
 </RULE>
```

ogni vari_tempi

```
 <RULE name="durante" targ_sg="TIMEX[TYPE='DATE']">
<REL match="W/#~^durante$"></REL>
  <REL match="&IL;"></REL>
   <REL type="GROUP" match="DISJF">
<REL type="REF" match="mom"></REL>
<REL type="REF" match="ep"></REL>
<REL type="REF" match="tn"></REL>
<REL type="REF" match="vari_tempi"></REL>
<REL match="W/#~^settimana$"></REL>
        </REL>
 </RULE>
```

durante IL mom||ep||tn||vari_tempi||settimana

```
 <RULE name="per_tempo" targ_sg="TIMEX[TYPE='DATE']">
<REL match="W/#~^per$"></REL>
<REL type="REF" match="vari_tempi"></REL>
 </RULE>
```

per vari_tempi

```
<RULE name="nell_arco" targ_sg="TIMEX[TYPE='DATE']">
<REL match="&IN;"></REL>
   <REL type="GROUP" match="DISJF">
<REL match="W/#~^arco$"></REL>
<REL match="W/#~^corso$"></REL>
<REL match="W/#~^mezzo$"></REL>
        </REL>
  <REL match="&DI;" m_mod="QUEST"></REL>
<REL type="GROUP" match="DISJF">
<REL type="REF"  match="vari_tempi"></REL>
<REL type="REF"  match="di_tempo"></REL>
        </REL>
 </RULE>
```

IN arco||corso||mezzo [DI] vari_tempi||di_tempo

```
<RULE name="all_inizio" targ_sg="TIMEX[TYPE='DATE']">
<REL match="&A;"></REL>
   <REL type="GROUP" match="DISJF">
<REL match="W/#~^metà$"></REL>
<REL match="W/#~^fine$"></REL>
<REL match="W/#~^termine$"></REL>
<REL match="W/#~^inizio$"></REL>
<REL match="W/#~^principio$"></REL>
        </REL>
  <REL match="&DI;" m_mod="QUEST"></REL>
<REL type="GROUP" match="DISJF">
<REL type="REF"  match="vari_tempi"></REL>
<REL type="REF"  match="di_tempo"></REL>
```

```
        </REL>
 </RULE>
```

A metà||fine||termine||inizio||principio [DI] vari_tempi||di_tempo

```
<RULE name="all" type="DISJ">
<REL type="REF" match="avv_loc"></REL>
<REL type="REF" match="agg"></REL>
<REL type="REF" match="prima_dopo"></REL>
<REL type="REF" match="fra_tempo"></REL>
<REL type="REF" match="da_tempo"></REL>
<REL type="REF" match="entro_tempo"></REL>
<REL type="REF" match="tempo_fa"></REL>
<REL type="REF" match="fino_a"></REL>
<REL type="REF" match="verso_tempo"></REL>
<REL type="REF" match="in_tempo"></REL>
<REL type="REF" match="dopo_tempo"></REL>
<REL type="REF" match="di_tempo"></REL>
<REL type="REF" match="ogni_tempo"></REL>
<REL type="REF" match="durante"></REL>
<REL type="REF" match="per_tempo"></REL>
<REL type="REF" match="g_m_a"></REL>
<REL type="REF" match="nell_arco"></REL>
<REL type="REF" match="all_inizio"></REL>
<REL type="REF" match="ore"></REL>
</RULE>
```

avv_loc|agg|prima_dopo|fra_tempo|da_tempo|entro_tempo|tempo_fa|fino_a|
verso_tempo|in_tempo|dopo_tempo|di_tempo|in_tempo|ogni_tempo|durante|
per_tempo|g_m_a|nell_arco|all_inizio|ore

3.3 Confronto fra il programma originale e la nuova versione

3.3.1 Prestazioni a confronto

Il calcolo delle prestazioni è stato rilevato in termini di riconoscimenti e di errori compiuti.

I due programmi sono stati testati su una dozzina di documenti medio-brevi di diversi generi (tematiche storiche, biografie, racconti di vita) dai quali, in una fase preliminare, sono state estratte manualmente le espressioni temporali. Successivamente, sono stati esaminati i file in uscita, raccogliendo i dati relativi alle espressioni riconosciute correttamente e a quelle erroneamente etichettate. Si precisa che, in certi casi, sono stati considerati validi anche i riconoscimenti parziali (e.g. sono considerate come riconosciute le espressioni del tipo "in quei ***primi giorni***", in cui sia stato estratto il sintagma nominale che costituisce il nucleo dell'espressione). I dati in questo modo estratti sono riportati nelle tabelle sottostanti. Essi consentono di ricavare i seguenti parametri (analoghi a quelli valutati nelle competizioni B.4):

recall = rapporto tra il numero di espressioni *correttamente individuate* dal programma e il numero totale delle espressioni temporali effettivamente presenti nel testo;

precisione = rapporto tra il numero delle espressioni *correttamente individuate* e il numero totale delle espressioni individuate.

Com'è evidente, i risultati sono tanto migliori quanto maggiori sono *recall* e *precisione*.

Riportiamo, di seguito, i risultati conseguiti da ciascuno dei programmi nelle varie prove.

1.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	1/14	0/13	7%	100%
modificato	12/14	1/13	86%	92%

2.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	9/22	0/9	40.9%	100%
modificato	21/22	0/21	95.45%	100%

3.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	5/6	0/5	83.33%	100%
modificato	6/6	0/6	100%	100%

4.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	16/47	1/16	31.9%	93.75%
modificato	47/47	4/47	91.48%	91.5%

5.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	4/21	0/4	19%	100%
modificato	19/21	2/19	90.47%	89.48%

6.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	1/6	0/1	16.66%	100%
modificato	3/6	1/3	50%	66.66%

7.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	0/10	0/10	0%	100%
modificato	9/10	0/9	90%	100%

8.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	12/20	0/12	60%	100%
modificato	17/20	0/17	85%	100%

9.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	9/21	2/11	42.85%	81.82%
modificato	19/21	2/19	90.47%	90.48%

10.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	10/31	1/31	32.25%	91%
modificato	29/31	5/10	93.54%	85.3%

11.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	19/46	1/19	41.3%	94.8%
modificato	39/46	1/40	84.78%	97.5%

12.

Programma	*recall*	*precisione*	*recall*(%)	*precisione*(%)
originale	13/41	0/13	31.7%	100%
modificato	38/41	1/39	92.68%	97.44%

Le prestazioni dei due programmi vengono valutate, infine, come medie percentuali complessive dei valori di *recall* e *precisione* raggiunti nelle singole prove:

Programma	*recall*(%)	*precisione*(%)
originale	34%	97%
modificato	87%	93%

3.4 Verifica degli obiettivi

Gli obiettivi inizialmente prefissati prevedevano un miglioramento del programma in termini di quantità di espressioni temporali individuate e un contenimento del numero delle RULEs definite. Si verificano, di seguito, i risultati ottenuti.

