Grammatiche locali per il riconoscimento automatico e la classificazione delle FAQ sull'Informazione Comunitaria Europea

Annibale Elia, Fabiola Bocchino, Alberto Maria Langella, Mario Monteleone, Daniela Vellutino

Università di Salerno, Dipartimento di Scienze della Comunicazione

Riassunto

Questo studio riguarda l'analisi testuale di un corpus di testi del dominio specialistico dell'informazione comunitaria, in relazione alle opportunità di cofinanziamento offerte dai Fondi strutturali. Sono state analizzate le FAQ dei principali siti istituzionali dedicati all'argomento. Per l'analisi testuale è stato utilizzato NOOJ, un software di trattamento automatico dei testi che si basa sull'uso di dizionari elettronici, grammatiche locali ed automi a stati finiti. Il software NOOJ utilizza il sistema dei dizionari elettronici, sviluppato dal Dipartimento di Scienze della Comunicazione dell'Università di Salerno, che integra al momento cinque dizionari elettronici (sistema DELA). Ai dizionari del sistema DELA è stato affiancato un dizionario elettronico di dominio sull'informazione comunitaria al fine di rilevare le unità lessicali contenute nel corpus delle FAQ. Le FAQ sono state descritte attraverso un trasduttore a stati finiti. Le FAQ con struttura simile sono state accorpate in un'unica grammatica locale.

Abstract

This study concerns textual analysis of a given text corpus in which the main subject coped with is European Community information related to the opportunities within Structural Funds co-financing. We analyzed FAQ extracted by common Web sites of main establishments dedicated to that domain. We used NOOJ for textual analysis; an automated parser based on electronics dictionaries, local grammars and finite states automata. NOOJ uses electronic dictionaries system developed by Department of Communication Science of University of Salerno. The package implemented in that software contains five dictionaries (DELA system). In this system we added a specialized electronics dictionary on domain of European Community information in order to localizing lexical units in the FAQ corpus. Finally FAQ are been profiled using a finite state transducer. Similar structure FAQ are been collected in a same local grammar.

Keywords: local grammars, natural language processing, multiword expressions

1. Introduzione

Il presente studio [1] descrive l'analisi testuale automatica effettuata su un corpus relativo al dominio specialistico dell'Informazione Comunitaria Europea, ed in relazione alle opportunità di cofinanziamento con Fondi Strutturali. L'analisi effettuata ha riguardato le Frequently-Asked

[1] Riguardo i contenuti di questo articolo, Annibale Elia ha realizzato le sezioni 1 e 2. Fabiola Bocchino ha realizzato la sottosezione 3.3. Alberto Maria Langella ha realizzato la sezione 3 e la sottosezione 3.4. Mario Monteleone ha realizzato la sottosezione 2.1. e la sezione 4. Daniela Vellutino ha realizzato le sottosezioni 3.1. e 3.2.

Questions (da qui in poi, FAQ) estratte da siti Web dedicati a questo specifico settore informativo. Per la realizzazione dell'analisi testuale automatica è stato utilizzato il pacchetto software NOOJ [2], principalmente basato sull'uso di dizionari elettronici, grammatiche locali e automi e trasduttori a stati finiti (da qui in poi FST). Nel corso delle applicazioni del software, sono stati utilizzati i dizionari elettronici sviluppati dal Dipartimento di Scienze della Comunicazione dell'Università di Salerno, ovvero cinque dizionari del sistema DELA (dizionari elettronici di parole semplici e composte) ai quali è stato affiancato un dizionario elettronico specialistico del settore dell'Informazione Comunitaria Europea. Infine, grazie alla creazione ed all'applicazione di specifici FST, è stato possibile effettuare la lettura automatica del corpus prescelto per poi riconoscere, etichettare e formalizzare gli specifici pattern testuali contenuti dalle FAQ.

Il procedimento di formalizzazione effettuato sarà utile, nelle ricerche future, a verificare l'esistenza di modelli linguistici ricorsivi nella produzione di FAQ online, quindi a produrre un sistema di risposta semi-automatico di cui potranno dotarsi i siti Web di Enti Locali e Amministrazioni Pubbliche.

