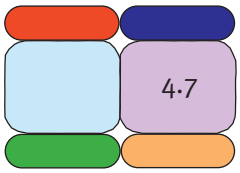
LINGUISTICA COMPUTAZIONALE

STRUMENTI E RISORSE PER IL TRATTAMENTO AUTOMATICO DELLA LINGUA

Nicoletta Calzolari
Alessandro Lenci

Le ricerche sul TAL hanno aperto nuove prospettive per la creazione di applicazioni per l'accesso intelligente al contenuto documentale. Sviluppi significativi riguardano i sistemi per l'analisi "robusta" del testo, i metodi per l'acquisizione automatica di conoscenza dai documenti e le infrastrutture per lo sviluppo e gestione di risorse linguistiche di grandi dimensioni, grazie ai quali è oggi possibile realizzare modelli e strumenti per il trattamento della lingua utilizzabili in contesti operativi reali.



1. IL TRATTAMENTO AUTOMATICO DELLA LINGUA

Nella società dell'informazione differenti categorie di utenti (professionisti, amministratori pubblici e comuni cittadini) devono confrontarsi con la necessità quotidiana di accedere a grandi quantità di contenuti digitali *semi-strutturati o non strutturati*, all'interno di basi documentali in linguaggio naturale disponibili sul Web o su Intranet locali. Un'alta percentuale delle conoscenze e processi che regolano le attività di gruppi di lavoro, istituzioni e imprese risiede, infatti, all'interno di documenti dalle forme e tipologie più varie (testi normativi, manuali, agenzie stampa, rapporti tecnici, *e-mail* ecc.), talvolta in lingue diverse e, sempre più di frequente, accompagnati da materiale multimediale. La natura non strutturata di tale informazione richiede due passi fondamentali per una sua gestione efficace: ovvero, la selezione dei documenti rilevanti rispetto alle necessità specifiche dell'utente e l'estrazione dell'informazione dai testi, per garantire il suo impiego in altre applicazioni o per compiti specifici. La facilità di tale accesso, la capacità di recuperare l'informazione adeguata in tempi rapidi, la sua gestione e usabilità sono, dunque, parametri chiave per garantire il successo di imprese economiche, lo sviluppo imprenditoriale, la competitività professionale, così come anche l'integrazione sociale e occupazionale e la formazione permanente.

Gli sviluppi più recenti della *linguistica computazionale* e del *natural language engineering* hanno creato soluzioni tecnologiche dalle enormi potenzialità per migliorare la ricerca e gestione intelligente dell'informazione contenuta nei documenti testuali. Le nuove tecnologie della lingua, infatti, permettono ai sistemi informatici di accedere ai contenuto digitale attraverso il *Trattamento Automatico della Lingua* (TAL) o *Natural Language Processing* (NLP). Il problema di come acquisire e gestire la conoscenza depositata nei documenti testuali dipende dal suo essere codificata all'interno della rete di strutture e relazioni grammaticali e lessicali che costituiscono la natura stessa della comunicazione linguisti-

ca. Sono il lessico e le regole per la combinazione delle parole in strutture sintatticamente complesse che nel linguaggio si fanno veicoli degli aspetti multiformi e creativi dei contenuti semantici. Attraverso l'analisi linguistica automatica del testo, gli strumenti del TAL sciolgono la tela del linguaggio per estrarre e rendere espliciti quei nuclei di conoscenza che possono soddisfare i bisogni informativi degli utenti. Dotando il computer di capacità avanzate di elaborare il linguaggio e decodificarne i messaggi, diventa così possibile costruire automaticamente rappresentazioni del contenuto dei documenti che permettono di potenziare la ricerca di documenti anche in lingue diverse (*Crosslingual Information Retrieval*), l'estrazione di informazione rilevante da testi (*Information Extraction*), l'acquisizione dinamica di nuovi elementi di conoscenza su un certo dominio (*Text Mining*), la gestione e organizzazione del materiale documentale, migliorando così i processi di elaborazione e condivisione delle conoscenze.

2. UN PO' DI STORIA: IL TAL IERI E OGGI

Nata come disciplina di frontiera, di fatto ai margini sia del mondo umanistico che delle applicazioni informatiche più tradizionali, la linguistica computazionale in poco più di 50 anni è riuscita a conquistare una posizione di indiscussa centralità nel panorama scientifico internazionale. In Italia, alla storica culla pisana rappresentata dall'Istituto di Linguistica Computazionale del CNR – fondato e diretto per lunghi anni da Antonio Zampolli – si sono affiancati molti centri e gruppi di ricerca attivi su tutto il territorio nazionale. Sul versante applicativo, le numerose iniziative imprenditoriali nel settore delle tecnologie della lingua testimoniano l'impatto crescente della disciplina (sebbene con ritmi molto più lenti che nel resto dell'Europa, come risulta dal rapporto finale del progetto comunitario Euromap [12]) al di fuori dello specifico ambito accademico, prova del fatto che i tempi sono diventati maturi perché molti dei suoi risultati affrontino la prova del mercato e della competizione commerciale.
Quali i motivi di questa crescita esponenziale? Sebbene facilitato dai progressi nel settore informatico e telematico, unitamente all'effetto catalizzante di Internet, sarebbe improprio spiegare lo sviluppo della disciplina solo in termini di fattori meramente tecnologici. In realtà, la linguistica computazionale possiede, oggi, una sua maturità metodologica nata dalla conquista di un preciso spazio di autonomia disciplinare anche rispetto alle sue anime originarie, l'indagine umanistica e la ricerca informatica. Questa autonomia si contraddistingue per un nuovo e delicato equilibrio tra lingua e computer. Le elaborazioni computazionali sono, infatti, chiamate a rispettare la complessità, articolazione, e multidimensionalità della lingua e delle sue manifestazioni testuali. Al tempo stesso, i documenti testuali emergono come una risorsa di conoscenza che può essere gestita ed elaborata con le stesse tecniche, metodologie e strumenti che rappresentano lo stato dell'arte nella tecnologia dell'informazione.
A tale proposito è utile ricordare come la linguistica computazionale affondi le sue radici in due distinti paradigmi di ricerca. Da un lato, è possibile trovare i temi caratteristici dell'applicazione di metodi statistico-matematici e informatici allo studio del testo nelle scienze umane, di cui Padre Roberto Busa e Antonio Zampolli rappresentano i pionieri nazionali. Il secondo paradigma fondante è rappresentato dall'*Intelligenza Artificiale* (IA) e, in particolare, dall'ideale delle "macchine parlanti", che hanno promosso temi di ricerca rimasti "classici" per il settore, come la traduzione automatica, i sistemi di dialogo uomo-macchina ecc..
Il TAL si è sviluppato alla confluenza di queste due tradizioni promuovendo il faticoso superamento di alcune forti dicotomie che hanno caratterizzato le anime della linguistica computazionale ai suoi esordi, dicotomie riassumibili proprio in diverse, e a tratti ortogonali, concezioni della lingua e dei metodi per le sue elaborazioni computazionali. Da un lato, la lingua, come prodotto complesso e dinamico realizzato nella variabilità delle sue tipologie testuali, si è a lungo opposta alla lingua in *vitro* di esperimenti da laboratorio troppo spesso decontestualizzati e riduttivi rispetto alle sue reali forme e usi. A questo bisogna unire anche la prevalenza dei metodi statistici per lo studio delle regolarità distribuzionali delle pa-

role tipico di molta linguistica matematica applicata al testo, in forte contrasto col prevalere di tecniche simboliche che hanno costituito, per lungo tempo, il modello dominante per la progettazione dei primi algoritmi per il TAL. Il superamento di tale dicotomia è stato reso possibile grazie al radicale mutamento di paradigma avvenuto nel TAL, a partire dalla seconda metà degli anni '80, caratterizzato dal diffondersi, e poi dal netto prevalere, di un'epistemologia neo-empirista. Questo cambiamento si è concretizzato nella diffusione dei metodi statistico-quantitativi per l'analisi computazionale del linguaggio [19], e nella rinnovata centralità dei dati linguistici.