3.4.1 Contenimento del numero delle RULEs

Un semplice conteggio permette di constatare che il numero totale delle RULEs definite nel file timex2.gr ristrutturato (pari a 102) supera di pochissimo quello delle RULEs originali (pari a 100).

Uno dei risultati desiderati, dunque, è stato ottenuto.

3.4.2 Prestazioni del programma

Dai risultati ottenuti nelle prove si può notare che il programma modificato guadagna, rispetto a quello originale, in termini di *recall*, ma presenta un minore grado di *precisione*. Un maggior *recall* non corrisponde necessariamente a un effettivo miglioramento e il valore "reale" del parametro; nel complesso, pertanto, è da considerarsi leggermente inferiore a quello precedentemente calcolato. D'altra parte, l'aumento percentuale dei fallimenti è molto inferiore a quello dei riconoscimenti corretti, perciò si ritiene possa essere considerato "assorbito" da quest'ultimo. I valori percentuali medi ottenuti sono: *recall* = 87%, *Precisione* = 93%.

Si può ritenere conseguito anche il secondo obiettivo.

Nel paragrafo seguente vengono esposte alcune osservazioni che possono essere utili per future revisioni.

3.4.3 Osservazioni...

... sulla stesura del programma. La fase di progettazione del programma ha evidenziato il seguente fatto: seguire con un certo rigore una metodologia di sviluppo può essere una scelta efficace, in quanto, se condotta in modo adeguato, può permettere di ottenere un'implementazione caratterizzata da un buon grado di *modularità*, quindi più *flessibile* e più *ordinata* (perciò anche leggibile con minor sforzo). Uno svantaggio potrebbe essere, tuttavia, la generazione di molte RULEs che potrebbe creare difficoltà di gestione delle stesse. La decisione -qui presa per motivi di tempo- di non abbracciare rigorosamente una specifica metodologia porta alla realizzazione di un programma dotato di minore precisione: il numero delle RULEs è limitato, ma le stesse sono più complesse e tendono a presentarsi sovrapposizioni tra i risultati delle loro elaborazioni; ciò comporta il rischio di "conflitti" difficilmente controllabili (e.g. in certi casi, più di una RULE è in grado di estrarre dal testo la stessa sequenza di termini; questo complica l'individuazione e la gestione di eventuali problemi).

... sulle espressioni riconosciute. Com'è evidente, il file lessicale è stato ampliato di molto; è stato scelto di introdurre, in particolare, avverbi e locuzioni temporali che nel complesso sono presenti in numero considerevole. Lo standard TIMEX2 di cui si è parlato nei paragrafi 1.4 e 1.3.3 (MITRE) effettua, in certi casi, scelte diverse (e.g. non considera come etichettabile l'avverbio *quando*). In questa versione del programma si sceglie di etichettare ogni sintagma dotato di una connotazione temporale.

... sulle espressioni non riconosciute. Gli estratti delle elaborazioni mostrano che entrambi i programmi non sono in grado di riconoscere espressioni dei tipi seguenti:

- *1304-07*: il programma originale riconoscerebbe, invece, 1304-1307, mentre quello nuovo ne perde la capacità;

- *due estati*: il lessico non contiene, infatti, la forma plurale dei nomi delle stagioni;
- *all'inizio*: la RULEs che ne specifica il riconoscimento, richiede che esso sia seguito da un complemento di specificazione (e.g. *all'inizio del secolo*);
- *nel medioevo* , *nella modernità*, *fin dall'antichità* : anche in questo caso, si tratta della mancanza di termini nel file lessicale (*modernità*, *antichità* e *medievo*);
- *in vita mia*, *per tutta la vita*: il lessico non contiene né aggettivi/pronomi possessivi, né la parola vita; potrebbero essere introdotte nel file le intere espressioni come "frasi fatte";
- *per due notti* la motivazione non è chiara poiché ci si aspetterebbe il riconoscimento da parte della RULE per_tempo; il caso necessita diun'indagine.

. . . sul confronto fra i due programmi. Può essere interessante confrontare le prestazioni dei programmi non solo in termini di *recall* e *precisione* dell'estrazione di informazione, ma anche in relazione alle tipologie di espressioni individuate. In generale, il nuovo programma riconosce un maggior numero di espressioni grazie al lessico più esteso e all'implementazione più completa e strutturata delle regole grammaticali inerenti alla lingua italiana.

. . . su eventuali sviluppi futuri. Come visto, il nuovo programma presenta alcuni problemi da risolvere relativi all'individuazione delle espressioni. Un eventuale sviluppo futuro, potrebbe prevedere la risoluzione di questi, mediante un ulteriore ampliamento del lessico e una strutturazione dele RULEs caratterizzata da maggiore modularità.

Sarebbe anche molto interessante definire una classificazione delle espressioni temporali (e.g. suddividendole in tre classi, per la rappresentazione di: date, intervalli di tempo, indicazione sulla frequenza degli eventi) e una modalità di etichettatura basata sulla classificazione stessa.

3.5 Esempio di elaborazione

Per concludere la presentazione del progetto sviluppato, si illustrano i risultati relativi a un'applicazione concreta del programma, facendo riferimento ad una fase del test precedente. Saranno elencati e riprodotti (a titolo di esempio):

- il file di testo in ingresso;

- la lista delle espressioni temporali *individuate manualmente*;

- la visualizzazione del file HTML che il programma fornisce in uscita;

- il file di testo etichettato -file HTML sorgente- (le espressioni temporali sono indivuate dalle etichette.

3.5.1 Il file in ingresso

Il file d'ingresso del programma è un semplice file di testo di nome *paragra*, contenuto in una directory denominata *works* esterna alla directory del sistema TTT. Nella stessa *works* dev'essere posto anche il programma costituito dalla pipeline di comandi dedicati all'elaborazione; questo file eseguibile, nel nostro caso, è chiamato *gt*. Per elaborare un testo scelto a piacimento è necessario copiare e salvare il testo stesso nel file *paragra*, e mandare poi in esecuzione il programma dalla directory in cui è collocato.

Viene riportato, come esempio, una parte di un documento utilizzato nei test effettuati sui programmi.