2. Quadro teorico di riferimento

La metodologia linguistica descrittiva di riferimento per lo studio da noi effettuato è il Lessico-Grammatica, originariamente introdotto nella comunità scientifica europea da Maurice Gross all'Università Parigi 7 dopo il 1960. Scopo principale del Lessico-Grammatica è descrivere dettagliatamente tutti i meccanismi combinatori indissolubilmente legati alle concrete entrate lessicali. Uno dei risultati più recenti ed importanti della descrizione lessico-grammaticale è proprio la costruzione di tre software e lingware orientati all'analisi testuale automatica, ovvero UNITEX [3] e INTEX [4], oltre il già citato NOOJ.

Oggi, il quadro metodologico lessico-grammaticale è applicato da gruppi di ricercatori che lavorano in diverse università, laboratori e strutture di ricerca mondiali, e che si riconoscono nella rete di ricerca scientifica che va sotto il nome di RELEX.

Il Dipartimento di Scienze della Comunicazione dell'Università di Salerno ha avuto una lunga collaborazione scientifica con il LADL (Laboratoire d'Automatique Documentaire et Linguistique) dell'Università Parigi 7, diretto per oltre venti anni da Maurice Gross. Questa stretta collaborazione scientifica ha sempre avuto il fine di raffinare il lingware prodotto nel corso degli anni e di adattarlo ad applicazioni software.

2.1. Il lessico-grammatica e l'analisi testuale automatica

NOOJ, così come gli altri due software lessico-grammaticali menzionati, utilizzano due specifici tools lingware per la realizzazione dell'analisi testuale automatica: i DELA e le grammatiche locali in forma di FST.

I DELA, e nello specifico quelli dell'italiano, sono basi di dati [5] lessicali formalizzate, strutturate omogeneamente ed in cui le caratteristiche morfo-grammaticali delle entrate (genere, numero e flessione) sono indicate da etichette alfanumeriche univoche e non ambigue. I DELA hanno

[2] Per ulteriori indicazioni sulle funzionalità di NOOJ, si veda http://www.nooj4nlp.net/pages/nooj.html.
[3] Per ulteriori indicazioni sulle funzionalità di UNITEX, si veda http://www-igm.univ-mlv.fr/~unitex/.
[4] Per ulteriori indicazioni sulle funzionalità di INTEX, si veda http://intex.univ-fcomte.fr/.
[5] Il termine *basi di dati* va qui inteso nella più comune accezione informatica, sia dal punto di vista teorico che da quello di strutturazione pratica. Inoltre, un dizionario elettronico non può essere usato come un dizionario cartaceo, né può sostituirne il ruolo.

la funzione di motori linguistici basici per i software di analisi testuale automatica già citati, e sono di due tipi, ovvero:

- di parole semplici (denominati DELAS-DELAF), ed includono tutte quelle parole semanticamente autonome e composte da sequenze di lettere non interrotte e delimitate da spazi bianchi (ad esempio le parole *casa*, *battello*);
- di parole composte (denominati DELAC-DELACF), ed includono tutte quelle sequenze formate da due o più parole e che costruiscono congiuntamente singole unità di significato (ad esempio, le sequenze *casa di cura*, *battello a vapore*). Il dizionario elettronico specialistico del settore dell'Informazione Comunitaria Europea, utilizzato insieme ai DELA per lo studio in oggetto, è stato configurato come dizionario di parole composte.

Tale suddivisione è necessaria sia dal punto di vista formale e morfologico, sia da quello semantico [6]. Infatti, in fase di compilazione di un software di query, recupero informazioni o analisi testuale automatica, come ad esempio i già citati INTEX e UNITEX, l'assenza di separatori all'interno delle parole semplici e la presenza di questi all'interno delle parole composte, saranno fattori discriminanti e comporteranno impostazioni e scelte differenti per quanto riguarda gli automatismi nel trattamento dei dati. Nel caso delle parole semplici, sarà infatti necessario prevedere e trattare solo dati alfabetici o numerici; con le parole composte, la presenza di separatori inserirà un livello aggiuntivo di dati, ai quali si dovranno assegnare funzioni univoche, non ambigue, e di valore diverso da quelli alfabetici e numerici. Diverso è invece il caso di NOOJ, il più recente dei software precedentemente citati, compilato in modo da gestire congiuntamente basi di dati di struttura diversa; in NOOJ, infatti, è possibile realizzare un unico dizionario elettronico contenente sia parole semplici che composte.

