LE PRINCIPALI ATTIVITÀ DELL'ISTITUTO DI LINGUISTICA COMPUTAZIONALE: *IL PUNTO DI VISTA DEL DIRETTORE*

Antonio Zampolli

1. BREVE EXCURSUS STORICO

1.1. GLI INIZI

Fra la fine degli anni '50 e l'inizio dei '60, P.R.Busa S.J., universalmente riconosciuto come il pioniere degli spogli elettronici, da lui iniziati nel 1948, si era posto l'obiettivo di produrre gli spogli elettronici dell'intero corpus di scritti di S. Tommaso di Aquino o a lui attribuiti, per un totale di 10 milioni di occorrenze, presso il Centro per l'Automazione dell'Analisi Linguistica (CAAL) di Gallarate, creato principalmente con finanziamenti della IBM Italia. Per quei tempi, l'impresa era di dimensioni eccezionali. Fu lì che cominciai a lavorare nel 1960, all'indomani della discussione della tesi di laurea, dal titolo *Studi di statistica linguistica eseguiti con impianti IBM*. Il mio compito era quello di assistente del direttore P.R.Busa S.J., con attività di coordinamento e organizzazione delle operazioni elettroniche: input dei testi su schede perforate (attività che coinvolgeva circa 60 persone) ed elaborazione delle schede con macchine Unit Record (condotta da 30 operatori). Al tempo stesso, ero il responsabile delle specifiche linguistico-computazionali dello spoglio dei testi latini (metodi di contestualizzazione, sistema di lemmatizzazione e di analisi morfologica, ecc.) e della loro traduzione in algoritmi per i calcolatori allora in uso (IBM 7090 e IBM 1401). Questo mio lavoro portò, tra l'altro, alla creazione di un dizionario di macchina latino (basato sul lemmario del Forcellini - *Lexicon Totius Latinitatis*) e allo sviluppo degli algoritmi per la sua consultazione.

Le procedure e i programmi realizzati presso il CAAL vennero ben presto adottati in progetti di altri Enti (tra i quali, gli spogli per l'Enciclopedia Dantesca e l'archivio lessicale del costituendo

Istituto di Documentazione Giuridica), tanto che la IBM mi assunse presso il Centro Scientifico di Pisa con il compito di fornire consulenza e assistenza ai progetti che utilizzavano gli elaboratori elettronici nel settore degli studi linguistici ed umanistici in genere. Uno dei punti di riferimento per questa tipologia di attività fu il CNUCE (Centro Nazionale Universitario di Calcolo Elettronico) di Pisa, inaugurato nel 1965 dal Presidente della Repubblica Giuseppe Saragat, al quale la IBM offrì, per l'occasione, il volume di Concordanze e Indici della Divina Commedia. Nel 1969 il numero di progetti italiani che si avvalevano di metodi, procedure software e standard per la rappresentazione dei testi e dell'analisi linguistica erano diventati così numerosi, che la Direzione del CNUCE accettò, dietro mio suggerimento, di istituire una Divisione Linguistica e di affidarmene la direzione.

Il 1969 fu anche l'anno del mio primo corso di Linguistica Computazionale (il primo in assoluto in Italia). Tra i frequentatori del corso, l'allora direttore dell'ufficio di elaborazione elettronica della Camera dei Deputati, Beniamino Placido, si convinse dell'opportunità di arricchire le procedure di *information retrieval* sulle leggi al servizio dei Parlamentari, con un "dizionario di macchina" che permettesse di "proiettare" in modo più efficace i termini delle domande su quelli dei testi. A tal fine il dizionario doveva contenere non solo le informazioni necessarie per l'analisi morfologica automatica, ma anche informazioni di tipo semantico: definizioni, sinonimi, ecc. Il contributo, stanziato dalla Camera dei Deputati per finanziare questo progetto, permise al CNUCE di bandire 15 borse di studio, per la compilazione del Dizionario Macchina dell'Italiano (DMI). Queste borse, trasformate poi in posti di ricercatore in occasione del passaggio del CNUCE al CNR (1974), portarono ad oltre 20 gli effettivi della Divisione Linguistica, cui la Direzione del CNUCE aveva nel frattempo destinato prima 2 e poi 5 unità di personale. Le informazioni allora codificate nel DMI sono state successivamente (e sono ancora oggi) il punto di partenza di alcuni importanti progetti dell'Istituto quali ACQUILEX e ITALWORDNET (v. infra).

1.2. LA DIVISIONE LINGUISTICA DEL CNUCE (1969-1978)

Nella mia attività di Direzione della Divisione Linguistica del CNUCE, mi sforzai di orientare programmaticamente le ricerche nel settore a svilupparsi congiuntamente nei due filoni della linguistica computazionale che, dopo un periodo di frequenti e interessanti collaborazioni nel decennio 1955-65, si erano venuti progressivamente allontanando, fino a perdere ogni contatto tra loro. Da un lato lo *Humanistic Text Processing* (HTP) e cioè l'impiego di mezzi computazionali a supporto di ricerche umanistiche su testi e documenti, in particolare attraverso la produzione di indici e concordanze. Dall'altro il *Natural Language Processing* (NLP) (sviluppatosi a partire dalle prime pionieristiche attività di *Machine Translation*), volto essenzialmente ad applicare modelli formali, per lo più elaborati da scuole linguistiche generativo-trasformazionali, al "calcolo" (o meglio all'analisi, intesa sia come riconoscimento che come rappresentazione) di strutture linguistiche soggiacenti ai testi, o viceversa, per la generazione dei testi a partire dalla rappresentazione di tali strutture.

In particolare, il filone HTP, mentre da un lato seguiva attentamente gli sviluppi tecnologici, incorporando rapidamente le possibilità progressivamente offerte (fotocomposizione, terminali, numeri di caratteri, collegamenti on-line, ecc.), dall'altro limitava le proprie analisi alle unità grafiche dei testi, senza utilizzare le conoscenze e i metodi di NLP (o *Trattamento Automatico del Linguaggio*, TAL) per introdurre nei testi la individuazione e rappresentazione delle unità e proprietà linguistiche di diverso livello: morfologico, lessicale, sintattico, ecc. Pochissimi centri di ricerca al mondo potevano dirsi attivi in entrambi i filoni: nella mia attività di direzione, ho orientato le attività della Divisione Linguistica, dapprima, e dell'Istituto di Linguistica computazionale (ILC) poi, a sviluppare ricerche congiunte in entrambi i settori, dichiarando esplicitamente che operavamo programmaticamente in entrambi, utilizzando tecniche, metodi, conoscenze comuni.