La disponibilità crescente di *risorse linguistiche*, in particolari *corpora* testuali e lessici computazionali, ha costituito un fattore determinante in questa svolta metodologica e tecnologica nel TAL. La disponibilità di *corpora* di grandi dimensioni è diventata una variabile fondamentale in ogni fase di sviluppo e valutazione degli strumenti per l'elaborazione dell'informazione linguistica. Gli strumenti per il TAL sono, infatti, ora chiamati a confrontarsi non con pseudolinguaggi di laboratorio, ma con testi di grande complessità e variabilità linguistica e strutturale. A sua volta, questo ha portato al diffondersi di tecniche di elaborazione linguistica più "robuste" di quelle simboliche tradizionali, in grado di affrontare la variabilità lessicale e strutturale del linguaggio, e anche quel suo continuo resistere ai vincoli grammaticali che è così evidente in molte sue manifestazioni, prima fra tutte, la lingua parlata. La possibilità di accedere a quantità sempre crescenti di dati linguistici digitali ha indubbiamente facilitato tale innovazione metodologica, fornendo i dati linguistici necessari per un uso intensivo dei metodi statistici, che hanno incominciato a ibridare le architetture e gli algoritmi più tradizionali. Un ulteriore fattore di accelerazione è stato fornito dalla necessità della tecnologia della lingua di passare da prototipi di laboratorio a sistemi funzionanti in grado di offrirsi agli utenti come affidabili strumenti per la gestione dell'informazione linguistica. Il banco di prova del *World Wide Web*, per sua natura risorsa di informazione documentale multiforme e magmatica, ha imposto ai sistemi per il TAL di acquisire una capacità di adeguarsi alle complessità della lingua reale, prima impensabile.

3. DAL TESTO ALLA CONOSCENZA

All'interno dell'ampio spettro di attività del TAL, che coinvolgono quasi tutti i domini dell'*Information Technology*[1], di particolare interesse e impatto sono le possibilità offerte dalle più recenti tecnologie della lingua per *trasformare i documenti testuali in risorse di informazione e conoscenza*. Alla base di questo processo di accesso e analisi del contenuto digitale risiedono tre tipi di tecnologie, fondamentali per ogni sistema basato sul TAL:

1. *strumenti per l'analisi linguistica di testi e l'acquisizione dinamica di conoscenza* – analizzatori morfologici, parser sintattici[2], acquisitori automatici di terminologia e informazione semantica dai testi ecc.;

2. *risorse linguistiche* – lessici computazionali, reti semantico-concettuali multilingui, *corpora* testuali anche annotati sintatticamente e semanticamente per lo sviluppo e la valutazione di tecnologia del linguaggio;

3. *modelli e standard per la rappresentazione dell'informazione linguistica* – ontologie per il *knowledge sharing* e la codifica lessicale, modelli per la rappresentazione e interscambio di dati linguistici.

Grazie anche alle nuove opportunità offerte dalla tecnologia XML (*eXtensible Markup Language*) è possibile realizzare una maggiore integrazione tra i diversi moduli per l'elaborazione della lingua, e la standardizzazione della rappresentazione dei dati, necessaria per assicurare la loro interscambiabilità

[1] Questi vanno dal riconoscimento automatico del parlato alla traduzione automatica, dallo sviluppo di interfacce uomo-macchina multimodali ai sistemi di *question-answering* che permettono di interrogare una base documentale formulando la richiesta come una domanda in linguaggio naturale. Un'ampia rassegna delle varie applicazioni del TAL è disponibile in [11, 13].

[2] *Il parsing* è il processo di analisi linguistica attraverso cui viene ricostruita la struttura sintattica di una frase, rappresentata dall'articolazione dei costituenti sintagmatici e dalle relazioni di dipendenza grammaticale (esempio soggetto, complemento oggetto ecc.).

e la coerenza del trattamento dell'informazione. Strumenti di analisi, risorse linguistiche e standard di rappresentazione vengono, dunque, a costituire un'infrastruttura per il TAL che attraverso l'analisi linguistica automatica dei documenti testuali permette di estrarre la conoscenza implicitamente contenuta in essi, trasformandola in conoscenza esplicita, strutturata e accessibile sia da parte dell'utente umano che da parte di altri agenti computazionali (Figura 1).
È importante sottolineare l'aspetto di stretta interdipendenza tra i vari componenti per il TAL, illustrata in maggior dettaglio in figura 2. Gli strumenti di analisi linguistica costruiscono una rappresentazione avanzata del contenuto informativo dei documenti attraverso elaborazioni del testo a vari livelli di complessità: analisi morfologica e lemmatizzazione, analisi sintattica, interpretazione e disambiguazione semantica ecc.. I moduli di elaborazione sono solitamente interfacciati con *database* linguistici, che rappresentano e codificano grandi quantità di informazione terminologica e lessicale, morfologica, sintattica e semantica, che ne permettono sofisticate modalità di analisi. Le analisi linguistiche forniscono l'*input* per i moduli di estrazione, acquisizione e strutturazione di conoscenza. La conoscenza estratta costituisce una risorsa per l'utente finale, e permette allo stesso di popolare ed estendere i repertori linguistico-lessicali e terminologici che sono usati in fase di analisi dei documenti. Si realizza, così, un ciclo virtuoso tra strumenti per il TAL e risorse linguistiche. Le risorse linguistiche lessicali e testuali permettono di costruire, ampliare, rendere operativi, valutare modelli, algoritmi, componenti e sistemi per il TAL, sistemi che sono, a loro volta, strumenti necessari per alimentare dinamicamente ed estendere tali risorse.
Un esempio di architettura per il trattamento automatico dell'Italiano è *Italian NLP*, sviluppato dall'Istituto di Linguistica Computazionale – CNR in collaborazione con il Dipartimento di Linguistica – Sezione di Linguistica Computazionale dell'Università di Pisa. *Italian NLP* è un ambiente integrato di strumenti e risorse che consentono di effettuare analisi linguistiche incrementali dei