Per quanto invece riguarda il secondo dei tools lingware utilizzati nell'analisi testuale automatica, si indica con grammatica locale, o anche grammatica dipendente dal contesto, un algoritmo trasduttore a stati finiti che include istruzioni di carattere lessicale e morfo-grammaticale e che viene usato da un software specifico per leggere, comprendere ed etichettare sequenze di parole all'interno di testi informatizzati. Le grammatiche locali si configurano quindi sia come dei tools per la semplice analisi testuale automatica, ovvero per l'information retrieval, sia come degli strumenti basilari per le operazioni di parsing. La loro denominazione deriva dal fatto che ciascun algoritmo viene elaborato per descrivere un solo aspetto della grammatica di una data lingua, anche se complesso e dettagliato. Per cui, sarà possibile avere ad esempio una grammatica locale che descriva le possibili trasformazioni passive del verbo "mangiare", o le possibili frasi interrogative realizzabili a partire dal verbo "trovare".

I ricercatori italiani che si sono occupati e chi si occupano oggi di grammatiche locali hanno come fini ultimi:

1) la descrizione di tutti gli aspetti morfo-grammaticali e morfo-sintattici dell'italiano;
2) la costruzione di una grammatica locale per ogni aspetto specifico;
3) l'interconnessione e l'applicazione congiunta di tutte le grammatiche elaborate.

La realizzazione del punto 3) permetterebbe perciò di creare un'unica grammatica omnicomprensiva, in grado di analizzare ogni testo scritto in lingua italiana. Tale grammatica sarebbe

[6] Aggiungiamo che le parole semplici hanno sempre un tasso di polisemia (leggi ambiguità) più elevato di quelle composte, che ne hanno uno molto prossimo allo zero. Per quanto riguarda il recupero informazioni e l'analisi testuale automatica, ciò implica indirettamente la necessità di prevedere ed impostare modalità analitiche diverse per i due tipi di parole.

quindi applicabile come un'unica grammatica indipendente dal contesto, di carattere universale, formata tuttavia da tante singole grammatiche locali, dipendenti dal contesto.

3. L'analisi testuale automatica delle FAQ sull'Informazione Comunitaria Europea

3.1. Glossari istituzionali e corpus di testi scritti di dominio: strumenti per estrarre conoscenza lessicale

È stato svolto uno studio delle fonti istituzionali dell'informazione comunitaria, al fine di ricercare glossari utili per gli studi linguistici lessicografici sulle parole semplici e sulle parole composte del dominio terminologico "Fondi Strutturali". Sono stati individuati, esaminati e repertoriati 16 glossari dell'informazione comunitaria tratti da fonti istituzionali.

Al contempo, per la creazione del corpus da analizzare, è stata definita una classificazione tipologica dei testi dei documenti dei domini investigati in base alle finalità pragmatiche dei differenti processi di comunicazione di cui i testi sono oggetto. Pertanto, il corpus presenta la seguente ripartizione:

- testi dei documenti delle fonti normative primarie e secondarie per la comunicazione normativa;
- testi dei documenti di programmazione economica per la comunicazione della trasparenza amministrativa;
- testi per l'informazione comunitaria dei siti web istituzionali e dei materiali informativi scritti finalizzati alla comunicazione pubblica e veicolati su supporto cartaceo e multimediale.

La classificazione dei testi è stata finalizzata alla costruzione di un corpus stratificato utilizzabile per il rilevamento della diffusione delle unità lessicali terminologiche registrate nei glossari istituzionali dell'informazione comunitaria, nonché per acquisire informazioni sui comportamenti distribuzionali, sulle co-occorrenze di tali unità lessicali repertoriate come fisse dagli stessi glossari.

