

Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio

FELICE DELL'ORLETTA, ALESSANDRO LENCI, SIMONE MARCHI,
SIMONETTA MONTEMAGNI, VITO PIRRELLI, GIULIA VENTURI

The paper focuses on the automatic extraction of domain knowledge from Italian legal texts and presents a fully-implemented ontology learning system (T2K, Text-2-Knowledge) that includes a battery of tools for Natural Language Processing, statistical text analysis and machine learning. Evaluated results show the considerable potential of systems like T2K, exploiting an incremental interleaving of NLP and machine learning techniques for accurate large-scale semi-automatic extraction and structuring of domain-specific knowledge.

Keywords: Natural Language Processing – Machine Learning – knowledge extraction from texts – ontology learning – legal ontologies

1. Introduzione

La necessità quotidiana di accedere a grandi quantità di conoscenza digitale disponibile in forma *non strutturata*, cioè convogliata attraverso testo libero all'interno di basi documentali anche molto vaste e variegata per stile e argomento, ha dato grande impulso allo sviluppo di tecnologie per l'acquisizione, la classificazione e la gestione automatica dell'informazione testuale e al loro sempre più diffuso impiego in una miriade di contesti applicativi (si vedano a questo proposito, tra gli altri, Brin, 1998; Kogut e Holmes, 2001; Vargas-Vega *et alii*, 2002; Dingli, Ciravegna e Wilks, 2003; Popov *et alii* 2003; Dill *et alii*, 2003; Giovannetti *et alii*, 2007). Nonostante gli evidenti successi conseguiti, tuttavia, le tecnologie esistenti si scontrano ancora oggi con un limite fondamentale: l'insufficiente attenzione prestata all'analisi linguistica dei testi. Un reale salto tecnologico verso l'accesso avanzato all'informazione testuale richiede di superare i limiti rappresentati da una capacità solo rudimentale di accedere al contenuto semantico codificato dalla struttura linguistica di un testo. Affinare questa capacità significa dotare i sistemi per l'estrazione di informazione da testi di un'adeguata *intelligenza linguistica*. Questa può variare dalla semplice integrazione di conoscenze lessicali e terminologiche all'annotazione di livelli più avanzati di informazione sintattico-semantica, essenziali per aumentare la quantità e la qualità dell'informazione recuperata e per filtrare il "rumore di fondo" dell'informazione irrilevante.

A nostro avviso, l'esigenza di analizzare grandi repertori di documenti in cui la lingua si manifesta in tutta la sua complessità e variabilità d'uso è destinata ad avere un notevole impatto non solo sulle fasi di progettazione e sviluppo delle tecnologie per il Trattamento Automatico del Linguaggio (TAL), ma anche sul nostro stesso modo di studiare il linguaggio umano e di concettualizzare i rapporti tra linguaggio, cognizione e contesti comunicativi concreti. Da una parte, si va sempre più avvertendo la necessità di affiancare ai tradizionali componenti di analisi linguistica (grammatiche formali sviluppate su base introspettiva) nuove tipologie di strumenti per l'acquisizione dinamica di conoscenza, basati sull'impiego di algoritmi di *machine learning*, in grado di adattarsi con rapidità ed efficienza a diversi domini applicativi e terminologici e alla variabilità linguistica offerta da tipologie testuali anche radicalmente differenti. D'altro canto, gli attuali sistemi di rappresentazione formale della conoscenza di dominio (ontologie) offrono l'inedita opportunità di sviluppare procedure di integrazione dinamica tra livelli di conoscenza linguistica ed extra-linguistica, consentendo di superare gli attuali limiti delle rispettive tecnologie. Ad esempio, l'estrazione di informazioni di dominio da testi liberi può utilmente integrare le attività di sviluppo e popolamento manuale di un modello ontologico, per loro natura lente, ripetitive e soggette a errori. Per converso, la natura fortemente implicita, personalizzata e fraseologizzata dell'informazione estraibile da testi può beneficiare in larga misura dell'insieme di concetti, proprietà e relazioni offerte da una rappresentazione formale e strutturata di uno specifico dominio di conoscenza.

Text-to-Knowledge (T2K), una piattaforma *software* sviluppata congiuntamente dall'Istituto di Linguistica Computazionale (CNR) e dal Dipartimento di Linguistica dell'Università di Pisa finalizzata all'acquisizione di tipi diversi di informazione semantico-lessicale da documenti testuali, ci consente di illustrare concretamente la portata di questi cambiamenti metodologici e di delinearne le prospettive future. Attraverso l'uso combinato di tecniche statistiche e di strumenti avanzati per il TAL, T2K è in grado di analizzare il contenuto linguistico dei documenti, individuare i termini potenzialmente più significativi, ricostruire una "mappa" multidimensionale dei concetti espressi da questi termini, sviluppare un'ontologia del dominio di interesse.

2. Il paradosso dell'acquisizione

La costruzione semi-automatica di ontologie di dominio, intese come *repertori strutturati di concetti rilevanti per la descrizione e organizzazione di un certo dominio di conoscenza* (Gruber, 1995), è un ambito di ricerca particolarmente attivo in diversi settori specialistici quali la bio-informatica, il campo della pubblica amministrazione, quello della gestione documentale aziendale e giuridico-legislativa. In questo senso sono state

proposte diverse metodologie finalizzate all'estrazione automatica di informazione da basi testuali e alla loro strutturazione in ontologie di dominio (per una rassegna, cfr. Buitelaar et alii, 2005). Gli sforzi in tale direzione, tuttavia, sono tipicamente inficiati da un classico paradosso. Stabilire una corrispondenza adeguata tra la rappresentazione linguistica del dominio di conoscenza fornita da un insieme di documenti rilevanti e la rappresentazione ontologica sottostante dello stesso dominio *presuppone* la disponibilità di una notevole quantità di conoscenza rilevante rispetto all'ambito trattato. Ad esempio, acquisire conoscenza relativa agli obblighi a cui è soggetta una particolare entità giuridica (ad es. il "datore di lavoro") richiede la capacità preliminare di identificare il "datore di lavoro" come un tipo di entità giuridica, nonché la capacità di localizzare, nel contesto di un testo legislativo, gli obblighi a cui tale entità è soggetta (ad es. garantire la sicurezza personale del lavoratore). Estrarre informazione da un testo richiede dunque altra informazione. Più tecnicamente, "popolare" i nodi di un'ontologia a partire da informazione testuale richiede che la struttura dell'ontologia sia già in piedi.

È nostra convinzione che l'unione sinergica di tecnologie linguistiche e tecniche di *machine learning* possa rappresentare una strategia metodologica vincente per far fronte a questo paradosso. Da questo punto di vista, T2K offre un esempio interessante di strumento *ibrido* per l'analisi automatica del contenuto testuale, al cui interno si integrano aspetti tradizionali di analisi linguistica e funzionalità per l'accesso a livelli sempre più astratti e strutturati di conoscenza. Riteniamo che questa strategia possa portare a coniugare in modo non banale due esigenze complementari del mondo della comunicazione: l'esigenza di una rappresentazione esplicita, normalizzata e condivisa del contenuto, cui fanno fronte i modelli più o meno recenti di rappresentazione formale della conoscenza, e quella derivante dal bisogno di personalizzare questo contenuto, secondo prospettive soggettive condizionate dal contesto e dal punto di vista dell'utente, rispetto alla quale il linguaggio rappresenta uno strumento insostituibile. Ritourneremo su questo aspetto metodologico al paragrafo 4 del presente contributo.