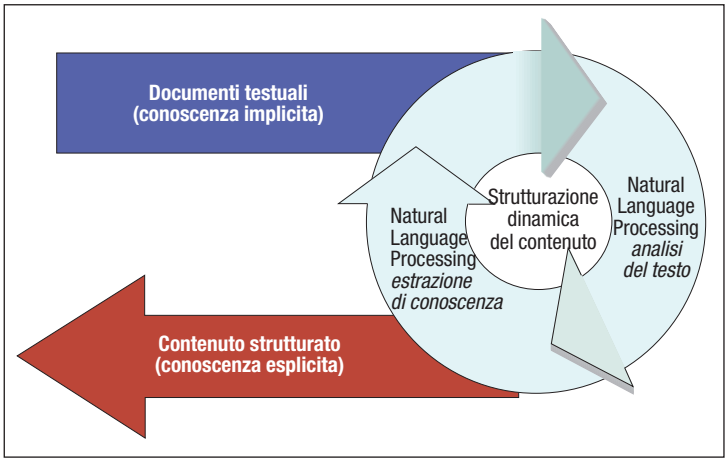


FIGURA 1
Dalla conoscenza implicita alla conoscenza esplicita

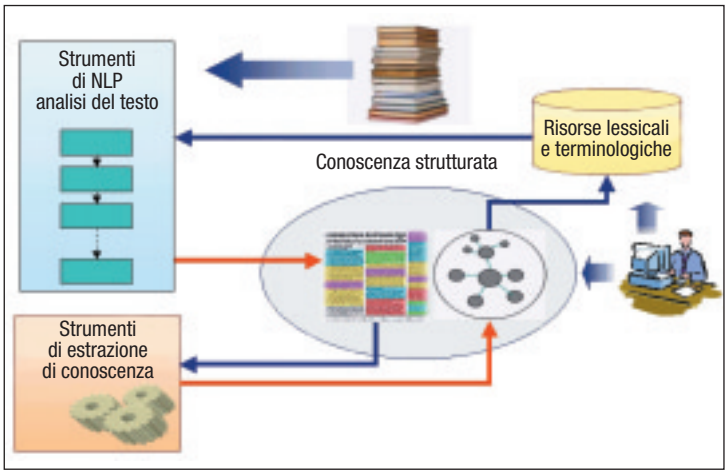


FIGURA 2
Un'architettura per l'estrazione di conoscenza dai testi basata sul TAL

testi. Ciascun modulo di *Italian NLP* procede all'identificazione di vari tipi di unità linguistiche di complessità strutturale crescente, ma anche utilizzabili singolarmente come fonte di informazione sull'organizzazione linguistica dei testi.
Come si vede in figura 3, un aspetto significativo di *Italian NLP* è il carattere ibrido della sua architettura. Moduli simbolici di *parsing* (basati su metodologie consolidate nella linguistica computazionale, come le tecnologie a stati finiti) sono affiancati a strumenti statistici che sono usati per operare disambiguazioni sintattiche e semantiche, filtrare "rumore" dalle analisi e anche arricchire le risorse lessicali con informazioni direttamente estratte dai testi oggetto di analisi, permettendo l'aggiornamento e specializzazione continua delle risorse linguistiche, e garantendo una maggiore robustezza e portabilità

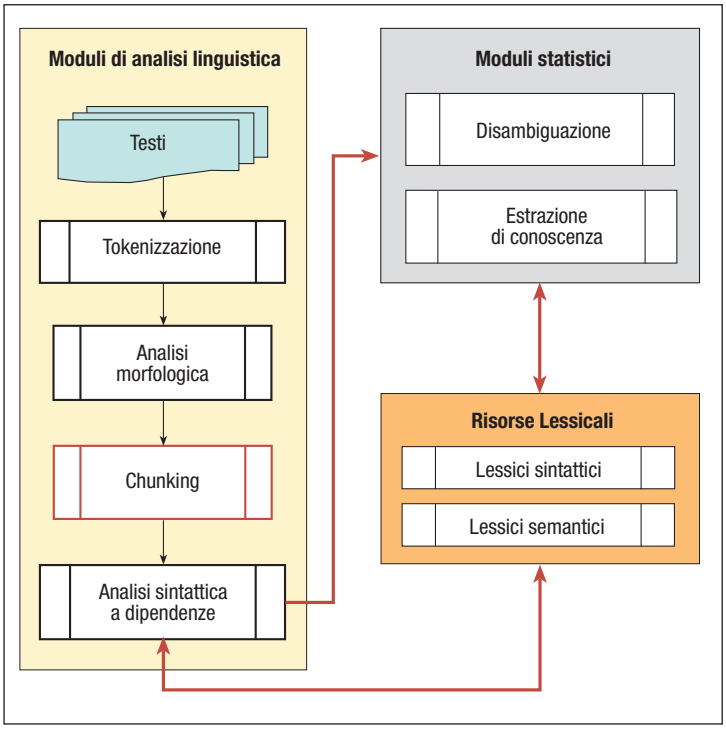


FIGURA 3
Strumenti di analisi e risorse linguistiche in Italian NLP

degli strumenti di analisi del linguaggio su domini e registri linguistici diversi.

Uno dei livelli di analisi linguistica più impegnativi è l'analisi sintattica automatica. In *Italian NLP*, questa è realizzata in due fasi successive. Dopo un processo di tokenizzazione[3] e analisi morfologica, viene effettuato un *parsing* "leggero" del testo (*shallow parsing*), in cui un *chunker* realizza contemporaneamente la disambiguazione morfosintattica delle parole, cioè l'identificazione della categoria sintattica con cui una forma occorre in un dato contesto linguistico, e la segmentazione del testo in sequenze di gruppi sintattici non ricorsivi (*chunk*) di cui vengono individuati il tipo (nominale, verbale ecc.) e la testa lessicale [15]. Per esempio, la frase "Il Presidente della Repubblica ha visitato la capitale della Francia" viene segmentata dal chunker nel modo seguente[4]:

$[_{N\_C}$ *Il Presidente*$][_{P\_C}$ *della Repubblica*$][_{FV\_C}$ *ha visitato*$][_{N\_C}$ *la capitale*$][_{P\_C}$ *della Francia*$]$

Come risultato del *chunking*, si ottiene dunque una strutturazione del testo in unità linguisticamente rilevanti sia per processi di estrazione dell'informazione e text mining, sia come input per la seconda fase di parsing in cui il testo segmentato è analizzato a livello sintattico-funzionale, per identificare relazioni grammaticali tra gli elementi nella frase come soggetto, oggetto, complemento, modificatore ecc.. In *Italian NLP* questo tipo di analisi è realizzato da IDEAL, *Italian DEpendency AnaLyzer* [1, 2], un compilatore di grammatiche a stati finiti definite su sequenze di chunk. Le regole della grammatica fanno uso di test sulle informazioni associate ai chunk (per esempio, informazioni morfosintattiche, tratti di accordo) e su informazioni lessicali esterne (il lessico che viene usato a questo fine comprende circa venticinquemila *frame sintattici di sottocategorizzazione*)[5]. L'output di IDEAL è costituito da relazioni grammaticali binarie tra una testa lessicale e un suo dipendente che forniscono una rappresentazione della struttura sintattica come la seguente[6]:

sogg	(visitare, presidente)
comp	(presidente, repubblica.<intro=di>)
ogg	(visitare, capitale)
comp	(capitale, Francia.<intro=di>)

Simili rappresentazioni della struttura linguistica del testo forniscono l'input fondamentale per processi di estrazione della conoscenza. Un esempio di applicazione di questo tipo è l'acquisizione semi-automatica di on-

[3] La *tokenizzazione* consiste nella segmentazione del testo in unità minime di analisi (parole). In questa fase l'input è sottoposto a un processo di normalizzazione ortografica (esempio separazione di virgolette e parentesi della parole, riconoscimento dei punti di fine frase ecc.), nell'ambito del quale vengono anche identificate le sigle, gli acronimi e le date.

[4] N_C, P_C e FV_C stanno rispettivamente per chunk di tipo nominale, preposizionale e verbale

[5] Un f*rame di sottocategorizzazione* specifica il numero e tipo di complementi che sono selezionati da un termine lessicale. Per esempio, il verbo *mangiare* seleziona per un complemento oggetto opzionale (cfr. *Gianni ha mangiato un panino; Gianni ha mangiato*), mentre il verbo *dormire*, in quanto intransitivo, non può occorrere con un complemento oggetto.