A conclusione dell'analisi, le FAQ individuate e classificate sono risultate 345.

3.2. Creazione del dizionario di dominio

Attraverso una prima fase di analisi testuale automatica, in base al numero di tokens rilevati nelle liste di frequenza sono state individuate informazioni lessicali utili per definire i contesti di distribuzione all'interno della frase: ristretta, mediamente fissa, invariabile.

Al fine di selezionare i termini candidati a costituire le entrate dei dizionari elettronici di dominio, è stato svolto un lavoro di selezione dei termini, in base sia alle specifiche terminologiche sia alle specifiche desunte dal rilevamento del grado di co-occorrenza.

Di conseguenza, per il dominio in questione, sono stati definiti i valori di una base di dati terminologica comprensiva di tutte le varianti ortografiche dei lemmi. Per ogni entrata, sono state perciò riportate le varianti tutte minuscole, tutte maiuscole, con una maiuscola iniziale, e con la presenza sia di maiuscole sia di minuscole, come nel protogramma "AdG" e nella sequenza lessicale fissa "Autorità di Gestione". Non sono state tuttavia riportate le variazioni di tipo tipografico.

Quindi, la base di dati terminologica del dominio "Fondi strutturali" include le seguenti informazioni linguistiche:

- frequenza delle parole terminologiche nell'intero corpus di dominio;
- frequenza delle parole terminologiche in relazione ai differenti tipi di testo classificati;
- frequenza di distribuzione delle co-occorrenze delle unità lessicali sintagmatiche rilevate come fisse dai glossari;
- riconoscimento e catalogazione dei contesti grammaticali d'uso delle parole terminologiche;
- riconoscimento e catalogazione delle varianti espressive classificabili come equivalenze semantiche oppure aventi con il termine relazioni di sinonimia/iperonimia/iponimia.

Tale studio è stato ritenuto necessario per selezionare il materiale lessicale da includere nei dizionari elettronici da sviluppare.

Inoltre, per il dominio in questione, il materiale lessicale da inserire nel dizionario è stato selezionato attraverso un procedimento che ha compreso le seguenti fasi:

- ricerca dei glossari di dominio dai siti istituzionali dell'informazione comunitaria;
- catalogazione dei glossari di dominio;
- confronto tra i glossari al fine d'individuare termini comuni e varianti lessicali;
- registrazione delle varianti lessicali per sinonimia, per iponimia, per iperonimia e varianti ortografiche (utilizzo degli acronimi e delle differenti scritture, ad esempio POR e P.O.R.), varianti sociolinguistiche (internazionalismi ed i corrispondenti italiani, ad esempio "gender mainstreaming", "prospettiva di genere", "ottica di genere", ma anche termine di uso comune e termine di uso non comune);
- depurazione dei lemmi relativi a nomi propri;
- confronto tra i termini contenuti nei testi dei documenti istituzionali comunitari e le entrate dei glossari istituzionali dei siti d'informazione comunitaria.

Il dizionario elettronico di dominio contiene attualmente 1949 lemmi flessi, ed è stato testato su un ulteriore corpus costituito dai testi di documenti comunitari della dimensione di 6 Megabyte (pari a 3.071.610 occorrenze tokenizzate). In questo corpus, compaiono testi rappresentativi del dominio analizzato, ovvero del lessico terminologico individuato per il dominio dell'informazione comunitaria, e si riferiscono ad uno specifico periodo di rilevamento sincronico (luglio 2008-giugno 2009). Le tipologie di testo in esso racchiuse sono le seguenti:

- testi dei documenti delle fonti normative primarie e secondarie per la comunicazione normativa (674.746 occorrenze tokenizzate);
- testi dei documenti di programmazione economica per la comunicazione per la trasparenza amministrativa (2.390.093 occorrenze tokenizzate);
- testi i testi per la comunicazione pubblico-istituzionale comprendono 345 frasi di FAQ estratte da 51 siti istituzionali in materia relative alle richieste d'informazioni sugli interventi di finanziamento comunitario (6.771 occorrenze tokenizzate).