Il funzionamento di T2K sarà illustrato con i risultati di alcuni esperimenti di estrazione e strutturazione di terminologia condotti su due corpora di testi giuridici italiani in materia di legislazione ambientale e di protezione del consumatore. Ad oggi, il dominio giuridico costituisce un'area attiva di studio, particolarmente aperta alla necessità di dotare le tecnologie di gestione dell'informazione di un'adeguata "intelligenza linguistica". Le difficoltà intrinseche al linguaggio naturale, in generale, e al Trattamento Automatico del linguaggio dei testi giuridico-amministrativi, in particolare, hanno infatti fatto sì che le ricerche in materia di costruzione di ontologie giuridiche siano state condotte per lo più *in modo manuale* da esperti del dominio e in una prospettiva *top-down* rivolta soprattutto alla strutturazione della dottrina giuridica (per uno stato dell'arte cfr. Valente, 2005). Riteniamo, pertanto, che la metodologia ibrida seguita da T2K nell'acquisizione di conoscenza di dominio a partire da categorie di analisi interne

ai testi possa costituire un promettente punto di partenza per sviluppare in modo semi-automatico un'ontologia giuridica in una prospettiva *bottom-up*. Finora, pochi tentativi sono infatti stati fatti in questa direzione per indurre in modo automatico ontologie giuridiche a partire da corpora testuali (per uno stato dell'arte cfr. Lenci *et alii*, 2008).

3. Obiettivi

Obiettivo generale di T2K è trasformare le conoscenze implicitamente codificate all'interno di un corpus di testi in conoscenza esplicitamente strutturata. Il risultato finale è un glossario terminologico arricchito con informazione semantico-concettuale. Per arrivare a identificare i concetti rilevanti e più caratterizzanti i documenti di un certo dominio di interesse, T2K si avvale di strumenti di analisi in linea con lo stato dell'arte nella ricerca linguistico-computazionale partendo da un'ipotesi di lavoro molto semplice: i concetti e i temi rilevanti nel testo sono veicolati dai termini statisticamente più significativi. Questi ultimi possono essere unità lessicali monorematiche come *accordo*, *produttore* o *presidente* oppure unità lessicali polirematiche come *procedimento amministrativo*, *Ministro dell'ambiente*, *incenerimento dei rifiuti pericolosi*, *assistenza reciproca*, *contratto di multiproprietà*, ecc. 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.

I termini che formano il glossario terminologico acquisito possono essere a loro volta raggruppati secondo diverse relazioni di similarità. Ad esempio, *tutela ambientale*, *tutela dei consumatori*, *tutela dell'ozono stratosferico*, *tutela del paesaggio* e simili condividono il concetto più generale di TUTELA cui sono tutti ricondotti attraverso la relazione di iponimia (o ISA). Oltre questa strutturazione concettuale di tipo gerarchico, T2K è anche in grado di identificare classi di termini semanticamente correlati come ad esempio {*disposizioni*, *norme*, *decisione*, *atto*, *prescrizioni*}, {*legge*, *regolamento*, *protocollo*, *accordo*, *statuto*}, {*inquinamento*, *danno ambientale*, *effetti nocivi*, *conseguenza*}. L'organizzazione e la strutturazione dei termini secondo relazioni gerarchiche e di quasi-sinonimia rappresenta il risultato della successiva fase operativa di T2K sulla base della quale è possibile condurre un'indicizzazione concettuale dei testi.

Un sistema di conoscenza non è tuttavia costituito solo da concetti che si riferiscono a entità del dominio, ma anche da processi, azioni ed eventi che vedono coinvolte queste entità secondo ruoli e funzioni diverse. Ad esempio, un *decreto legislativo* così come un *articolo* o un *comma* sono tipicamente *abrogati*, *sostituiti*, *modificati* così come possono essere *emanati*, *recepiti*, *applicati*; la *qualità dell'aria* insieme al *livello di protezione ambientale* e all'*ecosistema* sono *garantiti*, *protetti* e *salvaguardati* così come posso-

no essere *pregiudicati* e *inquinati*. Gli sviluppi più recenti di T2K vanno nella direzione appena delineata, cercando di identificare le relazioni più tipiche che legano le entità e i concetti identificati con il fine ultimo di arrivare a ricostruire dai testi una “mappa” semantica del dominio esplorato.

Nel suo complesso, il risultato finale di T2K si presenta dunque come una rete multi-dimensionale di unità terminologico-concettuali con significativi punti di contatto sia con l'architettura classica dei thesauri sia con reti semantico-lessicali come WordNet (Fellbaum, 1998). La rete di conoscenza acquisita da T2K è articolata sui seguenti livelli:

- *glossario terminologico*, costituito da una lista di termini mono o polirematici, estratti automaticamente dai testi arricchiti con vari livelli di annotazione linguistica, in particolare morfo-sintattica e sintattica. I termini vengono selezionati con criteri statistici come i più significativi o salienti per la caratterizzazione dei documenti. Il glossario terminologico finale include anche indicazione delle varianti ortografiche, morfologiche e strutturali associate ad ogni unità terminologica acquisita;
- *tassonomia concettuale* – i termini del glossario sono strutturati attraverso relazioni gerarchiche di iponimia/iperonimia ricostruite a partire dalla loro struttura linguistica interna;
- *famiglie di termini concettualmente affini* – i termini del glossario sono organizzati in classi di termini semanticamente simili sulla base della loro distribuzione all'interno di contesti lessico-sintattici;
- *rete semantico-concettuale* – i concetti che si riferiscono ad entità del dominio sono messi in relazione attraverso le azioni e gli eventi che li vedono tipicamente coinvolti secondo ruoli e funzioni diverse. A partire dal testo viene ricostruita una rete semantica costituita dalle relazioni più tipiche che coinvolgono i concetti selezionati.

Ad oggi, il risultato di T2K è circoscritto a quanto descritto ai punti 1-3, mentre la costruzione della rete semantico-concettuale (4) rappresenta un'area attiva di studio e di sperimentazione.

4. Architettura funzionale di T2K

L'architettura e i processi di elaborazione di T2K sono organizzati in due fasi successive, strettamente legate ai livelli di organizzazione dell'ontologia descritti nel paragrafo 3:

Fase 1: creazione del glossario dei termini;

Fase 2: strutturazione concettuale del repertorio terminologico acquisito. Il diagramma in Figura 1 schematizza l'architettura e i processi di elaborazione

di T2K. Nella colonna centrale del diagramma sono riportate le varie fasi del processo estrattivo, dove livelli di analisi linguistica si alternano con fasi di elaborazione statistica del testo linguisticamente annotato. Il risultato di questo processo è costituito dall'ontologia di dominio (rappresentata nel riquadro a destra) articolata su diversi livelli: glossario terminologico, tassonomia concettuale, famiglie di termini semanticamente affini e rete semantico-concettuale. I componenti di analisi linguistica integrati in T2K sono parte di AnIta, una batteria di strumenti ad ampio spettro per il trattamento automatico dell'italiano (si veda il diagramma in Figura 2), disponibile anche in versione web <foxdrake.ilc.cnr.it/webtpls>. Ritorniamo più avanti sul loro ruolo all'interno di T2K.