[6] sogg = soggetto; comp = complemento; ogg = oggetto diretto

tologie (*ontology learning*) da testi come supporto avanzato alla gestione documentale [8, 18]. Un'*ontologia* [9, 22] è un sistema strutturato di concetti e relazioni tra concetti che viene a costituire una "mappa" della conoscenza di un certo dominio od organizzazione. Gli strumenti e le risorse del TAL permettono di trasformare le conoscenze implicitamente codificate all'interno dei documenti testuali in conoscenza esplicitamente strutturata come un'ontologia di concetti. Attraverso il TAL è possibile, dunque, dotare i sistemi informatici di una chiave di accesso semantica alle basi documentali, consentendo agli utenti di organizzare e ricercare i documenti su base concettuale, e non solo attraverso l'uso di parole chiave. Le ontologie estratte dinamicamente dai testi vengono a costituire un ponte tra il bisogno di informazione degli utenti - rappresentato da idee, concetti o temi di interesse - e i documenti in cui l'informazione ricercata rimane nascosta all'interno dell'organizzazione linguistica del testo, che spesso ne ostacola il recupero. Anche in un linguaggio tecnico e apparentemente controllato, infatti, lo stesso concetto può essere espresso con una grande variazione di termini, e la scelta di uno di questi da parte dell'utente in fase di ricerca o indicizzazione, può impedire il recupero di documenti ugualmente rilevanti, ma in cui lo stesso concetto appare sotto forme linguistiche diverse. Le tecnologie della lingua rendono possibile lo sviluppo di *un ambiente per la creazione dinamica di ontologie a partire dall'analisi linguistica dei documenti*. Diventa così possibile velocizzare il processo di gestione dell'indicizzazione e della classificazione della base documentale, e ridurre il grado di arbitrarietà dei criteri di classificazione. La questione è, infatti, come fare a determinare i concetti rilevanti e più caratterizzanti per i documenti di un certo dominio di interesse. Per affrontare questo problema le tecniche linguistico-computazionali si basano su un'ipotesi molto semplice: i documenti sono estremamente ricchi di termini che con buona approssimazione veicolano i concetti e i temi rilevanti nel testo. Termini sono nomi propri, nomi semplici come *museo* o *pinacoteca,* oppure gruppi nominali strutturalmente complessi come *museo archeologico*, *ministero dei beni culturali, soprintendenza archeologica* ecc.. I termini possono essere a loro volta raggruppati, in quanto esprimono concetti molto simili. Per esempio, *scultura, affresco* e *quadro* condividono tutti un concetto più generico di "opera artistica" a cui possono essere ricondotti a un certo grado di astrazione. Attraverso l'uso combinato di tecniche statistiche e di strumenti avanzati per l'analisi linguistica come quelli di *Italian NLP* è possibile analizzare il contenuto linguistico dei documenti appartenenti a un dato dominio di conoscenza, individuare i termini potenzialmente più significativi e ricostruire una "mappa" dei concetti espressi da questi termini, ovvero costruire un'ontologia per il dominio di interesse. Come si vede nella figura 4, alla base dell'ontologia risiede un glossario di termini (semplici e complessi) estratti dai testi dopo una fase di analisi linguistica, effettuata con moduli di parsing. I termini estratti vengono successivamente filtrati con criteri statistici per selezionarne i più utili per caratterizzare una certa collezione di documenti. I termini sono organizzati e strutturati come in un Thesaurus di tipo classico, sulla base di alcune relazioni semantiche di base. L'ontologia viene, dunque, a essere composta di unità concettuali definite come insiemi di termini semanticamente affini. I concetti possono, inoltre, essere organizzati secondo la loro maggiore o minore specificità articolando l'ontologia come una tassonomia. Dal momento che un sistema di conoscenza non è fatto solo di concetti che si riferiscono a entità del dominio, ma anche di processi, azioni ed eventi che vedono coinvolte queste entità secondo ruoli e funzioni diverse, uno stadio più avanzato di estrazione può puntare anche all'identificazione di relazioni non tassonomiche tra concetti (per esempio, la funzione tipica di una certa entità, la sua locazione ecc.). È importante sottolineare che il processo di *ontology learning* attraverso l'analisi linguistica dei documenti avviene generalmente in stretta cooperazione con gli utenti, che sono chiamati a intervenire nelle varie fasi di estrazione della conoscenza per validarne i risultati. Come in altri settori di applicazione del TAL, anche in questo caso le tecnologie della lingua utilmente contribuiscono alla gestione dei contenuti di

informazione a *supporto* dell'esperto umano, senza pretendere di sostituirsi ad esso.
Gli strumenti di *Italian NLP* sono usati in molteplici contesti applicativi, in cui hanno dimostrato l'ampiezza e rilevanza delle opportunità pratiche offerte dal TAL. Tra gli esempi più significativi a livello nazionale è possibile citare i moduli linguistico-computazionali SALEM (*Semantic Annotation for LEgal Management*) [3] - sviluppato nell'ambito del progetto *Norme in Rete* (NIR) del Centro Nazionale per l'Informatica nella Pubblica Amministrazione (CNIPA) - e T2K (Text-2-Knowledge) - realizzato nell'ambito del progetto *TRAGUARDI* del Dipartimento della Funzione Pubblica - FORMEZ[7]. SALEM è un modulo per l'annotazione automatica della struttura logica dei documenti legislativi, integrato nell'editore normativo *NIREditor*, sviluppato dall'Istituto di Teoria e Tecnica dell'Informazione Giuridica - CNR. Attraverso l'analisi computazionale del testo, SALEM rende espliciti gli aspetti più rilevanti del contenuto normativo, individuando elementi quali il destinatario della norma, la sanzione prevista ecc.. Questi elementi di contenuto sono annotati esplicitamente sul testo con metadati XML, garantendo una migliore gestione e ricerca della documentazione legislativa. Il modulo T2K è, invece, finalizzato alla costruzione semi-automatica di thesauri di termini e di ontologie di metadati semantici per la gestione documentale nella pubblica amministrazione. A livello internazionale, gli strumenti per il TAL, illustrati sopra, sono stati applicati in numerosi progetti finanziati dall'Unione Europea, tra i quali si vogliono qui citare POESIA (*Public Open-source Environment for a Safer Internet Access*) [10], dedicato alla creazione di sistemi avanzati di *filtering* di siti web, e VIKEF (*Virtual Information and Knowledge Environment Framework*)[8], in cui gli strumenti di *Italian NLP* sono utilizzati per l'annotazione semantica di testi e la costruzione di ontologie, nell'ambito delle iniziative relative al Semantic Web. Questi sono solo alcuni dei numerosi esempi di progetti e iniziative in cui i prodotti del TAL la-

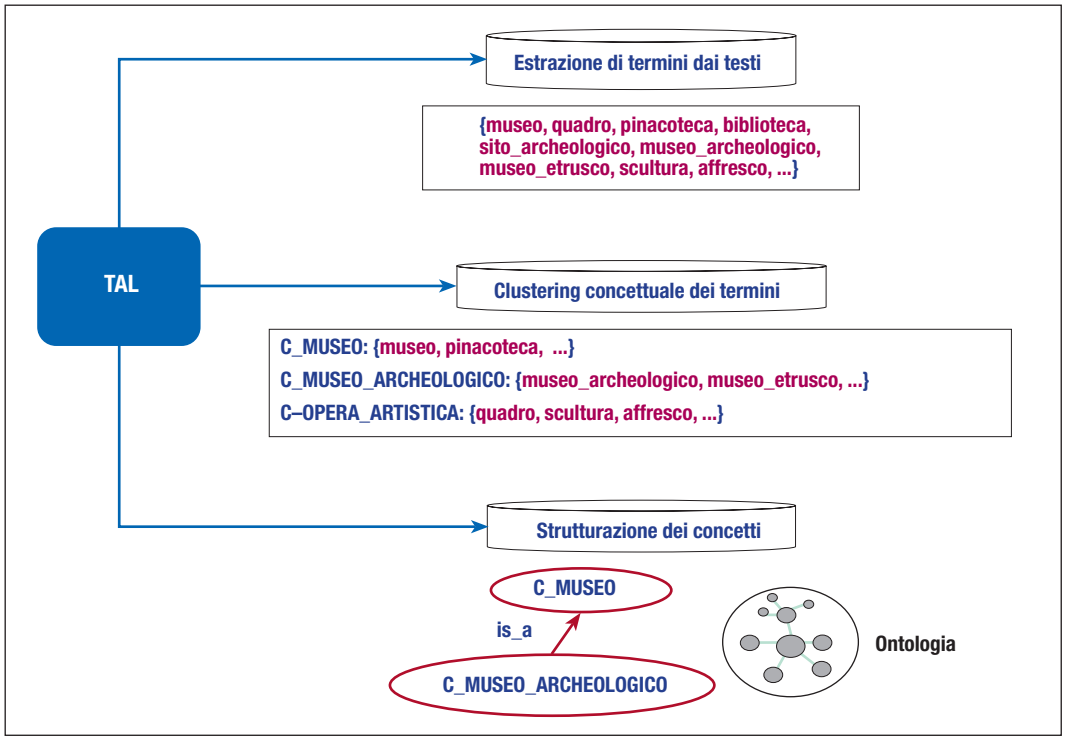


FIGURA 4
TAL e ontology learning

[7] Il progetto TRAGUARDI, di cui è responsabile la dott.ssa Anna Gammaldi di FORMEZ, è un'azione di sostegno alle pubbliche amministrazioni per la gestione dei fondi strutturali.