Le nostre ricerche si concentrarono in particolare sulla creazione e sull'adattamento di strumenti di NLP per potenziare le applicazioni di HTP, per esempio attraverso procedure per la

lemmatizzazione semiautomatica e l'etichettatura morfosintattica dei testi, attraverso l'impiego di basi di conoscenza lessicale come supporto nella consultazione dei testi, ecc. D'altro lato, l'esperienza acquisita nel trattamento di grandi quantità di testi di tipo umanistico, ci spinse ad estendere la copertura linguistica e la robustezza dei componenti di NLP, ad utilizzare sofisticate tecniche di analisi quantitativa, e promuovere la consapevolezza della necessità di formulare degli standard di rappresentazione. Gli sviluppi successivi, a partire dagli anni '90, e soprattutto quelli attuali sul trattamento del *digital content* hanno dato pienamente ragione a questa scelta strategica.

1.3. LE SCUOLE ESTIVE

Il periodo compreso tra gli anni '70 e '80 fu particolarmente denso di attività internazionali volte a promuovere questi obiettivi. Nel 1970 organizzai la prima *International Summer School* in *Computational and Mathematical Linguistics* di Pisa, dal titolo "L'elaborazione elettronica di dati linguistici e letterari", tra i cui docenti figuravano R. Dyer (Università di New York), M. Gross (Università di Vincennes), D.G. Hays (Università di Buffalo), O. Menchi (Università di Perugia), Ch. Muller (Università di Strasburgo) e J. Raben (Queens College New York). L'iniziativa fu replicata nella stessa sede due anni dopo, e poi ancora nel '74 e nel '77, sempre con docenti di calibro internazionale rappresentanti gli orientamenti più innovativi del settore. L'ultima (la 5ª) Scuola Estiva Internazionale di Linguistica Computazionale è stata organizzata nel 1988, principalmente per conto e richiesta della *European Science Foundation* (ESF). La scelta del tema *Computational Lexicology and Lexicography* è stata determinata dall'osservazione del crescente interesse di molte discipline per il lessico, e dalla percezione dell'uso dei calcolatori come punto di convergenza tra queste discipline e come "potenziale focus" di una collaborazione interdisciplinare innovativa.

Queste scuole hanno profondamente influenzato lo sviluppo della Linguistica Computazionale a livello nazionale e internazionale. Per riconoscimento universale, infatti, hanno

formato i quadri dirigenti della LC europea, promosso la collaborazione tra i due filoni della LC (HTP e NLP) e la convergenza dei metodi basati su regole con quelli basati su dati quantitativi. Inoltre le scuole hanno messo in contatto, istituzionalmente, ricercatori europei ed americani, e contribuito all'avanzamento dello stato dell'arte in settori cruciali della Linguistica Computazionale (LC), della Linguistica, della Intelligenza Artificiale. Per fornire solo alcuni esempi, nella 2ª scuola, vennero confrontati gli approcci *data driven* e *rule-driven*. Nella 3ª scuola, venne fondata a Pisa una nuova scuola linguistica (la *Lexical Functional Grammar*), e trovò ispirazione il primo dizionario (LDOCE - *Longman Dictionary Of Contemporary English*) elaborato sulla base dei principi della LC, la cui disponibilità diede luogo a una serie di studi che si svilupparono poi attraverso le ricerche del gruppo IBM di Yorktown, il progetto ACQUILEX, ecc. Nella 4ª scuola estiva, si formò un gruppo che diede origine, nella *bay-area*, al sorgere della *frame semantics*.

Per l'Italia, queste scuole hanno anche rappresentato l'occasione per importare metodi e tecniche di analisi che furono poi sviluppati in modo originale dall'ILC (es. parser sintattico in MAGMA-LISP - basato sul modello Augmented Transition Network (ATN), integrato e ottimizzato da dati statistici; metodi per la rappresentazione della conoscenza; sviluppo di grammatiche formali; approcci e metodi innovativi nell'ambito della lessicologia e lessicografia computazionale, ecc.).

2. LABORATORIO DI LINGUISTICA COMPUTAZIONALE DEL CNR (1978-1980)

Nei primi anni di attività del Laboratorio di Linguistica Computazionale ci fu lo sforzo consapevole di assicurare la priorità alle attività di ricerca rispetto a quelle di assistenza e consulenza che avevano assorbito la parte maggiore delle risorse umane, in considerazione della missione essenzialmente di servizio che caratterizzava il CNUCE cui fino ad allora eravamo appartenuti. In questa attività fu di grande aiuto lo spirito di collaborazione del Consiglio Scientifico, in particolare del Suo Presidente Prof. G.

Nencioni, membro del Comitato 08 e Presidente dell'Accademia della Crusca.
Le linee di ricerca attivate nel periodo iniziale del laboratorio furono:

- lo sviluppo del Dizionario di Macchina dell'Italiano, consistente nella digitalizzazione di circa 100.000 lemmi associati ad un algoritmo per la trascrizione fonologica e a un analizzatore/generatore morfologico;
- gli studi di fattibilità per l'applicazione delle metodologie sviluppate per l'italiano allo spagnolo e al latino;
- lo sviluppo di un Parser dell'italiano basato sulla strategia ATN;
- le ricerche di linguistica quantitativa nell'archivio di testi disponibili in Machine Readable Form (MRF) presso il laboratorio;
- la creazione di una rete internazionale di basi di dati testuali attraverso la definizione di procedure effettive di scambio, come fase sperimentale preliminare al disegno di uno standard internazionale.

3. LE INIZIATIVE INTERNAZIONALI

Questo paragrafo contiene una breve illustrazione delle iniziative ed esperienze internazionali di maggior rilievo avviate negli anni precedenti la nascita dell'ILC. La promozione e la partecipazione a queste, come a innumerevoli altre iniziative di carattere internazionale è stata determinante per la creazione dell'ILC, per la tipologia di attività svolte, e per il ruolo e il riconoscimento dell'ILC in ambito nazionale e internazionale. Esse hanno infatti permesso all'Istituto di promuovere e diffondere metodologie, principi strategici, paradigmi scientifici (nel senso di Kuhn) importanti per il progresso della nostra disciplina, in particolare, verso direzioni che corrispondevano alle esigenze strategiche della comunità nazionale.