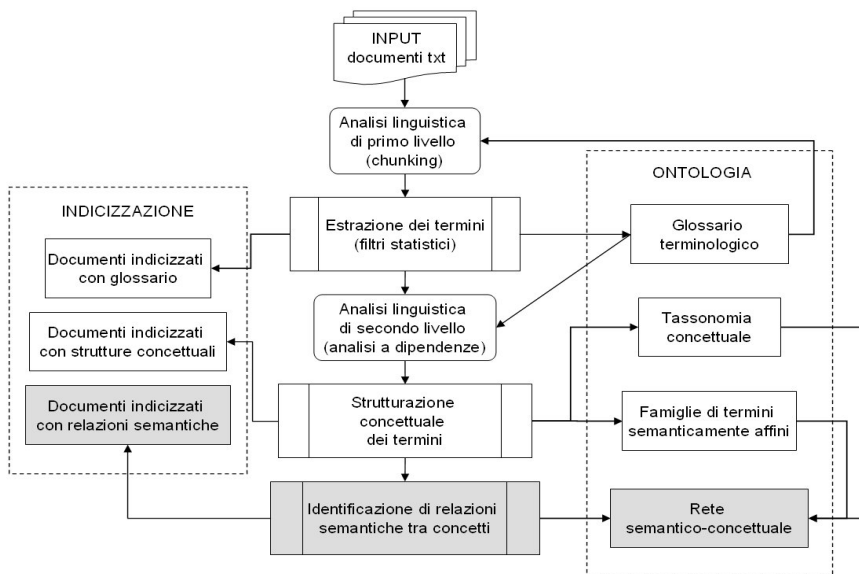


Figura 1 - L'architettura e i processi di elaborazione di T2K

Vale la pena di sottolineare che alcuni livelli di informazione ontologica estratti dal testo vengono reintegrati nel testo di partenza, stratificati come livelli di annotazione o glosse, per arricchirne il contenuto. Il testo così arricchito viene a sua volta usato come input nelle fasi successive del processo di acquisizione di conoscenza. Ad esempio, il glossario terminologico estratto dal testo viene riproiettato sul testo stesso in vista di una seconda fase di analisi linguistica a dipendenze. È a questo livello di analisi che diventa possibile recuperare le relazioni che ciascuno dei termini acquisiti intrattiene

contestualmente con altri termini all'interno della base documentale di partenza. In questo modo, si attiva un ciclo virtuoso di annotazione-acquisizione-annotazione, in cui il testo mantiene il suo ruolo centrale di deposito di informazioni progressivamente strutturate a livelli di crescente astrazione e complessità. Questa metodologia consente a nostro avviso di integrare al meglio i contributi di diverse tecnologie dell'informazione, affrontando in modo dinamico il problema di come acquisire la conoscenza implicitamente contenuta in basi documentali di domini specialistici, dove le strutture linguistiche del testo si intrecciano strettamente con aspetti di conoscenza del mondo reale. Soprattutto in campo giuridico, infatti, applicazioni sviluppate su ampie basi di conoscenza necessitano di ontologie che comprendano conoscenza di dominio continuamente aggiornata. Un processo testo-conoscenza organizzato in modo dinamico può dunque risultare estremamente efficace in questo ambito operativo. Da una parte, la necessità di disporre di rappresentazioni altamente strutturate, compatte e decontestualizzate dell'informazione acquisita è soddisfatta attraverso la costruzione di repertori informativi strutturati quali glossari, ontologie e reti semantiche. Usando queste strutture come chiavi di accesso al testo (in forma di "metadati" testuali) è possibile indicizzarne il contenuto per termini, concetti e relazioni. D'altra parte, la stratificazione di questi livelli di informazione sul testo è funzionale alla creazione di ulteriori livelli di analisi, attraverso un processo incrementale che da una parte distilla nuova conoscenza e dall'altra la mette a disposizione per successive elaborazioni. Torneremo su questi aspetti metodologici nelle conclusioni dell'articolo.

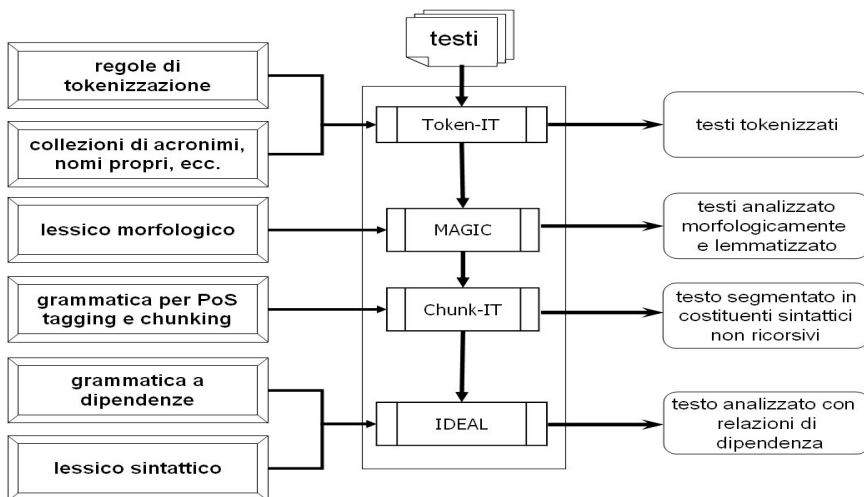


Figura 2 - Architettura di AnIta

5. Creazione del glossario terminologico

In T2K il processo di estrazione terminologica riguarda unità terminologiche monorematiche e polirematiche. Nel caso delle prime, il processo di acquisizione opera sul testo lemmatizzato ed etichettato a livello morfo-sintattico (Battista e Pirrelli, 2000) e avviene sulla base della frequenza dei lemmi nei documenti. Ciò equivale a dire che la frequenza di ogni termine del glossario è la somma delle frequenze delle diverse forme flesse riconducibili allo stesso lemma. Una volta definita la lista dei potenziali termini semplici, i termini candidati sono filtrati sulla base della loro frequenza all'interno del corpus di acquisizione.

Diverso è il caso delle unità terminologiche polirematiche la cui acquisizione si articola in più fasi. Innanzitutto, i potenziali termini complessi sono identificati all'interno di testi segmentati sintatticamente in costituenti sintattici elementari non ricorsivi detti "chunks" (Abney, 1991; Federici et alii, 1996; Lenci et alii, 2003). Un chunk è identificato come una sequenza continua di parole del testo che va dalla prima unità grammaticale (tipicamente, una preposizione, un ausiliare, un (pre)determinatore o un ausiliare) incontrata nel testo (detta "iniziatore di chunk") alla prima unità lessicale piena selezionata dall'iniziatore di chunk. Esempi di chunk nominale (N_C) sono *la mia prima casa* (con *la* iniziatore di chunk e *casa* testa lessicale piena selezionata) e *questo difficile problema* (con *questo* iniziatore di chunk e *problema* testa lessicale piena selezionata). *Per il medico e da sotto il tavolo* sono tipici chunk preposizionali (P_C), con *per* e *da* iniziatori di chunk e *medico* e *tavolo* teste lessicali. Infine nei chunks verbali di modo finito (FV_C) è *stato evidenziato* e *ho ripetutamente osservato* gli ausiliari *è* e *ho* iniziano il chunk e i participi passati *evidenziato* e *osservato* rappresentano le teste lessicali selezionate dagli ausiliari.

Il testo così segmentato viene analizzato da una mini-grammatica deputata al riconoscimento delle strutture linguistiche che formano potenziali termini complessi: ad esempio, una sequenza di chunk nominale (N_C) seguito da un chunk aggettivale (ADJ_C) (es. *organizzazione internazionale*), oppure una sequenza di chunk nominale seguito da un chunk preposizionale (P_C) (es. *presidente della repubblica*). L'assunto di base è che se due o più parole formano un termine complesso in un certo dominio, è molto probabile che nell'uso linguistico relativo a quel dominio esse tendano a ricorrere insieme in maniera statisticamente significativa. La soglia di significatività statistica è determinata a partire da una stima della probabilità che le parole in questione *possano ricorrere insieme per caso*, probabilità esprimibile in funzione della frequenza delle parole stesse all'interno di un corpus rappresentativo. Se una coppia di parole ricorre nel testo più spesso di quanto saremmo autorizzati ad aspettarci in base alle leggi del caso, allora il test di significatività è soddisfatto.

