Documenti digitali

a cura di Roberto Guarasci
in collaborazione con Antonietta Folino

ITER

Documenti digitali
a cura di Roberto Guarasci
in collaborazione con Antonietta Folino

EDITORE:
ITER Srl
Via F.lli Bressan 14
20126 Milano

ISBN: 978-88-903419-3-9

Finito di stampare: Aprile 2013

Indice

Estrazione Terminologica Automatica e Indicizzazione: Scenari Applicativi, Problemi e Possibili Soluzioni

SIMONETTA MONTEMAGNI*

1. Introduzione

Il ricorso a metodi e tecniche di estrazione automatica di terminologia[1] settoriale da corpora di dominio, ovvero da insiemi di documenti relativi a uno specifico settore della conoscenza, rappresenta una sempre più diffusa pratica di supporto al processo di indicizzazione di collezioni documentali, inteso come l'operazione volta all'individuazione delle voci indice che ne costituiscono il contenuto concettuale. L'obiettivo di questo contributo è una rivisitazione critica di esperienze condotte all'interno di diversi scenari applicativi in cui i risultati del processo di estrazione automatica di terminologia sono utilizzati per la costruzione di vocabolari controllati o di thesauri[2] sulla base dei quali è condotto il processo di indicizzazione.

* Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche.

1 Per estrazione terminologica si intende l'identificazione e il recupero da una collezione documentale di termini che sono ritenuti significativi rispetto al dominio al quale i documenti si riferiscono. Tale operazione può essere eseguita manualmente, semi-automaticamente o automaticamente: in quanto segue, ci concentreremo sulle modalità automatica e semi-automatica di estrazione terminologica.

2 Nel caso del thesaurus, il vocabolario controllato si arricchisce di relazioni semantiche tra i termini identificati. In particolare si distingue tra rela-

Nella sua forma più semplice, un vocabolario controllato è costituito dal lessico che rappresenta un sapere specialistico, per esempio un elenco dei termini specifici di una disciplina (arte, medicina, economia, ecc.). Un vocabolario controllato di questo tipo può essere costruito manualmente da esperti di dominio, oppure essere acquisito in modo semi-automatico facendo uso di metodi e tecniche di trattamento automatico del linguaggio. Nel secondo caso, si va dalla soluzione più elementare di scartare dal vocabolario sottostante a una collezione di testi, assunta come rappresentativa delle conoscenze relative a uno specifico settore del sapere, le cosiddette *stop-words* (ovvero, parole semanticamente *vuote* come articoli, preposizioni, pronomi ecc.) fino al ricorso a tecniche avanzate di filtraggio dei termini rilevanti, ad esempio di estrazione terminologica.

Oltre agli inevitabili costi di costruzione di un vocabolario controllato, l'indicizzazione condotta in relazione a risorse costruite manualmente presenta numerosi problemi. Ad esempio, non è garantito l'allineamento tra il vocabolario controllato e la terminologia usata nei testi per convogliare i contenuti, con potenziali e non indifferenti ripercussioni a livello dell'indicizzazione se condotta in modo automatico. Un'altra difficoltà è connessa con la valutazione della terminologia complessa, ovvero costituita da sequenze di più parole: su che base un esperto decide se una sequenza di parole rappresenta un termine complesso oppure se il contenuto sottostante vada ricondotto alle singole parole che lo compongono? Nell'intuizione dell'esperto, per quanto radicata nella sua conoscenza approfondita del dominio, vi è inevitabilmente un margine di soggettività.

Consideriamo ora il caso dell'indicizzazione condotta in rela-

zioni di equivalenza, per la gestione dei sinonimi, delle varianti, ecc., gerarchiche, per correlare concetti legati da un rapporto genere-specie o parte-tutto, e associative, per definire i restanti tipi di relazioni che possono sussistere tra due o più concetti.

zione a vocabolari controllati costruiti in modo automatico con filtraggio di *stop-words*. Come (Chung, Nation, 2004) osservano, *«it seems that even after eliminating the stop words, the most frequent words from a specialized corpus are not all true terms but include many general words used across a wide range of subjects»*[3]. In questo caso, non tutte le unità di indicizzazione sono rilevanti rispetto al dominio. Un ulteriore problema connesso con questo tipo di approccio riguarda la composizione del vocabolario controllato che contiene soltanto termini singoli o *unità terminologiche monorematiche*, composte da un'unica parola. Ciò è in contrasto con quanto sappiamo della terminologia specialistica, che è prevalentemente costituita da termini complessi, o *unità terminologiche polirematiche* costituite da sequenze di più parole: si vedano in proposito (Jackendoff, 1997)[4], (Nakagawa, Mori, 2003)[5] e (Chung, Nation, 2004)[6].

L'ultima opzione è costituita dall'indicizzazione condotta in relazione a vocabolari controllati costruiti in modo semi-automatico mediante tecniche di estrazione di terminologia di dominio. In linea di principio, oggi questa rappresenta la soluzione ottimale che supera i limiti rilevati in relazione agli altri approcci: l'allineamento con i testi è garantito così come l'inclusione all'interno del vocabolario controllato di terminologia polirematica selezionata su base statistica; inoltre, il vocabolario così acquisito dovrebbe contenere solo i termini rilevanti per il dominio di indagine. Questa evoluzione è ben delineata da (Manning et

---

3 TERESA MIHWA CHUNG, PAUL NATION, *Identifying technical vocabulary*, in «System», vol. 32, 2004, p.259.

4 Cfr. RAY JACKENDOFF, *Twistin' the night away*, in «Language», vol. 73, 1997, pp. 534-559.

5 Cfr. HIROSHI NAKAGAWA, TATSUNORI MORI, *Automatic Term Recognition based on Statistics of Compound Nouns and their Components*, in «Terminology», vol. 9, n. 2, 2003, pp. 201-209.

6 Cfr. CHUNG, T.M., NATION, P., *op. cit.*

alii, 2008) nel loro libro sul recupero di informazioni da testi, che affermano:

> *The general trend in Information Retrieval systems over time has been from standard use of quite large stop lists (200–300 terms) to very small stop lists (7–12 terms) to no stop list whatsoever. Web search engines generally do not use stop lists. Some of the design of modern IR systems has focused precisely on how we can exploit the statistics of language so as to be able to cope with common words in better ways*[7].

Il ricorso a tecniche di estrazione automatica di terminologia di dominio da collezioni documentali si presenta dunque come la direzione da seguire nella selezione dei termini di indicizzazione da includere in un vocabolario controllato: ciò può riguardare la costruzione ex novo di un vocabolario controllato, oppure la sua estensione a partire da una base compilata in modo manuale, o l'aggiornamento di una versione precedente.

Se l'utilizzo di tecniche di estrazione terminologica appare assodato, rimane da valutare se le tecniche correnti siano sempre adeguate ed efficaci per il trattamento di linguaggi settoriali di diversi domini del sapere. Per quanto gli obiettivi dell'indicizzazione e dell'estrazione terminologica coincidano parzialmente, vi sono importanti differenze che possono rendere il risultato dell'estrazione terminologica non del tutto adeguato ai fini del processo di indicizzazione. Se l'obiettivo dell'indicizzazione è quello di trovare termini in grado di discriminare un documento da un altro, quello dell'estrazione di terminologia è l'identificazione di termini settoriali che designano concetti di un dominio specifico: ne consegue che un termine di indicizzazione può non

[7] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008, p. 27.

costituire un termine di dominio così come, nel caso di alcune collezioni documentali, i termini settoriali possono non rappresentare unità di indicizzazione utili.

Partendo dall'analisi dei risultati ottenuti con sistemi di estrazione terminologica in diversi scenari applicativi in relazione a diversi domini del sapere, il presente contributo indaga questi interrogativi e illustra alcune soluzioni avanzate per il superamento – o almeno il ridimensionamento – delle criticità rilevate.

L'articolo è organizzato come segue. Dopo un inquadramento della nozione di linguaggio settoriale con particolare riguardo al rapporto con la lingua comune e sue variazioni d'uso (sezione 2) segue una breve rassegna delle tecniche correnti per l'estrazione automatica di terminologia di dominio da collezioni documentali (sezione 3). Una volta inquadrata la tipologia di problemi che si trovano a fronteggiare i sistemi di estrazione terminologica in relazione a collezioni documentali rappresentative di diverse varietà di linguaggi settoriali (sezione 4), la seconda parte di questo contributo discute possibili soluzioni messe a punto per il superamento dei problemi enucleati e i risultati raggiunti all'interno di diversi scenari applicativi (sezione 5).

2. Il linguaggio della comunicazione specialistica: rapporto con la lingua comune e varietà d'uso

Nella ricerca linguistica italiana si osserva un alto grado di variazione nella designazione della nozione di *language for special purposes*: le denominazioni spaziano da *lingue speciali*, *lingue per scopi speciali*, *lingue di specializzazione*, *linguaggi settoriali*, *micro lingue*, *sottocodici*, *linguaggi specialistici*, e sono spesso usate con accezioni non del tutto equivalenti. Ad esempio, (Sobrero, 1993)[8] distingue tra *lingue specialistiche*, caratterizza-

[8] Cfr. ALBERTO SOBRERO, *Lingue speciali*, in Introduzione all'italiano con-

te da un alto grado di specializzazione come il linguaggio della medicina, della fisica, dell'informatica o del diritto, e *lingue settoriali*, che mostrano un minor grado di specializzazione come quella della pubblicità, della politica o della burocrazia: la differenza tra i due tipi si colloca principalmente a livello del lessico (regole di formazione delle parole, quantità di terminologia specialistica all'interno del vocabolario del testo). (Dardano, 1987)[9], invece, opta per un'unica classe dei cosiddetti *linguaggi settoriali* all'interno della quale distingue tra linguaggi *forti*, altamente organizzati e stabili dal punto di vista lessicale (come il linguaggio della matematica), e linguaggi *deboli*, caratterizzati da organizzazioni lessicali meno strutturate, come il linguaggio giuridico. Se da un lato questo estremo grado di variabilità nella denominazione di tale nozione rappresenta senza dubbio un indizio della mancanza, nella terminologia linguistica italiana odierna, di una definizione univoca della nozione di lingua utilizzata nella comunicazione specialistica, dall'altro mette in evidenza che la tipologia dei linguaggi usati all'interno dei domini del sapere è varia. In quanto segue, ci riferiremo a questa nozione con il termine di *linguaggi settoriali*.

Ai fini del presente contributo, due sono gli aspetti rilevanti del dibattito sui linguaggi settoriali: il rapporto tra questi e la lingua comune e l'esistenza di varietà d'uso all'interno di uno stesso linguaggio settoriale. Si tratta di due questioni parzialmente correlate, in quanto il rapporto con la lingua comune diventa un fattore che contribuisce significativamente all'identificazione di diverse varietà d'uso all'interno dello stesso linguaggio settoriale. Questi due aspetti sono ampiamente dibattuti nella letteratura

temporaneo. La variazione e gli usi, Sobrero A. (a cura di), Roma-Bari, Laterza, 1993, pp. 237-277.

9 Cfr. MAURIZIO DARDANO, *Linguaggi settoriali e processi di riformulazione*, in Parallela 3. Linguistica contrastiva / Linguaggi settoriali / Sintassi generativa, Dressler W. et alii (a cura di), Tubinga, Narr, 1987, pp. 134-145.

linguistica ma, come vedremo in seguito, non sono stati affrontati in modo sistematico nella definizione dei correnti metodi e tecniche di estrazione automatica di terminologia settoriale da corpora di dominio.

