Francesca Bonin°
Felice Dell'Orletta*
Simonetta Montemagni*
Giulia Venturi*
(*Istituto di Linguistica Computazionale "Antonio Zampolli" - ILC-CNR, °Dipartimento di Informatica, Università di Pisa)

Lessico settoriale e lessico comune nell'estrazione di terminologia specialistica da corpora di dominio

1 IL PROBLEMA

Sebbene, come ricordato in Cortelazzo (1990), il lessico fornisca «elementi distintivi che individuano una lingua speciale sia rispetto ad altre lingue speciali sia rispetto alla lingua comune», tuttavia, la definizione dei confini tra lessico settoriale e lessico comune presenta non pochi problemi di delimitazione. Ciò è dovuto, da un lato, all'«escursione terminologica» di ogni linguaggio settoriale, tendenza connaturata al lessico che lo caratterizza e legata, secondo Beccaria (1973), alla «crescente forza espansiva», al «prestigio reale nell'uso parlato e scritto» di ogni lessico settoriale. D'altro canto, le difficoltà di definire confini netti tra lessico settoriale e comune sono riconducibili al fatto che nel lessico di una lingua «si manifestano sia il carattere di continuum nella scala dei registri sia i punti di contatto e di transizione fra sottocodici». È infatti in questi termini che Rovere (1989), affrontando la dibattuta questione, mette in luce come essa riguardi non solo la dimensione 'orizzontale', le cui variazioni non sempre nette «rendono labili i confini disciplinari» tra lingue settoriali (o sottocodici), ma anche quella 'verticale', dal momento che all'interno di uno stesso linguaggio settoriale le varie «tipologie comunicative diverse per grado di tecnicità e formalità» (o registri) non sempre sono ben delineate.

Dunque, riguardo sia alla varietà dei contenuti (variazione orizzontale) sia alla variazione situazionale (verticale) si può parlare di un continuum all'interno del quale il lessico di una lingua varia tra due estremi rappresentati, da un lato, dalla lingua comune, dall'altro, da linguaggi caratterizzati da un lessico altamente specialistico (Rondeau *et al.*, 1984).

Ciò è dimostrato dal fatto che le forme del rapporto tra lessico settoriale e comune cambiano a seconda del linguaggio settoriale. Ricorda Mortara Garavelli (2001): una cosa è la «formalizzazione delle lingue speciali scientifiche», altra cosa «è la condizione condivisa dalle varietà di lingua che differiscono dalla matrice comune per l'impiego di tecnicismi lessicali e per una formalità di registri». Nel secondo caso, l'allusione è alla lingua del diritto, una lingua alla quale la «qualifica di lingua tecnica sta un po' stretta» (Fiorelli, 2008), che si caratterizza al contrario come «un sottoinsieme, distinto ma non separato dal linguaggio generale o comune» (Cassese, 1992). Essa, dunque, ben esemplifica i noti e non lievi problemi di delimitazione rispetto sia alla dimensione 'orizzontale' sia 'verticale' di variazione tra linguaggi settoriali e lingua comune. La

lingua del diritto infatti «più delle altre fa ricorso a risemantizzazioni del lessico comune, […] diffonde nel lessico comune i propri termini, e […] contemporaneamente è impegnata in scambi comunicativi cui partecipano anche parlanti non specialistici» (Cortelazzo, 1995). Inoltre, essendo essa finalizzata a «dar norma alla vita comune e ad attività specialistiche di ogni genere in mille diversi aspetti», vi si intrecciano la «ricchezza del linguaggio comune e dei vari linguaggi specialistici» (Scarpelli, 1959), oltre al lessico tecnico-giuridico.

A partire da tali considerazioni, questo studio è finalizzato a suggerire alcune possibili soluzioni a diversi ordini di problemi connessi con l'estrazione automatica di terminologia specialistica da corpora di dominio. La questione riguarda 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 dimensione di variazione lessicale sia 'orizzontale' sia 'verticale'. In particolare, i problemi sono connessi con la difficoltà di estrarre lessico settoriale a partire a) da corpora rappresentativi di sottocodici caratterizzati da diversi livelli di specializzazione, e b) da collezioni di testi appartenenti a diversi tipi di registri. Un ulteriore e tuttavia centrale problema riguarda la necessità di distinguere all'interno di un corpus rappresentativo di un unico sottocodice i *termini* appartenenti a più di un lessico settoriale: è il caso, ad esempio, di corpora rappresentativi della lingua del diritto, nei quali si intrecciano il lessico del dominio giuridico e quello proprio della materia legislata.