La dimensione attuale del corpus è a 3.071.610 occorrenze tokenizzate.

3.3. Creazione degli automi a stati finiti

Per creare gli FST utili a riconoscere ed etichettare le FAQ precedentemente individuate, si è partiti dall'attività di controllo e di normalizzazione dei testi contenuti nel corpus di riferimento, con la finalità di eliminare e/o ridurre tutti gli elementi testuali che potessero, in qualche modo, compromettere l'analisi con il software NOOJ, e che potessero creare ritardi nell'elaborazione delle grammatiche locali. Successivamente ogni singolo testo del corpus è stato sottoposto ad analisi testuale automatica, cominciando con la fase del pre-processing che ha come obiettivi principali la individuazione e la classificazione delle entrate lessicali (parole semplici e/o composte) nonché la rilevazione della frequenza dei tokens presenti all'interno del testo, e utilizzando poi la funzione Construct FST –Text che consente di descrivere ogni singola frase del testo per mezzo di un trasduttore a stati finiti. Ad esempio, la frase "*Quali sono i principali obiettivi del FEP?*" è rappresentata dal trasduttore illustrato in Fig. 1:

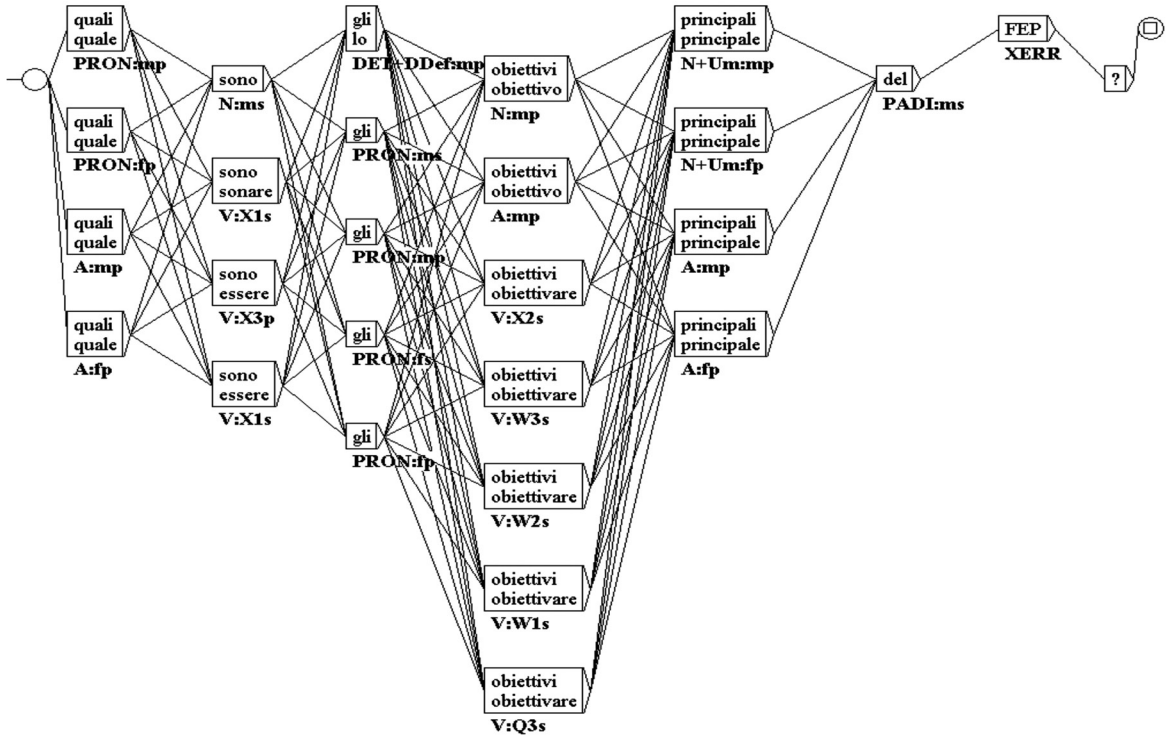

Figura 1

Su ogni trasduttore è stato necessario intervenire manualmente per eliminare gli stati con descrizioni linguistiche non pertinenti [7], nonché per realizzare una grammatica locale che garantisse un'interpretazione corretta ed immediata della frase analizzata. Per l'FST precedente, la procedura di eliminazione dei nodi non pertinenti ha consentito di ottenere il grafo illustrato in Fig. 2:

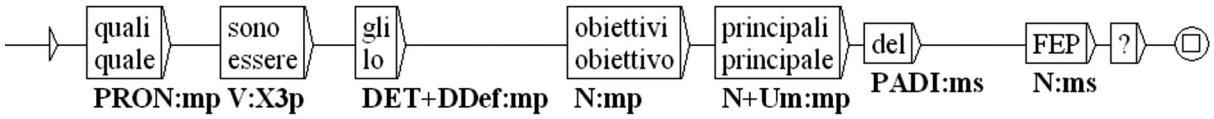

Figura 2

Per non moltiplicare in modo esponenziale il numero delle grammatiche, le frasi con struttura simile sono state accorpate in un unico grafo. Si considerino, in proposito, le frasi seguenti:

Posso presentare un progetto per estendere il mio certificato ad altre attività aziendali?
Posso comunque iniziare i lavori?
Ci sono probabilità che venga finanziata in un prossimo futuro?
È possibile inviare più istanze di contributo all'interno della stessa busta?
È possibile variare la società di consulenza/responsabile di progetto in una domanda già presentata?

in cui sono presenti diverse realizzazioni del verbo modale *potere*, di cui alcune in forma aggettivale precedute dal verbo supporto *essere*. Le frasi sono quindi tutte riconducibili ad una

[7] In questa fase di analisi, gli FST realizzati senza automaticamente dal software elencano tutti i possibili usi di una singola parola lessicale, come dimostrato dal grafo della Fig. 1.

struttura base. La soluzione più logica e vantaggiosa si è rivelata, quindi, quella di raggrupparle in un'unica grammatica locale, come mostra il grafo in Fig. 3:

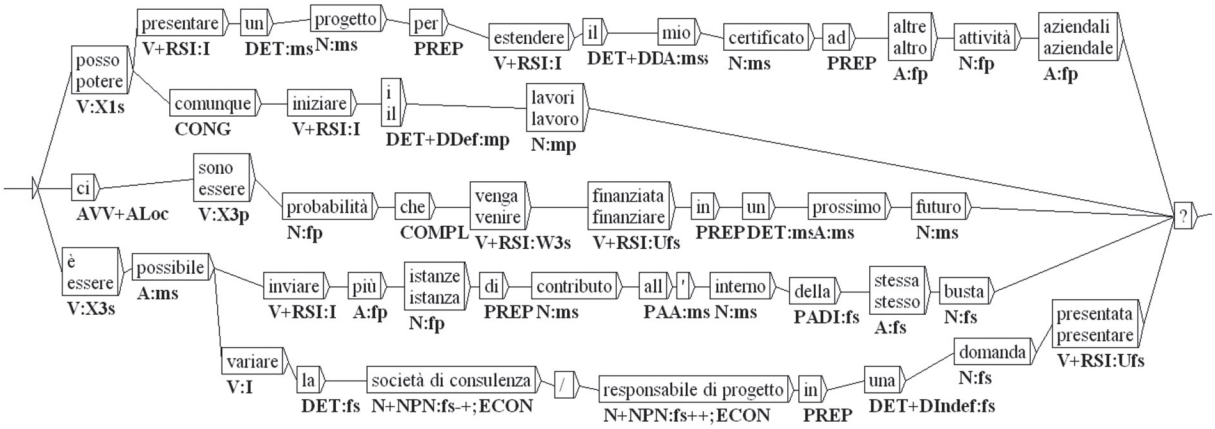

Figura 3

Le operazioni illustrate per le Figg. 1, 2 e 3 sono state applicate ad ogni singolo testo del corpus di riferimento ed hanno condotto all'elaborazione di 142 grafi. Per ognuno di essi è stata effettuata una verifica puntuale allo scopo di testarne il funzionamento. A tal fine, è stato utilizzato il pannello "Locate Pattern" di NOOJ per stabilire se ogni grafo potesse essere correttamente caricato nel programma e potesse essere poi utilizzato per il riconoscimento delle domande presenti nel corpus di riferimento.