Il rapporto tra la lingua comune e i diversi linguaggi settoriali è stato ampiamente dibattuto nella letteratura italiana e internazionale[10]. Si tratta di una distinzione che sfugge a una caratterizzazione univoca. (Varantola, 1986)[11] osserva al riguardo quanto segue: *«Definitions of LSP [Language for Special Purposes] or SL [Specialized Language] versus GL [General Language] abound; none is universally applicable, for obvious reasons. Basically we are dealing with two intuitively correct assumptions that are good as working concepts but which resist a clear-cut definition and delimitation»*. Piuttosto che mirare a definire un confine che separi nettamente la lingua comune dal linguaggio settoriale, si è andata affermando l'idea che i due tipi di linguaggio rappresentano i due poli di un continuum che si estende dalla lingua comune ai linguaggi settoriali e caratterizzato da una gamma di livelli intermedi[12]. I due estremi di questo continuum presentano divergenze significative dal punto di vista linguistico, pragmatico e funzionale (Cabré, 1999)[13]: per quanto riguarda il

[10] Per una rassegna del dibattito sul problema si rinvia a: sul versante nazionale italiano, Cfr. STEFANIA CAVAGNOLI, *La comunicazione specialistica*, Roma, Carocci, 2007; a livello internazionale, Cfr. MARIA TERESA CABRÉ, *Terminology: Theory, Methods, and Applications*, Amsterdam, John Benjamins, 1999.

[11] KRISTA VARANTOLA, *Special Language and General Language: Linguistic and Didactic Aspects*, in «Unesco ALSED-LSP Newsletter», vol. 9, n. 2, dicembre 1986, p.10.

[12] Cfr. in proposito, tra gli altri: GUY RONDEAU, JUAN SAGER, *Introduction à la terminologie*, ed. 2, Chicoutimi, Gatan Morin, 1984; VARANTOLA, K., *op. cit.*; GIANFRANCO PORCELLI, *Principi di Glottodidattica*, Brescia, La Scuola, 1994.

[13] Cfr. CABRÉ, M.T., *op. cit.*

piano linguistico, i tratti divergenti spaziano tra i diversi livelli di descrizione linguistica, da quello lessicale, per il quale si registrano le differenze più significative, a quelli morfologico e sintattico (ad esempio, vi sono costruzioni sintattiche o formazioni morfologiche che sono parte del linguaggio comune ma non si registrano, se non in modo del tutto sporadico, all'interno dei linguaggi settoriali).

Nella transizione dalla lingua comune al linguaggio settoriale si osservano varietà linguistiche intermedie. Questa prospettiva sul rapporto tra lingua comune e linguaggi settoriali introduce il secondo problema che intendiamo affrontare in questa sede, ovvero che all'interno della comunicazione specialistica si osservano dimensioni di variazione che danno luogo a diversi tipi di linguaggi settoriali, anche all'interno dello stesso dominio del sapere. La Figura 1 riporta il grafico proposto da (Rondeau, 1983)[14] e riprodotto in (Cabré, 1999)[15] che visualizza il complesso rapporto che lega lingua comune (LC) e linguaggi settoriali (LS) di cui si distinguono diverse varietà a seconda del grado di specializzazione.

Nella figura ciascun settore corrispondente a un linguaggio settoriale presenta variazioni di tipo verticale, corrispondenti a diversi livelli di comunicazione, che vanno da un livello altamente specialistico e specializzato, in cui gli attori della comunicazione sono esclusivamente esperti di dominio, a livelli comunicativi più vicini all'utente comune come quello della divulgazione o della comunicazione didattica, in cui gli attori sono esperti e non esperti (nel caso della divulgazione) o futuri esperti (nella comunicazione di tipo didattico). Questo tipo di variazione verticale all'interno della comunicazione specialistica è vi-

[14] GUY RONDEAU, *Introduction à la terminologie*, Québec, Gaëtan Morin éditeur, 1983.

[15] CABRÉ, M.T., *op. cit.*, p. 69.

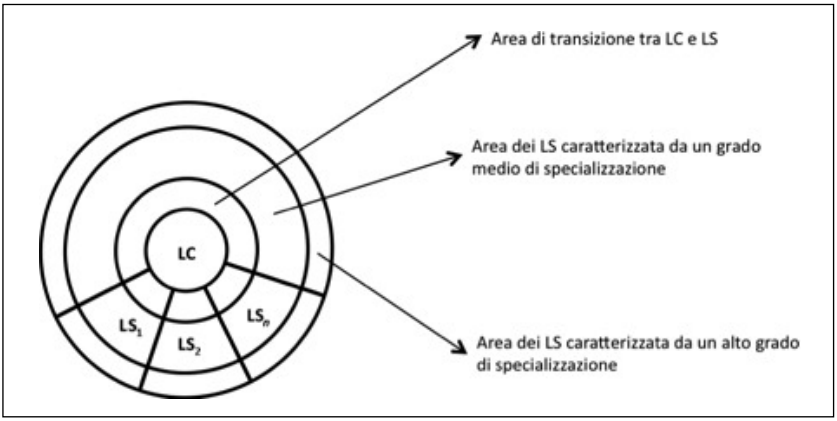

Figura 1. Rapporto tra lingua comune e linguaggi settoriali.

sta da (Lavinio, 2004)[16] come una variazione di tipo diastratico in quanto legata al livello socio-culturale degli interlocutori. Questa variazione di natura verticale si intreccia e sovrappone parzialmente con una variazione di tipo diafasico, ovvero legata alle modifiche che avvengono in una situazione comunicativa, sulla base del contesto, degli interlocutori, degli scopi della comunicazione e connessa ai diversi registri della lingua: a seconda della situazione all'interno della quale si colloca la comunicazione specialistica e delle sue finalità, si possono riconoscere sottovarietà di uno stesso linguaggio settoriale.

Sulla base di quanto detto finora, appare difficile marcare confini netti nella descrizione dei linguaggi settoriali, sia quando si vada a rapportarli alla lingua comune, sia quando se ne vadano a identificare sottovarietà definite sulla base della situazione comunicativa all'interno della quale sono usati. Ciò nonostante, i linguaggi settoriali rappresentano un'unica varietà che si diffe-

16 Cfr. CRISTINA LAVINIO, *Comunicazione e linguaggi disciplinari. Per un'educazione linguistica trasversale*, Roma, Carocci, 2004.

renzia nettamente dalla lingua comune: come afferma (Cabré, 1999), *«from a theoretical and methodological standpoint, it is important to establish the concept of special language in the singular»*[17]. Tuttavia, continuando con Cabré, la nozione astratta di linguaggio settoriale può essere suddivisa in sottovarietà a seconda della situazione comunicativa all'interno della quale sono usate, con importanti ripercussioni a livello dell'uso terminologico. A questo punto, la nozione unitaria di linguaggio settoriale si frammenta: a seconda che si tratti di comunicazione legata all'elaborazione del sapere o all'apprendimento o alla sua applicazione o alla sua divulgazione, il tipo di terminologia usata per convogliare gli stessi contenuti varierà in modo significativo, presentando diverse *miscele* di terminologia specialistica e lessico comune.

Finora, si è parlato di linguaggi settoriali e relative sottovarietà in termini molto astratti. Tuttavia, in questa sede il nostro interesse è legato al fatto che tali linguaggi si manifestano nei testi specialistici, che contengono – oltre agli elementi specialistici – quelli della lingua comune. Quella di linguaggio settoriale è un'astrazione costruita a partire dai testi. E' con i testi che ci confronteremo nel prosieguo di questo articolo, in quanto costituiscono il punto di partenza del processo di estrazione terminologica.

3. Estrazione automatica di terminologia specialistica da corpora di dominio

Sempre più centrali per lo sviluppo di applicazioni reali di gestione della conoscenza (o Knowledge Management), i sistemi di estrazione automatica di terminologia sono finalizzati all'identificazione e all'estrazione di unità terminologiche monorematiche (come *accordo*, *produttore* o *presidente*) e polirematiche (co-

[17] CABRÉ, M.T., *op. cit.*, p.76.

me *procedimento amministrativo*, *Ministro dell'ambiente*, *incenerimento dei rifiuti pericolosi*, *assistenza reciproca*, *contratto di multiproprietà*) da corpora di dominio. Questo compito, le cui denominazioni spaziano nella letteratura sull'argomento da *terminology extraction* a *automatic term recognition* fino a *terminology mining*, rappresenta il primo e ormai consolidato passo nel processo incrementale di estrazione di conoscenza ontologica (denominato *Ontology Learning*[18]) da collezioni documentali (Buitelaar et alii, 2005)[19]. Si parte dall'assunto di base che i termini costituiscono la rappresentazione linguistica dei concetti specifici di un dominio e per questo motivo il compito di estrazione terminologica rappresenta il primo e fondamentale passo verso l'accesso al contenuto di collezioni documentali.

Il processo di estrazione terminologica si articola in due passi fondamentali:

1) identificazione delle potenziali unità terminologiche, siano esse monorematiche oppure polirematiche;
2) filtraggio della lista dei termini candidati al fine di discriminare la terminologia di dominio da non-termini (o parole comuni).

Queste due fasi del processo estrattivo possono essere basate su diversi tipi di evidenza, ovvero linguistica, statistica oppure una combinazione dei due: quest'ultimo rappresenta il caso più frequente.

---

18 Per *Ontology Learning* si intende il processo di supporto automatico o semi-automatico nello sviluppo di ontologie di dominio, attraverso l'acquisizione di conoscenza a partire dai testi. Un'ontologia è una rappresentazione formale e condivisa di un dato dominio di conoscenza.

19 Cfr. PAUL BUITELAAR, PHILIPP CIMIANO, BERNARDO MAGNINI, *Ontology Learning from Text: An Overview*, in Ontology Learning from Text: Methods, Evaluation and Applications, Buitelaar P. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications Series», vol. 123, IOS Press, 2005, pp. 3-12.

L'identificazione dei potenziali termini avviene in relazione al testo linguisticamente annotato, ovvero arricchito con informazione relativa alla struttura linguistica sottostante: tipicamente, a tal fine viene utilizzato il testo annotato morfo-sintatticamente oppure segmentato sintatticamente in costituenti sintattici elementari non ricorsivi detti *chunk* (Abney, 1991), (Federici et alii, 1996)[20]. Al livello di annotazione morfo-sintattica, a ogni parola (*token*[21]) del testo viene associata informazione relativa alla categoria grammaticale che la parola ha nel contesto specifico (ad es. sostantivo, verbo, aggettivo); nell'annotazione sintattica non ricorsiva, il testo viene segmentato in *chunk*, ovvero sequenze di parole del testo che vanno da un'unità grammaticale (tipicamente, una preposizione, un ausiliare, un [pre]determinatore o un ausiliare) fino alla prima unità lessicale semanticamente *piena* selezionata dall'unità grammaticale (esempi di *chunk* di tipo nominale sono: *un difficile problema* oppure *la mia prima casa*).