2 I SISTEMI DI ESTRAZIONE TERMINOLOGICA

I sistemi di estrazione automatica di terminologia da corpora di dominio non sempre riescono a fronteggiare in modo adeguato i diversi ordini di problemi delineati sopra. All'interno della comunità di ricerca impegnata a sviluppare metodi e strumenti per estrarre in modo automatico terminologia da corpora specialistici, Cabré (1999) ricorda come le maggiori difficoltà siano dovute proprio al confine non sempre così netto tra linguaggi settoriali e lingua comune, nonché al costante scambio biunivoco che li lega.

Centrali per lo sviluppo di applicazioni reali quali la costruzione di sistemi di organizzazione della conoscenza di dominio e di accesso al testo su base semantica, i sistemi di estrazione terminologica sono finalizzati all'identificazione e all'estrazione di unità terminologiche mono- e polirematiche da corpora di dominio. A questo scopo, vengono utilizzate una serie di misure statistiche finalizzate a determinare la probabilità per un'unità lessicale di essere un termine rilevante per il dominio. In particolare, l'estrazione di unità monorematiche è tipicamente realizzata sulla base della distribuzione di frequenza di occorrenza nel corpus, oppure su misure di rilevanza statistica quali la TF/IDF (Term Frequency/Inverse Document Frequency, Salton *et al.*, 1988). Per le unità polirematiche, si parte dall'assunto di base che se due o più parole formano un termine è molto probabile che nell'uso reale esse tendano a ricorrere insieme in maniera statisticamente 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, "Mutual Information" (Church

e Hanks, 1990), "Log-likelihood" (Dunning, 1993) per arrivare al più recente "C-NC Value" (Frantzi e Ananiadou, 1999).

Sebbene tali misure riescano a identificare con successo le unità terminologiche rilevanti per il corpus di estrazione, tuttavia esse non sono sempre sufficienti a discriminare tra *termini* settoriali e parole comuni (o *non-termini*). Le difficoltà riguardano soprattutto i casi di estrazione terminologica a partire da testi caratterizzati da un linguaggio settoriale non altamente specialistico e rivolti a un pubblico di non esperti di dominio. Il lessico di tali tipologie di testi presenta infatti confini non sempre netti rispetto al lessico comune.

Sino ad oggi, i migliori risultati dei sistemi di estrazione 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. Tale tipologia di testi è infatti un caso esemplare di netta separazione tra lessico settoriale e lessico comune. Al contrario, nei casi di estrazione automatica di terminologia specialistica da corpora rappresentativi di domini non altamente specialistici e/o composti da testi rivolti ad un ampio pubblico i risultati sono meno soddisfacenti. Sono questi infatti testi nei quali il lessico di dominio non è nettamente distinto dal lessico comune anche per il fatto di essere destinati a un pubblico più vasto.

Una soluzione operativa a queste difficoltà è fornita da un secondo tipo di sistemi di estrazione terminologica. Sono i sistemi che si basano sul cosiddetto 'approccio contrastivo'. L'estrazione di unità terminologiche monorematiche e polirematiche è cioè condotta a partire dal confronto della distribuzione delle unità terminologiche monorematiche e polirematiche 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 (ovvero 'per contrasto') al corpus di riferimento. A questo scopo sono state sviluppate una serie di metodologie in grado di computare la misura della diversa rilevanza di unità terminologiche all'interno dei due corpora che vengono confrontati. La possibilità di discriminare *termini* e *non-termini* è così empiricamente realizzata sulla base di un'analisi 'contrastiva' della loro distribuzione in un corpus di dominio (il corpus di acquisizione) rispetto a un corpus rappresentativo della lingua comune (usato come 'corpus di contrasto').