3.4. Applicazione e verifica degli automi a stati finiti

Dopo questa fase di controllo del funzionamento delle grammatiche locali è stato necessario procedere ad una generalizzazione dei risultati ottenuti. In particolare, si è ritenuto opportuno procedere alla individuazione ed alla classificazione degli elementi linguistici presenti nei grafi elaborati, con l'obiettivo di costruire grafi più generali, applicabili non soltanto al corpus di riferimento, ma anche a corpora di altra provenienza. Ad ogni parola contenuta nei nodi delle grammatiche locali è stata perciò sostituita la corrispondente etichetta morfo-grammaticale, per cui verbi come *presentare*, *inviare* e *variare* sono stati tutti sostituiti dalla categoria <V> che, in NOOJ, consente il riconoscimento di tutte le entrate verbali flesse. Allo stesso modo, le preposizioni semplici sono state etichettate con <PREP>, gli aggettivi con <A>, gli avverbi con <AVV>, utilizzando un sistema di descrizione condiviso dagli utenti di NOOJ e, in generale, di dizionari elettronici come il DELAS, il DELAF e così via.

Il processo appena descritto ci ha consentito di ottenere grammatiche locali non più legate ad uno specifico testo, ma applicabili a testi differenti. Si veda, in proposito, il grafo di Fig. 4, in cui ogni singola parola è stata sostituita dalla corrispondente categoria sintattica:

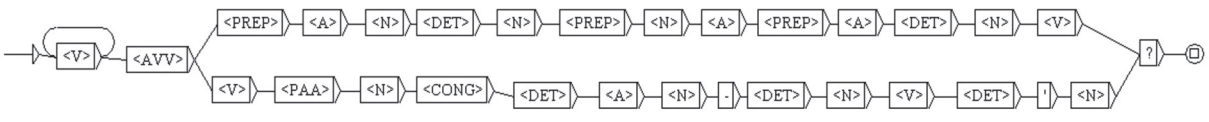

Figura 4

Effettuata questa generalizzazione dei risultati ottenuti, si è ritenuto opportuno procedere ad un raggruppamento delle grammatiche locali riferite ad uno stesso testo del corpus. Tali grammatiche, utilizzando l'apposita funzione di NOOJ, sono state inserite all'interno di un unico grafo, la cui struttura generale è rappresentata in Fig. 5:

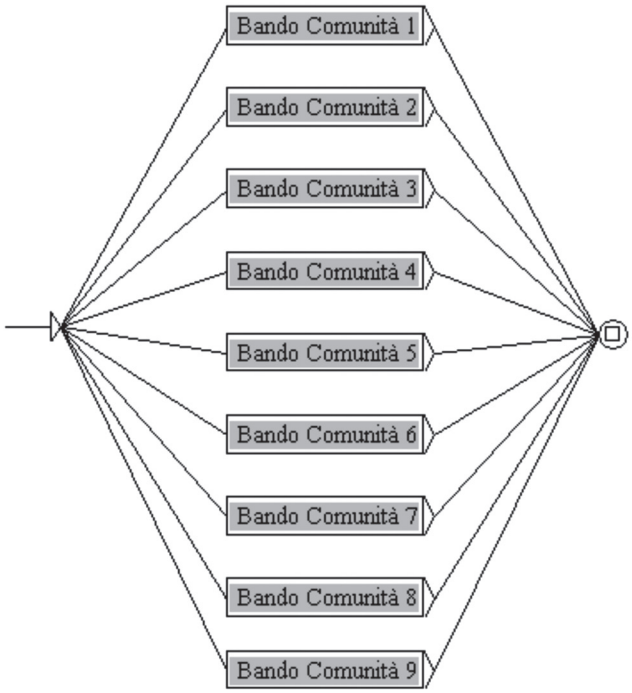

Figura 5

Ogni singolo metanodo del grafo dà la possibilità di accedere ad un altro grafo.