Il testo arricchito con informazione linguistica viene analizzato da una mini-grammatica deputata al riconoscimento delle strutture linguistiche che formano potenziali termini. Nel caso di unità terminologiche monorematiche, si farà riferimento alle categorie grammaticali: tipicamente, verranno identificati come candidati tutti i *token* etichettati come nomi, anche se in linea di principio la terminologia di un dominio specifico include anche aggettivi o verbi che designano proprietà o eventi tipici del dominio. In questa sede, ci limiteremo a considerare terminologia

---

20 Cfr. STEVEN ABNEY, *Parsing by chunks*, in Principle-based Parsing: Computation and Psycholinguistics, Berwick R.C. et alii (a cura di), Dordrecht, Kluwer, 1991, pp. 257-278. Per questo tipo di annotazione in relazione alla lingua italiana, cfr. STEFANO FEDERICI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, *Shallow Parsing and Text Chunking: a View on Underspecification in Syntax*, in Proceedings of Workshop On Robust Parsing and Eight Summer School on Language, Logic and Information, Praga, Repubblica Ceca, 12-16 agosto 1996, pp. 35-44.

21 Ogni parola unità distinta del testo.

di tipo nominale. L'identificazione delle unità terminologiche polirematiche può avvenire in relazione al testo morfo-sintatticamente annotato o segmentato in costituenti sintattici non ricorsivi. In entrambi i casi, la mini-grammatica sarà finalizzata al riconoscimento di sequenze di categorie grammaticali o di *chunk* che corrispondono a potenziali unità polirematiche: si va da semplici casi di modificazione aggettivale (es. *inquinamento atmosferico*), a casi di modificazione con un complemento preposizionale (es. *diritto di recesso*), fino a strutture più complesse che combinano diversi tipi di modificatori (es. *encefalopatia spongiforme bovina*, *somatotropina bovina di ricombinazione*, *agenzia europea di valutazione dei medicinali* oppure *trattamento degli oli usati*).

La seconda fase del processo estrattivo è volta a verificare se e in quale misura un termine candidato rappresenti un termine valido per il dominio considerato. A questo scopo, nella letteratura sull'estrazione terminologica vengono utilizzate una serie di misure statistiche finalizzate a discriminare la terminologia di dominio da parole comuni, caratterizzate da una designazione generica. In particolare, l'estrazione di unità monorematiche è tipicamente realizzata sulla base della distribuzione di frequenza all'interno del corpus, oppure su misure di rilevanza statistica tipiche dell'Information Retrieval quali la *TF/IDF* (*Term Frequency/Inverse Document Frequency*) (Salton, Buckley, 1998), (Baeza-Yates, Ribeiro-Neto, 1999)[22]. Per le unità polirematiche, oltre alle misure già menzionate, vi sono metodi e tecniche che si basano sull'assunto di base che se due o più parole formano un termine è molto probabile che nell'uso linguistico relativo a quel dominio esse tendano a ricorrere insieme in maniera statistica-

[22] Cfr. GERARD SALTON, CHRIS BUCKLEY, *Term-Weighting Approaches in Automatic Text Retrieval*, in «Information Processing and Management», vol. 24, n. 5, 1988, pp. 513-523; Cfr. RICARDO BAEZA-YATES, BERTHIER RIBEIRO-NETO, *Modern Information Retrieval*, New York, ACM Press,1999.

mente significativa. La significatività del legame sussistente tra le parole che formano il termine viene calcolata attraverso il ricorso a misure di associazione che considerano la frequenza di co-occorrenza delle parole che compongono l'unità terminologica polirematica in relazione alle occorrenze totali delle singole parole che la formano: per menzionarne alcune, la misura della *Mutual Information* (Church, Hanks, 1990)[23] o della *Log-likelihood* (Dunning, 1993)[24] per arrivare al più recente e più sofisticato metodo denominato *C/NC-value* (Frantzi et alii, 2000)[25] che rappresenta uno standard *de facto* nel settore dell'estrazione terminologica (Vu et alii, 2008)[26]. Seguendo (Kageura, Umino, 1996)[27], la varietà di misure utilizzate per l'estrazione di terminologia settoriale da corpora di dominio può essere ricondotta a due classi fondamentali, a seconda che valutino:

a) l'unità di un termine (*unithood*), ovvero la forza di associazione che lega le parole che formano un'unità terminologica polirematica. Ovviamente, questo tipo di misura si applica solo nel caso di unità terminologiche formate da più parole;
b) la pertinenza rispetto al dominio (*termhood*), espressa co-

[23] Cfr. KENNETH WARD CHURCH, PATRICK HANKS, *Word association norms, mutual information, and lexicography*, in «Computational Linguistics», vol. 16, n. 1, 1990, pp. 22-29.

[24] Cfr. TED DUNNING, *Accurate Methods for the Statistics of Surprise and Coincidence*, in «Computational Linguistics», vol. 19, n. 1, 1993, pp. 61-74.

[25] Cfr. KATERINA FRANTZI, SOPHIA ANANIADOU, HIDEKI MIMA, *Automatic recognition of multi-word terms*, in «International Journal of Digital Libraries», vol. 3, n. 2, 2000, pp.117-132.

[26] Cfr. THUY VU, AITI AW, MIN ZHANG, *Term Extraction Through Unithood and Termhood Unification*, in Third International Joint Conference on Natural Language Processing. Proceedings of the Conference, Hyderabad, India, 07-12 gennaio 2008, pp. 631-636.

[27] Cfr. KYO KAGEURA, BIN UMINO, *Methods of automatic term recognition: a review*, in «Terminology», vol. 3, n. 2, 1996, pp. 259-289.

me misura del grado di rilevanza di una parola all'interno del dominio considerato, ovvero di quanto un termine candidato costituisca un'unità rappresentativa del contenuto della base documentale. Diversamente dal caso precedente, questa classe di misure si applica a unità terminologiche sia monorematiche sia polirematiche.

Mentre le misure come la frequenza grezza o *TF/IDF* così come il *C/NC-value* sono riconducibili alla seconda classe finalizzata alla quantificazione della rilevanza rispetto al dominio di un termine candidato, le misure di riconducibili alla prima classe, che includono la *Log-likelihood* e la *Mutual Information*, catturano piuttosto la forza e la stabilità dell'associazione che lega le parole che formano un termine polirematico. Data la complementarità di questi due tipi di misure, recentemente si registrano vari tentativi di combinarli ai fini dell'acquisizione di termini di dominio (Vu et alii, 2008)[28].

Nonostante le differenze rilevate, le misure viste finora si basano tutte sulla distribuzione dei termini candidati all'interno di uno stesso dominio, studiato attraverso una collezione documentale che ne convoglia i contenuti tipici. Un'altra classe di approcci all'estrazione terminologica si basa ancora su evidenza di tipo distribuzionale, ma rilevata attraverso un'analisi contrastiva inter-dominio: in questo caso, l'estrazione di unità terminologiche monorematiche e polirematiche è condotta a partire dal confronto della distribuzione dei termini candidati nel corpus di acquisizione rispetto a un corpus di riferimento (detto anche *corpus di contrasto*). In questo modo, la lista finale di unità terminologiche estratte conterrà quelle unità che sono maggiormente rilevanti nel corpus di acquisizione rispetto al corpus di riferimento. A questo scopo è stata sviluppata una serie di metodi in grado di computare la misura della diversa rilevanza di unità ter-

[28] Cfr. VU, T. et alii, *op. cit.*

minologiche all'interno dei due corpora oggetto dell'analisi contrastiva. La possibilità di discriminare termini e non-termini è così empiricamente realizzata sulla base del confronto della loro distribuzione in un corpus di dominio (il corpus di acquisizione) rispetto a un altro corpus: il corpus di riferimento è generalmente rappresentativo della lingua comune, ma a seconda del tipo di analisi contrastiva che si vuole condurre potrebbe anche essere relativo ad un altro dominio specialistico (cfr. sezione 5.2). Questo tipo di approccio può essere usato direttamente per l'identificazione dei termini all'interno di collezioni documentali di dominio (Basili et alii, 2001)[29], così come può essere utilizzato in combinazione con le misure precedenti (Bonin et alii, 2010a)[30].

Qualsiasi sia la tecnica adottata, il risultato del processo di estrazione automatica di terminologia da corpora di dominio dovrà essere validato e filtrato da parte di esperti che saranno supportati nelle decisioni finali non solo dalla loro competenza del dominio analizzato, ma anche da evidenza statistica che riflette la significatività dei termini acquisiti, sia essa costituita dalla rilevanza rispetto al dominio (intra-dominio o inter-dominio), oppure dalla forza di associazione che lega le parole all'interno di termini poliremiatici. Per questo motivo, il processo di costruzione di vocabolari controllati basato su questo approccio viene definito complessivamente come semi-automatico.

---

29 Cfr. ROBERTO BASILI, ALESSANDRO MOSCHITTI, MARIA TERESA PAZIENZA, FABIO MASSIMO ZANZOTTO, *A contrastive approach to term extraction*, in Atti della «4th Conference on Terminology and Artificial Intelligence (TIA-2001)», Nancy, 3-4 maggio 2001; Cfr. CHUNG, T.M., NATION, P., *op. cit.*; Cfr. ANSELMO PENAS, FELISA VERDEJO, JULIO GONZALO, *Corpus-Based Terminology Extraction Applied to Information Access*, in Proceedings of the Corpus Linguistics 2001 Conference, Università di Lancaster, 29 marzo – 2 aprile 2001, Rayson P., Wilson A., McEnery T., Hardie A., Khoja S. (ed.), pp. 458-465.

30 Cfr. FRANCESCA BONIN, FELICE DELL'ORLETTA, SIMONETTA MONTEMAGNI, GIULIA VENTURI (a), *A Contrastive Approach to Multi-word Extraction*

4. Questioni aperte nel processo di estrazione terminologica da corpora

Al termine della breve rassegna sui metodi e le tecniche correntemente usati per l'estrazione di terminologia da corpora di dominio riportata nella precedente sezione, appare legittimo chiedersi se la loro affidabilità ed efficacia possano essere influenzate dal tipo di linguaggio settoriale usato nel corpus impiegato per l'acquisizione. Concludendo la loro illustrazione del metodo *C/NC-value* testato su un corpus di dominio biomedico, (Frantzi et alii, 2000)[31] affermano che «*although we have shown that the method performs well for this text type of corpora, we are cautious in making this claim for other types of special language corpora, before conducting appropriate experiments*»: ciò lascia chiaramente intravedere che l'efficacia del metodo possa variare in relazione al tipo di corpus di acquisizione.

I sistemi di estrazione terminologica sono nati in relazione a collezioni di testi caratterizzati da un lessico altamente specialistico e rivolti a un pubblico di esperti, come ad esempio la letteratura biomedica. Se i risultati raggiunti su questa tipologia di testi sono ormai più che soddisfacenti, rimane una questione aperta: il loro *rendimento scientifico* in relazione a corpora rappresentativi di domini non altamente specialistici e/o composti da testi rivolti ad un ampio pubblico.

Gli approcci correnti al problema non sembrano aver affrontato in modo sistematico i diversi ordini di difficoltà connessi con la varia tipologia di linguaggi settoriali descritta nella sezione 2. In relazione a ciò, (Cabré, 1999)[32] ricorda come le maggiori

from Domain-specific Corpora, in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 17-23 maggio 2010, pp. 3222–3229.