Per quanto forniscano una risposta positiva al problema di discriminare *termini* da parole comuni, tali sistemi (per menzionarne alcuni, Penas *et al.*, 2001; Chung *et al.*, 2004; Basili *et al.*, 2001) presentano a nostro avviso un limite fondamentale per quanto riguarda il modo con cui vengono acquisite le unità terminologiche polirematiche la cui estrazione è subordinata alla precedente acquisizione di unità monorematiche. Ciò causa, almeno in linea di principio, due ordini di problemi, ovvero il risultato finale del processo estrattivo può: a) includere unità polirematiche non rilevanti ma lessicamente "governate" da una testa che è stata identificata come specifica per il dominio; b) non includere unità polirematiche rilevanti che non sono state acquisite perché la loro testa lessicale non è stata selezionata come specifica per il dominio. Più concretamente, in un esperimento di estrazione terminologica condotta a partire da un corpus di articoli scientifici sul cambiamento climatico, l'unità terminologica polirematica *effetto serra* è acquisita solo sulla base della precedente identificazione dell'unità monorematica *effetto*. Di conseguenza, 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 estratte come termini di dominio.

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. Abbiamo visto in precedenza che un ulteriore problema riguarda la necessità di distinguere all'interno di un corpus rappresentativo di un unico sottocodice i termini appartenenti a più di un lessico settoriale: a nostra conoscenza, 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'. A nostro avviso, questo rappresenta un aspetto che non può essere ignorato nel processo di estrazione terminologica automatica.

3 LA NOSTRA SOLUZIONE: LINEE GUIDA

Tenute in considerazione, da una parte, le difficoltà connesse con la definizione di confini chiari e ben definiti tra lessico settoriale e comune e, dall'altra, le soluzioni sin ad oggi adottate da chi ha sviluppato sistemi di estrazione automatica di terminologia specialistica da corpora, questo studio ha l'obiettivo di proporre una nuova strategia di estrazione terminologica automatica. In particolare, la proposta qui descritta è finalizzata a suggerire una possibile soluzione ai diversi ordini di problemi delineati nel Paragrafo 1 per i quali abbiamo visto che i sistemi di estrazione terminologica automatica correnti non forniscono risposte adeguate.

In primo luogo, la metodologia qui proposta, basata su un approccio di tipo 'contrastivo', suggerisce una strategia di estrazione terminologica in grado di discriminare in modo automatico *termini* da *non-termini* a partire da un corpus di dominio. Tale approccio è stato applicato a un caso particolarmente spinoso, quello cioè dell'estrazione di terminologia settoriale da corpora di testi scritti in linguaggi che occupano una posizione intermedia nel continuum tra linguaggi altamente specialistici e lingua comune. A questo scopo i due domini scelti sono stati quello della storia dell'arte e il dominio giuridico, entrambi caratterizzati da un lessico non altamente specialistico. Ciò ha permesso di dimostrare come la metodologia di estrazione adottata offra una possibile soluzione ai problemi a) dell'acquisizione di lessico settoriale a partire da collezioni di testi rivolti ad un pubblico di non esperti di dominio (cfr. Paragrafo 5.1) e b) della distinzione tra più lessici settoriali all'interno di un corpus rappresentativo di un unico sottocodice (cfr. Paragrafo 5.2).

In secondo luogo, in questo studio è proposto un metodo innovativo per estrarre unità terminologiche polirematiche. A differenza infatti dei precedenti studi su base contrastiva, la metodologia di estrazione automatica di polirematiche qui proposta ne considera la rilevanza di dominio sulla base della loro settorialità come elementi 'unici' e non rispetto alla rilevanza della monorematica che ne costituisce la testa. Ad esempio, nel caso del corpus di testi legislativi in materia ambientale, la strategia di estrazione adottata consente di acquisire come polirematiche rilevanti solo *principio attivo* (rilevante per il lessico ambientale) e *principio di sussidiarietà* (rilevante per il lessico del diritto), a prescindere dalla rilevanza del termine monorematico *principio*; ciò

permette al contempo di escludere dal risultato finale polirematiche quali *principio generale* e *principio fondamentale*, non rilevanti per il dominio in questione ma presenti nel corpus di acquisizione.

Basandosi sulla considerazione che le unità terminologiche polirematiche rappresentano più della metà del vocabolario di un madre-lingua (Jackendoff, 1997), tale approccio trova conferma nello studio di De Mauro e Voghera (1996). Gli autori conducendo un'analisi dei lessemi complessi (LC) presenti nel *Lessico di frequenza dell'italiano parlato* (LIP), rispetto al grado di composizionalità del loro significato, a proposito dei LC appartenenti a linguaggi settoriali, concludono che «non sempre la settorialità di un LC è connessa con l'esistenza di accezioni speciali dei membri componenti, ma può derivare dal fatto che il LC assume in determinati contesti un significato globale speciale». Ciò comporta che la settorialità di un LC non è necessariamente funzione della rilevanza di dominio delle unità monorematiche di cui il LC si compone.