3.1. TRADUZIONE AUTOMATICA

Nel 1976, venni incaricato dal Governo Italiano di far parte di un gruppo di esperti che aveva il compito di valutare la opportunità, da parte della Comunità Europea, di acquisire il sistema americano di *Machine Translation* (MT) SYSTRAN. Con la cooperazione del delegato tedesco H. Zimmermann riuscii a convincere la CE ad affiancare al sistema SYSTRAN un esperimento di sistema di traduzione avanzato basato sulla tecnologia europea. Negli stessi anni, in qualità di Presidente del Comitato Scientifico del COLING '78 (Bergen), rilevai la necessità da parte del settore della LC di dimostrarne le possibilità applicative e di stabilire contatti con le Agenzie di Ricerca Internazionali quali la CE.

Le attività preparatorie all'avvio di quello che sarà poi conosciuto come il progetto europeo EUROTRA per la traduzione automatica richiesero una delicata mediazione tra le scuole di traduzione francese (Grenoble) e quella tedesca (Saarbrücken), ciascuna delle quali voleva imporre le proprie tecnologie di analisi e di generazione. Riuscii a risolvere questo elemento di conflittualità proponendo una organizzazione generale della architettura del sistema di traduzione basata sul concetto di "struttura di interfaccia" (IS), una struttura di rappresentazione semantico-sintattica del testo da tradurre, specificata secondo regole comuni accettate e adottate da tutti i gruppi linguistici partecipanti (dapprima 7, poi 9 lingue europee) e indipendente dalle tecniche di analisi adottate. Ogni gruppo avrebbe potuto scegliere quale tecnologia utilizzare per l'analisi di testi nella propria lingua (considerata come lingua *source*) purché avesse prodotto la IS secondo le regole comuni, e per trasferire la IS ricevuta da altri gruppi nazionali in una IS dalla quale fosse più agevole generare i testi nella propria lingua (considerata come lingua *cible*). La formulazione della struttura di interfaccia costituisce uno dei maggiori risultati scientifici di EUROTRA e il concetto di IS da me proposto e la organizzazione di base furono mantenuti fino alla fine del progetto.

EUROTRA non produsse un sistema effettivo di traduzione multilingue, ma contribuì in modo determinante a creare un network di Istituti europei di LC, promuovendo la creazione e la

organizzazione di équipes specializzate in paesi nei quali la LC era stata, fino ad allora, assente, e producendo come fall-out una serie di conoscenze e di esperienze determinanti per gli sviluppi successivi della LC in Europa. Per quanto concerne l'ILC, la partecipazione al gruppo italiano di EUROTRA ha creato competenze soprattutto nel settore delle basi di dati lessicali e delle cosiddette grammatiche di transfer.

3.2. LE "RISORSE LINGUISTICHE"

I lessicografi "accademici" erano stati i primi a mostrare interesse per il potenziale offerto dall'impiego dei calcolatori negli spogli lessicografici. Questo interesse si spostò progressivamente dalla raccolta dei testi in MRF e dalla produzione di indici e concordanze, alla assistenza del calcolatore nella scelta e nell'ordinamento delle citazioni, e nella redazione delle voci. La maggioranza dei progetti "accademici" di lessicografia richiedeva tuttavia tempi molto lunghi e costituiva un peso notevole nei bilanci degli Enti di ricerca. Nel 1980 la *European Science Foundation* (ESF) mi chiese di condurre una rassegna dei progetti lessicologici e lessicografici europei supportati da fondi pubblici. L'analisi dei dati mostrò che un gran numero di questi progetti faceva uso del calcolatore, ma che le tecnologie disponibili non erano sfruttate convenientemente.

Alla luce di questi dati, la ESF mi spinse ad organizzare a Pisa, nel 1981, un Workshop sul tema *The possibilities and limits of the computers in producing and publishing dictionaries*. Il compito centrale consisteva nel verificare lo stato dell'arte, identificare le priorità di ricerca, formulare delle raccomandazioni strategiche, proponendo metodologie innovative per contribuire al miglioramento dei prodotti lessicografici e, se possibile, alla riduzione dei tempi e dei costi dei progetti finanziati dagli enti pubblici di ricerca. Dalle discussioni, cui parteciparono i delegati degli Enti di ricerca europea federati nella ESF e della NEH (*National Endowment for the Humanities*) statunitense, direttori di grandi imprese lessicografiche, ed esperti di LC da me scelti per il loro potenziale contributo innovativo, emersero alcune idee chiave, alcune delle quali sono ancora oggi attuali. In particolare, poiché le

varie scuole di pensiero in linguistica non offrivano una metodologia adeguata per la descrizione dei significati di insiemi estesi di entrate lessicali, la semantica lessicale avrebbe dovuto sviluppare nuovi approcci teorici che potessero collocare l'identificazione, la definizione, la strutturazione dei significati dei dizionari su una base più coerente e teoricamente motivata.

Oggi assistiamo al convergere, per la realizzazione di questo compito, degli sforzi della semantica generativa, della psicolinguistica (WordNet), della rappresentazione della conoscenza (ontologie). Si sarebbe dovuto inoltre riflettere sull'uso innovativo della tecnologia nel disegnare nuovi tipi di dizionari, in particolare per il cosiddetto *electronic publishing*.

Per la costruzione degli strumenti che aiutassero i lessicografi nel loro lavoro, era necessaria un'analisi del *modus operandi* del lessicografo accademico. (Questa raccomandazione è stata realizzata, per es., attraverso la costruzione della stazione lessicografica dell'ILC). Emerse inoltre la necessità di stabilire degli standard per codificare e scambiare dati testuali, analisi linguistiche, descrizioni lessicali. (Questa raccomandazione viene considerata come il primo passo concreto verso iniziative di standardizzazione poi affermatesi internazionalmente, quali TEI (*Text Encoding Iniziative*), e EAGLES (*Expert Advisory Group on Language Engineering Standards*).