Per ciascuno dei potenziali termini complessi identificati dalla mini-grammatica viene dunque quantificata su base statistica la forza di associazione tra le parole che lo compongono. In T2K, questa forza è stimata applicando la misura associativa detta “log-likelihood” (Dunning, 1993). Una serie di esperimenti preliminari ha infatti evidenziato che tale misura produce risultati sensibilmente migliori rispetto ad altre misure statistiche, quali “pointwise mutual information”, “t-score”, ecc., in quanto appare più robusta nel caso di dati linguistici con bassa frequenza di occorrenza. Infatti, a differenza di altre misure associative quali la “mutual information”, la log-likelihood non privilegia l'associazione di parole rare (Manning e Schütze, 1999). Non escludiamo tuttavia che risultati diversi potrebbero essere ottenuti lavorando su corpora di grandi dimensioni, dove il problema della “sparsità” dei dati risulta notevolmente ridimensionato.

La lista dei potenziali termini complessi viene poi ordinata sulla base del grado di significatività della loro associazione all'interno del dominio dei documenti. Vale la pena qui ricordare che in T2K la misura di associazione si applica alle teste lessicali piene dei costituenti sintattici che lo compongono. Ad esempio, nel caso di un'unità terminologica polirematica come *adempimento dell'obbligo* viene misurata la forza di associazione tra le teste lessicali dei due chunks *adempimento* (N_C) e *dell'obbligo* (P_C): rispettivamente *adempimento* e *obbligo*. Il vantaggio derivante da questo approccio è duplice: da un lato permette di condurre il processo di estrazione terminologica facendo astrazione da variazioni di natura ortografica, morfologica così come strutturale, dall'altro rende possibile l'acquisizione delle varianti terminologiche associate a ciascun termine acquisito (vedi infra).

La lista dei termini candidati così ottenuta può essere ulteriormente estesa con termini complessi di “ordine” superiore al primo. Infatti, a questo livello i termini acquisiti sono tipicamente caratterizzati da una struttura sintattica binaria: ad es. N_C-ADJ_C, N_C-P_C, N_C-N_C, ecc. Per arrivare ad acquisire termini più complessi, la procedura di estrazione di termini appena descritta viene applicata iterativamente riproiettando sul testo segmentato in chunks i termini composti identificati durante la prima fase di estrazione. Ad esempio, se al primo passo è stato estratto il termine composto *comunità economica*, al secondo passo è possibile estrarre un nuovo termine composto, *comunità economica europea*, che include il termine individuato al passo precedente. Riportiamo di seguito alcuni esempi di unità terminologiche polirematiche ottenute mediante l'applicazione di questo processo di acquisizione incrementale. In grassetto sono evidenziati i termini complessi acquisiti allo stadio di analisi precedente:

- *acquisizione di un diritto di godimento a tempo parziale*
- *violazione delle disposizioni nazionali*
- *direttiva del parlamento europeo*
- *dispositivo automatico di chiamata*.

Abbiamo verificato che questo approccio incrementale all'estrazione terminologica presenta numerosi vantaggi rispetto a un'acquisizione condotta in un passo unico: innanzitutto la tipologia di strutture sintattiche da tenere in considerazione viene ridotta a un numero contenuto di strutture di base. L'acquisizione di termini più complessi è vincolata al fatto che almeno uno dei componenti del termine più complesso sia già stato selezionato come tale al passo precedente di acquisizione, con il risultato di ridurre il potenziale rumore nei risultati. In sintesi, questo processo incrementale di acquisizione terminologica, che rappresenta uno dei tratti caratterizzanti di T2K, fornisce maggiori garanzie sull'effettiva rilevanza dei termini acquisiti (Bartolini et alii, 2005).

Il processo di acquisizione terminologica produce due insiemi di termini candidati, ovvero una lista di potenziali unità terminologiche monorematiche ordinate per frequenze decrescenti, e una lista di potenziali unità terminologiche polirematiche ordinate per forza di associazione decrescente. Il glossario terminologico finale è costruito stabilendo diverse soglie riguardanti a) la frequenza di occorrenza minima dell'unità terminologica nella collezione documentale, e b) la percentuale di termini selezionati nelle liste ordinate di termini potenziali. Tali soglie possono essere definite in modo interattivo dall'utente sulla base delle caratteristiche della collezione documentale (tipicamente la sua estensione) e del tipo di risultato atteso (in termini di accuratezza e copertura). I termini così selezionati vengono a comporre il glossario di base, di cui riportiamo un estratto in Tabella 1 dove la colonna "termine" riporta la forma del termine che è stata selezionata come prototipica all'interno del corpus di acquisizione, la colonna "valore" indica la frequenza di ciascun termine nell'intera collezione di documenti analizzati e la colonna "lemma" riporta i lemmi delle teste lessicali dei chunks corrispondenti al termine in questione. Infine, la colonna "stop" è usata per segnalare i termini potenzialmente spuri che sono contrassegnati da un valore diverso da NULL (per maggiori dettagli sul ruolo di questo campo v. infra). Alcune precisazioni sono necessarie per una corretta interpretazione delle colonne "termine" e "lemma" della Tabella 1. Per quanto riguarda la prima, abbiamo deciso di rappresentare il termine acquisito attraverso la sua forma prototipica attestata all'interno del corpus piuttosto che mediante un esponente lessicale o lemma selezionato con criteri astratti (ad es. la forma maschile singolare); questa scelta deriva dal fatto che spesso un termine di dominio è legato a una forma specifica, ad esempio quella plurale come nel caso del termine *operazioni di smaltimento* riportato nel frammento di glossario in Tabella 1. La colonna "lemma" riporta invece una codifica astratta del termine formulata come sequenza delle teste lessicali dei chunks che lo costituiscono (es. *ordinanza presidente* come codifica astratta del termine *ordinanza del presidente*); questa codifica è condivisa da tutte le varianti morfologiche e strutturali del termine in questione.

KWID	Termine	Valore	Lemma	Stop
544	OPERATORE ECONOMICO	18	OPERATORE ECONOMICO	NULL
920	OPERAZIONE DI RECUPERO	8	OPERAZIONE RECUPERO	NULL
888	OPERAZIONE DI RICICLAGGIO	8	OPERAZIONE RICICLAGGIO	NULL
201	OPERAZIONI	109	OPERAZIONE	NULL
1098	OPERAZIONI DI SMALTIMENTO	5	OPERAZIONE SMALTIMENTO	NULL
1443	ORDINAMENTI DEGLI STATI MEMBRI	5	ORDINAMENTO STATO MEMBRO	NULL
240	ORDINAMENTO	91	ORDINAMENTO	NULL
360	ORDINAMENTO GIURIDICO	39	ORDINAMENTO GIURIDICO	NULL
1252	ORDINAMENTO GIURIDICO INTERNO	11	ORDINAMENTO GIURIDICO INTERNO	NULL
609	ORDINAMENTO NAZIONALE	15	ORDINAMENTO NAZIONALE	NULL
1043	ORDINAMENTO OLANDESE	6	ORDINAMENTO OLANDESE	NELL'
288	ORDINANZA	69	ORDINANZA	NULL
730	ORDINANZA DEL PRESIDENTE	11	ORDINANZA PRESIDENTE	NULL
921	ORDINANZA DI RINVIO	8	ORDINANZA RINVIO	NULL
825	ORGANI AMMINISTRATIVI	10	ORGANO AMMINISTRATIVO	NULL
398	ORGANI GIURISDIZIONALI	32	ORGANO GIURISDIZIONALE	NULL

Tabella 1 - Un estratto del glossario di base automaticamente acquisito da T2K

Un'ulteriore precisazione è necessaria in relazione al campo "stop" che, quando diverso da NULL, contiene la preposizione che introduce il termine in questione in tutte le sue attestazioni nel corpus di apprendimento. Questa informazione è stata introdotta per poter identificare semi-automaticamente la presenza di termini spuri, corrispondenti a locuzioni preposizionali (del tipo *ai sensi di*, *in materia di*) o locuzioni avverbiali (come *in linea di massima* o *in tempo utile*). Si è deciso di segnalare piuttosto che di eliminare i termini potenzialmente spuri in quanto essi possono corrispondere a termini caratterizzati da una bassa frequenza, come nel caso del termine *ordinamento olandese* in Tabella 1. La Tabella 2 riporta un insieme di termini spuri correttamente segnalati come tali:

KWID	Termine	Valore	Lemma	Stop
12	SENSI	691	SENSO	AI
42	SENSI DELL' ART.	332	SENSO ART.	AI
380	SENSI DELL' ARTICOLO	35	SENSO ARTICOLO	AI
377	SENSI DELLA DIRETTIVA	37	SENSO DIRETTIVA	AI
1366	SENSI DELLE DISPOSIZIONI DEL CAPITOLO	6	SENSO DISPOSIZIONE CAPITOLO	AI

Tabella 2 - Un estratto del glossario di T2K con termini marcati come potenzialmente spuri

Nell'estratto riportato, la colonna "stop" registra la preposizione *ai* che sistematicamente introduce le varie attestazioni di questa famiglia di termini spuri. Sulla base di questa informazione è possibile rimuovere dal glossario tali locuzioni preposizionali in modo automatico o – preferibilmente – attraverso una selezione e valutazione manuale. Il glossario terminologico finale è affiancato da una tabella che registra le varianti terminologiche attestate nel corpus di acquisizione. L'acquisizione di varianti terminologiche rappresenta una questione centrale nel processo di costruzione di risorse terminologiche (Nenadic et alii, 2004). Secondo Jacquemin (2001), in media un terzo delle occorrenze di un termine sono varianti. Ne consegue che un compito di riconoscimento automatico di terminologia debba tenere in considerazione non solo la forma attestata più frequente, ma la debba mettere in relazione alle varianti terminologiche corrispondenti. Ciò se da un punto di vista teorico permette di far luce sulla natura stessa di un termine (ad esempio scoprendo se si tratta di un'unità polirematica "monolitica" o flessibile), da un punto di vista applicativo consente di migliorare i risultati di compiti di indicizzazione e recupero di documenti. La tipologia di varianti acquisite da T2K nell'ambito del processo di estrazione terminologica include:

- varianti ortografiche: *tasso* [*d'interesse* vs *di interesse*]
- varianti morfologiche:
 - singolare vs plurale: *bene* vs *beni*
 - preposizione semplice vs preposizione articolata: *esercizio* [*del diritto* vs *dei diritti* vs *di un diritto*]
- varianti strutturali: *fornitore* [*per il servizio finanziario* vs *di servizi finanziari*]
- varianti con modificatori: *ostacoli* [*al funzionamento* vs *al buon funzionamento*]
- varianti che combinano diversi tipi di variazione: *contratto* [*di fornitura* vs *contratti di fornitura* vs *contratti per la fornitura*]

Per ogni unità terminologica, mono- o polirematica, T2K estrae le varianti che coprono almeno il 5% delle occorrenze del termine nel corpus di partenza. Il risultato di questo processo è esemplificato in Tabella 3 dove ogni riga rappresenta una variante: la colonna "var_ID" contiene un identificativo della variante, la colonna "Termine" specifica il termine a cui si riferisce la variante, la colonna "Forma_variante" contiene la variante stessa, e la colonna "Frequenza" riporta la frequenza di occorrenza della variante descritta nella base documentale.

Alla tipologia di varianti discussa sopra va aggiunta un'altra classe di varianti terminologiche che sono acquisite da T2K nel corso della fase successiva di strutturazione concettuale dei termini (descritta nel paragrafo 6): si tratta di varianti lessicali che possono includere, oltre a sinonimi, acronimi e abbreviazioni, questi ultimi forme di varianti molto comuni in linguaggi settoriali, come ad esempio *PBC* abbreviazione di *policlorobifenile* o *l.r.* acronimo di *legge regionale*.

var_ID	Termine	Forma_variante	Frequenza
680	ACQUISTO DEL BENE IMMOBILE	ACQUISITO DELL'IMMOBILE	5
681	ACQUISTO DEL BENE IMMOBILE	ACQUISTO DEL BENE IMMOBILE	12
682	ACQUISTO DEL BENE IMMOBILE	ACQUISITO DI UN BENE IMMOBILE	10
1050	ADEMPIMENTO DELL'OBBLIGO	ADEMPIMENTO DI OBBLIGHI	2
1051	ADEMPIMENTO DELL'OBBLIGO	ADEMPIMENTO DEL DETTO OBBLIGO	2
1052	ADEMPIMENTO DELL'OBBLIGO	ADEMPIMENTO DEGLI OBBLIGHI	6
1053	ADEMPIMENTO DELL'OBBLIGO	ADEMPIMENTO DELL'OBBLIGO	6
1631	ADOZIONE DI MISURE	ADOZIONE DELLA MISURA	2
1632	ADOZIONE DI MISURE	ADOZIONE DI MISURE	6
1629	ADOZIONE DI MISURE	ADOZIONE DI TUTTE LE MISURE	2
1630	ADOZIONE DI MISURE	ADOZIONE DELLE MISURE	3
1633	ADOZIONE DI MISURE	ADOZIONE DI UNA MISURA	3
937	AGENTE	AGENTE	87
938	AGENTE	AGENTI	67
378	CONTRATTI DI FORNITURA	CONTRATTO DI FORNITURA	3
379	CONTRATTI DI FORNITURA	CONTRATTI PER LA FORNITURA	4
380	CONTRATTI DI FORNITURA	CONTRATTI DI FORNITURA	17
680	ACQUISTO DEL BENE IMMOBILE	ACQUISITO DELL'IMMOBILE	5
681	ACQUISTO DEL BENE IMMOBILE	ACQUISTO DEL BENE IMMOBILE	12
682	ACQUISTO DEL BENE IMMOBILE	ACQUISITO DI UN BENE IMMOBILE	10

Tabella 3 - Varianti terminologiche acquisite automaticamente da T2K

6. Strutturazione concettuale dei termini

Dopo la fase di estrazione terminologica, T2K procede alla fase di organizzazione concettuale. Essa consiste nell'identificazione a) di relazioni di iponimia e iperonimia e b) di relazioni di affinità semantica tra i termini del glossario. A partire dai frammenti di strutture concettuali identificate, è possibile procedere alla loro validazione e riorganizzazione in base alle relazioni individuate.

Ad oggi, l'identificazione delle relazioni tassonomiche tra termini si basa su una relazione di inclusione lessicale. Due unità terminologiche polirematiche che condividano la medesima testa lessicale (e talora anche gli stessi modificatori) al livello della loro rappresentazione in chunks vengono interpretati come iponimi del termine corrispondente alla struttura condivisa. Da queste singole relazioni iponimiche è possibile ricostruire catene tassonomiche articolate su diversi livelli di profondità come esemplificato di seguito:

PROTEZIONE

PROTEZIONE DEI CONSUMATORI

PROTEZIONE DEGLI INTERESSI

PROTEZIONE DEGLI INTERESSI ECONOMICI

PROTEZIONE DEI MINORI

È in corso di valutazione il potenziamento di T2K con un componente per l'estrazione di relazioni tassonomiche tra termini che non condividano la testa lessicale (ad esempio, relazioni iperonimiche come quella che lega *ministro dell'ambiente* a *autorità*) secondo tecniche già ampiamente sperimentate per l'acquisizione di informazione tassonomica da dizionari (cfr. Calzolari, 1984; Montemagni, 1996).