"La teoria dell'evoluzione, una delle scoperte scientifiche che hanno influito più profondamente sulla cultura moderna e sulla concezione del mondo dell'uomo contemporaneo, fu concepita e messa a punto, nelle sue linee essenziali, da Charles Darwin nel corso dell'Ottocento, in un periodo di grandi progressi nelle scienze della natura. Nel XVIII secolo diversi scienziati e filosofi avevano cominciato a mettere in discussione la concezione di un mondo immutabile: ad esempio Kant nel 1775 aveva formulato l'ipotesi secondo cui il sistema solare trarrebbe origine dal moto vorticoso di una nebulosa primitiva; quest'ipotesi sarà ripresa alla fine del secolo da Pierre Simon de Laplace. Nella seconda metà del Settecento viaggi, spedizioni scientifiche sistematiche ed esplorazioni, seppur motivate principalmente da scopi commerciali, avevano dato un forte impulso alla ricerca in campo biologico e fatto nascere la paleontologia e la geologia che, con gli studi di Charles Lyell e Georges Cuvier, avevano rivelato strati geologici formatisi in tempi successivi, che incorporavano i resti di specie animali e vegetali ormai scomparse da tempo dalla Terra."
(http://www.cicap.org/enciclop/at100257.htm)

3.5.2 Le espressioni temporali

Per poter valutare le prestazioni dei programmi, durante la fase dei test è necessario estrarre manualmente le espressioni temporali dei testi in esame. Nel caso in questione, si possono individuare le seguenti voci:

nel corso dell'Ottocento	in un periodo
Nel XVIII secolo	nel 1775
alla fine del secolo	Nella seconda metà del Settecento
in tempi successivi	ormai
da tempo	

3.5.3 Il testo etichettato

Il file fornito in ingresso, una volta elaborato, viene restituito in uscita in formato HTML, nella directory *$TTT\OUTPUT\HTML*, con il nome indicato nello stesso programma *gt*. Per quanto ci riguarda, il nome assegnato è *out.html*. L'aspetto del file sorgente HTML è del tipo rappresentato sotto. Le uniche etichette individuabili nel file sono (o, meglio, dovrebbero essere) le sole etichette relative alle espressioni di tempo, dunque quelle in cui all'attributo CLASS sono assegnati valori DATE, TIME o DURATION. Si noti che, all'inizio del file, è riportato il contenuto di *generaltrans*, di cui si è detto sopra.

Sono riportati in corsivo alcuni commenti.

<HTML>
<HEAD>
<TITLE>TTT Output</TITLE>
<STYLE>
P {white-space: pre}
H2 {color:black}
SPAN.PHR-CD {background:FFFFFF}
SPAN.WRD-CD {background:FFFFFF}
SPAN.PHR-ORD {background:CCCCFF}
SPAN.WRD-ORD {background:FFFFFF}
SPAN.PHR-FRAC {background:CCCCFF}
SPAN.WRD-FRAC {background:CCCCFF}
SPAN.PHR-FRACORD {background:CCCCFF}
SPAN.WRD-FRACORD {background:FFFFFF}
SPAN.PHR-RANGE {background:FFFFFF}
SPAN.PHR-QUANT {background:CCCCFF}
SPAN.DATE {font-weight:bolder; font-style: italic}
*definisce lo stile **grassetto e corsivo** per le espressioni temporali*
SPAN.TIME {font-weight:bolder; font-style: italic}
questo elemento e il precedente potrebbero essere utilizzati per ottenere aspetti di-

versi per diverse etichettature, qualora venisse definita una classificazione delle espressioni
SPAN.DURATION {font-weight:bolder; font-style: italic; background:FFCCCC}
questo elemento e il precedente potrebbero essere utilizzati per ottenere aspetti diversi per diverse etichettature, qualora venisse definita una classificazione delle espressioni
SPAN.MONEY {background:FFFFFF}
SPAN.PERCENT {background:FFFFFF}
</STYLE>
</HEAD>
<BODY>

<P>La teoria dell'evoluzione, una delle scoperte scientifiche che hanno influito più profondamente sulla cultura moderna e sulla concezione del mondo dell'uomo contemporaneo, fu concepita e messa a punto, nelle sue linee essenziali, da Charles Darwin <SPAN CLASS='DATE'>nel corso dell'Ottocento</SPAN>, in un <SPAN CLASS='DATE'>periodo</SPAN> di grandi progressi nelle scienze della natura. <SPAN CLASS='DATE'>Nel XVIII secolo</SPAN> diversi scienziati e filosofi avevano cominciato a mettere in discussione la concezione di un mondo immutabile: ad esempio Kant <SPAN CLASS='DATE'>nel 1775</SPAN> aveva formulato l'ipotesi <SPAN CLASS='DATE'>secondo</SPAN> cui il sistema solare trarrebbe origine dal moto vorticoso di una nebulosa primitiva; quest'ipotesi sarà ripresa <SPAN CLASS='DATE'>alla fine del secolo</SPAN> da Pierre Simon de Laplace. Nella <SPAN CLASS='DATE'>seconda metà del Settecento</SPAN> viaggi, spedizioni scientifiche sistematiche ed esplorazioni, seppur motivate principalmente da scopi commerciali, avevano dato un forte impulso alla ricerca in campo biologico e fatto nascere la paleontologia e la geologia che, con gli studi di Charles Lyell e Georges Cuvier, avevano rivelato strati geologici formatisi <SPAN CLASS='DATE'>in tempi successivi</SPAN>, che incorporavano i resti di specie animali e vegetali <SPAN CLASS='DATE'>ormai</SPAN> scomparse <SPAN CLASS='DATE'>da tempo</SPAN> dalla Terra.</P>
</BODY>
</HTML>

3.5.4 Il file in uscita

Visualizzando il file *out.html* mediante un browser si ottiene il risultato finale dell'elaborazione presentato nell'esempio 3.1: le espressioni temporali trovate sono evidenziate in ***grassetto corsivo***.

E' opportuno considerare che è possibile personalizzare la presentazione dei risultati apportando adeguate modifiche al sistema. Il file utilizzato per reindiriz-

James Augustine Aloysius Joyce, uno dei più grandi autori di narrativa ***di questo secolo***, nasce a Rathgar, una frazione di Dublino, il ***2 febbraio 1882***. Appartiene ad una famiglia della buona società di Dublino, le cui condizioni finanziarie vanno però via via declinando fino al punto che l'indigenza lambisce la famiglia Joyce in modo preoccupante. I suoi genitori lo iscrivono ad una scuola cattolica, precisamente presso un istituto di gesuiti, il Clongowes Wood College (ma studierà anche al Belvedere College, ***sempre*** di proprietà dei gesuiti). ***Successivamente***, iscrittosi all'università di Dublino, si laurea in lingue moderne. ***In questi anni*** inizia a manifestare un carattere anticonformista e ribelle. Difende con articoli e conferenze il teatro di Ibsen, considerato ***ai tempi*** immorale e sovversivo e, trascinato dalla sua foga idealista, pubblica "Il giorno del Volgo", un pamphlet nel quale si scaglia contro il provincialismo della cultura irlandese.