31 FRANTZI, K. et alii, *op. cit.*, p. 130.

32 Cfr. CABRÉ, M.T., *op. cit.*

difficoltà siano dovute proprio al confine non sempre così netto tra linguaggi settoriali e lingua comune, nonché al costante scambio biunivoco che li lega. La questione riguarda in particolare la difficoltà di estrarre terminologia rilevante di dominio, ovvero di distinguere tra termini del dominio (lessico settoriale) e non-termini (lessico comune), tenendo in considerazione la varie dimensioni di variazione all'interno di questa classe di linguaggi. In particolare, i problemi sono connessi con la difficoltà di estrarre la terminologia a partire da corpora rappresentativi di linguaggi settoriali caratterizzati da diversi livelli di specializzazione, oppure da collezioni di testi appartenenti a diversi tipi di registri.

La discriminazione tra termini settoriali e parole comuni non è tuttavia l'unico aspetto che non trova una risposta adeguata nei sistemi correnti di estrazione terminologica automatica. Un ulteriore problema riguarda il trattamento di testi all'interno dei quali si osserva una commistione di tipologie di termini, non sempre nettamente distinguibili tra di loro. Questo è il caso, ad esempio, dei testi giuridici: una peculiarità della lingua del diritto, che contribuisce alla sua eterogeneità dal punto di vista terminologico, è dovuta agli stretti e biunivoci rapporti con la lingua comune da un lato e con i linguaggi settoriali dall'altro (Cortelazzo, 1997), (Venturi, 2011)[33]. Ne consegue che nei corpora rappresentativi della lingua del diritto si intrecciano il lessico del dominio giuridico, quello proprio della materia legislata così come del linguaggio comune. Ai fini del presente studio, possiamo de-

[33] Cfr. MICHELE CORTELAZZO, *Lingua e diritto in Italia. Il punto di vista dei linguisti*, in La lingua del diritto. Difficoltà traduttive. Applicazioni didattiche, Atti del primo Convegno Internazionale, Milano, 5-6 ottobre 1995, Schena L. (a cura di), Roma, CISU (Centro d'Informazione e Stampa Universitaria), 1997, pp. 35-50; Cfr. GIULIA VENTURI, *Lingua e diritto: una prospettiva linguistico-computazionale*, Tesi di Dottorato, Università degli Studi di Torino, Scuola di Dottorato in Studi Umanistici, 2011.

finirli corpora *multi-dominio*: si tratta di una tipologia di testi che trova nei corpora giuridici un esempio prototipico, ma che non è circoscritta ad essi. Altri esempi di tale tipologia di corpora includono: il corpus dei testi che descrivono le linee di attività del CNR (Guarasci, 2006)[34] in cui si intreccia terminologia relativa alla gestione delle linee di attività con la terminologia del settore disciplinare a cui tali attività si riferiscono; oppure corpora della Pubblica Amministrazione, come quello trattato in (Taverniti, 2008)[35], in cui la terminologia relativa all'oggetto della comunicazione (i beni e i servizi informatici oggetto di acquisto) si combina con la terminologia propria del linguaggio burocratico.

Ai fini dell'indicizzazione di tali corpora, è molto importante poter discriminare tra i diversi tipi di contenuti. (Francesconi et alii, 2010)[36] hanno recentemente proposto un approccio alla rappresentazione formalizzata della conoscenza giuridica basato sulla distinzione tra conoscenza tecnico-giuridica e conoscenza del mondo regolato: il modello suggerito prevede infatti due distinti livelli di organizzazione, ovvero il *Domain Independent Legal Knowledge level* (DILK) che include concetti giuridici, e il *Domain Knowledge level* (DK) all'interno del quale sono resi espliciti i principali concetti rappresentativi di un determinato dominio di conoscenza regolato dalle norme. Questa doppia ar-

34 Cfr. ROBERTO GUARASCI, *Estrazione terminologica e gestione della conoscenza*, in «iged.it», n. 3, 2006, pp. 46-51.

35 Cfr. MARIA TAVERNITI, *Tra terminologia e documentazione: estrazione automatica di voci indice da corpora documentali della Pubblica Amministrazione*, in «Aida Informazioni», a. 26, n. 1-2, gennaio-giugno 2008, pp. 239-250.

36 Cfr. ENRICO FRANCESCONI, SIMONETTA MONTEMAGNI, WIM PETERS, DANIELA TISCORNIA, *Integrating a Bottom-Up and Top-Down Methodology for Building Semantic Resources for the Multilingual Legal Domain*, in Semantic Processing of Legal Texts, Francesconi E. et alii (a cura di), in «LNCS/LNAI», Springer-Verlag, vol. 6036, 2010, pp. 95-127.

ticolazione, riproposta al livello della terminologia acquisita, crea i presupposti per una indicizzazione multi-dimensionale del testo, dove sicuramente la prospettiva del dominio specialistico legislato rappresenta forse la chiave di accesso privilegiata. A conoscenza di chi scrive, le tecniche e i metodi di estrazione terminologica automatica correnti non si sono mai confrontati con casi di acquisizione di terminologia rilevante da corpora *multi-dominio* con il duplice fine di acquisire la terminologia rilevante e di classificarla secondo il dominio di appartenenza.

Nel momento in cui i sistemi di estrazione terminologica da corpora di dominio stanno diventando uno strumento sempre più usato in compiti di Knowledge Management quali l'indicizzazione automatica di testi, la costruzione di ontologie di dominio ecc., le sfide delineate sopra diventano aspetti che non possono essere ignorati nel processo di estrazione terminologica automatica e per i quali vanno trovate soluzioni operative appropriate. In quanto segue, partendo dall'esperienza condotta in diversi scenari applicativi con una piattaforma per l'estrazione da testi di conoscenza terminologico-ontologica (descritta nella sezione 5) verranno illustrati i problemi rilevati e le soluzioni proposte.

5. T2K: una piattaforma per l'estrazione di conoscenza ontologica da collezioni documentali

Text-to-Knowledge, in breve T2K, è una piattaforma software progettata e sviluppata congiuntamente dall'Istituto di Linguistica Computazionale *Antonio Zampolli* del CNR di Pisa e dal Dipartimento di Linguistica dell'Università di Pisa, che si propone di offrire una batteria integrata di strumenti avanzati di analisi linguistica del testo, analisi statistica e apprendimento automatico del linguaggio, destinati a offrire una rappresentazione accurata del contenuto di una base documentale non strutturata, per scopi di indicizzazione avanzata e navigazione intelligente (Del-

l'Orletta et alii, 2008)[37]. T2K trasforma le conoscenze implicitamente codificate all'interno di un corpus di testi in conoscenza esplicitamente strutturata: il risultato finale di questo processo interpretativo spazia dall'acquisizione di conoscenze lessicali e terminologiche complesse alla loro organizzazione in strutture proto-concettuali.

Per arrivare a identificare i concetti rilevanti e più caratterizzanti i documenti di un certo dominio di interesse, T2K impiega lo stato dell'arte della ricerca in linguistica computazionale. I termini acquisiti da T2K possono essere unità lessicali monorematiche come *monitoraggio* o *audit* oppure unità lessicali polirematiche come *Quadro Comunitario di Sostegno*, *obiettivi specifici*, *progetto integrato*, *autorità di gestione*, *autorità di pagamento*, ecc. Per quanto riguarda le unità monorematiche, il processo estrattivo opera sul testo annotato a livello morfo-sintattico[38] e lemmatizzato[39] e avviene sulla base della loro frequenza all'interno del corpus di acquisizione. Diverso è il caso delle unità terminologiche polirematiche, la cui estrazione si articola in due fasi: la prima finalizzata all'identificazione dei potenziali

37 Cfr. FELICE DELL'ORLETTA, ALESSANDRO LENCI, SIMONE MARCHI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, GIULIA VENTURI, *Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio*, in «Aida Informazioni», a. 26, n. 1-2, gennaio-giugno 2008, pp. 185-206.

38 Lo scopo dell'annotazione morfo-sintattica è l'assegnazione a ogni parola (o *token*) del testo dell'informazione relativa alla *categoria grammaticale* (o *parte del discorso*) che la parola ha nel contesto specifico (ad es. nome, verbo, aggettivo). Questa informazione viene talora integrata da ulteriori specificazioni morfologiche (ad es. riguardanti categorie flessionali come persona, genere, numero, ecc.).

39 Il processo di lemmatizzazione consiste nel ricondurre ogni parola del testo al relativo esponente lessicale o *lemma* (tipicamente l'infinito per i verbi, oppure il singolare maschile per gli aggettivi, corrispondente approssimativamente all'esponente lessicale delle voci di un dizionario).

termini sulla base di una mini-grammatica operante sul testo segmentato sintatticamente in *chunk*, la seconda in cui la forza di associazione tra le parole che compongono il termine candidato viene stimata applicando la misura associativa detta *Log-likelihood*, che si è dimostrata produrre risultati sensibilmente migliori rispetto ad altre misure statistiche in quanto più robusta nel caso di dati linguistici con bassa frequenza di occorrenza. La compilazione di un repertorio di terminologia di dominio sulla base delle concrete attestazioni nei testi costituisce il risultato della prima fase operativa di T2K sulla base del quale è possibile condurre un'indicizzazione terminologica dei documenti. Si noti che in T2K è possibile nonché auspicabile validare il risultato del processo automatico di estrazione terminologica, in modo che il glossario di termini automaticamente acquisito possa diventare una risorsa di riferimento (ovvero rappresentativa dei termini di un dominio) sulla base della quale condurre l'indicizzazione dei testi.

I termini che formano il glossario terminologico automaticamente acquisito e validato dall'esperto di dominio sono a loro volta raggruppati secondo diverse relazioni di similarità semantica, che vanno dalle relazioni gerarchiche di iperonimia/iponimia (denominate anche BT *Broader Term* e NT *Narrower Term* nella terminologia dei thesauri per far riferimento rispettivamente al concetto più generico e a quello più specifico) a classi di termini semanticamente correlati (o RT, *Related Term* secondo la terminologia dei thesauri), ovvero termini genericamente correlati al termine di partenza da rapporti di implicazione e/o associazione semantica[40]. L'organizzazione e la strutturazione dei termini secondo le relazioni appena delineate rappresenta il risultato della seconda fase operativa di T2K, al termine della qua-

[40] In tal caso si parla anche di *quasi-sinonimi*: si tratta di una relazione di sinonimia relativa, nel senso che i termini sono considerati *sinonimi* essenzialmente ai fini dell'indicizzazione.

le è possibile condurre un'indicizzazione concettuale dei testi. Anche in questo caso, il risultato del processo automatico di estrazione di strutture proto-concettuali dovrà essere validato dall'esperto di dominio che costruirà l'ontologia di riferimento sulla base della quale condurre l'indicizzazione concettuale dei testi.