A nostro avviso, ciò risulta particolarmente significativo nel caso dell'estrazione di terminologia da corpora di testi giuridici caratterizzati, com'è noto, da una lingua alquanto 'formulaica'. In un'ottica di indagine lessicale condotta a partire da corpora testuali, le ricerche svolte in Nystedt (2000) e Eklund-Braconi (2000), attraverso l'interrogazione automatica di collezioni di documenti normativi europei, offrono una dimostrazione empirica di tale giudizio. In particolare, le ricerche condotte da Eklund-Braconi dimostrano come «l'analisi della singola parola non sia sufficiente a fornire il quadro semantico completo e reale» del corpus di normativa europea in materia ambientale esaminato. Al contrario, risultati più significativi per il dominio si ottengono dall'esame di quelle parole che «sono spesso legate tra loro in formule più o meno fisse» così da costituire «unità semantiche complete» dotate di un «significato finito e specialistico».

4 LA METODOLOGIA DI ESTRAZIONE

La metodologia di estrazione proposta, illustrata in dettaglio in Bonin *et al.* (2010), si articola in tre fasi:

- Fase 1: annotazione linguistica del testo condotta con strumenti di Trattamento Automatico del Linguaggio;
- Fase 2: identificazione all'interno del testo linguisticamente annotato di unità terminologiche monorematiche e polirematiche candidate all'estrazione;
- Fase 3: confronto della distribuzione dei termini candidati identificati nel corpus di acquisizione rispetto un corpus di riferimento.

Nella prima fase, il corpus di acquisizione viene lemmatizzato ed etichettato a livello morfo-sintattico (Dell'Orletta, 2009). Dal testo così annotato, attraverso l'uso di filtri linguistici e statistici, vengono estratte due liste di potenziali unità terminologiche, monorematiche e polirematiche. I filtri linguistici consentono di individuare all'interno del corpus di acquisizione:

i) le potenziali unità monorematiche, sulla base della categoria morfo-sintattica assegnata ('sostantivo');

ii) le potenziali unità polirematiche, sulla base di una serie di sequenze di categorie morfo-sintattiche rappresentative di diversi tipi di modificazione nominale. Ad

esempio, da una sequenza come 'sostantivo+aggettivo' sono individuate polirematiche quali *arte contemporanea*, *rifiuto pericoloso*, *norma nazionale*; da una sequenza 'sostantivo+preposizione+sostantivo' sono individuati potenziali termini quali *opera d'arte*, *limite di emissione*, *licenza d'importazione*; per arrivare a sequenze complesse come 'sostantivo+aggettivo+aggettivo+preposizione+aggettivo+sostantivo' sulla base della quale è individuato un termine come *inquinamento atmosferico transfrontaliero a grande distanza*.

I filtri statistici consentono di ordinare i termini potenziali individuati sulla base della loro rilevanza all'interno del corpus di acquisizione, attribuendo loro un valore di significatività. In particolare, la significatività delle unità monorematiche viene stabilita sulla base della loro frequenza di occorrenza all'interno del corpus di acquisizione; mentre le unità polirematiche sono ordinate sulla base del C-NC Value, una delle misure più utilizzate nei sistemi di estrazione terminologica. Il risultato di questa fase è rappresentato da una lista di unità monorematiche e polirematiche, costituite sia da *termini* (specialistici per il dominio) sia da *non-termini* (o parole comuni). Si noti che l'ordinamento ottenuto sulla base dei filtri statistici utilizzati non permette ancora di discriminare in modo preciso tra lessico settoriale e lessico comune.

Ciò avviene nella successiva fase di confronto con un corpus di riferimento, all'interno della quale la distribuzione di una selezione di termini candidati, effettuata sulla base dei valori di significatività ad essi assegnati, viene confrontata con la distribuzione delle medesime unità in un corpus usato come riferimento. Questo passaggio permette di riorganizzare la selezione di termini candidati all'estrazione rispetto ad un valore di contrasto calcolato statisticamente sulla base del confronto con corpus di riferimento (per maggiori dettagli sulla misura cfr. Bonin *et al.* 2010). Ne risulta che, ai termini più significativi per il dominio di appartenenza del corpus di acquisizione sarà associato un valore di contrasto maggiore, mentre a quelli meno significativi saranno attribuiti valori più bassi. Ciò permette di discriminare, nel glossario finale, tra *termini*, rilevanti per il dominio, e *non-termini*.