L'identificazione di relazioni di affinità semantica tra i termini acquisiti segue un approccio completamente diverso. Si parte dall'assunto di base che la semantica di una parola si correla alle sue proprietà distribuzionali nel testo, ovvero due parole sono semanticamente simili se sono reciprocamente sostituibili in un numero significativo di contesti sintattici. Questo assunto è condiviso da un fertile filone di studi della letteratura linguistico-computazionale che a partire dalla lezione di Firth (1957) e di Harris (1968) cerca di indurre automaticamente aspetti del contenuto semantico delle parole sulla base della loro distribuzione nel testo (si veda, tra gli altri, Pereira et alii, 1993; Grefenstette, 1994; Lin, 1998; Rooth et alii, 1999; per una rassegna dei metodi e delle tecniche di acquisizione automatica di rappresentazioni semantico-lessicali a partire da testi cfr. Lenci et alii, 2005).

In T2K, l'acquisizione di famiglie di termini semanticamente affini è condotta sulla base di misure di similarità semantica basate su proprietà distribuzionali come illustrato in Allegrini et alii (2000a e 2000b). Secondo questo approccio, due termini sono correlati semanticamente se sono reciprocamente sostituibili all'interno di un numero significativo di relazioni di dipendenza sintattica tra teste lessicali. Ad esempio, l'evidenza fornita da contesti come *abrogare un decreto* e *abrogare una direttiva* così come di *integrare un decreto* e *integrare una direttiva* suggerisce che *decreto* e *direttiva* rappresentano termini semanticamente simili perché si correlano con la stessa funzione sintattica a due verbi, *abrogare* e *integrare*. Ovviamente, non tutti i verbi che co-occorrono con i

termini sono ugualmente significativi; si pensi a contesti del tipo *scrivere un decreto* dove il termine *decreto* si accompagna con un verbo che è poco informativo riguardo al suo significato in quanto la classe dei possibili oggetti di *scrivere* è alquanto vasta. Ne consegue che la similarità tra termini che ricorrono con verbi più selettivi (nei riguardi dei loro complementi) deve essere intuitivamente più alta di quella tra termini che dipendono da verbi meno selettivi. *Decreto* e *direttiva* sono dunque molto più simili tra loro (per il fatto di essere entrambi *abrogati* o *integrati*) di quanto non lo siano – mettiamo – *decreto* e *libro* a causa del fatto che entrambi possono essere *scritti*.

Il tipo di similarità semantica catturata attraverso il meccanismo inferenziale delineato sopra è basato sul reale contesto di uso delle parole. Come illustrato in Allegrini et alii (2002, 2003), sono possibili diverse misure di similarità semantica: in particolare, viene fatta distinzione tra una misura relativizzata di similarità semantica e una assoluta (dove la scelta tra le due dipende essenzialmente dall'informazione disponibile in partenza e dal tipo di risultato atteso). Mentre nel primo caso la similarità semantica di due parole w_1 e w_2 viene valutata in relazione a specifici contesti di uso (che costituiscono la prospettiva rispetto alla quale viene formulato il giudizio di prossimità), nel secondo caso la similarità semantica di w_1 e w_2 viene valutata in termini assoluti, ovvero senza alcuna indicazione riguardante i contesti d'uso rispetto ai quali deve essere valutata la similarità delle due parole.

L'induzione di classi di termini semanticamente affini in T2K avviene sulla base di una misura di similarità semantica "ibrida" che combina le due nozioni di similarità semantica relativizzata e assoluta appena enunciate. Il corpus viene prima analizzato da un componente per l'analisi delle relazioni di dipendenza funzionale (IDEAL, Bartolini et alii, 2002, 2004). All'interno del risultato di questo processo di analisi sintattica a dipendenze, vengono rintracciate le dipendenze grammaticali principali (soggetti, oggetti) che interessano i termini del glossario all'interno del corpus di acquisizione, ad esempio, *ogg_d(garantire, protezione dell'ambiente)* oppure *sogg(garantire, ministero)*. L'insieme delle coppie verbo-termine funzionalmente annotate rappresenta la base di conoscenza distribuzionale sulla base della quale vengono identificati i raggruppamenti di termini semanticamente affini all'interno del dominio analizzato. Per evitare il rumore introdotto da contesti verbali poco selettivi (es. *essere, prendere, fare, ecc.*) e al contempo per focalizzarsi sugli eventi più salienti a cui ciascun termine si correla all'interno del dominio esaminato, per ciascun termine del glossario è stato selezionato un sottoinsieme della base di conoscenza generale ritagliato a partire dall'identificazione dei verbi più rilevanti rispetto al dominio ed escludendo a priori verbi semanticamente "vuoti" o semplicemente più "leggeri" quali gli ausiliari o i verbi supporto. I verbi più salienti in relazione a ciascun termine sono stati identificati ricorrendo nuovamente alla misura associativa della log-likelihood (vedi sopra); ad esempio, nella base di conoscenza distribuzionale acquisita, tra i verbi più salienti associati al termine *decreto* troviamo

abrogare, istituire, applicare, regolare, stabilire e simili. In questo modo, per ciascun termine del glossario è stata estratto lo spazio delle sue distribuzioni sul piano sintagmatico, a partire dalla quale vengono ricostruite le sue relazioni di similarità a livello semantico-paradigmatico. Il risultato finale di questa fase di elaborazione si presenta in forma tabellare, come esemplificato in Tabella 4 dove per ciascun termine del glossario (prima colonna) viene riportato l'insieme dei termini identificati come semanticamente affini (seconda colonna), ordinati secondo valori decrescenti di similarità (vale a dire che, ad esempio, il termine *norma* è più strettamente correlato a *regolamento* di quanto non lo sia *art.*).

KWID	Termine	Termine correlato
40	REGOLAMENTO	NORMA
41	REGOLAMENTO	MODIFICA
42	REGOLAMENTO	PARAGRAFO
43	REGOLAMENTO	DIRETTIVA
44	REGOLAMENTO	ART.
676	REGOLE	NORMA
677	REGOLE	PRINCIPIO
678	REGOLE	REQUISITI
679	REGOLE	DIVIETO
680	REGOLE	PROCEDURA COMUNITARIA

Tabella 4: Raggruppamenti di termini semanticamente affini acquisiti da T2K

7. Un esperimento di acquisizione terminologica a partire da corpora di testi giuridici

In questa sezione riportiamo i risultati di un esperimento condotto su due corpora di testi giuridici diversificati rispetto al dominio legislativo (legislazione ambientale e tutela del consumatore) e all'ente emittente (Unione Europea, Stato italiano, Regione Piemonte).

7.1. I Corpora

Abbiamo applicato T2K a due corpora, un *Corpus Ambientale* (AMB) e un *Corpus sulla Tutela del Consumatore* (CONS). In dettaglio, AMB contiene 824 testi normati-

vi (es. legge, decreto, direttiva, ecc...) e amministrativi (es. ordinanza, deliberazione, circolare, ecc...), emanati dall'Unione Europea, dallo Stato italiano e dalla Regione Piemonte, per un totale di 1.399.617 parole, reperiti dal Bollettino Giuridico Ambientale edito dall'Assessorato all'ambiente della regione Piemonte. CONS è invece costituito dalla versione italiana di 16 direttive comunitarie e 42 sentenze, per un totale di 292.609 parole.

I due *corpora* sono stati pre-analizzati con il duplice obiettivo a) di estendere il lessico morfologico ad ampia copertura (*general purpose*) sottostante ad AnIta, aggiornandolo con i termini dei domini d'interesse, e b) di adattare la mini-grammatica per il riconoscimento di unità polirematiche alle specificità del linguaggio dei testi giuridico-amministrativi.