Come risultato del Workshop di Pisa, la ESF creò dapprima un gruppo di esperti per la lessicografia computazionale (1983, Zampolli, Quemada, Zimmermann, Van Sterkenburg, Weiner) e quindi, nel 1985, venni cooptato come membro dello *Standing Committee for the Humanities* (SCH) della ESF, come *subject representative* per la linguistica e la LC, al posto di J. Lyons, allora indisposto, per il quinquennio 1985-1990.

Il Workshop *On automating the lexicon*, da me organizzato in collaborazione con D. Walker, J. Sager, L. Rolling, N. Calzolari, nel maggio 1986, è universalmente riconosciuto come il punto d'inizio del processo che ha portato a stabilire il settore delle risorse linguistiche (RL) quale è oggi. Il Workshop passò in rassegna le ricerche, le pratiche correnti, gli sviluppi potenziali delle attività su lessici e corpora, con particolare riguardo all'ambiente multilingue. Le raccomandazioni finali, da me

trasmesse (1987) alla CE, diedero origine a tutta una serie di progetti europei (ACQUILEX, ET-7, MULTILEX, MULTEXT, GENELEX, DELIS, ecc.), e di attività organizzative e di ricerca (per es. la Scuola Estiva dell'88 *Computational Lexicology and Lexicography*).

L'ILC ha rivestito un ruolo decisivo nel formarsi e nel diffondersi del nuovo paradigma che caratterizza oggi l'intero campo disciplinare della LC, il cosiddetto *data-driven approach*, che è fondato sull'utilizzo e sullo studio di estese raccolte di dati linguistici e delle loro descrizioni, le cosiddette "risorse linguistiche", costituite essenzialmente dalla documentazione sulla quale si basa lo studio di una lingua e dalla registrazione analitica dei risultati di tale studio: in particolare corpora rappresentativi e annotati di grandi dimensioni, lessici il più possibile completi, grammatiche a larga copertura linguistica.

In un incontro organizzato nel settembre del 1991 a Torino da ESPRIT e dalla National Science Foundation (NSF), tra 10 rappresentanti della ricerca europea e 10 rappresentanti della ricerca nord-americana, corpora e lessici orali e scritti vennero indicati come esigenze prioritarie e comuni di entrambe le comunità. Come rappresentante della ricerca europea, introdussi (per quanto mi risulta, per la prima volta) il termine risorse linguistiche (RL) per sottolineare il ruolo infrastrutturale di questi componenti, paragonandolo a quello delle risorse di base (per es. acquedotti, elettricità, strade) necessarie per lo sviluppo industriale di una nuova area geografica. Proposi poi il termine "risorse linguistiche" nella mia relazione al panel Danzin (1992). Il rapporto finale del panel Danzin accolse sia il termine, che da allora entrò nella terminologia della Commissione e da questa nella letteratura corrente del settore, sia l'affermazione della natura infrastrutturale delle RL.

La costruzione di RL estese ha un costo molto elevato, e impone uno sforzo organizzativo rilevante. Nell'ultima decade, la nostra comunità è stata però indotta ad intraprendere la costruzione e l'utilizzo di RL adeguate, da spinte molto forti di ordine sia scientifico-tecnico sia economico-organizzativo.

Per comune consenso, le attività essenziali da svolgere per le risorse linguistiche si articolano in quattro sotto-settori, che ho chiaramente individuato e descritto per la prima volta in una comunicazione tenuta con N. Calzolari al Workshop organizzato dalla CE a Santorini nel 1993 sul futuro della *language industry* (LI) in Europa:

a) elaborazione di standard consensuali;
b) creazione delle RL necessarie;
c) distribuzione e condivisione delle risorse;
d) creazione delle sinergie tra progetti nazionali, progetti comunitari e internazionali, iniziative industriali.

L'avvio delle attività in ciascuno di questi 4 sotto-settori è stato uno dei temi focali delle ricerche dell'ILC, sia in campo nazionale (v. infra), sia in campo internazionale, soprattutto attraverso la proposta, il coordinamento e la esecuzione di progetti internazionali per lo più comunitari dei quali elenco qui alcuni esempi[1]. Questi progetti erano e sono intesi sia a diffondere nei diversi paesi la consapevolezza della centralità e dell'importanza delle RL per la LC, sia a rispondere - sia pure in modo parziale - ai bisogni di RL della comunità di Ricerca e Sviluppo (RS).

3.3. L'INDUSTRIA DELLE LINGUE

Grazie anche ad alcuni spettacolari successi di alcuni sistemi, dovuti alla coincidenza tra la natura delle particolari operazioni linguistiche da svolgere e la maturità raggiunta dalle tecnologie richieste per automatizzarle, verso la metà degli anni '80 è apparso chiaro che la LC avrebbe potuto offrire la possibilità, attraverso

[1] I progetti comunitari qui descritti cui mi riferisco sono: TEI (contratti assegnati in versioni successive), EAGLES, EAGLES/ISLE, NERC1, NERC2, MLAP-PAROLE, LE-PAROLE, LE-SIMPLE, LRE-RELATOR, ELRA. Altri progetti diretti da me o da altri colleghi dell'ILC come coordinatore europeo, o nazionale, o ai quali l'ILC ha partecipato con attività scientifiche e/o manageriali di vario tipo, sono elencati in Appendice 1.

sviluppi opportunamente indirizzati, di rispondere ad esigenze profonde ed impellenti della emergente Società dell'Informazione.

Mi adoperai per promuovere presso l'ILC ricerche volte alla creazione e sperimentazione di prototipi di sistemi applicativi o di loro componenti, e proposi al CNR il progetto "Metodi e strumenti per l'Industria delle Lingue nella cooperazione Internazionale", successivamente approvato. Ricercatori, sviluppatori, agenzie nazionali e internazionali per la ricerca, divennero sempre più consapevoli del potenziale strategico, industriale, culturale dell'Industria delle Lingue, che emerge come un settore autonomo nelle industrie dell'informazione.

Iniziammo così a parlare del paradigma dell'industria delle lingue (LI)[2], indicando con questo termine applicazioni, di interesse economico e commerciale, basate su sistemi computazionali capaci di compiere automaticamente, sulla lingua, operazioni e compiti che sono parti essenziali dell'applicazione.