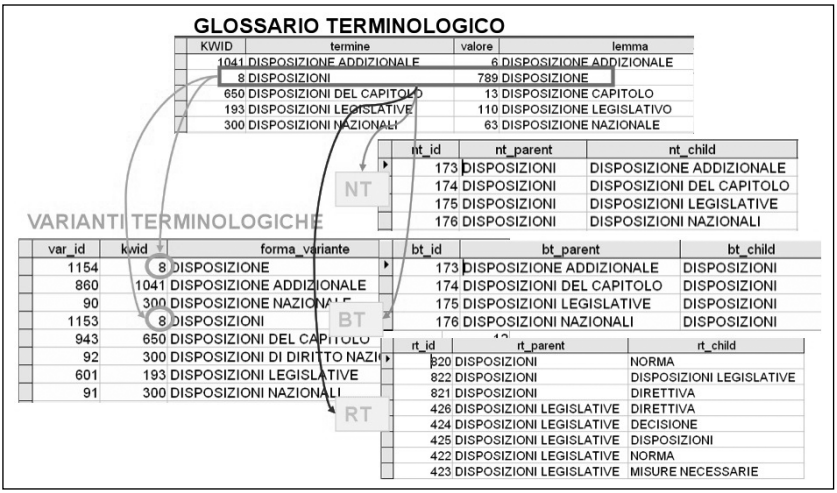

Figura 2. Frammento della base di conoscenza di T2K costruita dinamicamente a partire dai testi.

La Figura 2 riporta un frammento della base di conoscenza costruita dinamicamente a partire dai testi con la piattaforma software T2K. Le unità terminologiche monorematiche e polirematiche acquisite – integrate da informazione riguardante le relative varianti testuali (morfologiche ma anche strutturali) – sono organizzate all'interno di strutture proto-concettuali. Si vedano in particolare le tabelle contrassegnate come BT e NT, che contengono relazioni gerarchiche rispettivamente di iperonimia e iponimia e RT, che include relazioni tra termini genericamente correlati corrispondenti a rapporti di implicazione e/o associazione semantica.

La piattaforma T2K nella versione descritta sopra, d'ora in avanti denominata T2K_v.1, è stata utilizzata in diversi scenari applicativi: per la gestione documentale nella Pubblica Amministrazione (PA) (progetti *Traguardi* e *Pubblicamente.it* del *Formez*); per l'indicizzazione di contenuti didattici multimediali nell'E-learning (progetto PEKITA *Personalized Knowledge In The Air* in collaborazione con Università della Calabria e Siemens Italdata); per l'acquisizione di conoscenza ontologica da cataloghi di prodotti nell'ambito del progetto europeo VIKEF (*Virtual Information and Knowledge Environment Framework*, IP 507173); per lo sviluppo di risorse ontologiche a supporto del *drafting* legislativo nel progetto europeo DALOS (*Drafting Legislation with Ontology-based Support*, eParticipation project n. 2006/01/024). Inoltre, T2K è stato oggetto di sperimentazione con basi documentali di varia natura: per menzionarne alcune, la documentazione scientifica del CNR (Guarasci, 2006)[41]; corpora di testi giuridico-legislativi (Venturi, 2006)[42] e di letteratura linguistico-computazionale (Montemagni, 2007)[43]; per la costruzione di un vocabolario di indicizzazione per la gestione normalizzata e l'estrazione di conoscenza di documenti relativi ai pareri obbligatori sugli acquisti di beni e servizi informatici proposti dalle pubbliche amministrazioni centrali (Taverniti, 2008)[44] o sui temi dell'efficienza energetica e dell'applicazione di fonti rinnovabili agli usi finali civili (Oliveri et alii, 2010)[45].

41 Cfr. GUARASCI, R., *op. cit.*

42 Cfr. GIULIA VENTURI, *L'ambiente, le norme, il computer. Studio linguistico-computazionale per la creazione di ontologie giuridiche in materia ambientale*, Tesi di Laurea Specialistica, Università di Pisa, dicembre 2006.

43 Cfr. SIMONETTA MONTEMAGNI, *Acquisizione automatica di termini da testi: primi esperimenti di estrazione e strutturazione di terminologia metalinguistica*», in Lessicologia e metalinguaggio: Atti del Convegno, Macerata, 19-21 dicembre 2005, Poli D. (a cura di), Roma, Il Calamo, 2007.

44 Cfr. TAVERNITI, M., *op. cit.*

45 Cfr. ELISABETTA OLIVERI, CONCETTA BARONIELLO, ANTONIETTA FOLINO,

Nel corso degli ultimi due anni, la piattaforma T2K_v.1 è stata utilizzata in scenari applicativi che hanno portato alla luce nuove sfide per affrontare le quali sono state progettate e sperimentate soluzioni software innovative. Le principali novità della nuova piattaforma software T2K, d'ora innanzi denominata T2K_v.2, riguardano i componenti utilizzati per l'annotazione linguistica del testo e per l'estrazione di terminologia di dominio.

In T2K_v.2, l'annotazione linguistica è condotta mediante componenti di analisi del testo basati su metodi statistico-quantitativi in linea con il paradigma dominante nel settore della linguistica computazionale che è rappresentato da sistemi basati su algoritmi di apprendimento automatico supervisionato. Secondo questo approccio, il compito di *annotazione linguistica* viene modellato come un compito di classificazione probabilistica: a ogni passo di computazione il sistema sceglie l'annotazione più probabile data la parola in input, i suoi tratti descrittivi, il contesto e le annotazioni linguistiche già identificate. A partire da un corpus di addestramento, annotato con informazione morfo-sintattica e sintattica, viene costruito un modello probabilistico per l'annotazione linguistica del testo. In particolare, per quanto riguarda l'annotazione morfo-sintattica il componente utilizzato (Dell'Orletta, 2009)[46] risulta tra gli strumenti più precisi e affidabili secondo la campagna di valutazione di strumenti per l'analisi automatica dell'italiano EVALITA-2009[47].

ROSSELLA SCAIOLI, *Terminologia, lessici specialistici e strutture tassonomiche nel dominio dell'efficienza energetica e dell'applicazione di fonti rinnovabili agli usi finali civili*, Contributo alla «VI Giornata Scientifica della Rete Panlatina di Terminologia», Università dell'Algarve, Faro, Portogallo, 14 maggio 2010.

46 Cfr. FELICE DELL'ORLETTA, *Ensemble system for Part-of-Speech tagging*, in Atti della «11th Conference of Evaluation of NLP and Speech Tools for Italian (EVALITA) 2009», Reggio Emilia, 12 dicembre 2009.

47 Cfr. EVALITA, *Poster and Workshop Proceedings of the 11th Conference*

Il componente di estrazione terminologica all'interno di T2K_v.2 implementa una nuova strategia che opera sul testo morfo-sintatticamente annotato e procede in due fasi, la prima volta all'identificazione all'interno del corpus di acquisizione di unità terminologiche rilevanti per il dominio, la seconda basata sul confronto della distribuzione inter-dominio dei termini estratti nella fase precedente per una validazione della loro pertinenza.

Per quanto riguarda la prima fase, la maggiore novità riguarda l'acquisizione delle unità terminologiche polirematiche, la cui estrazione è basata sul metodo denominato *C/NC-value* (Frantzi et alii, 2000)[48] che appartiene alla classe delle misure di rilevanza rispetto al dominio (*termhood*). Questa misura tiene conto simultaneamente di quattro aspetti caratterizzanti il termine candidato, ovvero a) la sua frequenza di occorrenza all'interno del corpus di acquisizione, b) la sua frequenza di occorrenza come sottostringa di altri termini candidati, c) il numero di diversi termini candidati che lo contengono come sottostringa, e d) il numero di parole di cui si compone il termine candidato. Questa misura, denominata *C-value*, risulta particolarmente utile per il trattamento di terminologia complessa che include al suo interno altri termini. La lista di termini candidati definita sulla base del *C-value* viene ulteriormente rivista prendendo in considerazione informazione relativa ai contesti di occorrenza (*NC-value*).

I risultati ottenuti al termine della fase 1 per entrambe le tipologie di termini estratti (ovvero unità monorematiche e polirematiche) vengono filtrati sulla base di una funzione, chiamata *funzione di contrasto*, che valuta dal punto di vista quantitativo quanto un termine della lista estratta al passo precedente sia specifico di un certo dominio. Per calcolare la specificità del termine, sulla base della quale viene definito un nuovo ordinamento

of the Italian Association for Artificial Intelligence, Reggio Emilia, 12 Dicembre 2009. <http://www.evalita.it/2009/proceedings>.

48 Cfr. FRANTZI, K. et alii, *op. cit.*

dei termini in base alla pertinenza rispetto al dominio, viene considerata la distribuzione del termine sia nel corpus di acquisizione sia in un corpus differente, detto *corpus di contrasto*. La funzione di contrasto utilizzata, chiamata *Contrastive Selection of multi–word terms (CSmw)*, si è rivelata particolarmente adatta per l'analisi di variazioni distribuzionali di eventi a bassa frequenza (come appunto l'occorrenza di un termine polirematico). Se per una descrizione dettagliata del metodo si rinvia a (Bonin et alii, 2010a)[49], in questa sede vale la pena ricapitolare quali siano i principali elementi di novità dell'approccio proposto.

Contrariamente a (Penas et alii, 2001)[50], (Chung, Nation, 2004)[51] e (Basili et alii, 2001)[52], la fase di analisi contrastiva viene condotta in relazione alle unità terminologiche polirematiche acquisite nel corso della precedente fase: ciò è possibile grazie alla nuova funzione di contrasto che può essere applicata anche a eventi caratterizzati da basse frequenze. Questo previene potenziali problemi quali l'inclusione, nel risultato finale, di unità polirematiche non rilevanti ma lessicamente *governate* da una testa che è stata identificata come unità monorematica specifica per il dominio, oppure l'esclusione di unità polirematiche rilevanti che non sono state acquisite perché la loro testa lessicale non è stata selezionata come specifica per il dominio (Bonin et alii, 2012)[53]. Illustriamo quanto detto finora con un esempio, tratto da un esperimento di estrazione terminologica condotto su

49 Cfr. Bonin, F. et alii, *op. cit.*, 2010(a).
50 Cfr. Penas, A. et alii, *op. cit.*
51 Cfr. Chung, T.M., Nation, P., *op. cit.*
52 Cfr. Basili, R. et alii, *op. cit.*
53 Cfr. Francesca Bonin, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi, *Lessico settoriale e lessico comune nell'estrazione di terminologia specialistica da corpora di dominio*, in Lessico e lessicologia. Atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI 2010), Viterbo, 27-29 settembre 2010, Ferreri S. (a cura di), 2012, pp. 207–220.

un corpus di articoli scientifici sul cambiamento climatico. Seguendo un approccio contrastivo di tipo tradizionale, l'acquisizione dell'unità terminologica polirematica *effetto serra* è subordinata alla identificazione dell'unità monorematica *effetto* come rilevante per il dominio: nel caso in cui l'unità monorematica *effetto* non sia stata selezionata come rilevante per il corpus di acquisizione, neanche l'unità polirematica, di cui essa è la testa, sarà estratta, sebbene essa sia significativa per il dominio. Ma se l'unità monorematica *effetto* è stata selezionata come rilevante, allora anche polirematiche come *effetto domino*, se ricorrenti nel testo, potranno essere qualificate come termini di dominio. Nell'approccio proposto, ciò non si verifica in quanto la funzione di contrasto opera direttamente sulla lista delle unità terminologiche polirematiche acquisite al passo precedente.

La nuova strategia di estrazione terminologica è stata verificata in diversi contesti applicativi e con risultati incoraggianti in relazione a corpora di leggi e sentenze (Venturi, 2011)[54], corpora web di cultura italiana (relativi ai domini di letteratura, arte e linguistica), corpora di referti relativi ad esami radiologici (Pirrelli et alii, 2010)[55]. In quanto segue, si riportano i risultati di esperimenti condotti in questi scenari applicativi che illustrano in dettaglio come la soluzione proposta sia in grado di fornire risultati più precisi e affidabili in relazione alle situazioni problematiche descritte nella sezione 4.