[8] http://www.vikef.net

sciano i centri di ricerca per entrare a diretto contatto con l'utenza e il mercato. Inoltre è importante notare come i contesti applicativi riguardino tipologie di testi completamente diverse, che vanno dai documenti legislativi alla documentazione della pubblica amministrazione, fino al linguaggio dei siti web. Questo testimonia la versatilità della ricerca attuale sul TAL nella sua capacità di affrontare il linguaggio naturale nella complessità delle sue più diverse e varie manifestazioni.

4. RISORSE LESSICALI PER IL TAL

Gli strumenti e le applicazioni del TAL hanno bisogno di poter interpretare il significato delle parole, porta di accesso al contenuto di conoscenza codificato nei documenti. I *lessici computazionali* hanno lo scopo di fornire una rappresentazione esplicita del significato delle parole in modo tale da poter essere direttamente utilizzato da parte di agenti computazionali, come, per esempio, parser, moduli per Information Extraction ecc.. I lessici computazionali multilingui aggiungono alla rappresentazione del significato di una parola le informazioni necessarie per stabilire delle connessioni tra parole di lingue diverse.
Nell'ultimo decennio numerose attività hanno contribuito alla creazione di lessici computazionali di grandi dimensioni. All'esempio più noto, la rete semantico-concettuale WordNet [7] sviluppata all'università di Princeton, si sono affiancati anche altri repertori di informazione lessicale, come PAROLE [21], SIMPLE [14] e EuroWordNet [23] in Europa, Comlex e FrameNet negli Stati Uniti, ecc.. Per quanto riguarda l'italiano, è importante citare i lessici computazionali ItalWordNet e CLIPS, entrambi sviluppati nell'ambito di due progetti nazionali finanziati dal MIUR e coordinati da Antonio Zampolli.
ItalWordNet è una rete semantico-lessicale per l'italiano, strutturata secondo il modello di WordNet e consiste in circa 50.000 entrate. Queste sono costituite da uno o più sensi raggruppati in *synset* (gruppi di sensi sinonimi tra loro). I synset sono collegati tra loro principalmente da relazioni di *iperonimia*[9], che permettono di strutturare il lessico in gerarchie tassonomiche. I nodi più alti delle tassonomie sono a loro volta collegati agli elementi di una ontologia (*Top Ontology*), indipendente da lingue specifiche, che ha la funzione di organizzare il lessico in classi semantiche molto generali. Infine, ogni *synset* della rete è collegato, tramite una relazione di equivalenza, a synset del WordNet americano. Questo collegamento costituisce l'indice interlingue (*Interlingual Index* – ILI) e attraverso di esso ItalWordNet viene a essere integrata nella famiglia di reti semantiche sviluppata nel progetto europeo EuroWordNet, diventando così una vera e propria risorse lessicale multilingue (Figura 5). L'ILI è anche collegato alla *Domain Ontology*, che contiene un'ontologia di domini semantici. Oltre all'iperonimia, il modello ItalWordNet comprende anche una grande varietà di altre *relazioni semantiche* che, collegando sensi di lemmi anche appartenenti a categorie morfosintattiche differenti, permettono di evidenziare diverse relazioni di significato, operanti sia a livello paradigmatico sia a livello sintagmatico.
Il progetto SIMPLE (*Semantic Information for Multipurpose Plurilingual LExica*) ha portato alla definizione di un'architettura per lo sviluppo di lessici computazionali semantici e alla costruzione di lessici computazionali per 12 lingue europee (Catalano, Danese, Finlandese, Francese, Greco, Inglese, Italiano, Olandese, Portoghese, Spagnolo, Svedese, Tedesco). I lessici di SIMPLE rappresentano un contributo estremamente innovativo nel settore delle risorse lessicali per il TAL, offrendo una rappresentazione articolata e multidimensionale del contenuto semantico dei termini lessicali. Il modello di rappresentazione semantica di SIMPLE è stato usato anche per la costruzione di CLIPS, che include 55.000 entrate lessicali con informazione fonologica, morfologica, sintattica e semantica.
Il modello SIMPLE costituisce un'architettura

[9] Un termine lessicale *x* è un *iperonimo* di un termine lessicale *y* se, e solo se, *y* denota un sottoinsieme delle entità denotate da *x*. Per esempio, *animale* è un iperonimo di *cane*. La relazione simmetrica è quella di *iponimia*, per cui *cane* è un iponimo di *animale*.

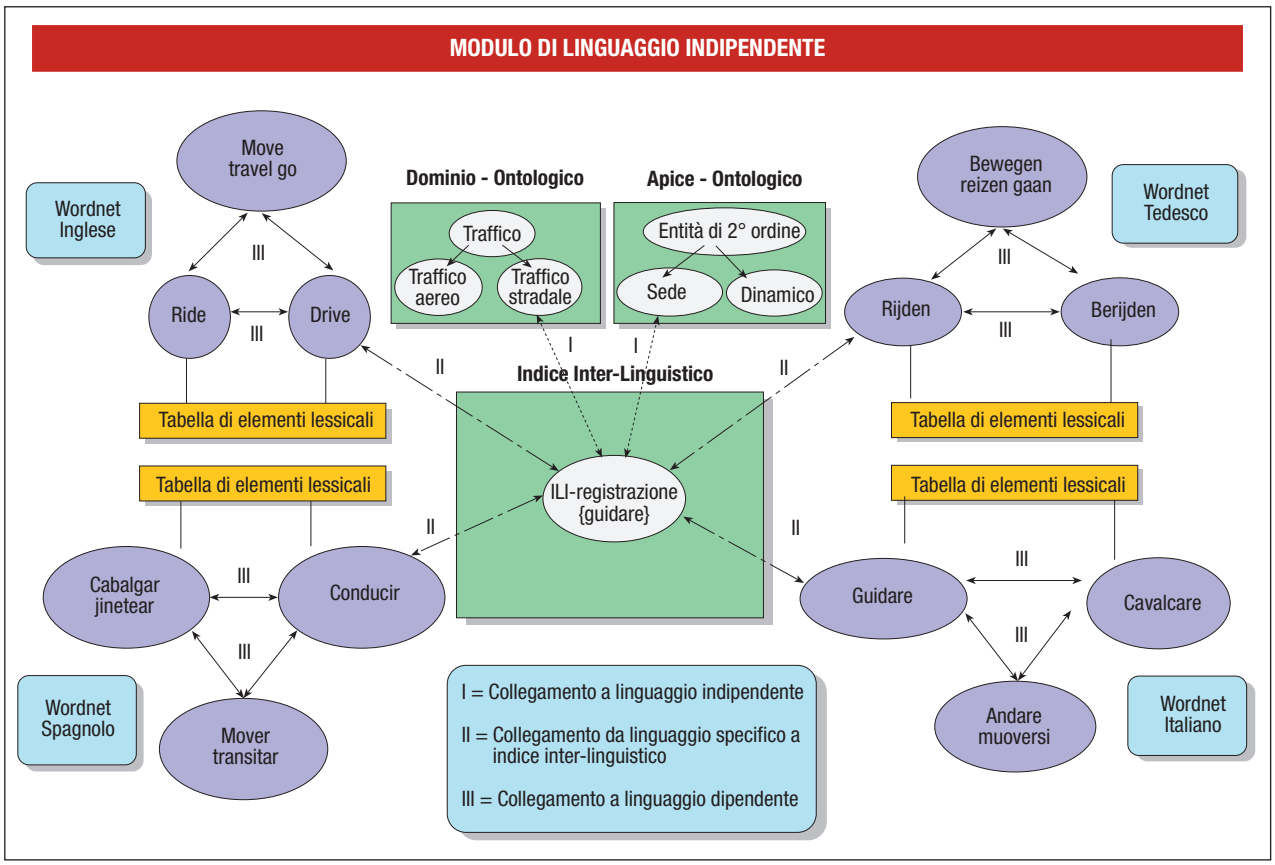


FIGURA 5

L'architettura di EuroWordNet

TABELLA 1

Un frammento della Core Ontology di SIMPLE

1. TELIC
2. AGENTIVE
2.1. *Cause*
3. CONSTITUTIVE
3.1. *Part*
3.1.1. Body_part
3.2. *Group*
3.2.1. Human_group
3.3. *Amount*
4. ENTITY
4.1. *Concrete_entity*
4.1.1. Location
...