A questo termine se ne affiancarono rapidamente altri, scelti per attirare la attenzione del pubblico su particolari aspetti del settore. Per esempio, nel 3° Programma Quadro Comunitario di Ricerca, il settore era chiamato *Language Research and Engineering*, e poi, nel 4°, *Language Engineering* tout-court, mettendo in rilievo la necessità di ingegnerizzare la tecnologia per renderla utilizzabile in applicazioni concrete. Nel 5° Programma Quadro, e nei più recenti programmi governativi americani, si parla di *Human Language Technology*, per evidenziare il ruolo che le tecnologie della LC possono sostenere a favore dello sviluppo di una società *user-friendly*, centrata sull'uomo.

Nella fase di preparazione del 6° Programma Quadro (2002-2006) attualmente in corso, che sembra destinato in particolare a

[2] Il termine *industries de la langue* venne lanciato al Convegno Internazionale organizzato a Tours dal Consiglio di Europa nel 1986. Esso copre sia le attività nelle quali il calcolatore è utilizzato essenzialmente per assistere le professioni "tradizionali" della linguistica applicata (per es. lessicografia, traduzione, insegnamento delle lingue) sia le attività dirette a sviluppare nuovi tipi di applicazioni (per es., sistemi di interfaccia uomo-macchina in lingua naturale, traduzione automatica, estrazione dell'informazione, ecc.).

potenziare la ricerca europea affrontando temi altamente innovativi e prioritari per lo sviluppo futuro della società, il trattamento del linguaggio è considerato all'interno di azioni diverse, quali:

- la gestione della conoscenza (in particolare, il *Semantic Web* multilingue);
- la interattività multimediale e multisensoriale.

3.4. NETWORKING

Ho operato, sin dal principio, per promuovere la collaborazione scientifico-tecnica e organizzativa tra le diverse comunità nelle quali si articola il panorama della LC. Tra le iniziative promosse a tal fine, ne ricordo qui alcune.

3.4.1. *Fondazione e management di ELSNET*

Nel 1991, nell'ambito del programma comunitario ESPRIT, ho fondato con E. Klein (Edimburgo) ELSNET, *European Network of Excellence in Speech and Human Language Technologies*, un forum su base europea dedicato alle tecnologie del linguaggio umano. Esso opera in un contesto internazionale, riunendo circa 150 nodi, tra pubblici e privati, e considera tutte le aree di ricerca della comunicazione umana relative al linguaggio e al discorso. Esso mira a migliorare la ricerca e lo sviluppo delle tecnologie del linguaggio umano in Europa mettendo insieme gli operatori chiave del settore e fornendo un forum aperto e propositivo che serva da piattaforma:

- per lanciare azioni innovative;
- per fare analisi del presente e sviluppare visioni future;
- per incoraggiare un ambiente comune comprendente risorse, standard e valutazioni;
- per costruire, condividere e sfruttare conoscenze ed esperienze.

Task-force "Risorse Linguistiche" di ELSNET

Dalla sua fondazione, ELSNET ha istituito una task-force nel settore delle RL e della valutazione i cui obiettivi sono:

- fornire una piattaforma per lo scambio di informazioni e di competenze nell'area delle risorse linguistiche, della valutazione, e degli standard, sia in Europa sia in contatto con altri paesi emergenti o di alta tecnologia;
- diffondere metodi e competenze sulle RL generate da altri progetti europei;
- favorire le sinergie tra progetti nazionali e azioni multinazionali;
- promuovere iniziative limitate, ma ben focalizzate per:
    - definire meglio i bisogni degli utilizzatori;
    - promuovere la creazione di piccoli prototipi/modelli iniziali di risorse innovative che implicano un rischio metodologico o tecnico;
    - esplorare la fattibilità di organizzare le attività necessarie per costruire, mantenere ed aggiornare risorse che implicano un certo grado di incertezza;
    - dare un supporto iniziale a questo tipo di iniziative, o ad altre che richiedono azioni tempestive "a rischio";
    - stimolare la evoluzione del settore attraverso questa funzione.

Queste iniziative possono poi seguire un cammino indipendente, ma le varietà e la complessità delle conoscenze rappresentate da ELSNET e la relativa flessibilità e rapidità di decisione hanno già mostrato di essere fattori essenziali.

ELSNET ha iniziato recentemente, anche sotto il mio impulso, un'azione di estensione internazionale attraverso contatti con altri continenti, e il disegno di una *roadmap* delle ricerche e delle applicazioni prevedibili/auspicabili del settore per i prossimi 10 anni.

3.4.2. *Distribuzione delle risorse*

Uno dei maggiori problemi da risolvere per assicurare il riutilizzo delle RL, con il conseguente risparmio di sforzi, di tempi e di costi, è organizzare un meccanismo di distribuzione delle RL. Le Agenzie di ricerca del governo americano si preoccuparono per prime di questo problema, e indissero una prima riunione nel 1988 alla quale partecipai come esperto europeo. Fu decisa così la costituzione del *Linguistic Data Consortium* (LDC) presso l'Università di Philadelphia. Esso dipende da finanziamenti pubblici (ARPA e NSF) e dalle sottoscrizioni annuali degli utenti. Consapevole del significato strategico della iniziativa, proposi alla CE la costituzione di una organizzazione equivalente europea.

RELATOR (1993-1995)

La CE lanciò un bando per uno studio di fattibilità, cui risposi con una proposta nel quale avevamo come partner LIMSI (CNRS-Parigi), DFKI (Saarbrücken), l'Università di Edimburgo, il CST di Copenhagen. Coordinai il progetto che ne seguì, RELATOR, il quale, attraverso l'utilizzo delle competenze scientifiche del nostro Istituto e dei partner, l'opera di un pubblicista americano per le pubbliche relazioni, l'assistenza di uno *Steering Committee* (formato dalle principali Industrie europee del settore e presieduto dal Direttore Generale della DGXIII), l'aiuto degli Uffici Giuridici della Commissione e dell'Istituto di Diritto Internazionale del CNRS, e il monitoraggio di un comitato di saggi,[3] propose la fondazione di un'Associazione Europea per le Risorse Linguistiche, che venne registrata in Lussemburgo (ELRA - *European Language Resources Association*) nel 1996.