5.1 *Lessico settoriale vs lessico comune*[56]

Partiamo dal problema della discriminazione tra lessico settoriale e lessico comune. Riportiamo di seguito i risultati di un espe-

54 Cfr. VENTURI, G., *op. cit.*

55 Cfr. VITO PIRRELLI, ALESSANDRO LENCI, SIMONETTA MONTEMAGNI, FELICE DELL'ORLETTA, EMILIANO GIOVANNETTI, SIMONE MARCHI, *Connect To Life (modulo semantico): Rapporto Finale*, Rapporto Tecnico, 2010.

56 Parti di quanto segue sono riprese da: BONIN, F. et alii, *op. cit., 2012*, sezione 5.1; FRANCESCA BONIN, FELICE DELL'ORLETTA, GIULIA VENTURI, SI-

rimento condotto nell'ambito del progetto *Panorama FIRB: arte, lingua e letteratura italiana* (n. RBNE07C4R9, finanziato dal Ministero dell'Istruzione, dell'Università e della Ricerca) con T2K_v2 su un corpus di testi del settore della storia dell'arte. La sfida, nel caso specifico, è rappresentata dall'acquisizione di lessico settoriale da un corpus di testi caratterizzati da un livello non particolarmente alto di specializzazione. In particolare, l'estrazione di unità terminologiche monorematiche e polirematiche è stata condotta a partire da un corpus di testi di storia dell'arte estratti da siti di cultura italiana sul web (per un totale di 326.066 parole) costruito da esperti di dominio. Se tale corpus (denominato da ora in avanti ARTE) si presenta omogeneo rispetto al dominio, esso appare alquanto eterogeneo per quanto riguarda la tipologia di registri linguistici in esso testimoniati in ragione della natura variegata del web: in ARTE sono contenuti testi specialistici, così come testi divulgativi rivolti a un pubblico più vasto.

Per la fase di analisi *contrastiva* è stato selezionato un corpus di riferimento rispetto al quale confrontare la distribuzione delle unità terminologiche estratte dal corpus di acquisizione ARTE. Dato l'obiettivo di filtrare dal risultato finale il lessico comune, in questo esperimento come *corpus di contrasto* è stato usato il corpus PAROLE, un corpus di italiano contemporaneo di circa 3 milioni di parole (Marinelli et alii, 2003)[57].

MONETTA MONTEMAGNI (b), *Contrastive filtering of domain specific multiword terms from different types of corpora*, in Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications, Beijing, Cina, Coling 2010 Organizing Committee, agosto 2010, pp. 76-79, sezione 3.

57 Cfr. RITA MARINELLI, LISA BIAGINI, REMO BINDI, SARA GOGGI, MONICA MONACHINI, PAOLA ORSOLINI, EUGENIO PICCHI, SERGIO ROSSI, NICOLETTA CALZOLARI, ANTONIO ZAMPOLLI, *The Italian PAROLE corpus: an overview*, in «Linguistica Computazionale», Special Issue in «Computational Linguistics in Pisa», Zampolli A. et alii (a cura di), voll. 16-17, t. 1, Pisa-Roma, IEPI, 2003, pp. 401–421.

La Tabella 1 esemplifica il risultato della prima fase di estrazione, riportando le prime 10 unità terminologiche monorematiche e polirematiche delle rispettive liste ordinate per valori decrescenti di frequenza (unità monorematiche) e *C/NC-value* (unità polirematiche). Come si può notare, le liste risultanti da questa fase includono sia termini come *artista,* appartenenti evidentemente al lessico specialistico del settore artistico, sia voci come *anno* appartenenti piuttosto al lessico comune (marcate in corsivo nella tabella).

Ordinamento sulla base del filtro statistico (frequenza vs *C/NC-value*)			
	Unità monorematiche		**Unità polirematiche**
1	Arte	1	*Punto di vista*
2	Opera	2	Opera d'arte
3	Artista	3	Storia dell'arte
4	*Anno*	4	Arte contemporanea
5	Mostra	5	Figura umana
6	Parte	6	Bene culturale
7	Pittura	7	Storico dell'arte
8	*Secolo*	8	Movimento artistico
9	Forma	9	Produzione artistica
10	*Tempo*	10	*Anno scorso*

Tabella 1. Frammento delle liste di unità monorematiche e polirematiche estratte dopo la prima fase di estrazione terminologica. In corsivo i *non-termini.*

La seconda fase estrattiva, basata sulla distribuzione inter-dominio dei termini estratti, consiste nel riordinamento della lista (a tal fine vengono selezionati i primi 600 termini) sulla base della significatività dei termini di ARTE rispetto al corpus di contrasto[58]. È nel corso di questa fase di confronto della distribuzio-

[58] La soglia è stata stabilita su base sperimentale.

ne dei termini nei due corpora che il lessico settoriale viene distinto da quello comune. Grazie all'analisi contrastiva, le unità terminologiche precedentemente individuate come rilevanti per il corpus di acquisizione, ma non necessariamente per il dominio di acquisizione, vengono riordinate sulla base di un valore di contrasto. Da questa lista, vengono selezionati i termini risultanti alle prime 300 posizioni[59].

La Tabella 2 illustra il risultato della fase di analisi contrastiva che, come si può notare, ha consentito di filtrare nelle posizioni più alte della lista termini particolarmente specifici non solo per il corpus di acquisizione in sé, ma anche per il dominio trattato. Ad esempio, l'unità polirematica *anno scorso,* di pertinenza del lessico comune ma che occupava la decima posizione nella lista dei termini in Tabella 1, viene filtrata dopo la fase di confronto con il corpus di riferimento, scendendo oltre la trecentesima posizione.

Ordinamento sulla base delle funzione di contrasto (confronto PAROLE)			
	Unità monorematiche		**Unità polirematiche**
1	Artista	1	Opera d'arte
2	Pittura	2	Figura umana
3	Pittore	3	Movimento artistico
4	Scultura	4	Produzione artistica
5	Arte	5	Arte contemporanea
6	Mostra	6	Pittore italiano
7	Dipinto	7	Percorso espositivo
8	Affresco	8	Elemento architettonico
9	Architettura	9	Storia dell'arte
10	Museo	10	Storico dell'arte

Tabella 2. Frammento della lista finale di unità monorematiche e polirematiche estratte.

[59] La soglia è stata stabilita su base sperimentale.

La valutazione dei risultati raggiunti è stata condotta confrontando il glossario ottenuto con un Thesaurus di dominio[60], seguita da una fase di validazione da parte di esperti. Da questa duplice valutazione è emerso un aumento significativo dei termini di dominio estratti, che sono passati da 61,33% al termine della fase 1 al 79,40% a conclusione dell'analisi contrastiva, con un incremento relativo[61] registrato di +29,34%.

La discriminazione tra lessico settoriale e lessico comune può rappresentare un problema, sebbene di portata più limitata, anche nel caso di letteratura specialistica. In quanto segue, riportiamo i risultati di due esperimenti condotti con due corpora su tematiche di tipo ambientale, ovvero: a) letteratura scientifica (costituito da articoli scientifici sul cambiamento climatico per un totale di 397.297 *token*) e b) le voci di Wikipedia riconducibili al settore *Ecologia e Ambiente* per un totale di 174.391 *token*. Come nel precedente esperimento, ai fini dell'analisi contrastiva è stato usato il corpus PAROLE.

L'estrazione terminologica è stata condotta in due fasi, focalizzandosi sulla terminologia complessa. In particolare, i primi 2.000 termini risultanti dalla prima fase di analisi basata sul *C/NC-value* sono stati ordinati sulla base della funzione di contrasto *CSmw*. Da entrambe le liste ordinate di termini risultanti dalla prima e seconda fase di analisi sono stati estratti i primi 300 termini che sono stati sottoposti a valutazione. La valutazione è stata condotta semi-automaticamente: prima, i termini estratti sono stati confrontati con il *Thesaurus EARTh*[62] contenente 12.398

60 Il Thesaurus è stato fornito dal Dipartimento di Storia delle Arti dell'Università di Pisa.

61 Per tenere sotto controllo l'effetto della fase di analisi contrastiva, si è fatto ricorso all'incremento relativo (IR) ottenuto dividendo l'incremento assoluto osservato nel risultato della seconda fase per la numerosità dei termini estratti al termine della prima fase.

62 *Environmental Applications Reference Thesaurus*. <http://uta.iia.cnr.it/earth.htm#EARTh%202002>.

termini ambientali; i termini che non hanno trovato una corrispondenza all'interno della risorsa di riferimento selezionata sono stati manualmente validati da esperti di dominio. L'incremento relativo osservato nella selezione di terminologia rilevante rispetto al dominio al termine della seconda fase di analisi contrastiva è di +11,30% nel caso di Wikipedia e di +12,82% nel caso del corpus di articoli scientifici. Confrontando questo dato con l'incremento relativo osservato nel precedente esperimento in relazione a testi di storia dell'arte (+29,34%), possiamo concludere che questa strategia di analisi è particolarmente promettente con corpora caratterizzati da un linguaggio non altamente specialistico come quello della storia dell'arte ma ottiene miglioramenti significativi, anche se inferiori rispetto al caso precedente, anche nel caso di corpora di letteratura specialistica.

5.2 *Estrazione di terminologia multi-dominio*[63]

Come già accennato nella sezione 4, non si verifica sempre il caso che i corpora specialistici siano caratterizzati da un lessico espressione di un unico dominio di conoscenza e nettamente separato da quello comune. Un esempio di questo tipo è costituito dal dominio giuridico contraddistinto da una commistione di tipologie di termini, non sempre nettamente distinguibili tra di loro. (Agnoloni et alii, 2009)[64] così come (Lenci et alii, 2009)[65] ri-

[63] Parti di quanto segue sono riprese da: BONIN, F. et alii, *op. cit.*, 2010b, sezione 5.2; BONIN, F. et alii, *op. cit.*, 2012, sezione 3.

[64] Cfr. TOMMASO AGNOLONI, LORENZO BACCI, ENRICO FRANCESCONI, WIM PETERS, SIMONETTA MONTEMAGNI, GIULIA VENTURI, *A two-level knowledge approach to support multilingual legislative drafting*, in Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood, Breuker J. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications», Springer, vol. 188, 2009, pp. 177-198.

[65] Cfr. ALESSANDRO LENCI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, GIULIA VENTURI, *Ontology learning from Italian legal texts*, in Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood,

portano che i metodi tradizionali di estrazione terminologica su questo tipo di testi arrivano ad acquisire repertori terminologici nei quali le diverse tipologie di termini sono mescolate. Questi ultimi notano anche che il repertorio terminologico acquisito contiene un maggior numero di termini giuridici rispetto a quelli relativi alla materia legislata. Ciò viene ricondotto alla bassa frequenza (e alto rango[66]) di quest'ultimo tipo di termini nel corpus di testi giuridici di partenza, in accordo con la legge di Zipf, secondo la quale la frequenza di una parola è inversamente proporzionale al suo rango.