per lo sviluppo di lessici computazionali nel quale il contenuto semantico è rappresentato da una combinazione di diversi tipi di entità formali [14] con i quali si cerca di catturare la multidimensionalità del significato di una parola. In tal modo, SIMPLE tenta di fornire una risposta a importanti questioni che coinvolgono la costruzione di ontologie di tipi lessicali, facendo emergere allo stesso tempo problemi cruciali relativi alla rappresentazione della conoscenza lessicale. Al cuore del modello SIMPLE è possibile trovare un repertorio di *tipi semantici* di base e un insieme di informazioni semantiche che devono essere codificate per ciascun senso. Tali informazioni sono organizzate in *template*, ovvero strutture schematiche che rappresentano formalmente l'articolazione interna di ogni tipo semantico, specificando così vincoli semantico-strutturali per gli oggetti lessicali appartenenti a quel tipo. I tipi semantici formano la *Core Ontology* di SIMPLE (Tabella 1), uno dei cui modelli ispiratori è la *Struttura Qualia* definita nella teoria del Lessico Generativo [5, 20]. I tipi semantici sono, infatti, organizzati secondo principi ortogonali, quali la funzione tipica delle entità, la loro origine o costituzione mereologica ecc., nel tentativo di superare i limiti quelle ontologie che troppo spesso ap-

Lemma:	Violino
SEMU_ID:	#V1
POS:	N
GLOSS:	Tipo di strumento musicale
DOMAIN:	**MUSIC**
SEMANTIC_TYPE:	**Instrument**
FORMAL_ROLE:	**Isa** *strumento_musicale*
CONSTITUTIVE_ROLE:	**Has_as_part** *corda*
	Made_of *legno*
TELIC_ROLE:	**Used_by** *violinista*
	Used_for *suonare*
Lmma:	**Guardare**
SEMU_ID:	#G1
POS:	V
GLOSS:	Rivolgere lo sguardo verso qualcosa per osservarlo
SEMANTIC_TYPE:	**Perception**
EVENT_TYPE	**Process**
FORMAL_ROLE:	**Isa** *percepire*
CONSTITUTIVE_ROLE:	**Instrument** *occhio*
	Intentionality = yes
PRED_REPRESENTATION:	Guardare (**Arg0: aniùate**) (**Arg1: entity**)
SYN_SEM_LINKING:	**Arg0 = subj_NP**
	Arg1 = obj NP

TABELLA 2
Entrate lessicali di SIMPLE per violino e guardare

piattiscono la ricchezza concettuale sulla sola dimensione tassonomica.
Il modello di SIMPLE fornisce le specifiche per la rappresentazione e la codifica di un'ampia tipologia di informazioni lessicali, tra le quali il tipo semantico, l'informazione sul dominio, la struttura argomentale per i termini predicativi, le preferenze di selezione sugli argomenti, informazione sul comportamento azionale e aspettuale dei termini verbali, il collegamento delle strutture predicative semantiche ai *frame* di sottocategorizzazione codificati nel lessico sintattico di PAROLE, informazioni sulle relazioni di derivazione tra parole appartenenti a parti del discorso diverse (per esempio, *intelligente – intelligenza; scrittore – scrivere* ecc.). In SIMPLE, i sensi delle parole sono codificati come *Unità Semantiche o SemU*. Ad ogni *SemU* viene assegnato un tipo semantico dall'ontologia, più altri tipi di informazioni specificate nel *template* associato a ciascun tipo semantico. La tabella 2 fornisce una rappresentazione schematica di due entrate lessicali (per il nome *violino* e il verbo *guardare*) codificate secondo le specifiche del modello SIMPLE. Il potere espressivo di SIMPLE è costituito da un ampio insieme di relazioni organizzate lungo le quattro dimensioni della Struttura Qualia proposta nel Lessico Generativo come assi principali della descrizione lessicale, cioè *Formal Role, Constitutive Role, Agentive Role e Telic Role*. Le dimensioni Qualia vengono usate per cogliere aspetti diversi e multiformi del significato di una parola. Per esempio il Telic Role riguarda la funzione tipica di un'entità o l'attività caratteristica di una categoria di individui (esempio, la funzione prototipica di un professore è insegnare). L'Agentive Role riguarda, invece, il modo in cui un'entità è creata (esempio, naturalmente o artificialmente), mentre il Constitutive Role rappresenta la composizione o struttura interna di un'entità (per esempio, le sue parti o il materiale di cui è composta). In SIMPLE, è possibile discriminare fra i vari sensi delle parole calibrando l'uso dei diversi tipi di informazione resi disponibili dal modello. Per esempio, la figura 6 mostra una possibile caratte-

rizzazione di una porzione di spazio semantico associato alla parola "ala" il cui contenuto può essere articolato in quattro SemU che hanno in comune lo stesso tipo semantico (ovvero *PART*), ma che possono comunque essere distinte attraverso le relazioni che esse hanno con altre unità semantiche. Per esempio, se da una parte la SemU_1 e la SemU_3 sono simili per quanto concerne la dimensione della funzionalità (entrambe si riferiscono a entità usate per volare), sono distinte per quanto riguarda gli aspetti costitutivi, poiché la SemU_1 si riferisce a una parte di un aereo e la SemU_3 alla parte di un uccello ecc.. Nonostante si sia ancora lontani dal poter fornire rappresentazioni veramente soddisfacenti del contenuto di una parola, l'architettura di SIMPLE tenta di avvicinarsi alla complessità del linguaggio naturale fornendo un modello altamente espressivo e versatile per descrivere il contenuto linguistico.

I lessici computazionali devono essere concepiti come sistemi dinamici il cui sviluppo si integra strettamente con processi di acquisizione automatica di informazione dai testi. Dal momento che i significati delle parole vivono, crescono e mutano nei contesti linguistici in cui occorrono, la loro rappresentazione nei repertori lessicali deve tenere necessariamente in considerazione le modalità con le quali l'informazione lessicale emerge dal materiale testuale e come quest'ultimo contribuisce alla creazione e alla variazione del significato. Conseguentemente, i lessici computazionali – anche di grandi dimensioni – non possono essere mai concepiti come repertori statici e chiusi. Al contrario, i lessici computazionali sono in grado al più di fornire nuclei di descrizione semantica che comunque devono essere costantemente personalizzati, estesi e adattati a diversi domini, applicazioni, tipologie di testo ecc.. In questo senso, il processo di creazione di risorse semantico-lessicali si deve accompagnare allo sviluppo di strumenti e metodologie per il *lexical tuning*, ovvero per l'adattamento dell'informazione semantica ai concreti contesti d'uso [4]. Questa sembra essere una condizione essenziale affinché le risorse linguistiche possano diventare strumenti versatili e adattativi per l'elaborazione del contenuto semantico dei documenti.