7.2. Valutazione del glossario terminologico acquisito

Dato l'ambito settoriale dei due *corpora* di partenza, in entrambi gli esperimenti di acquisizione il glossario terminologico estratto comprende termini che appartengono sia al dominio giuridico sia al dominio legislativo. Dal momento che le due tipologie di termini hanno frequenze di distribuzione piuttosto diverse nelle rispettive basi documentali di partenza, abbiamo deciso di condurre la valutazione dei glossari acquisiti da T2K tenendo in considerazione il fatto che a) a basse soglie percentuali di acquisizione corrispondeva un incremento di precisione nell'estrazione di terminologia appartenente al dominio legislativo e b) viceversa, imponendo soglie di selezione più alte, la precisione generale del glossario terminologico acquisito aumentava a discapito della terminologia ambientale e in materia di tutela del consumatore.

La valutazione è stata condotta su un glossario di 4.685 unità terminologiche mono- e polirematiche estratte da AMB e su un glossario di 1.443 termini estratti da CONS confrontando i risultati ottenuti con risorse terminologiche di riferimento sia di tipo giuridico sia specifiche dei domini legislativi. In considerazione della natura eterogenea dei glossari acquisiti, le risorse di riferimento selezionate come "gold standard" (GS) di riferimento sono state: sul versante giuridico, il *Dizionario Giuridico* (Edizioni Simone), *JurWordNet* (JWN) e la lista di parole chiave usate per la ricerca in rete dell'*Archivio DoGi* (Dottrina Giuridica); sul versante ambientale, il *Glossario dell'Osservatorio Nazionale sui Rifiuti* (rilasciato dal Ministero dell'Ambiente) e il *Thesaurus EARTH* (Environmental Applications Reference Thesaurus). La valutazione è stata condotta in termini di precisione, calcolata come la percentuale di termini acquisiti in modo corretto da T2K rispetto a tutti i termini estratti. A causa della diversa copertura delle risorse di riferimento selezionate rispetto alle basi documentali di partenza, una valutazione in termini di "recall" (calcolato come la percentuale di termini acquisiti corretta-

mente rispetto a tutti i termini presenti in GS) è stata condotta solo in relazione a un sottoinsieme di concetti selezionati come particolarmente rilevanti (v. infra).

Per quanto concerne i criteri di valutazione, oltre ai casi di corrispondenza piena tra i termini estratti automaticamente da T2K e quelli presenti nella risorsa di riferimento, sono stati considerati diversi tipi di corrispondenza parziale (per una descrizione dettagliata dei criteri di valutazione cfr. Montemagni et alii, 2007 e Lenci et alii, 2008) distinti nei seguenti casi:

- termine che nei due glossari si presenta con diverse forme prototipiche, ad esempio:
 - *accantonamento* (T2K) vs *accantonamenti* (GS_giuridico);
 - *acquisizione dati* (T2K) vs *acquisizione di dati* (GS_ambientale);
 - *abbandono di rifiuti* (T2K) vs *abbandono dei rifiuti* (GS_ambientale);
- termine che si presenta con diversa estensione di significato nei due glossari:
 - la risorsa di riferimento contiene il termine nella sua accezione più vasta, mentre T2K ha acquisito uno dei suoi iponimi, ad esempio *abrogazione di norme* (T2K) vs *abrogazione* (GS_giuridico);
 - la risorsa di riferimento contiene il termine nella sua accezione più ristretta, mentre T2K ha acquisito il suo iperonimo, ad esempio *agente di polizia* (T2K) vs *agente di polizia giudiziaria* (GS_giuridico);
- il termine estratto da T2K è un co-iponimo di un termine presente nel *gold standard*; è il caso ad esempio del termine estratto automaticamente *direttiva del consiglio*, i cui co-iponimi in *DoGi* e *JurWordNet* sono *direttiva comunitaria* e *direttiva amministrativa*.

Con l'esperimento condotto su AMB si è raggiunta una precisione del 75,4% considerando come risorse di riferimento sul versante giuridico il *Dizionario Giuridico* e le parole chiave dell'*Archivio DoGi* (Dottrina Giuridica) e sul versante ambientale il *Glossario dell'Osservatorio Nazionale sui Rifiuti* e il *Thesaurus EARTH*. Si noti che al 75,4% di precisione hanno contribuito sia i casi di corrispondenza piena sia i casi di corrispondenza parziale di cui ai punti 1) e 2) sopra. Per i rimanenti termini estratti è stata condotta una valutazione manuale che ne ha stabilito la rilevanza rispetto ai domini trattati: ad esempio, sono stati recuperati come rilevanti termini come *anidride carbonica* per il dominio ambientale, oppure *beneficiari* per il dominio legale. In questo modo, la percentuale di termini rilevanti è salita a 83,7%. Ancora migliori sono stati i risultati della valutazione del glossario estratto da CONS in relazione a *DoGi* e *JurWordNet*: i casi di corrispondenza sia totale sia parziale (casi 1), 2) e 3) sopra) costituiscono l'85,38%.

Nel caso degli esperimenti condotti con CONS, è stato inoltre possibile condurre una valutazione in termini di "recall" rispetto a un sottoinsieme di 56 concetti EULG (*European Union Legal Concepts*) selezionati come rilevanti rispetto al dominio legislativo (cfr. Peters et alii, 2005 per la lista completa): in relazione a questo sottoinsieme si è raggiunto un *recall* dell'80,69%.

8. Conclusioni

T2K si presenta come uno strumento versatile e personalizzabile a vari livelli (sulla base delle caratteristiche quantitative dei repertori documentali così come delle peculiarità linguistiche dei testi e delle finalità dell'utente) per l'individuazione di tipologie diverse di informazione testuale, che vanno dall'estrazione di terminologia tecnica di dominio all'acquisizione di diversi tipi di strutture concettuali per l'indicizzazione di banche dati documentali. L'interesse di T2K non è a nostro avviso limitato al versante puramente applicativo della gestione documentale (Bourigault et alii, 2001; Jacquemin e Bourigault, 2002), ma rappresenta un importante strumento di ausilio esplorativo per l'indagine terminologica e per la strutturazione ontologica di un dominio di conoscenze.

I risultati degli esperimenti di estrazione condotti su corpora di testi giuridici dimostrano infatti che uno degli aspetti più innovativi dell'architettura di T2K sia proprio l'interazione tra livelli di annotazione della struttura linguistica della base documentale di partenza e livelli di strutturazione semantico-lessicale della conoscenza di dominio acquisita. Tale visione dinamica e incrementale del processo di accesso al contenuto dimostra quanto il preteso confine tra conoscenza linguistica e conoscenza di dominio sia inesistente in reali contesti d'uso, laddove strutture linguistiche e aspetti di conoscenza del mondo sono uniti in modo inestricabile. Il ciclo annotazione-estrazione-annotazione alla base di T2K rappresenta, a nostro avviso, una sfida metodologica importante: estraendo basi di conoscenza di dominio direttamente dal testo, utilizzando strumenti di analisi linguistici "poveri" a priori di tale conoscenza, possiamo raggiungere livelli di rappresentazione e strutturazione del contenuto progressivamente sempre più "ricchi".