Figura 3.1: *esempio di output.*

zare l'output, ad esempio, può essere sostituito. In particolare, attraverso il file *generaltrans* (\$TTT\OUTPUT\HTML\) è possibile variare l'aspetto delle espressioni temporali individuate: e.g. possono essere evidenziate con uno sfondo colorato. Nel nostro caso, la scelta dello stile ***grassetto corsivo*** è dovuta all'utilizzo di browser che non riconoscono l'attributo di *background* per l'evidenziatura; evidenziare, tuttavia, può essere molto utile, in quanto consente di verificare se le espressioni selezionate sono riconosciute ciascuna nel suo complesso o, piuttosto, in maniera "spezzettata" (i.e. come più espressioni di minore estensione separate da spazi).

Maggiori informazioni relative alle opzioni possibili sono fornite nel sito del Language Technology Group di Edimburgo, nella pagina contenente la documentazione del sistema LT TTT:
http://www.ltg.ed.ac.uk/software/ttt/tttdoc.html

3.6 Esempi di espressioni temporali riconosciute dal programma

Viene riportata una lista esemplificativa delle espressioni temporali che il programma è in grado di riconoscere.

3 o 4 giorni fa	a dicembre
30 ottobre 1990	a Pasqua
alcuni anni prima	alla fine degli anni ’80
alla fine del secolo	alle ore 15:30
all’alba del giorno dopo	al mattino presto
attorno all’anno 2000	da alcuni giorni
di giorno in giorno	diversi anni fa
ieri pomeriggio all’ora di merenda	durante il mese di maggio
entro e non oltre il 2004	entro qualche giornata
fino a qualche giorno fa	fino allo scorso aprile
fra il 15 aprile e il 16 maggio	dopo alcuni minuti
negli anni ’30 e ’40	l’estate prossima
negli ultimi decenni del Settecento	nel corso del XX secolo
nella prossima stagione	ogni giorno dell’anno
parecchio tempo prima	pochi giorni dopo
sabato 4 ottobre 1970	tanto tempo prima
tre o quattro giorni fa	verso il Natale prossimo
fra l’inizio del pomeriggio e la fine della giornata	fra tre ore
fra il secolo V e il VI secolo d.c.	tra ieri e oggi
dal sabato sera alla domenica mattina	da ieri a oggi
dalle 15 di sabato alla domenica mattina	fra un mese e dodici giorni
dalle ore tre alle ore sette	da Natale a Pasqua
verso gli anni ’50	verso Capodanno
nel 450 a.c.	un’ora fa
tanto tempo prima	subito
quella volta	qualche volta
prima o poi	prima degli anni ’30
per tre o quattro mesi	per domani
ogni volta	ogni ora della notte
ogni 5 anni	oggi a mezzanotte
non prima di marzo	nello stesso tempo
nell’arco del mese	nel fine settimana prossimo

3.7 Conclusioni

In questo lavoro è stato affrontato il problema dell'individuazione di espressioni temporali in testi italiani. Dopo una panoramica relativa all'informazione temporale e all'interesse ad essa rivolto nel campo scientifico, è stata introdotta una formalizzazione dei concetti di *tempo* e di *espressione temporale*.
Sono stati illustrati la caratterizzazione delle espressioni di tempo e i problemi della granularità e dell'indeterminatezza ad esse legati.
E' stato presentato un programma per l'individuazione automatica delle espressioni temporali stesse, il quale consiste in un riadattamento di un altro elaborato; sono state analizzate e confrontate le prestazioni dei due programmi. Si è mostrato che il programma rinnovato, progettato sulla base delle regole grammaticali della lingua italiana, ottiene prestazioni migliori rispetto all'originale.
Futuri sviluppi potrebbero prevedere la risoluzione di alcuni problemi riscontrati e la classificazione delle espressioni temporali.

Appendice A

Richiami di grammatica italiana

Vengono forniti specifici richiami alla grammatica italiana, riguardanti le modalità con cui vengono comunemente espresse le indicazioni tempo. Si farà riferimento alla morfologia e all'analisi sintattica.

A.1 Morfologia

La morfologia è lo "studio della forma" delle parole (dal greco 'morphé', *forma* e 'logos', *studio*): fornisce una classificazione e una descrizione delle forme della parole mediante l'analisi delle variazioni che esse presentano a seconda del significato assunto e della funzione svolta nel contesto in cui sono inserite (*analisi grammaticale*); consente di ottenere una preparazione del testo adeguata ad una successiva analisi di tipo semantico e funzionale.

Le espressioni temporali individuabili in questa fase appartengono all'insieme delle *parti invariabili* del discorso (parti che non variano mai forma, essendo dotate di una forma unica). Appartengono a questa categoria gli avverbi e le locuzioni avverbiali, i quali, nell'analisi logica, assumono valore e significato di *complementi avverbiali di tempo* (rispondono alla domanda: "Quando?"); da un punto di vista sintattico, possono, perciò, essere sostituiti col corrispondente complemento.

A.1.1 Avverbi

La parola *avverbio* (detto anche *modificante*) deriva dal latino dotto *adverbium* (composto di 'ad', *presso*, e di 'verbum', *parola*) e assume il significato di 'parola che sta accanto a un'altra parola'. L'avverbio modifica, determina o precisa il verbo o le altre parti del discorso cui si riferisce.

Gli avverbi di tempo fanno parte degli avverbi determinativi: esprimono una determinazione di tempo e possono indicare il momento, il periodo, la circostanza in cui si verifica un evento.

Esempi:

- ora, ancora, allora, oggi, ieri, domani, già, immediatamente, poi, presto, tardi, stamani, stamane, stasera, stanotte, subito, tosto, talora, talvolta, infine, finalmente, adesso, ormai, oramai, poscia, sempre, spesso, sovente, tuttora, finora, mai, dopodomani, testé e dianzi (di livello letterario);

- *dopo* e *prima* sono avverbi di tempo quando determinano un verbo (in altri casi sono preposizioni, locuzioni preposizionali o congiunzioni);

- avverbi in *-mente*: precedentemente, successivamente, recentemente;

- avverbi interrogativi *quando?*, *da quando?*, *per quando?* sono avverbi relativi se introducono una subordinata interrogativa (=interrogativa indiretta), congiunzioni subordinanti se mettono in relazione due proposizioni introducendo una subordinata temporale.

Alcuni avverbi ammettono il comparativo di maggioranza, di minoranza e di uguaglianza e il superlativo relativo e assoluto; ad esempio: *presto*, più presto, meno presto, prestissimo; *tardi*, più tardi, meno tardi, tardissimo; *spesso*, più spesso, spessissimo; *subito*, subitissimo; *raramente*, più raramente, rarissimamente, meno raramente.