[3] Il comitato comprendeva: Danzin, esperto del governo francese e della CE per le industrie della lingua, B. Quemada, vicepresidente dell'Alto Comitato per la Lingua Francese e B. Oakley, past-president di Logica, ed esperto della CE per l'IST.

ELRA (European Language Resources Association)

ELRA è guidata da un Executive Board di 12 persone. Lo scopo dell'Associazione (che è formata da 3 collegi: scritto, parlato, terminologia) è promuovere le RL in tutte le loro forme, e coordinare e portare avanti la loro validazione, distribuzione e riutilizzazione in un contesto europeo. Lo statuto di ELRA specifica le seguenti attività:

- valutare, selezionare, implementare i mezzi necessari per distribuire le RL. Dove appropriato, organizzare e gestire l'acquisizione di RL da parte dei produttori e sviluppare i quadri tecnici e legali per validare e distribuire queste risorse agli utenti interessati;
- su richiesta delle organizzazioni europee che finanziano programmi per la creazione di RL, dare consigli sulla distribuzione e sulla validazione di quelle risorse;
- fungere da fonte di informazioni concernenti i contenuti e la disponibilità di RL per tutte le parti interessate in Europa;
- identificare i bisogni di RL ancora inappagati e stimolare le organizzazioni appropriate a creare RL adatte per soddisfare questi bisogni.

A differenza di LDC, ELRA, la cui attività era stata inizialmente sostenuta da un contratto CE, si è ora resa autosufficiente, grazie al volume di risorse distribuite (per ogni risorsa ELRA ha diritto a una percentuale) e alle quote di iscrizione degli organismi membri. ELRA infatti è una Associazione non di persone fisiche, ma di Enti aventi personalità giuridica[4]:

[4] Tra gli oltre 100 Enti iscritti figurano i più prestigiosi centri di ricerca europei, americani ed asiatici (LIMSI, DFKI, ILSP, CST, SPEX, CLIF, EAFL, GMD, INALF, Cervantes, Real Academia de la Lengua Española, Carnegie Mellon, INL, Università di Taiwan, ecc.) e le principali industrie del settore: IBM, MICROSOFT, Philips, Siemens, Xerox, Sony, Panasonic, Thompson Multimedia, Dragon System, Eriksonn, Harper Collins, SRI, Daimler Chrysler AG, Telefunken, varie Compagnie telefoniche, ecc. ELRA è inoltre supportata dal Ministero della Ricerca Francese e dalla *Délégation Générale de la Langue Française*. Recentemente

Una delle principali attività di ELRA è l'organizzazione di LREC: *International Conference on Resources and Evaluation*.

3.4.3. *LREC (International Conference on Language Resources and Evaluation)*

Scopo della conferenza è fornire un panorama dello stato dell'arte, scambiare informazioni su attività in corso o previste, discutere le risorse linguistiche e loro applicazioni, discutere metodi e dimostrare strumenti per la valutazione, esplorare possibilità e promuovere iniziative per la cooperazione internazionale, ecc. L'esigenza di una conferenza come LREC muove dalla constatazione che numerosi attori, i quali lavorano in settori diversi, su aspetti diversi delle RL, concentrandosi su argomenti di particolare rilevanza per il loro interesse professionale (linguisti, linguisti computazionali, case editrici, ingegneri del software, operatori culturali, industrie multimediali, delle telecomunicazioni e dei calcolatori, tecnici della lingua e dell'insegnamento linguistico, ingegneri della conoscenza, fornitori di servizi sulla rete) appartenendo a comunità diverse, che hanno le proprie organizzazioni e conferenze specifiche, raramente hanno l'occasione di trovarsi per scambiare informazioni, ed esplorare possibili sinergie e cooperazioni. Il numero di partecipanti nelle tre edizioni di LREC sembra confermare che questa constatazione risponde a un bisogno diffuso tra gli operatori del settore.

3.4.4. *ENABLER*

Il network ENABLER, da me auspicato fin dai tempi del NERC, e che recentemente è diventato un progetto IST finanziato dalla CE, è stato lanciato alla luce delle seguenti considerazioni. Negli ultimi anni la maggior parte dei programmi nazionali europei nel campo del HLT (*Human Language Technology*) hanno identificato le RL come la prima area che deve essere supportata dai finanziamenti nazionali, sulla base di consultazioni che coinvolgono ricercatori,

stiamo svolgendo un contratto finanziato in parte dalla CE in parte da NSF, il cui scopo è armonizzare i cataloghi e le procedure di ELRA con quelle del LDC.

industrie, fornitori di servizi, ecc. La disponibilità di RL è anche un argomento "sensibile", che tocca direttamente la sfera dell'identità linguistica e culturale, ed è una precondizione cruciale per la partecipazione di una lingua, e dei cittadini che parlano questa lingua, nella società dell'informazione. Il costo della produzione di RL è comunque molto alto, e questo può ostacolare - e in parte lo ha fatto - il coinvolgimento degli sviluppatori industriali - in particolare le piccole e medie imprese - nel settore del NLP (*Natural Language Processing*). Inoltre, è emerso chiaramente in recenti discussioni che il programma quadro di ricerca della Commissione Europea così come delle Agenzie americane quali i programmi NSF e DARPA, non sono adatti ad offrire un completo e continuo supporto ad iniziative infrastrutturali. Dall'altro lato, è chiaro che soltanto i) combinando le forze di diverse iniziative e diversi compiti istituzionali, ii) sfruttando al meglio il "modus operandi" delle Autorità nazionali in diverse situazioni nazionali, iii) rispondendo ai bisogni e alle priorità di ciascuna comunità industriale e di ricerca e sviluppo, sulla base di una chiara distinzione di compiti e ruoli per attori diversi nella scena del HLT, possiamo produrre le sinergie, l'economia di scala, la convergenza e l'accumulazione di sforzi necessari a fornire le RL infrastrutturali necessarie a realizzare il pieno potenziale di una società dell'informazione globale multilingue.