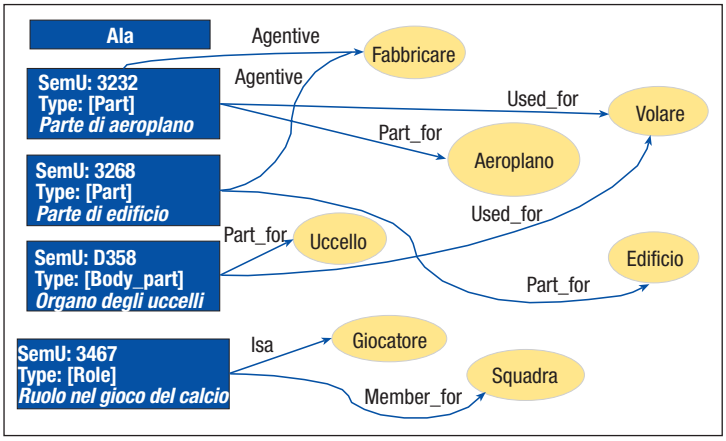


FIGURA 6

Rappresentazione dei significati di ala in SIMPLE

Gli strumenti per affrontare questo problema vengono dalla ricerca sull'*acquisizione automatica della conoscenza* e, più in generale, dall'uso di tecniche di *apprendimento automatico*, sia supervisionato che non supervisionato. Molti di questi metodi sono basati su un modello distribuzionale del significato, secondo il quale il contenuto semantico di una parola o termine è derivabile dal modo in cui esso si distribuisce linguisticamente, ovvero dall'insieme dei contesti in cui è usato [16]. Secondo questo approccio, a ciascuna parola di un testo viene associata una rappresentazione in forma di vettore distribuzionale. Le dimensioni dei vettore sono date dalle dipendenze grammaticali del termine con altri termini lessicali (verbi, nomi, aggettivi ecc.) nei documenti, oppure più semplicemente dalle parole che occorrono con il termine all'interno di una certa finestra di contesto. I vettori distribuzionali vengono generalmente estratti dai testi in maniera automatica con gli strumenti del TAL. Attraverso l'applicazione di algoritmi di *clustering* alle rappresentazioni vettoriali è possibile ricostruire spazi di similarità semantica tra i termini, ovvero classi di termini o parole semanticamente simili [17]. Infatti, il grado di similarità semantica tra due termini è proporzionale al grado di similarità della loro distribuzione grammaticale nei testi. In questo modo, è possibile arricchire ed estendere le risorse lessicali con nuove informazioni sul comportamento semantico delle parole e che direttamente rispecchiano il loro l'uso nei testi. Nuovi sensi o usi specifici di un certo domi-

nio o registro linguistico sono, quindi, derivabili automaticamente attraverso l'uso combinato del TAL e di algoritmi di apprendimento. Una maggiore comprensione dei problemi riguardanti le profonde interrelazioni tra rappresentazione e acquisizione del significato dei termini lessicali potrebbe avere importanti ripercussioni su come le risorse linguistiche verranno in futuro costruite, sviluppate e usate per le applicazioni.

5. STANDARD PER LE RISORSE LINGUISTICHE

Un altro aspetto di fondamentale importanza per il ruolo delle risorse lessicali (e più in generale linguistiche) nel TAL è come ottimizzare la produzione, mantenimento e interscambio tra le risorse linguistiche, così come il processo che porta alla loro integrazione nelle applicazioni. La precondizione essenziale per raggiungere questi risultati è stabilire una struttura comune e standardizzata per la costruzione dei lessici computazionali che possa garantire la codifica dell'informazione linguistica in maniera tale da assicurare la sua riutilizzazione da parte di applicazioni diverse e per compiti diversi. In questo modo, si può rafforzare la condivisione e la riusabilità delle risorse lessicali multilingui promuovendo la definizione di un linguaggio comune per la comunità degli sviluppatori e utilizzatori di lessici computazionali. Un'importante iniziativa internazionale in questa direzione è stata rappresentata dal progetto ISLE (*International Standards for Language Engineering*) [6], continuazione di EAGLES (*Expert Advisory Group for Language Engineering Standards*), ambedue ideati e coordinati da Antonio Zampolli. ISLE è stato congiuntamente finanziato dall'Unione Europea e dal *National Science Foundation* (NSF) negli USA e ha avuto come obiettivo la definizione di una serie di standard e raccomandazioni in tre aree cruciali per le tecnologie della lingua:

1. lessici computazionali multilingui,
2. interattività naturale e multimedialità,
3. valutazione.

Per quanto riguarda il primo tema, Il *Computational Lexicon Working Group* (CLWG) di ISLE si è occupato di definire consensualmente un'infrastruttura standardizzata per lo sviluppo di risorse lessicali multilingui per le applicazioni del TAL, con particolare riferimento alle specifiche necessità dei sistemi di traduzione automatica e di *Crosslingual Information Retrieval*. Nel corso della sua attività, ISLE ha fatto suo il principio metodologico secondo il quale il processo di standardizzazione, nonostante per sua natura non sia intrinsecamente innovativo, deve comunque procedere a stretto contatto con la ricerca più avanzata. Il processo di standardizzazione portato avanti da ISLE ha, infatti, perseguito un duplice obiettivo:

1. la definizione di standard sia a livello di contenuto che di rappresentazione per quegli aspetti dei lessici computazionali che sono già ampiamente usati dalle applicazioni;
2. la formulazione di raccomandazioni per le aree più di "frontiera" della semantica computazionale, ma che possono comunque fornire un elevato contributo di innovazione tecnologica nel settore del TAL.

Come strumento operativo per raggiungere questi obiettivi, il CLWG di ISLE ha elaborato MILE (*Multilingual ISLE Lexical Entry*), un modello generale per la codifica di informazione lessicale multilingue.

MILE è uno schema di entrata lessicale caratterizzata da un'architettura altamente *modulare e stratificata* [6]. La modularità riguarda l'organizzazione "orizzontale" di MILE, nella quale moduli indipendenti ma comunque correlati coprono diverse dimensioni del contenuto lessicale (monolingue, multilingue, semantico, sintattico ecc.). Dall'altro lato, al livello "verticale" MILE ha adottato un'organizzazione stratificata per permettere vari gradi di granularità nelle descrizioni lessicali. Uno degli scopi realizzativi di MILE è stato quello di costruire un ambiente di rappresentazione comune per la costruzioni di risorse lessicali multilingui, allo scopo di massimizzare il riutilizzo, l'integrazione e l'estensione dei lessici computazionali monolingui esistenti, fornendo al tempo stesso agli utilizzatori e sviluppatori di risorse linguistiche una struttura formale per la codifica e l'interscambio dei dati. ISLE ha, dunque, cercato di promuovere la creazione di un'infrastruttura per i lessici computazionali intesi come risorse di dati linguistici aperte e distribuite. In questa prospettiva, MILE agisce come un meta-modello lessicale per facilitare l'intero-

perabilità a livello di contenuto in due direzioni fondamentali: interoperabilità tra risorse linguistiche, per garantire la riusabilità e integrazione dei dati e interoperabilità tra risorse linguistiche e sistemi del TAL che devono accedere ad esse.
Il *ruolo infrastrutturale* delle risorse linguistiche nell'ambito del TAL richiede che esse vengano armonizzate con le risorse di altre lingue, valutate con metodologie riconosciute a livello internazionale, messe a disposizione della intera comunità nazionale, mantenute e aggiornate tenendo conto delle sempre nuove esigenze applicative. All'interno di questo contesto si inserisce, oggi, il disegno, promosso da chi scrive di un "cambiamento di paradigma" nella produzione e uso di una nuova generazione di risorse e strumenti linguistici, concepiti come *Open Linguistic Infrastructure*, attraverso l'utilizzo di metadati e di standard che permettono la condivisione di tecnologie linguistiche sviluppate anche in ambiti diversi, e il loro uso distribuito in rete. Questa nuova concezione è anche determinante per realizzare appieno la visione del *Semantic Web*, ovvero l'evoluzione del web in uno spazio di contenuti effettivamente "comprensibili" dal calcolatore e non solo da utenti umani e con accesso multilingue e multiculturale.