5 DUE ESPERIMENTI DI ESTRAZIONE TERMINOLOGICA

La metodologia di estrazione terminologica illustrata nei precedenti paragrafi è stata testata attraverso due esperimenti basati su due corpora caratterizzati da linguaggi che occupano una posizione intermedia nel continuum tra linguaggi altamente specialistici e lingua comune: quello della storia dell'arte e quello giuridico. Tali corpora presentano sfide e problematiche diverse: nel primo caso, l'acquisizione di lessico settoriale ha riguardato corpora caratterizzati da un livello non particolarmente alto di specializzazione, nel secondo caso si ha la compresenza di terminologia appartenente a due domini diversi all'interno dello stesso corpus.

5.1 Estrazione terminologica a partire da un corpus di testi di storia dell'arte

L'estrazione di unità terminologiche monorematiche e polirematiche è stata condotta a partire da un corpus di testi di storia dell'arte (326.066 parole), costruito da esperti di dominio con documenti tratti da pagine web di contenuto artistico. Tale corpus (da ora in avanti ARTE) si presenta dunque omogeneo rispetto al dominio, ma piuttosto

eterogeneo per quanto riguarda la tipologia di registri dei testi in esso contenuti in ragione della natura variegata del web: in ARTE sono contenuti testi specialistici, così come testi rivolti ad un pubblico più vasto.

Sulla base della metodologia 'contrastiva' di estrazione terminologica è stato selezionato un corpus di riferimento rispetto al quale confrontare la distribuzione delle unità terminologiche di ARTE (corpus di acquisizione). In questo esperimento è stato usato il corpus PAROLE, un corpus di italiano contemporaneo di circa 3 milioni di parole rappresentativo del lessico comune (Marinelli *et al.*, 2003).

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

Ordinamento sulla base del filtro statistico (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 seconda fase di estrazione terminologica. In corsivo i *non-termini.*

A partire da tali liste ordinate, si procede alla terza fase di estrazione selezionando i primi 600 termini, che vengono riordinati sulla base della loro significatività rispetto al corpus di contrasto[1]. È in questa fase di confronto della distribuzione dei termini nel corpus di acquisizione ARTE e nel corpus di contrasto (PAROLE), 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[2].

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

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

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

il corpus di riferimento, scendendo oltre la trecentesima posizione.

In fase di valutazione, il glossario ottenuto è stato prima di tutto confrontato con un Thesaurus di dominio (fornito dal dipartimento di Storia delle Arti dell'Università di Pisa), quindi validato da esperti. Tale valutazione ha registrato un incremento significativo dei termini di dominio estratti, che sono passati da 61,33% al termine della fase 2 al 79,40% a conclusione dell'analisi contrastiva.

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.

5.2 Estrazione terminologica a partire da un corpus di testi giuridici

In questo secondo esperimento è stato usato come corpus di acquisizione una collezione di direttive europee in materia ambientale per un totale di 394.088 parole (da ora in avanti AMB), reperito dalla versione disponibile on-line del Bollettino Giuridico Ambientale[3]. 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. Come illustrato nel Paragrafo 1, entrambe le tipologie di lessico sono infatti da considerarsi rilevanti per il dominio. 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). A differenza dell'esperimento precedente, qui ci si è concentrati sull'estrazione di unità terminologiche polirematiche; a questa scelta ha contribuito la loro particolare significatività per il dominio giuridico (cfr. Paragrafo 3).

Anche in questo caso, dopo la fase di annotazione linguistica automatica, è stata estratta una lista di 600 unità polirematiche ordinate per valori decrescenti sulla base dei valori del C-NC Value; in questo caso si osserva la co-occurrenza 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*).