Alcuni avverbi possono altresì presentarsi in forma alterata, ottenuta aggiungendo alla radice dell'avverbio le desinenze del diminutivo (-ino, -etto, -ettino), del

vezzeggiativo (-uccio), dell'accrescitivo (-one) o del peggiorativo (-accio): *tardi*, tardino, tardetto, tarduccio, presto, prestino.

A.1.2 Locuzioni avverbiali

Le locuzioni avverbiali di tempo sono costituite da gruppi di parole usati come frasi fatte che svolgono funzioni avverbiali.

Esempi:

- una volta, un tempo, per sempre, per l'addietro, per ora, d'ora in avanti, da ora in poi, poco fa, fra poco, in futuro, or ora, a volte, di buon'ora, di quando in quando, di ora in ora, di giorno in giorno, di tanto in tanto, di volta in volta, di punto in bianco, all'improvviso;
- locuzioni avverbiali interrogative: da dove?.

Anche le locuzioni avverbiali possono presentare forme alterate (e.g.: fra pochino).

A.2 Sintassi della proposizione

Il secondo livello di analisi di un testo è costituito dall'*analisi sintattica*, la quale riguarda la sintassi delle frasi, ossia l'ordine in cui sono distribuite le parole; viene indicata anche come analisi *logica* (da 'logos', discorso) poiché riguarda il discorso e tenta di cogliere i meccanismi che lo regolano.
A questo livello, la frase viene scomposta in tutti i suoi elementi e viene indagata la funzione sintattica di ciascuno di essi, in modo da evidenziare i rapporti da cui sono tra loro legati.

A.2.1 La frase

Sensini, nella sua grammatica [20], definisce la *frase* come "sequenza unitaria di parole, dotata di significato compiuto, compresa tra due segni di interpunzione forte e caratterizzata dalla presenza di un verbo di forma compiuta."

La frase, in base alla propria struttura, può essere:

semplice (detta anche *proposizione*), quando contiene una sola forma verbale;

complessa (detta anche *periodo*), quando contiene più forme verbali (i.e. è costituita da più frasi semplici).

La frase semplice può presentarsi nella sua forma di base (*frase minima*), costituita da soggetto e predicato, oppure arricchita da *espansioni* che forniscono precisazioni relative al soggetto e/o al predicato. Queste espansioni (cosiddette perché *espandono* il significato degli elementi di base) sono dette anche *determinanti* (perché *determinano* tale significato) o *complementi* (in quanto *completano* il significato tra soggetto e verbo).

Le informazioni temporali sono contenute negli elementi accessori della frase: sono espresse dai *complementi di tempo*.

I sintagmi

Gli elementi che formano una frase possono essera costituiti da singole parole o da insiemi di parole. Ognuno di questi insiemi di parole forma un'*unità sintattica* denominata *sintagma*. I sintagmi si suddividono in due gruppi, in base alla parola che li caratterizza e che fa da nucleo al gruppo:

- *sintagma o gruppo nominale* (GN), che può essere formato da nome (*lunedì*), nome + articolo (*un giorno*), nome + aggettivo (*quel giorno*), nome + aggettivo + articolo(*il mese scorso*), aggettivo sostantivato (*il presente*), avverbio sostantivato(*il dopo*), infinito sostantivato (il tardare);

- *sintagma o gruppo verbale* (GV), che può essere formato: nel caso del predicato verbale, da una voce verbale semplice o composta (ho mangiato, man-

gio), nel caso nominale da una voce del verbo essere + nome/aggettivo (*sono affamata*);

- *sintagma o gruppo preposizionale* (GP), formato da una preposizione + un sintagma nominale (*(*per il futuro).

Il gruppo nominale e il gruppo verbale sono i *costituenti* della frase; i gruppi preposizionali si collocano come espansioni attorno ad essi, arricchendo, determinando o precisando, il significato della frase.

A.2.2 I complementi di tempo

I *complementi di tempo* esprimono le circostanze temporali in cui si svolge un'azione o in cui si verificano le condizioni espresse dal verbo.
Si suddividono in due gruppi principali:

- i complementi di tempo determinato,

- i complementi di tempo continuato.

Il complemento di tempo determinato

Il complemento di tempo *determinato* risponde alla domanda: *quando?*. Indica il momento o il periodo in cui si attua l'azione o la condizione espressa dal verbo; può presentarsi come complemento diretto circostanziale (senza preposizione, come nel caso di date indicanti giorni, mesi, anni), oppure può essere introdotto da preposizioni (e.g.: in, a, di) o da locuzioni preposizionali (e.g.: al tempo di). Complementi di tempo che differiscano per la sola presenza o assenza di un articolo possono assumere significati molto diversi; e.g.: *Il sabato non lavoro*, *Sabato non lavoro*. Esprimono un tempo determinato anche:
- le espressioni che rispondono a specifiche domande, quali ad esempio: dopo quanto tempo?, quanto tempo prima?, fra quanto tempo?, entro quando?,
- alcuni avverbi e locuzioni avverbiali, e.g.: ieri, stasera, un tempo (complemento *avverbiale* di tempo).

Il complemento di tempo continuato

Il complemento di tempo *continuato* risponde alla domanda: *per quanto tempo?*. Esprime la durata temporale dell'azione o della situazione indicata dal verbo; può presentarsi senza preposizione (*Durante l'estate ho studiato oltre un mese*), oppure può essere introdotto da preposizioni (e.g. per, in, durante, oltre). Esprimono un tempo determinato anche:
- le determinazioni temporali che rispondono a specifiche domande, quali ad esempio: da quanto tempo?, per quanto tempo?, fino a quando?,
- alcuni avverbi e locuzioni avverbiali, e.g.: sempre, a lungo, da sempre (complemento *avverbiale* di tempo).

A.3 Sintassi del periodo

La sintassi del periodo indaga sulla struttura della frase complessa (o *periodo*, da 'perìodos', *circuito*, *giro*), individuando i rapporti da/con cui sono legate le proposizioni che la compongono. Le proposizioni sono combinate tra loro sintatticamente in modo vario ma preciso: ogni proposizione assume una specifica funzione e può essere classificata come principale, reggente, coordinata o subordinata.

Le proposizioni temporali fanno parte della classe delle *subordinate avverbiali* (dette anche *complementari indirette*), aventi all'interno del periodo funzione analoga a quella svolta dai complementi di tempo nella frase semplice.
La subordinata temporale esprime una circostanza legata alla reggente da un determinato rapporto cronologico. Ad esempio:

- *dacché*, *da quando* e simili, seguite dal verbo all'indicativo, indicano da quale momento ha inizio l'azione o la situazione descritta nella reggente;

- *finché*, *fino a quando*, *fintanto che* e simili (con eventuale *non* pleonastico posposto), indicano fino a quale momento sussisterà l'azione o la situazione espressa nella reggente;

- *ogni volta che*, *tutte le volte che* e simili, seguite dal verbo all'indicativo, indicano una circostanza che si ripete ogniqualvolta si verifica ciò che è espresso nella reggente;

- *man mano che* e *a man mano che*, seguite dal verbo all'indicativo, indicano un evento che si verifica gradualmente nel tempo, mentre si realizza quanto è espresso nella reggente.