Il procedimento fin qui descritto è stato replicato per tutti i testi del corpus ed ha condotto all'elaborazione di un unico grafo, i cui metanodi consentono di accedere alle grammatiche locali sottostanti, e che può essere utilizzato in un'unica soluzione per il riconoscimento automatico ricorsivo delle formule di cortesia. Inoltre, può rappresentare lo step iniziale verso la costruzione di un risponditore semi-automatico in cui è possibile associare una risposta non casuale ad ogni specifica FAQ riconosciuta.

4. Prospettive future di ricerca

Le prospettive di ricerca sugli argomenti qui esposti si rivolgono principalmente verso la creazione e la strutturazione di uno *Smart Information Retrieval System* in cui, rispetto al presente, vengano maggiormente utilizzati i metodi di formalizzazione del linguaggio indicati dal Lessico-Grammatica. Nello specifico, l'applicazione metodica e congiunta di DELA e FST permetterà di dare vita ad un motore di ricerca più "intelligente", più sensibile alla materia linguistica ed alle sue costrizioni e restrizioni, nonché in grado di recuperare le informazioni dai testi Web sia in base alle unità lessicali – semplici o complesse – da essi contenute, sia in base al parsing sulle strutture morfo-sintattiche dei predicati utilizzati. In tal modo, si riuscirà a ridurre in modo quasi definitivo la distanza che oggi ancora persiste fra le stringhe ricercate nelle finestre di dialogo dei motori Web ed i risultati testuali generalmente ottenuti.

Riferimenti bibliografici

De Bueriis G., Di Maio F., Elia A. and Monteleone M. (2008). Le poliriematiche dell'italiano. In De Bueriis, G. and Elia, A., editors, *Lessici elettronici e descrizioni lessicali, sintattiche, morfologiche ed ortografiche*, Plectica: Salerno, pp. 15-68.

Elia A., Martinelli M. and D'Agostino E. (1981). *Lessico e strutture sintattiche*. Napoli: Liguori.

Gross M. (1975). *Méthodes en syntaxe, régime des constructions complétives*. Paris: Hermann.

Gross M. (1989). La construction de dictionnaires électroniques. *Annales des Télécommunications*, vol. 44, 1-2: 4-19, CENT, Issy-les-Moulineaux/Lannion.

Monteleone M. (2002). *Lessicografia e dizionari elettronici. Dagli usi linguistici alle basi di dati lessicali*. Napoli: Fiorentino & New Technology.

Silberztein M. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Paris: Masson.

Vietri S. (2001a). *Navigare nei testi. Teorie e applicazioni informatiche per la linguistica testuale*. Napoli: Editoriale Scientifica Italiana.

Vietri S. and Elia A. (2001b). Analisi automatica dei testi e dizionari elettronici. In Burattini, E. and Cordeschi, R. (2001). *Intelligenza artificiale*. Roma: Carocci.

Vietri S. (2008). *Dizionari elettronici e grammatiche a stati finiti*. Salerno: Plectica.

Sitografia

Common agricultural policy (http://ec.europa.eu/agriculture/glossary/glossary_it.pdf).

CORDIS Research and Development (http://cordis.europa.eu/guidance/glossary_it.html).

Dipartimento per le Politiche di Sviluppo e Coesione (http://www.dps.tesoro.it/glossario.asp).

Environment glossary (http://ec.europa.eu/atoz_it.htm).

EU Competition Policy (http://ec.europa.eu/competition/publications/glossary_it.pdf).

Europa glossary (http://europa.eu/scadplus/glossary/index_it.htm).

European Convention Glossary (http://european-convention.eu.int/glossary.asp?lang=IT).

European Judicial Network in civil and commercial matters (http://ec.europa.eu/civiljustice/glossary/glossary_it.htm).

Gergo europeo (http://europa.eu/abc/eurojargon/index_it.htm).

Manuale interistituzionale di convenzioni redazionali (http://publications.europa.eu/code/it/it-390500.htm).

Regional Development Glossary (http://ec.europa.eu/regional_policy/glossary/glossary_it.htm).