È la successiva fase di confronto prima con il corpus PAROLE e poi con CONS che ha permesso di distinguere in primo luogo le unità polirematiche rilevanti per AMB dai *non-termini*, e in secondo luogo i termini del lessico del diritto da quelli del lessico ambientale. In particolare, la distribuzione delle prime 600 unità terminologiche

[3] http://extranet.regione.piemonte.it/ambiente/bga/index.htm

precedentemente estratte è stata confrontata con la loro occorrenza in PAROLE; ciò ha permesso di fare "emergere" i termini rilevanti per AMB, cioè sia giuridici sia ambientali. 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 prime due colonne le prime 10 unità terminologiche della lista estratta al termine della fase 2, nelle ultime due colonne le prime e ultime cinque posizioni della lista risultante dalla doppia analisi contrastiva (fase 3). Come si può vedere, mentre al termine della fase 2 i termini appartenenti al lessico del diritto (in corsivo) e al lessico ambientale (in grassetto) sono mischiati, nella lista finale i termini dei due lessici settoriali sono riordinati in modo da essere distinti (la testa della lista contiene i termini ambientali mentre nella coda si concentrano quelli del diritto).

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 2 e 3

La valutazione dei risultati conseguiti, condotta sulla base del "Dizionario Giuridico" (Edizioni Simone)[4] e del *Thesaurus EARTh* (Environmental Applications Reference Thesaurus)[5] seguita da una verifica manuale da parte di esperti, ha permesso di dimostrare come la metodologia seguita sia affidabile. Mentre, infatti, dopo l'estrazione sulla base del C-NC Value il 65,34% dei termini della lista di 300 termini 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 aumentano fino al 43,33% e quelle del lessico del diritto fino al 29,33% (con un incremento complessivo del 7,32%).

Ciò è anche chiaramente visibile nella Figura 1, che mostra la distribuzione dei termini del lessico ambientale e del lessico del diritto nella lista finale di 300 unità polirematiche estratte (suddivisa in gruppi di 30 termini). Come si può vedere, mentre nella prima parte della lista i termini ambientali sono in maggioranza su quelli appartenenti al lessico del diritto, nell'ultima parte la tendenza si inverte.

[4] http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index,5,A
[5] http://uta.iia.cnr.it/earth.htm\#EARTh\%202002

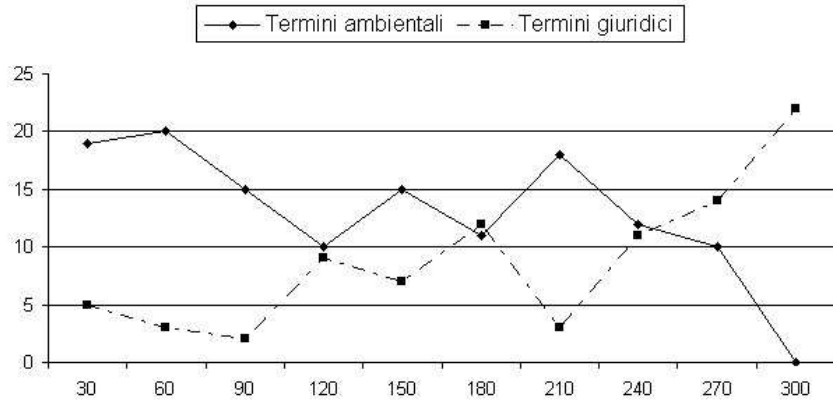

Figura 1: Distribuzione dei termini del lessico ambientale e del diritto nella lista finale estratta.

6 CONCLUSIONI

Attraverso una rivisitazione degli studi linguistici sul rapporto tra lessici settoriali e lessico comune abbiamo identificato diversi ordini di problemi ai quali i sistemi correnti di estrazione automatica di terminologia specialistica da corpora di dominio non forniscono, a nostro avviso, risposte adeguate. In particolare, abbiamo visto che la difficoltà di distinguere tra termini e non-termini varia in relazione al livello di specializzazione e al registro del corpus di acquisizione. Qui, la sfida è posta da testi che occupano una posizione intermedia nel continuum tra linguaggi altamente specialistici e lingua comune. Un'ulteriore ma non secondaria sfida riguarda la necessità di distinguere, all'interno di un corpus rappresentativo di un unico sottocodice, i termini appartenenti a diversi lessici settoriali; ad es. il lessico del diritto da quello proprio della materia legislata nel caso di corpora giuridici. Ad oggi, a nostra conoscenza nessun sistema automatico ha affrontato questo problema.

Il presente contributo ha cercato di colmare 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, mostrando al contempo un'interessante sinergia tra studi linguistici e applicazioni pratiche.