In particolare, la proposizione temporale può essere legata alla reggente da un rapporto di *anteriorità* (l'azione espressa dalla subordinata è anteriore a quella della reggente), *contemporaneità* (l'azione espressa dalla subordinata è contemporanea a quella della reggente), *posteriorità* (l'azione espressa dalla subordinata è posteriore a quella della reggente; assume funzione analoga al complemento di tempo continuato) e può presentarsi in forma *esplicita* (con verbo di modo finito -ad esclusione del modo imperativo-) o *implicita* (con verbo di modo indefinito - infinito, participio, gerundio).
La forma esplicita è introdotta da congiunzioni o da locuzioni, seguite da un modo verbale opportuno:

- nel caso di anteriorità, si utilizza la locuzione congiuntiva *prima che*, seguita da lverbo al congiuntivo;

- nel caso di contemporaneità, si usano congiunzioni come *quando* (per tempi determinati), *mentre* (per tempi continuati), *allorché* o locuzioni come *nel momento che* e il verbo è all'indicativo (quando il verbo si presenta al congiuntivo, il valore della subordinata è più propriamente ipotetico-condizionale);

- nel caso di posteriorità, si utilizzano le locuzioni *dopo che*, *una volta che*, *dopoché* e il verbo all'indicativo (anche in questo caso, quando il verbo si presenta al congiuntivo, il valore della subordinata è più propriamente ipotetico-condizionale).

La forma implicita è possibile solo se il soggetto della subordinata coincide con quello della reggente; è introdotta:

- nel caso di contemporaneità, dal verbo al gerundio presente o, talvolta, dalla preposizione *in* (articolata) seguita dal verbo all'infinito;

- nel caso di posteriorità, da *dopo* (o, a livello letterario, da *dopo di*) seguito dal verbo all'infinito passato, oppure dal solo verbo al participio passato (talvolta preceduto da locuzioni come *appena*, *non appena*, *una volta*; è anche possibile esprimere immediatezza introducendo la subordinata mediante le congiunzioni *quando*, *come* (nel senso di 'quando'), *(non) appena*, e.g.: *Non appena arriverò, andrò a dormire*;

- nel caso di anteriorità, dalla locuzione preposizionale *prima di* seguita dal verbo all'infinito.

Talvolta gli avverbi di tempo introducono proposizioni interrogative, e.g.: *Quando tornerai?* (interrogativa diretta), *Nessuno sapeva quando saresti partito.*

Appendice B

Siti Web

B.1 Elaborazione del Linguaggio Naturale

http://www.bmanuel.org
Guida ai corpora e alle risorse di linguistica computazionale basate sui corpora stessi disponibili in rete.

http://www.isi.edu/natural-language
Gruppo di ricerca sul Linguaggio Naturale.

http://www.xml.com/pub/a/98/10/guide0.html
Elaborazione del Linguaggio Naturale - estrazione dell'informazione; traduzione automatica; riconoscimento named entity; identificazione del linguaggio; elaborazione del linguaggio naturale; tagging part-of-speech; riconoscimento dei confini di frase; tokenizzazione.

http://www.itl.nist.gov/iaui/894.02
"Retrieval Group" della Divisione per l'Accesso all'Informazione; lavora per l'industria, le accademie e altre agenzie governative al fine di promuovere l'uso di tecniche efficaci ed efficienti per la manipolazione di informazione testuale non strutturata, in particolare: browsing, ricerca e presentazione.

http://www-csite.deis.unibo.it

(NUOVO SITO: *http://www.bo.ieiit.cnr.it*)
CSITE - Centro di Studio per l'Informatica e i Sistemi di Telecomunicazioni
Ricerca scientifica nel campo dell'informatica: gestione e trattamento della conoscenza, recupero dell'informazione (IR), basi di dati temporali, sistemi di basi di dati multimediali.

http://sds.colorado.edu/TExLab
Home Page of the Automatic Temporal Expression Labeler (ATEL).

http://www.dimi.uniud.it/~tasso/general.html
Laboratorio di Intelligenza Artificiale dell'Università di Udine - Dipartimento di Matematica e Computer Science.
Le aree principali di ricerca comprendono: estrazione intelligente dell'informazione, elaborazione del linguaggio naturale, temporal reasoning, basi di dati temporali e basi della conoscenza, granularità temporale, astrattezza, e indeterminazione.

http://www.ltg.ed.ac.uk/index.html
Home Page del Language Technology Group (LTG) di Edimburgo.
LTG è un gruppo di ricerca e sviluppo che lavora nell'area dell'ingegneria applicata al linguaggio naturale.

http://www.ltg.ed.ac.uk/ie/iegroup-index.html
Gruppo per l'estrazione dell'informazione.

home page: http://www.di.unito.it/~krr
Knowledge Representation and Reasoning Group - Department of Computer Science - Università di Torino

B.2 Strumenti per l'elaborazione testuale

http://www.ltg.ed.ac.uk/software/ttt/tttdoc.html
Strumenti per la tokenizzazione del testo.

http://sds.colorado.edu
Home Page dell'etichettatore automatico "ATEL" di espressioni temporali.

http://www.cl.cam.ac.uk/Research/NL/anlt.html
Strumenti per la ricerca nel campo dell'elaborazione del linguaggio naturale: analizzatore morfogico, strumenti di parsing, grammatica e lessico (ambiente di sviluppo per le grammatiche - sistema completo per l'analisi morfologica, sintattica e semantica della lingua inglese).

http://database.cs.wayne.edu/proj/langdl/links/2/Linguistic%20software%20tools.html
Strumenti software per la linguistica: annotazione, scrittura e analisi del testo, tagging, generazione e controllo di grammatiche, tokenization, markup testuale, parsing, riassunti automatici, elaborazione del testo.

http://www.bmanuel.org/clr2_tt.html#Tools
Software per l'elaborazione del linguaggio naturale orientato ai corpora (strumenti per il parsing, il tagging, il chunking, sistemi di interrogazione dei corpora, analizzatori di testo)

B.3 Basi di dati Temporali

http://db.cs.ualberta.ca/stdbm04
Secondo Workshop (il primo workshop è stato nel 1999 - STDBM'99)della durata di una giornata sulla gestione delle basi di dati Spatio-Temporali: incontro fra ricercatori e sviluppatori sullo stato dell'arte, sulle nuove idee e ricerche.