6. CONCLUSIONI E PROSPETTIVE

Una delle priorità a livello nazionale ed europeo è costruire una società basata sulla informazione e sulla conoscenza. La *lingua è veicolo e chiave di accesso alla conoscenza*, e oggi più che mai è urgente la realizzazione di una infrastruttura consolidata di tecnologie linguistiche. Gli sviluppi recenti nel TAL e la crescente diffusione di contenuti digitali mostrano che i tempi sono maturi per una svolta nella capacità di elaborare grandi quantità di documenti testuali al fine di renderli facilmente accessibili e usabili per un'utenza sempre più vasta e composita.
Alcuni temi su cui articolare il TAL per una società della conoscenza sono:
1. *accesso "intelligente" all'informazione multilingue e trattamento del "contenuto" digitale* - è urgente aumentare la disponibilità di strumenti e risorse capaci di automatizzare le operazioni linguistiche necessarie per produrre, organizzare, rappresentare, archiviare, recuperare, elaborare, navigare, acquisire, accedere, visualizzare, filtrare, tradurre, trasmettere, interpretare, utilizzare, in una parola *condividere* la conoscenza;
2. *interattività naturale e interfacce intelligenti* - si devono sviluppare sistemi che agevolino la naturalezza dell'interazione uomo-macchina e aiutare la comunicazione interpersonale mediando l'interazione tra lingue diverse;
3. *il patrimonio culturale e il contenuto digitale* - le tecnologie del TAL favoriscono la crescita dell'industria dei "contenuti", con ampie opportunità per un Paese, come l'Italia, tradizionale produttore di industria culturale:
4. *promozione della ricerca umanistica nella società dell'informazione* - le tecnologie del TAL forniscono nuovi strumenti anche per le scienze umanistiche, facilitando la produzione e fruizione dei contenuti culturali, e evidenziano il contributo potenziale anche delle ricerche umanistiche sul piano delle opportunità economiche e dello sviluppo sociale.
Per realizzare l'obiettivo di un accesso avanzato al contenuto semantico dei documenti è necessario affrontare la complessità del linguaggio naturale. L'attuale esperienza nel TAL dimostra che una tale sfida si può vincere solo adottando un approccio interdisciplinare e creando un ambiente altamente avanzato per l'analisi computazionale della lingua, l'acquisizione di conoscenze attraverso l'elaborazione automatica dei testi e lo sviluppo di una nuova generazione di risorse linguistiche basate sul rappresentazioni avanzate e standardizzate del contenuto lessicale.

Bibliografia

[1] Bartolini R., Lenci A., Montemagni S., Pirrelli V.: *Grammar and Lexicon in the Robust Parsing of Italian: Towards a Non-Naïve Interplay*. Proceedings of the Workshop on Grammar Engineering and Evaluation, COLING 2002 Post-Conference Workshop, Taipei, Taiwan, 2002.

[2] Bartolini R., Lenci A., Montemagni S., Pirrelli V.: *Hybrid Constraints for Robust Parsing: First Experiments and Evaluation*. Proceedings of LREC 2004, Lisbona, Portugal, 2004.

[3] Bartolini R., Lenci A., Montemagni S., Pirrelli, V., Soria C.: Semantic *Mark-up of Italian Legal Texts through NLP-based Techniques*. Proceedings of LREC 2004, Lisbona, Portugal, 2004.

[4] Basili R., Catizone R., Pazienza M-T., Stevenson M., Velardi P., Vindigni M., Wilks Y.: *An Empirical Approach to Lexical Tuning*. Proceedings of the LREC1998 Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, Granada, Spain, 1998.

[5] Busa F., Calzolari N., Lenci A., Pustejovsky J.: *Building a Semantic Lexicon: Structuring and Generating Concepts*. In Bunt H., Muskens R., Thijsse E. (eds.): Computing Meaning Vol. II. Kluwer, Dordrecht, 2001.

[6] Calzolari N., Bertagna F., Lenci A., Monachini M.: *Standards and best Practice for Multilingual Computational Lexicons and MILE (Multilingual ISLE Lexical Entry)*. ISLE deliverables D2.2 – D3.2 http://lingue.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm, 2003.

[7] Fellbaum C., (ed.): *WordNet. An Electronic Lexical Database*. MIT Press, Cambridge (MA), 1998.

[8] Gómez-Pérez A., Manzano-Macho D.: *A Survey of Ontology Learning Methods and Techniques*. Ontoweb Deliverable 1.5 http://ontoweb.aifb.uni-karlsruhe.de/About/Deliverables, 2003.

[9] Gruber T.R.: A Translation Approach to Portable Ontologies. *Knowledge Acquisition*, Vol. 5, 1993.

[10] Hepple M., Ireson N., Allegrini P., Marchi S., Montemagni S., Gomez Hidalgo J.M.: *NLP-enhanced Content Filtering within the POESIA Project*. Proceedings of LREC 2004, Lisbona, Portugal, 2004.

[11] Jackson P, Moulinier I.: *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*. John Benjamins, Amsterdam, 2002.

[12] Joscelyne A., Lockwood R.: *Benchmarking HLT Progress in Europe*. HOPE, Copenhagen, 2003.

[13] Jurafsky D., Martin J.H.: *Speech and Language Processing*. Prentice Hall, Upper Saddle River (NJ), 2000.

[14] Lenci A., Bel N., Busa F., Calzolari N., Gola E., Monachini M., Ogonowsky A., Peters I., Peters W., Ruimy N., Villegas M., Zampolli A.: SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, Vol. 13, 2000.

[15] Lenci A., Montemagni S., Pirrelli V.: CHUNK-IT. An Italian Shallow Parser for Robust Syntactic Annotation. *Linguistica Computazionale*, Vol. 16-17, 2003.

[16] Lenci A., Montemagni S., Pirrelli V., (eds.): *Semantic Knowledge Acquisition and Representation, Giardini Editori*. Pisa, in stampa.

[17] Lin D., Pantel P.: *Concept Discovery from Text*. Proceedings of the Conference on Computational Linguistics 2002, Taipei, Taiwan, 2002.

[18] Maedche A., Staab S.: Ontology Learning for the Semantic Web. *IEEE Intelligent Systems*, Vol. 16, 2001.

[19] Manning C.D., Sch_tze H.: *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge (MA), 1999.

[20] Pustejovsky J.: *The Generative Lexicon*. MIT Press, Cambridge (MA), 1995.

[21] Ruimy N., Corazzari O., Gola E., Spanu A., Calzolari N., Zampolli A.: *The European LE-PAROLE Project: The Italian Syntactic Lexicon*. Proceedings of the LREC1998, Granada, Spain, 1998.

[22] Staab S., Studer R. (eds.): *Handbook of Ontologies*. Springer Verlag, Berlin, 2003.

[23] Vossen P.: Introduction to EuroWordNet. *Computers and the Humanities*. Vol. 32, 1998.

Nicoletta Calzolari è direttore dell'Istituto di Linguistica Computazionale del CNR di Pisa. Lavora nel settore della Linguistica Computazionale dal 1972. Ha coordinato moltissimi progetti nazionali, europei e internazionali, è membro di numerosi Board Internazionali (ELRA, ICCL, ISO, ELSNET ecc.), Conference Chair di LREC 2004, *invited speaker* e membro di Program Committee dei maggiori convegni del settore.
glottolo@ilc.cnr.it

Alessandro Lenci è ricercatore presso il Dipartimento di Linguistica dell'Università di Pisa e docente di Linguistica Computazionale. Ha conseguito il perfezionamento alla Scuola Normale Superiore di Pisa e collabora con l'Istituto di Linguistica Computazionale del CNR. Autore di numerose pubblicazioni, i suoi interessi di ricerca riguardano la semantica computazionale, i metodi per l'acquisizione lessicale, e le scienze cognitive.
alessandro.lenci@ilc.cnr.it