Abbiamo provato ad affrontare questa situazione applicando il metodo contrastivo descritto in precedenza. Riportiamo di seguito i risultati di un esperimento condotto con T2K_v2 su una collezione di direttive europee in materia ambientale per un totale di 394.088 parole (d'ora in avanti AMB), reperito dalla versione disponibile on-line del Bollettino Giuridico Ambientale[67]. In questo caso, la metodologia contrastiva di estrazione terminologica ha svolto un duplice ruolo, finalizzato non solo a discriminare il lessico rilevante in AMB dal lessico comune, ma anche a distinguere il lessico del diritto da quello del dominio ambientale. A questo scopo sono stati usati due corpora di riferimento: il corpus PAROLE e un corpus di direttive europee in materia di protezione del consumatore (per un totale di 72.210 parole, d'ora in avanti CONS). In questo caso, l'analisi si è concentrata sull'estrazione di unità terminologiche poliematiche.

Analogamente agli esperimenti precedenti, è stata estratta una lista di 600 unità terminologiche poliematiche ordinate per va-

Breuker J. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications», vol. 188, Springer, 2009, pp. 75-94.

66 Nella lista delle parole di un testo ordinata per valori decrescenti di frequenza, il *rango* si riferisce alla posizione che una data parola occupa all'interno della lista.

67 <http://extranet.regione.piemonte.it/ambiente/bga/index.htm>.

lori decrescenti sulla base dei valori di *C/NC-value*; in questo caso si osserva la compresenza di unità appartenenti sia al lessico comune (es. *anno successivo*) sia al lessico del diritto (es. *norma nazionale*), sia a quello ambientale (es. *effetto serra*).

La fase di analisi contrastiva in questo caso è stata suddivisa in due passi, ciascuno condotto rispetto a un diverso corpus di riferimento: prima, con il corpus PAROLE per discriminare i termini rilevanti per AMB (sia giuridici sia ambientali) dai *non-termini*; in seconda battuta, con CONS per distinguere tra i termini del lessico del diritto e quelli del lessico ambientale. Il primo passo dell'analisi contrastiva ha riguardato le prime 600 unità terminologiche; da questa lista di unità riordinate sulla base della loro rilevanza per AMB, sono state selezionate le prime 300 su cui si è incentrata la seconda fase di analisi contrastiva basata sul confronto con CONS, volta a distinguere le unità proprie del lessico del diritto da quelle del dominio ambientale.

La Tabella 3 riporta nelle due colonne iniziali le prime 10 unità terminologiche della lista estratta al termine della fase 1, nelle ultime due colonne le prime e ultime cinque posizioni della lista risultante dalla doppia analisi contrastiva (fase 2). Come si può vedere, mentre a conclusione della fase 1 i termini appartenenti al lessico del diritto (in corsivo) si affiancano a termini ambientali (in grassetto) nelle prime posizioni della lista, nella lista finale i termini dei due lessici settoriali sono riordinati in modo da essere distinti (ovvero, la testa della lista contiene i termini ambientali mentre nella coda si concentrano quelli del diritto).

La valutazione è avvenuta in modo semi-automatico, analogamente ai precedenti esperimenti. Come risorse di riferimento per la valutazione dei risultati conseguiti, sono stati selezionati il *Dizionario Giuridico* (Edizioni Simone)[68] e il *Thesaurus EARTh* sopra citato. Anche in questo caso, i risultati raggiunti dimostra-

[68] <http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index,5,A>.

Ordinamento sulla base del filtro statistico (*C/NC-value*)	Unità polirematiche	Ordinamento sulla base della funzione di contrasto (confronto con CONS)	Unità polirematiche
1	*parlamento europeo*	1	**valore limite**
2	*autorità competente*	2	**sostanza pericolosa**
3	**valore limite**	3	**salute umana**
4	*valore limite di emissione*	4	**effetto serra**
5	*stato membro*	5	**sviluppo sostenibile**
6	**limite di emissione**	296	*diritto nazionale*
7	**sostanza pericolosa**	297	*testo della disposizione*
8	*destinatario della presente direttiva*	298	*disposizione essenziale del diritto interno*
9	*misura necessaria*	299	*disposizione nazionale*
10	**sviluppo sostenibile**	300	*funzionamento del mercato interno*

Tabella 3. Frammenti delle liste ordinate di unità polirematiche estratte al termine delle fasi 1 e 2.

no l'efficacia di questo approccio all'estrazione terminologica: mentre, infatti, dopo l'estrazione sulla base del *C/NC-value* il 65,34% dei primi 300 termini della lista ordinata era costituito da unità polirematiche del lessico ambientale (38,67%) e del lessico del diritto (26,67%), al termine della doppia analisi contrastiva le unità terminologiche ambientali sono aumentate fino al 43,33% e quelle del lessico del diritto fino al 29,33%. Per quanto riguarda il lessico ambientale l'incremento relativo è del 23,81%.

6. Conclusioni

I sistemi di estrazione automatica di terminologia da corpora di dominio sono oggi considerati maturi per poter essere integrati in applicazioni reali, per l'indicizzazione automatica di basi documentali e l'accesso su base semantica ai contenuti. I migliori risultati sono ottenuti nei casi di acquisizione di terminologia di dominio da testi caratterizzati da un lessico altamente specialistico e rivolti ad un pubblico di esperti, come ad esempio la letteratura biomedica. Il rendimento scientifico di tali sistemi decresce significativamente quando la collezione documentale usata come corpus di acquisizione non appartenga alla classe dei testi altamente specialistici. Infatti, l'analisi di testi che occupano una posizione intermedia nel continuum tra linguaggi altamente specialistici e lingua comune rappresenta una sfida tuttora aperta per tali sistemi. Un'ulteriore e non secondaria sfida riguarda la necessità di distinguere, all'interno di un corpus rappresentativo di un unico linguaggio settoriale, i termini appartenenti a diversi domini del sapere; ad es. il lessico del diritto da quello proprio della materia legislata nel caso di corpora giuridici. Ad oggi, a conoscenza di chi scrive, nessun sistema automatico ha affrontato questi due ordini di problemi in modo sistematico.

Partendo dall'analisi critica dei risultati ottenuti con un approccio *standard* all'estrazione terminologica in diversi scenari applicativi, il presente contributo raccoglie gli sforzi condotti per cercare di colmare le lacune e i limiti identificati nei sistemi correnti di estrazione terminologica, fornendo una risposta al problema dell'acquisizione di terminologia da corpora non altamente specialistici e da corpora *multi-dominio*. I risultati conseguiti, sebbene ancora preliminari, sono incoraggianti: gli scenari applicativi trattati sono vari, con un incremento relativo nella terminologia rilevante estratta che va dall'11/12% nel caso di testi specialistici (cfr. sezione 5.1), a più del 23% nel caso di corpora

giuridici (cfr. sezione 5.2), per arrivare fino al 29% registrato nel caso del corpus web di storia dell'arte (cfr. sezione 5.1). Altri esperimenti con risultati interessanti sono stati condotti con corpora di sentenze e corpora di referti diagnostici provenienti da reparti di Senologia Radiologica di diversi ospedali; mentre per quanto riguarda le sentenze la valutazione dei risultati è ancora in corso, nel secondo caso si è osservato un incremento altrettanto significativo rispetto a quanto riportato sopra.

Sulla base dei risultati raggiunti, si può affermare che T2K_v2, il prototipo software che implementa la nuova strategia estrattiva illustrata nelle precedenti pagine, sa far fronte in modo più che soddisfacente alle sfide poste da corpora non altamente specialistici o *multi-dominio*, fornendo così una prima risposta ai desiderata espressi in (Oliveri et alii, 2010)[69] che concludono il loro articolo auspicando *«una fase di estrazione terminologica tematica che recuperi solo i termini rappresentativi del contenuto concettuale dei documenti e al tempo stesso del dominio di riferimento»*.

Potenziali ed interessanti estensioni del metodo contrastivo per l'estrazione di terminologia di dominio includono il trattamento di variazioni di registro all'interno dello stesso linguaggio settoriale così come la ricostruzione dell'evoluzione diacronica di un lessico settoriale.

Per quanto riguarda le variazioni di registro, (Oliveri et alii, 2010)[70] nel loro studio sulla terminologia specialistica nel dominio dell'efficienza energetica e dell'applicazione di fonti rinnovabili agli usi finali civili si sono trovati a trattare con sottocorpora di testi caratterizzati da diversi livelli di specializzazione e da diverse finalità comunicative (ovvero articoli e rendicontazioni scientifiche, riviste di settore, leggi e norme, opuscoli e linee

69 Cfr. OLIVERI, E. et alii, *op. cit*.
70 Cfr. OLIVERI, E. et alii, *op. cit*.

guida). Il processo estrattivo è stato condotto separatamente per ciascun sottocorpus al fine di poter estrarre termini specialistici e termini appartenenti al linguaggio comune, tra i quali, nel thesaurus, sono state stabilite relazioni di equivalenza. Un'analisi contrastiva di tali collezioni di documenti, condotta con corpora di contrasto adeguatamente selezionati, dovrebbe poter rendere possibile l'estrazione della terminologia settoriale tipica di ogni registro, fornendo così all'esperto ulteriore evidenza utile per l'arricchimento del vocabolario controllato o del thesaurus.

Un altro aspetto importante riguarda l'evoluzione diacronica della terminologia settoriale, che va di pari passo con l'evolversi delle conoscenze all'interno di un dominio. L'estrazione terminologica non va ad operare su collezioni documentali chiuse che cambiano raramente se non mai, come ad esempio la produzione letteraria di un autore del passato. Un sistema di estrazione terminologica si trova tipicamente a trattare con collezioni documentali aperte e dinamiche, continuamente aggiornate con nuovi documenti, comprendenti nuovi contenuti e dunque nuova terminologia. La domanda è se sia possibile utilizzare il metodo estrattivo qui proposto anche per identificare termini in entrata e/o in uscita. Primi esperimenti in questa direzione, condotti sulla lingua comune, hanno fornito risultati interessanti (Montemagni, 2010)[71]. Il monitoraggio terminologico-lessicale alla ricerca di parole *in entrata* condotto su un corpus giornalistico tratto da *La Repubblica* degli anni 2002-2005 usando come corpus di contrasto un corpus della stessa testata di un periodo antecedente (1992-1995) ha identificato nelle unità monorematiche come *tsunami*, *devolution*, *web*, *info*, *sms*, *bipartisan*, *dvd*, ecc. o nelle

[71] Cfr. SIMONETTA MONTEMAGNI, *Tecnologie linguistico-computazionali per il monitoraggio della lingua italiana*, presentazione tenuta nell'ambito della Giornata di Studio «Lo stato della lingua. Il CNR e l'italiano nel terzo millennio», Roma, Consiglio Nazionale delle Ricerche - Dipartimento Identità Culturale, 8 marzo 2010.

unità polirematiche *milione di euro*, *influenza aviaria*, *cellule staminali*, *reality show* e *digitale terrestre* le parole caratterizzanti il corpus 2002-2005 rispetto a quello 1992-1995. Il monitoraggio terminologico-lessicale finalizzato alla ricerca delle parole *in via di recessione*, condotto analizzando il corpus del 1992-1995 in relazione a quello del 2002-2005 (usato per l'analisi contrastiva), ha fatto emergere voci quali *minimum tax*, *miliardo di marchi*, *svalutazione della lira*, *patto in deroga* oppure *quadripartito*, *pidiessini*, *rublo* come espressioni (polirematiche o monorematiche) che stanno scomparendo dall'uso linguistico. Seguendo questo approccio, sarebbe quindi possibile qualificare le voci di un vocabolario controllato o di thesaurus non solo in rapporto a un registro linguistico, ma anche sull'asse diacronico.