7 RIFERIMENTI BIBLIOGRAFICI

BASILI Roberto, MOSCHITTI Alessandro, PAZIENZA Maria Teresa, ZANZOTTO Fabio Massimo, *A contrastive approach to term extraction*, in «Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA-2001)», Nancy, 2001.

BECCARIA Gian Luigi, *Linguaggi settoriali e lingua comune*, in G.L. BECCARIA (a cura di), I linguaggi settoriali in Italia, Milano, Bompiani, pp. 7-59, 1973.

BONIN Francesca, DELL'ORLETTA Felice, MONTEMAGNI Simonetta, VENTURI Giulia, *A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora*, in «Proceedings di LREC'10 - Seventh International Conference on Language Resources and Evaluation», Valletta (Malta), 17-23 May 2010, pp. 3222 - 3229.

CABRÉ Maria Teresa, *The terminology. Theory, methods and applications*, John Benjamins Publishing Company, 1999.

CASSESE Sabino, *Introduzione allo studio della formazione*, in « Rivista trimestrale di diritto pubblico», 2, pp. 307-330, 1992.

CHUNG Teresa Mihwa., NATION Paul, *Identifying technical vocabulary*, in «System, 32», pp. 251-263, 2004.

CHURCH Kenneth Ward, HANKS Patrick, *Word association norms, mutual information, and lexicography*, in «Computational Linguistics», 16(1), pp. 22-29, 1990.

CORTELAZZO Michele, *Lingua e diritto in Italia. Il punto di vista dei linguisti*, in L. SCHENA (a cura di), *La lingua del diritto. Difficoltà traduttive. Applicazioni didattiche. Atti del primo Convegno Internazionale*, Milano, 5-6 ottobre, Roma, pp. 35-50, 1995.

CORTELAZZO Michele*, Lingue speciali. La dimensione verticale*, «Studi linguistici applicati», Padova, Unipress, 1990.

DE MAURO Tullio, VOGHERA Miriam, *Scala mobile. Un punto di vista sui lessemi complessi*, in P. BENINCÀ et al. (a cura di), *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio C. Lepschy*, Roma, Bulzoni, pp. 99-131, 1996.

DELL'ORLETTA Felice, *Ensemble system for Part-of-Speech tagging*, in «Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)», Reggio Emilia, dicembre, 2009.

DUNNING Ted, *Accurate Methods for the Statistics of Surprise and Coincidence,* in «Computational Linguistics», 19(1), 1993.

EKLUND-BRACONI Paola*, Il linguaggio normativo delle Comunità Europee. Studi quantitativi e semantici sul lessico con particolare riguardo al concetto di ambiente*, Dipartimento di francese e italiano – Università di Stoccolma, Stockolm, Graphium, 2000.

FIORELLI Piero, *Intorno alle parole del diritto*, Milano, Giuffrè, 2008.

FRANTZI Katerina, ANANIADOU Sophia*, The C-value / NC Value domain independent method for multi-word term extraction*, in «Journal of Natural Language Processing, 6(3)», pp. 145-179, 1999.

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

MARINELLI Rita et al., *The Italian PAROLE corpus: an overview*, in A. ZAMPOLLI et al. (eds.), *Computational Linguistics in Pisa*, Special Issue, XVI-XVII, Pisa-Roma, IEPI. Tomo I, pp. 401–421, 2003.

MORTARA GARAVELLI Bice, *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani,* Torino, Einaudi, 2001.

NYSTEDT Jane, *L'italiano nei documenti della CEE: le sequenze di parole*, in D. VERONESI (a cura di), *Linguistica giuridica italiana e tedesca: obiettivi, approcci, risultati, atti del Convegno di studi* (Bolzano, 1-3 ottobre 1998), Padova, Unipress, pp. 273-284, 2000.

PENAS Anselmo, VERDEJO Felisa, GONZALO Julio, *Corpus-Based Terminology Extraction Applied to Information Access*, in «Proceedings of the Corpus Linguistics 2001», pp. 458-465, 2001.

RONDEAU Guy, SAGER Juan, *Introduction à la terminologie (2nd ed.)*, Chicoutimi, Gatan Morin, 1984.

ROVERE Giovanni, *Sottocodici e registri in testi tecnici*, in «Rivista Italiana di Dialettologia», 13, pp. 135-160, 1989.

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

SCARPELLI Uberto, *Contributo alla semantica del linguaggio normativo*, Torino, Memoria dell'Accademia delle Scienze, 1959.