http://www.mm.di.uoa.gr/~toobis/Welcome.html
TOOBIS è un progetto con lo scopo di sviluppare un sistema di gestione di basi di dati temporali orientate agli oggetti; contribuisce ad un facile accesso all'informazione mediante l'estensione della funzionalità dei sistemi di gestione delle basi di dati. E' allo studio anche lo sviluppo di una metodologia innovativa di modellazione concettuale per applicazioni temporali.

http://www.dbnet.ece.ntua.gr/~choros/index.html
Rete di ricerca per i sistemi di basi di dati spazio-temporali (1996).

http://www.elet.polimi.it
Gruppo di Ricerca sulla rappresentazione del tempo nei sistemi informativi orientato alla modellazione di aspetti temporali di sistemi d'ufficio e di scenari multimediali e web, all'estensione delle basi di dati relazionali e orientate agli oggetti per il trattamento di dati temporali (con gestione dei problemi di granularità dei dati, dell'informazione temporale imprecisa, delle relazioni qualitative fra istanti temporali), al progetto di un linguaggio per estrarre informazione su oggetti in evoluzione.

http://www.scism.sbu.ac.uk/cios/paul/Research/tdb_links.html - 1997
Lista di link a siti Web e home page nel campo delle basi di dati temporali.

http://wwwdb.informatik.uni-rostock.de/~bg/tdb_res.html
Ricerca nel campo delle basi di dati temporali, software per sistemi temporali, link a basi di dati temporali.

B.4 Eventi: competizioni e convegni internazionali

SEBD2001 - Sistemi Evoluti Per Basi di Dati
Convegno italiano sui sistemi di basi di dati (SEBD), il maggior evento annuale della comunità di ricerca italiana nel campo delle basi di dati. E' pensato come forum per l'incontro, la discussione e lo scambio di esperienze tra coloro che, nell'accademia e nell'industria, sono interessati ai sistemi di

basi di dati e al loro vasto campo di applicazioni.
Il primo convegno SEBD fu organizzato nel 1993 come iniziativa del Sotto-progetto "Sistemi Evoluti per Basi di Dati" appartenente al Progetto Finalizzato "Sistemi Informatici e Calcolo Parallelo" del Consiglio Nazionale delle Ricerche.

http://www.di.unito.it/~krr/time.html
Temporal Reasoning: the LaTeR project (LAyered TEmporal Reasoning)
- intelligenza artificiale
- problemi di soddisfacimento dei vincoli
- temporal Reasoning
- basi di dati temporali

http://www.itl.nist.gov/iaui/894.02/related_projects/muc - 2001

http://www.cs.brandeis.edu/~jamesp/HLT
Human language technology conference 27 maggio-1 giugno 2003

http://www.sims.berkeley.edu:8000/research/conferences/hlt-naacl03/tutorials-details.html NAACL-HLT 2003 Tutorial

http://www.cs.brandeis.edu/~jamesp/arda/time/index.html - 2004
Markup Language for Temporal and Event Expressions
Project Funded by ARDA
Collegamenti utili, a partire da questo sito:
TimeML documents (http://www.cs.brandeis.edu/~jamesp/arda/time/timemldocs.html)
TimeML Specification (Version 1.2 – September 13, 2004) [HTML Document]
TimeML Annotation Guidelines (Version 1.2 – September 13, 2004) [PDF

Document]
TimeML Schema (Version 1.2 – September 20, 2004) [Schema Document]
TimeBank (Version 1.1 – April 2, 2004) [TimeBank 1.1]

http://time2005.cse.buffalo.edu
TIME 2005 - International Symposium on Temporal Representation and Reasoning
Burlington, USA, June 23-25, 2005
Temporal Databases
Temporal Logic in Computer Science
Temporal Representation and Reasoning in AI

http://www.ltg.ed.ac.uk/ie/IEDemos/IE-MUC
Message Understanding Conferences sponsored by DARPA in the U.S.; they were regular conferences where the participants competed to build IE systems.

B.5 Motori di ricerca

http://citeseer.ist.psu.edu/
CiteSeer si propone come libreria digitale di letteratura scientifica.
Riguardo ai documenti cercati, fornisce:

- link al documento, nei vari formati disponibili in rete (.pdf, .ps, etc);
- link a testi che citano il documento;
- link a documenti simili (a livello di frase);
- link a documenti simili sulla base del testo;
- documenti correlati da co-citazioni;

- link a testi citati nel documento;
- documenti presenti nello stesso sito.

http://www.scirus.com/
Scirus si propone come il più esaustivo motore di ricerca in rete orientato alle materie scientifiche.

Bibliografia

[1] Fabio Grandi, Federica Mandreoli *The Valid Web: it's Time to Go...*, Technical Report, TIMECENTER TR-46, 1999, http://www.cs.auc.dk/research/DP/tdb/TimeCenter/.

[2] Fabio Grandi, Federica Mandreoli *Effective Representation and Efficient Management of Indeterminate Dates*, TIME 2001, 164-169, http://www.computer.org/proceedings/time/1107/11070164abs.htm.

[3] Fabio Grandi, Federica Mandreoli *The "XML/Repetti Project: Encoding and Manipulation of Temporal Information in Historical Text Sources"*, ICHIM (2) 2001, 243-252.

[4] Fabio Grandi, Federica Mandreoli, et al. *Gestione di versioni temporali di risorse nel World Wide Web secondo il tempo di validità*, Dipartimento di Elettronica, Informatica e Sistemistica Universita' degli studi di Bologna, 9 agosto 1999 ftp://csite60.deis.unibo.it/pub/Report/T6-R13.ps.

[5] Fabio Grandi, Federica Mandreoli *The Valid Web: an XML/XSL Infrastructure for Temporal Management of Web Documents*, Proc. of 1st Intl' Conf. on Advances in Information Systems (ADVIS 2000), Izmir (September 2000), Turkey, LNCS Vol. 1909, Springer-Verlag, Heidelberg, pp. 294-303, 2000.

[6] Fabio Grandi *XML Representation and Management of Temporal Information for Web-based Cultural Heritage Applications*, Data Science Journal, Vol. 1, No. 1, pp. 68-83, April 2002.

[7] I. Mani, L. Ferro, B. Sundheim, G. Wilson *Guidelines for annotating temporal information*, In Notebook Proceedings of Human Language Technology Conference 2001, pages 299–302, San Diego, California, March 18-21

[8] A. Vasilakopoulos, M. Bersani, W.J. Black *A Suite of Tools for Marking Up Textual Data for Temporal Text Mining Scenarios*, Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, 24th-30th May 2004.