Bibliografia

- Abney S. (1991), *Parsing by chunks*. In: R.C. Berwick et al. (a cura di), *Principle-based Parsing: Computation and Psycholinguistics*, Kluwer, Dordrecht.
- Allegrini P., Montemagni S., Pirrelli V. (2000a), *Controlled Bootstrapping of Lexico-semantic Classes as a Bridge between Paradigmatic and Syntagmatic Knowledge: Methodology and Evaluation*. In: *Proceedings of Conference on Language Resources & Evaluation (LREC 2000)*, Atene, Grecia.
- Allegrini P., Montemagni S., Pirrelli V. (2000b), *Learning Word Clusters from Data Types*. In: *Proceedings of International Conference on Computational Linguistics (Coling 2000)*, Saarbruecken, Germania: 8-14.
- Allegrini P., Lenci A., Montemagni S., Pirrelli V. (2002), *Le Forme del Significato. Acquisizione e Rappresentazione dell'Informazione Semantica*. In: *Actas del Segundo*

- Seminario de la Escuela Interlatina de Altos Estudios en Linguistica Aplicada. Matematica y Tratamiento de Corpus*, Fundaciòn San Millàn de la Cogolla, Logroño: 245-268.
- Allegrini P., Montemagni S., Pirrelli, V. (2003), *Example-Based Automatic Induction Of Semantic Classes Through Entropic Scores*. In "Linguistica Computazionale", XVI-XVII: 1-45.
- Bartolini R., Lenci A., Montemagni S., Pirrelli V. (2002), *Grammar and Lexicon in the Robust Parsing of Italian. Towards a Non-Naïve Interplay*. In: *Proceedings of International Conference on Computational Linguistics (Coling 2002-Workshop on Grammar Engineering and Evaluation)*, Taipei.
- Bartolini R., Lenci A., Montemagni S., Pirrelli V. (2004), *Hybrid Constraints for Robust Parsing: First Experiments and Evaluation*. In: *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 26-28 May 2004, Centro Cultural de Belem, Lisbon, Portugal: 795-798.
- Bartolini R., Giorgetti D., Lenci A., Montemagni S., Pirrelli V. (2005), *Automatic Incremental Term Acquisition from Domain Corpora*. In: *Proceedings of 7th International conference on Terminology and Knowledge Engineering (TKE2005)*, Copenhagen Business School, 17-18 August 2005, Copenhagen, Denmark.
- Battista M., Pirrelli V. (2000), *Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane*. Rapporto Tecnico ILC-CNR-2000.
- Bourigault D., Jacquemin C., L'Homme M.C. (a cura di) (2001), *Recent Advances in Computational Terminology*. John Benjamins Publishing Company, Amsterdam-Philadelphia.
- Brin S. (1998), *Extracting Patterns and Relations from the World Wide Web*. In: *WebDB Workshop at 6th International Conference on Extending Database Technology*.
- Buitelaar P., Cimiano P., Magnini B. (Eds.) *Ontology Learning from Text: Methods, Evaluation and Applications*. Frontiers in Artificial Intelligence and Applications Series, Vol. 123, IOS Press, July 2005.
- Calzolari N. (1984), *Detecting Patterns in a Lexical Database*. In: *Proceedings of the 10th International Conference on Computational Linguistics (COLING-84)*, Stanford, California: 170-173.
- Dill S., Gibson N., Gruhl D., Guha R., Jhingran A., Kanungo T., Rajagopalan S., Tomkins A., Tomlin J.A., Zien J.Y. (2003), *SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation*. In: *Twelfth International World Wide Web Conference*. the Semantic Web, 2005.
- Dingli A., Ciravegna F., Wilks Y. (2003), *Automatic Semantic Annotation using Unsupervised Information Extraction and Integration*. In: *K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation*.
- Dunning, T. (1993), *Accurate Methods for the Statistics of Surprise and Coincidence*. In "Computational Linguistics", 19(1).

- Federici, S. Montemagni, S. Pirrelli, V. (1996), *Shallow Parsing and Text Chunking: a View on Underspecification in Syntax*. In: *Proceedings of the Workshop On Robust Parsing*, tenuto nell'ambito della European Summer School on Language, Logic and Information (ESSLLI-96), Praga, Repubblica Ceca, 12-16 Agosto 1996.
- Fellbaum C. (a cura di) (1998), *WordNet. An Electronic Lexical Database*. MIT Press, Cambridge MA.
- Firth J.R. (1957), *A synopsis of linguistic theory 1930-55*. In: *Studies in Linguistic Analysis (special volume of the Philological Society)*, Oxford, The Philological Society: 1-32.
- Giovannetti E., Marchi S., Montemagni S., Bartolini R. (2007), *Ontology-based Semantic Annotation of Product Catalogues*. In: *Proceeding of the 6th International Conference in Recent Advances in Natural Language Processing (RANLP 2007)*.
- Grefenstette G. (1994), *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston.
- Gruber T.R. (1995), *Toward principles for the design of ontologies used for knowledge sharing*. In "International Journal of Human and Computer Studies", XLIII, 1995: 907-928.
- Harris Z.S. (1968), *Mathematical structures of language*. Wiley.
- Jacquemin C. (2001), *Spotting and Discovering Terms through NLP*, MIT Press, Cambridge MA.
- Jacquemin C., Bourigault D. (2002), *Term extraction and automatic indexing*. In: R. Mitkov (a cura di), *Handbook of Computational Linguistics*. Oxford University Press, Oxford.
- Kogut P., Holmes W. (2001), *AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages*. In: *First International Conference on Knowledge Capture*.
- Lin D. (1998), *Automatic Retrieval and Clustering of Similar Words*. In: *Proceedings of COLING-ACL'98*, Montreal, Canada.
- Lenci A., Montemagni S., Pirrelli V. (2003), *Chunk-It. An Italian Shallow Parser For Robust Syntactic Annotation*. In "Linguistica Computazionale", XVI-XVII, 2003: 353-386.
- Lenci A., Montemagni S., Pirrelli V. (2005), *Acquiring and Representing Meaning: Computational Perspectives*. In: A. Lenci S., Montemagni V., Pirrelli (a cura di), *Acquisition and Representation of Word Meaning. Theoretical and computational perspectives*, Istituti Editoriali e Poligrafici Internazionali, Pisa/Roma, Italia: 19-66.
- Lenci A., Montemagni S., Pirrelli V., Venturi G., (2008), *Ontology learning from Italian legal texts*, in Breuker J. et al. (Eds.), *Legal Ontologies and the Semantic Web*, IOS-Press, (in corso di pubblicazione).
- Manning C.D., Schütze H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge MA.

- Montemagni S. (1996), *Architecture and Functioning of a System for the Acquisition of Taxonomical Information from Dictionary Definitions*. In: *Proceedings of the 4th Conference on Computational Lexicography and Text Research (COMPLEX '96)*, Budapest, Ungheria, 15-17 Settembre 1996.
- Montemagni S., Marchi S., Venturi G., Bartolini R., Bertagna F., Ruffolo P., Peters W., Tiscornia, D. (2007), *Report on Ontology learning tool and testing*. Progetto Europeo DALOS (Drafting Legislation with Ontology-Based Support), Deliverable 3.3, Dicembre 2007.
- Nenadic G., Ananiadou S., McNaught J. (2004), *Enhancing Automatic Term Recognition through Term Variation*, in *Proceedings of 20th International Conference on Computational Linguistics (Coling 2004)*, Geneva, Switzerland.
- Pereira F., Tishby N., Lee L. (1993), *Distributional Clustering Of English Words*. In: *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 1993: 183-190.
- Peters W., Sagri M-T., Tiscornia D. (2005), *The Structuring of Legal Knowledge in LOIS*, In *Proceedings of 10th International Conference of Artificial Intelligence and Law, (ICAIL 2005)*, Bologna, Italy, June 6th-11th.
- Popov B., Kiryakov A., Kirilov A., Manov D., Ognyanoff D., Goranov M. (2003), *KIM - Semantic Annotation Platform in 2nd International Semantic Web Conference*. In: *ISWC2003*.
- Rooth M., Riezler S., Prescher D., Carroll G., Beil F. (1999), *Inducing a Semantically Annotated Lexicon via EM-Based Clustering*. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, USA, June 1999: 104-111.
- Valente A. (2005), *Types and Roles of Legal Ontologie*, In Benjamins, V. R. et alii (eds.) *Law and the Semantic Web*. Springer: Berlin/Heidelberg, DE, pp. 65-76.
- Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A., Ciravegna F. (2002), *MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup*. In *the 13th International Conference on Knowledge Engineering and Management*.