Il Network mira ad attivare la progressiva realizzazione di questo quadro cooperativo urgentemente necessario, supportando collegamenti, stabilendo meccanismi di scambio, incoraggiando la cooperazione e l'interoperabilità dei risultati di progetti e attività nazionali che, in un sostanziale sottoinsieme di paesi membri dell'Unione Europea, sono stati recentemente finanziati da rilevanti autorità nazionali per fornire RL di diversi tipi alle rispettive lingue. Gli obiettivi centrali del Network sono:

- rinforzare e istituzionalizzare sotto l'ombrello della Commissione l'embrionale rete di iniziative nazionali esistente, forgiando appropriati collegamenti tra loro, creando un repertorio regolare, aggiornato, strutturato e pubblicamente disponibile di rilevanti informazioni organizzative e tecniche;

- fornire un forum ufficiale di discussione e meccanismi di coordinamento generale, scambio di informazioni, dati, pratiche ottimali, condivisione di strumenti, cooperazione multilaterale e bilaterale su specifici argomenti;
- ampliare gradualmente il Network iniziale, identificare e promuovere l'inclusione di rappresentanti di iniziative nazionali complementari;
- incoraggiare le sinergie tra le attività nazionali, e aumentare la compatibilità e l'interoperabilità dei loro risultati, in questo modo facilitando, tra le altre cose, il trasferimento di tecnologie tra le lingue;
- mantenere la compatibilità tra le varie RL, estendendo con fondi nazionali i nuclei iniziali prodotti secondo comuni specifiche tecniche nei progetti europei, in questo modo i) assicurando la realizzazione di una grande piattaforma infrastrutturale di RL europee standardizzate, prerequisito essenziale per qualsiasi futura e ampia costruzione di RL multilingue di larga scala, e pertanto ii) sfruttando in modo positivo il carattere multilingue dell'Unione Europea;
- migliorare la visibilità e l'impatto strategico - rispetto agli Stati Uniti, al Giappone, ecc. - delle varie attività nazionali, che sono il risultato di analoghe decisioni prese dai politici in diversi paesi sulla base di un simile apprezzamento delle priorità e dei bisogni;
- fornire un forum per discutere i bisogni dell'industria e formulare un'agenda comune di priorità industriali a medio e lungo termine;
- promuovere lo scambio di strumenti, di specifiche, di protocolli di validazione prodotti dai progetti nazionali, in questo modo evitando la duplicazione degli sforzi e la dispersione o la divergenza degli approcci;
- contribuire alla creazione di una organizzazione europea per l'armonizzazione della descrizione, sotto il profilo dei metadata, dei vari tipi di RL (discorso, testi, risorse multimediali e multimodali);
- promuovere lo sfruttamento industriale delle RL costruite in vari paesi;

- incoraggiare e contribuire al processo di progettazione e di progressiva implementazione di una rete cooperativa globale internazionalmente riconosciuta per la fornitura di RL.

I partecipanti al network sono Istituti che, nei rispettivi paesi, coordinano programmi nazionali governativi per le RL, o sono istituzionalmente incaricati di produrle.

Comitato Internazionale per le RL

Fra le attività del Network di ENABLER, abbiamo appena costituito un Comitato Internazionale per le Risorse Linguistiche, che riunisce i gruppi e le organizzazioni scientifiche internazionali interessati all'uso delle RL nella LC, e in particolare quelli che in diversi paesi o continenti operano nel settore delle RL scritte, parlate, multimodali. Scopi principali del Comitato sono:

- costituire un foro "propositivo" per lo scambio di informazioni tecnico-scientifiche sui progetti in corso;
- definire un insieme minimo di RL che dovrebbero essere disponibili per il maggior numero di lingue possibili, e trasmettere questo messaggio alle autorità nazionali e internazionali competenti;
- organizzare programmi congiunti di ricerca e di produzione di risorse multilingui comuni;
- facilitare la partecipazione al network di Progetti Nazionali Europei (ENABLER), di progetti nazionali di altri paesi (americani, asiatici, ecc.) che hanno espresso l'intenzione di collaborare.

4. L'ISTITUTO DI LINGUISTICA COMPUTAZIONALE - ILC (1980-2001)

Nel 1980 il Laboratorio venne trasformato in Istituto del CNR. Fu ben presto evidente che, sia per motivi scientifici - assicurare la collaborazione e le sinergie con i maggiori centri di ricerca stranieri -, sia per motivi organizzativi - completare la dotazione

finanziaria assegnata dal CNR con fondi provenienti da altre fonti, sia per svolgere l'azione di mediazione tra ricerca italiana e straniera prevista dal nostro Statuto:

- era necessario continuare a rafforzare la presenza dell'ILC in sede internazionale;
- era necessario svolgere una azione di profonda innovazione strategica e scientifica anche in campo internazionale, e far evolvere la nostra disciplina secondo la visione degli sviluppi necessari fin qui delineata, sia per farla uscire dall'impasse in cui sembrava caduta, sia per meglio rispondere alle esigenze prioritarie della nostra comunità nazionale.

La mia attività di Direttore, oltre ad assicurare la gestione scientifica e amministrativa dell'Istituto, si è quindi concentrata anche su azioni di tipo strategico e innovativo a livello internazionale e nazionale. Ciò comportava la assunzione di responsabilità dirette nella promozione di nuovi paradigmi scientifici e nella gestione e nell'indirizzo strategico di Strutture e Associazioni esistenti; la creazione di nuove Strutture e progetti internazionali e nazionali (ove necessario), nella cui gestione sono stato validamente assistito dai miei collaboratori; la partecipazione alle decisioni di Agenzie internazionali di ricerca, la discussione con i rappresentanti di Enti incaricati delle scelte strategiche fondamentali.

4.1. PRINCIPALI LINEE DI RICERCA DELL'ILC

Riporto qui, a grandi linee, gli obiettivi principali delle ricerche in corso presso l'ILC.[5] Nel loro insieme, per giudizio del Comitato

[5] Responsabili delle linee di ricerca - attive nel 2001 - cui si fa qui riferimento sono: Nicoletta Calzolari (Standard e risorse linguistiche computazionali); Eugenio Picchi (Metodi e tecnologie per basi di dati testuali e linguistici multifunzionali); Andrea Bozzi (Filologia computazionale); Irina Prodanof (Modelli e metodi per il trattamento delle lingue naturali); Giovanna Turrini (Tecnologia della lingua per la didattica e la disabilità); Antonio Zampolli (Coordinamento di attività nazionali e internazionali). A loro, e ad altri colleghi

Scientifico dell'ILC (confermato più volte anche dal Comitato di afferenza 08) esse "formano un insieme coordinato di attività allo stato dell'arte della Linguistica Computazionale, i cui meriti scientifici sono ampiamente riconosciuti in ambito internazionale e sono testimoniati dalla leadership conseguita dall'Istituto nei principali progetti europei del settore".