[9] Andrei Mikheev, Claire Grover, Marc Moens *XML Tools And Architecture for Named Entity Recognition*, Markup Languages, Volume 1, Number 3, pages 89–113, 1999

[10] Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, George Wilson *TIDES 2005 Standard for the Annotation of Temporal Expressions*, version 1 April 2005 http://timex2.mitre.org

[11] C.E. Dyreson,R.T. Snodgrass *Valid-Time Indeterminacy*, Proceedings of the International Conference on Data Engineering, Vienna, Austria, pgg 335–345 apr 1993

[12] C.E. Dyreson,R.T. Snodgrass *Supporting valid-time indeterminacy*, ACM Transactions on Database Systems, Vol.23 n.1 pgg 1–57, 1998

[13] J. Chomicki *Temporal query languages: a survey*, Temporal Logic: (ICTL'94) Vol 827 pgg 506-534 Springer-Verlag ed., 1994

[14] Benjamin Han, Alon Lavie *A framework for resolution of time in natural language*, ACM Transactions on Asian Language Information Processing Vol.3 Issue 1 pgg 11-32, 2004

[15] Claire Grover, Colin Matheson, Andrei Mikheev, Marc Moens *LT TTT - A Flexible Tokenization Tool*, Proceedings of. Second International Conference on Language Resources. and Evaluation, Athens, 2000, pp. 67-75 http://www.ltg.ed.ac.uk/papers/00tttlrec.pdf

[16] Federico Faccioni *Individuazione automatica di espressioni quantitative in testi italiani*, Tesi di Laurea, Università degli studi di Padova, A.A.2003/04

[17] Curtis E. Dyreson, Richard T. Snodgrass *Valid-time Indeterminacy*, Proceedings of the International Conference on Data Engineering, Vienna, Austria, april 1993 pp. 335-343 http://www.eecs.wsu.edu/ cdyreson/pub/temporal/papers/dis.pdf

[18] David Ahn, Sisay Fissaha Adafre, Maarten de Rijkee *Extracting Temporal Information from Open Domain Text: A Comparative Exploration*, J. Digital Information Management, 3(1):14-20, 2005 hhtp://www.science.uva.nl/ mdr/Publications/Files/jdim2005-tern.pdf

[19] Kalina Bontcheva, Hamish Cunningham *The Semantic Web: A New Opportunity and Challenge for Human Language Technology*, Proceedings workshop on Human Language Technology for the Semantic Web and Web Services at International Semantic Web Conference 2003. http://gate.ac.uk/sale/iswc03/iswc03.pdf

[20] Marcello Sensini *Le parole e il testo. Teoria e pratica della comunicazione linguistica*, A. Mondadori Editore

[21] I. Mani, G. Wilson *Robust temporal processing of news*, Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000), pages 69–76, New Brunswick, New Jersey, 2000. Association for Computational Linguistics. http://complingone.georgetown.edu/ linguist/papers/acl2000-tempex.pdf

[22] Sisay Fissaha Adafre, M. de Rijke *Feature Engineering and Post-Processing for Temporal Expression Recognition Using Conditional Random Fields*, Proceedings ACL-2005 Workshop on Feature Engineering, 2005 http://www.science.uva.nl/ mdr/Publications/Files/acl2005-fe-ws.pdf

[23] Massimo Franceschet, Angelo Montanari *Time Granularities in Databases, Data Mining, and Temporal Reasoning (Book re-*

view), The Computer Journal, vol. 45(6), pp. 683-685, 2002 http://staff.science.uva.nl/ francesc/pubs/cj02.pdf

[24] David Ahn, Sisay Fissaha Adafre, Maarten de Rijke *Towards Task-Based Temporal Extraction and Recognition*, Proceedings Dagstuhl Workshop on Annotating, Extracting, and Reasoning about Time and Events, 2005 http://www.science.uva.nl/ mdr/Publications/Files/dagstuhl2005-tern.pdf

[25] D. Ahn, S. Fissaha Adafre, M. de Rijke *Recognizing and Interpreting Temporal Expressions in Open Domain Texts*, S. Artemov, H. Barringer, A. S. d'Avila Garcez, L. C. Lamb, and J. Woods, editors, We Will Show Them: Essays in Honour of Dov Gabbay, Vol 1., College Publications, pages 31-50, 2005 http://www.science.uva.nl/ mdr/Publications/Files/gabbay-60-afadr.pdf

[26] Carlo Combi, Massimo Franceschet, Adriano Peron *Representing and reasoning about temporal granularities*, Journal of Logic and Computation 2004 14(1):51-77; doi:10.1093/logcom/14.1.51 Oxford University Press, 2004 http://logcom.oxfordjournals.org/cgi/reprint/14/1/51

[27] M. Franceschet, A. Montanari *Branching within Time: an Expressively Complete and Elementarily Decidable Temporal Logic for Time Granularity*, Research on Language and Computation, vol. 1, no. 3-4, Kluwer Academic Publishers, September 2003, pp. 229-263 http://staff.science.uva.nl/ francesc/pubs/rlc03.pdf

[28] James Pustejovsky, Inderjeet Mani *Annotation of Temporal and Event Expressions*, http://www.sims.berkeley.edu:8000/research/conferences/hltnaacl03/tutorials-details.html

[29] Benjamin Han, Alon Lavie *A Framework for Resolution of Time in Natural Language*, ACM TRans. Asian Lang. Inf. Process. 3(1): 11-32, 2004

[30] L. Ferro, L.Gerber, I. Mani, B. Sundheim, G. Wilson *2005 Standard for the Annotation of Temporal Expressions*, MITRE, 2005

[31] Hacioglu, Kadri, Chen, Ying and Douglas, Benjamin *Automatic Time Expression Labeling for English and Chinese Text*, Proceedings of CICLing-2005, Mexico City-Mexico, Feb. 13-19, 2005; http://sds.colorado.edu/TExLab/papers/hacioglu_CICLing-05.pdf

[32] Inderjeet Mani *Recent Developments in Temporal Information Extraction*, RANLP 2003, 45-60

[33] B. Oliboni, E. Quintarelli, L. Tanca, *Temporal aspects of semistructured data*, SEBD 2001, 215-222

[34] Graham Katz, Fabrizio Arosio *The Annotation of Temporal Information in Natural Language Sentences*, in Proceedings of the ACL Workshop on Spatial and Temporal Information Processing, Toulouse, France 2001, http://www.cogsci.uni-osnabrueck.de/~gkatz/Papers/ACL2001.pdf

[35] Graham Katz *(A)temporal complements*, Audiatur Vox Sapientiae C. Fery and W. Sternefeld (eds), Akademie-Verlag, Berlin. pp. 240-258, 2001, http://www.cogsci.uni-osnabrueck.de/ gkatz/Papers/StechowFestshrift.pdf

Ringraziamenti

Grazie a Michele Rubert per le ore di sonno che gli ho fatto perdere.
Grazie a Marco Brombin per la sua passione per l'informatica.