Ringraziamenti

Gli strumenti e le tecnologie illustrati in questo articolo sono il frutto del lavoro di collaborazione tra due gruppi di ricerca, rispettivamente dell'Istituto di Linguistica Computazionale *Antonio Zampolli* del Consiglio Nazionale delle Ricerche di Pisa, e del Dipartimento di Linguistica *Tristano Bolelli* dell'Università di Pisa, nell'ambito prima del laboratorio interistituzionale DY-LAN Lab, e oggi dell'ItaliaNLP Lab. In particolare, Felice Dell'Orletta e Giulia Venturi hanno contribuito in misura sostanziale al disegno e allo sviluppo di T2K_v2 per l'estrazione di terminologia settoriale da corpora di dominio di cui sono stati qui delineati i principi fondamentali.

Bibliografia

ABNEY, S. *Parsing by chunks*, in Principle-based Parsing: Computation and Psycholinguistics, Berwick R.C. et alii (a cura di), Dordrecht, Kluwer, 1991, pp. 257-278

AGNOLONI, T., BACCI, L., FRANCESCONI, E., PETERS, W., MONTEMAGNI, S., VENTURI, G. *A two-level knowledge approach to support multilingual legislative drafting*, in Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood, Breuker J. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications», Springer, vol. 188, 2009, pp. 177-198

BAEZA-YATES, R., RIBEIRO-NETO, B. *Modern Information Retrieval*, New York, ACM Press, 1999

BASILI, R., MOSCHITTI, A., PAZIENZA M.T., ZANZOTTO F.M., *A contrastive approach to term extraction*, in Atti della «4th Conference on Terminology and Artificial Intelligence (TIA-2001)», Nancy, 3-4 maggio 2001

BONIN, F., DELL'ORLETTA, F., MONTEMAGNI, S., VENTURI, G. (a) *A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora*, in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 17-23 maggio 2010, pp. 3222–3229

BONIN, F., DELL'ORLETTA, F., VENTURI, G., MONTEMAGNI, S. (b) *Contrastive filtering of domain specific multi-word terms from different types of corpora*, in Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications, Beijing, Cina, agosto 2010, Coling 2010 Organizing Committee, pp. 76-79

BONIN, F., DELL'ORLETTA, F., MONTEMAGNI, S., VENTURI, G., *Lessico settoriale e lessico comune nell'estrazione di terminologia specialistica da corpora di dominio*, in Lessico e lessicologia. Atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI 2010), Viterbo, 27-29 settembre 2010, Ferreri S. (a cura di), 2012, pp. 207–220

BUITELAAR, P., CIMIANO, P., MAGNINI, B., *Ontology Learning from Text: An Overview*, in Ontology Learning from Text: Methods, Evaluation and Applications, Buitelaar P. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications Series», vol. 123, IOS Press, 2005, pp. 3-12

CABRÉ, M.T., *Terminology: Theory, Methods, and Applications*, Amsterdam, John Benjamins, 1999

CAVAGNOLI, S., *La comunicazione specialistica*, Roma, Carocci, 2007

CHUNG, T.M., NATION P., *Identifying technical vocabulary*, in «System», vol. 32, 2004, pp. 251-263

CHURCH, K.W., HANKS, P., *Word association norms, mutual information, and lexicography*, in «Computational Linguistics», vol. 16, n.1, 1990, pp. 22-29

CORTELAZZO, M., *Lingua e diritto in Italia. Il punto di vista dei linguisti*, in La lingua del diritto. Difficoltà traduttive. Applicazioni didattiche. Atti del primo Convegno Internazionale, Milano, 5-6 ottobre 1995, Schena L. (a cura di) Roma, CISU (Centro d'Informazione e Stampa Universitaria), 1997, pp. 35-50

DARDANO, M., *Linguaggi settoriali e processi di riformulazione*, in Parallela 3. Linguistica contrastiva / Linguaggi settoriali / Sintassi generativa, Dressler W. et alii (a cura di), Tubinga, Narr, 1987, pp. 134-145

DELL'ORLETTA, F., *Ensemble system for Part-of-Speech tagging*, in Atti della «11th Conference of Evaluation of NLP and Speech Tools for Italian

(EVALITA) 2009», Reggio Emilia, 12 dicembre 2009

DELL'ORLETTA, F., LENCI, A., MARCHI, S., MONTEMAGNI, S., PIRRELLI, V., VENTURI, G., *Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio*, in «Aida Informazioni», a. 26, n. 1-2, gennaio-giugno 2008, pp. 185-206

DUNNING, T., *Accurate Methods for the Statistics of Surprise and Coincidence*, in «Computational Linguistics», vol. 19, n. 1, 1993, pp. 61-74

EVALITA, *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, 12 dicembre 2009
<http://www.evalita.it/2009/proceedings>

FEDERICI, S., MONTEMAGNI, S., PIRRELLI, V., *Shallow Parsing and Text Chunking: a View on Underspecification in Syntax*, in Proceedings of Workshop On Robust Parsing and Eight Summer School on Language, Logic and Information, Praga, Repubblica Ceca, 12-16 agosto 1996, pp. 35-44

FRANCESCONI, E., MONTEMAGNI, S., PETERS, W., TISCORNIA, D., *Integrating a Bottom-Up and Top-Down Methodology for Building Semantic Resources for the Multilingual Legal Domain*, in Semantic Processing of Legal Texts, Francesconi E. et alii (a cura di), in «LNCS/LNAI», Springer-Verlag, vol. 6036, 2010, pp. 95-127

FRANTZI, K., ANANIADOU, S., MIMA, H., *Automatic recognition of multi-word terms*, in «International Journal of Digital Libraries», vol. 3, n. 2, 2000, pp.117-132

GUARASCI, R., *Estrazione terminologica e gestione della conoscenza*, in «iged.it», n. 3, 2006, pp. 46-51

JACKENDOFF, R., *Twistin' the night away*, in «Language», vol. 73, 1997, pp. 534-559

KAGEURA, K., UMINO, B., *Methods of automatic term recognition: a review*, in «Terminology», vol. 3, n. 2, 1996, pp. 259-289

LAVINIO, C., *Comunicazione e linguaggi disciplinari. Per un'educazione linguistica trasversale*, Roma, Carocci, 2004

LENCI, A., MONTEMAGNI, S., PIRRELLI, V., VENTURI, G., *Ontology learning from Italian legal texts*, in Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood, Breuker J. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications», vol. 188, Springer, 2009, pp. 75-94

MANNING, C.D., RAGHAVAN, P., SCHÜTZE, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008

MARINELLI, R., BIAGINI, L., BINDI, R., GOGGI, S., MONACHINI, M., ORSOLINI, P., PICCHI, E., ROSSI, S., CALZOLARI, N., ZAMPOLLI, A., *The Italian PARO-*

LE corpus: an overview, in «Linguistica Computazionale», Special Issue in «Computational Linguistics in Pisa», Zampolli A. et alii (a cura di), voll. 16-17, Tomo I, Pisa-Roma, IEPI, 2003, pp. 401–421

MONTEMAGNI, S., *Acquisizione automatica di termini da testi: primi esperimenti di estrazione e strutturazione di terminologia metalinguistica*», in Lessicologia e metalinguaggio: Atti del Convegno, Macerata, 19-21 dicembre 2005, Poli D. (a cura di), Roma, Il Calamo, 2007

MONTEMAGNI, S., *Tecnologie linguistico-computazionali per il monitoraggio della lingua italiana*, presentazione tenuta nell'ambito della Giornata di Studio «Lo stato della lingua. Il CNR e l'italiano nel terzo millennio», Roma, Consiglio Nazionale delle Ricerche - Dipartimento Identità Culturale, 8 marzo 2010

NAKAGAWA, H., MORI, T., *Automatic Term Recognition based on Statistics of Compound Nouns and their Components*, in «Terminology», vol. 9, n. 2, 2003, pp. 201–209

OLIVERI, E., BARONIELLO, C., FOLINO, A., SCAIOLI, R., *Terminologia, lessici specialistici e strutture tassonomiche nel dominio dell'efficienza energetica e dell'applicazione di fonti rinnovabili agli usi finali civili*, Contributo alla «VI Giornata Scientifica della Rete Panlatina di Terminologia», Università dell'Algarve, Faro, Portogallo, 14 maggio 2010

PENAS, A., VERDEJO, F., GONZALO, J., *Corpus-Based Terminology Extraction Applied to Information Access*, in Proceedings of the Corpus Linguistics 2001 Conference, Università di Lancaster, 29 marzo – 2 aprile 2011, Rayson P., Wilson A., McEnery T., Hardie A., Khoja S. (ed.), pp. 458-465

PIRRELLI, V., LENCI, A., MONTEMAGNI, S., DELL'ORLETTA, F., GIOVANNETTI, E., MARCHI, S., *Connect To Life (modulo semantico): Rapporto Finale*, Rapporto Tecnico, CNR-ILC-Dylan LAB, TR-008, 2010

PORCELLI, G., *Principi di Glottodidattica*, Brescia, La Scuola, 1994

RONDEAU, G., *Introduction à la terminologie*, Québec, Gaëtan Morin éditeur, 1983

RONDEAU, G., SAGER, J., *Introduction à la terminologie*, ed. 2, Chicoutimi, Gatan Morin, 1984

SALTON, G., BUCKLEY, C., *Term-Weighting Approaches in Automatic Text Retrieval*, in «Information Processing and Management», vol. 24, n. 5, 1988, pp. 513-523

SOBRERO, A., *Lingue speciali*, in Introduzione all'italiano contemporaneo. La variazione e gli usi, Sobrero A. (a cura di), Roma-Bari, Laterza, 1993, pp. 237-277

TAVERNITI, M., *Tra terminologia e documentazione: estrazione automatica di voci indice da corpora documentali della Pubblica Amministrazione*, in «Aida Informazioni», a. 26, n. 1-2, gennaio-giugno 2008, pp. 239-250

VARANTOLA, K., *Special Language and General Language: Linguistic and Didactic Aspects*, in «Unesco ALSED-LSP Newsletter», vol. 9, n. 2, dicembre 1986, pp. 10-20

VENTURI, G., *L'ambiente, le norme, il computer. Studio linguistico–computazionale per la creazione di ontologie giuridiche in materia ambientale*, Tesi di Laurea Specialistica, Università di Pisa, dicembre 2006

VENTURI, G., *Lingua e diritto: una prospettiva linguistico-computazionale*, Tesi di Dottorato, Università degli Studi di Torino, Scuola di Dottorato in Studi Umanistici, 2011

VU, T., AW, A., ZHANG, M., *Term Extraction Through Unithood and Termhood Unification*, in Third International Joint Conference on Natural Language Processing. Proceedings of the Conference, Hyderabad, India, 07-12 gennaio 2008, pp. 631-636

Sitografia

<http://uta.iia.cnr.it/earth.htm#EARTh%202002>
<http://extranet.regione.piemonte.it/ambiente/bga/index.htm>
<http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index,5,A>