4.1.1. *Standard e Risorse Linguistico-computazionali*

Il termine Risorse Linguistico-computazionali (RL) designa, come abbiamo già visto, insiemi (di solito molto estesi) di dati linguistici, accompagnati o costituiti da annotazioni e rappresentazioni formalizzate, articolate a diversi livelli di descrizione linguistica, che sono usati nel costruire, ampliare, rendere operativi, valutare modelli, algoritmi, componenti e sistemi per il trattamento automatico del parlato e dello scritto. I tipi principali di risorse cui si fa di solito riferimento sono corpora testuali e orali, lessici computazionali, insiemi di informazioni grammaticali, raccolte terminologiche, ma sono inclusi anche strumenti software necessari per la creazione, l'acquisizione, la validazione, l'accesso, l'analisi di tali risorse, e in generale metodi e componenti software di base che assicurano le funzioni fondamentali per i sistemi di LC.

A partire dalla seconda metà degli anni '80, la mancanza di risorse linguistiche adeguate è stata riconosciuta come uno dei principali ostacoli al successo delle attività di ricerca e sviluppo nella LC, sia per la costruzione di modelli adeguati rispetto all'uso della lingua sia, in particolare, per il passaggio da prototipi a sistemi applicativi utilizzabili in contesti operativi reali.

Buona parte delle nostre ricerche in questo settore sono rappresentate dalla partecipazione alle diverse fasi (direzione, pianificazione, definizione, specifiche tecniche, realizzazione, valutazione, dimostrazione) di progetti comunitari/internazionali e/o di programmi di interesse nazionale da noi promossi e spesso da noi coordinati.[6]

dell'ILC coinvolti nelle varie ricerche, si deve anche parte delle descrizioni delle attività in questa sezione.

[6] Per es. i progetti europei ACQUILEX, DELIS, ET-7, ET-10, SPARKLE, MULTEXT, MULTILEX, NERC, EAGLES, EuroWordNet, ONOMASTICA, LS-GRAM, COLSIT,

Definizione di standard per lessici computazionali e per corpora testuali, orali, e multimodali

Il lavoro viene svolto principalmente in collaborazione con il progetto CE-NSF ISLE.[7] ISLE si avvale della collaborazione dei principali gruppi di esperti - sia accademici sia industriali, europei e americani - operanti nel settore. I risultati sono basati dunque su un ampio consenso dei principali attori del settore. La ricerca si propone la definizione di standard per *i*) la formalizzazione e codifica delle informazioni lessicali ai diversi livelli di descrizione linguistica (in particolare per la semantica), sia per lessici monolingui sia per lessici multilingui, con particolare riguardo alle esigenze di sistemi computazionali applicativi multilingui, *ii*) il trattamento di *multiwords*, sia a livello di codifica lessicale sia di annotazione nei testi, *iii*) il disegno di ontologie per lessici generici e specialistici, e per diversi tipi di applicazioni che richiedono la comprensione del contenuto dei testi, *iv*) l'annotazione di corpora testuali ai diversi livelli di descrizione linguistica e a livello concettuale, in funzione sia della valutazione di sistemi automatici di parsing, disambiguazione semantica, ecc., sia dell'annotazione richiesta da sistemi applicativi quali *Information Extraction*, *Machine Translation*, ecc., sia dell'utilizzo di corpora annotati per *Machine Learning*, *v*) l'annotazione del dialogo ai livelli morfo-sintattico, sintattico e pragmatico, *vi*) l'annotazione di corpora multimodali, *vii*) la definizione di criteri di validazione di risorse linguistiche.

PAROLE, SIMPLE, MATE, NITE, ISLE, MUSI, POESIA, il progetto americano NSF XMELLT, e i progetti nazionali in corso.

[7] ISLE è da me coordinato per l'Europa, e N. Calzolari è responsabile europea del Working Group sui Lessici Computazionali.

Creazione di risorse lessicali: Banca di Conoscenza Lessicale dell'Italiano e acquisizione "dinamica" di informazioni

a) Banca di Conoscenza Lessicale dell'Italiano (BCL-It)

Il tema di una Banca Computazionale di Conoscenza Lessicale (BCL) dell'Italiano, da utilizzarsi in diversi sistemi e applicazioni di LC, si inserisce in una tendenza sempre più evidente nell'ambito della LC (e della linguistica teorica) che pone il componente lessicale al centro di qualsiasi sistema di elaborazione e trattamento del linguaggio naturale. Ci si propone l'arricchimento del Database Lessicale dell'Italiano, esistente presso l'ILC, con ulteriori informazioni a livello morfologico, sintattico, semantico, collocazionale, e di continuare la sua trasformazione in una BCL contenente informazioni che permettano di effettuare inferenze. L'estensione è prevista sia in quantità (maggior numero di entrate), sia in profondità, cioè qualità e livelli di informazioni (aggiunta di entrate di diversi tipi, quali nomi propri, *multiwords*, terminologia, maggiori informazioni semantiche, ecc., aggiunta di collegamenti bilingui e/o multilingui). In particolare si faranno convergere in una struttura coordinata le informazioni - a livello morfologico, sintattico, semantico - formulate secondo il modello PAROLE/SIMPLE, con le relazioni tra sensi codificate secondo il modello Ital- / EuroWordNet.

Si è progettata inoltre, in funzione di utilizzi applicativi, la creazione di un software che gestisca in modo unificato l'insieme delle risorse lessicali disponibili, con possibilità di adattarle alle esigenze dell'utente o del sistema che le usa.

Va da sé che i modelli adottati per rappresentare le informazioni nelle basi di conoscenza lessicali sono parte fondamentale e centrale di qualsiasi prototipo o sistema di trattamento del linguaggio naturale. Tra i problemi più dibattuti, sia da parte nostra sia nei centri più avanzati di LC possiamo citare:

- quali siano i metodi per decidere quali informazioni lessicali (soprattutto semantiche) siano necessarie/richieste da un determinato sistema di analisi/generazione;

- quale sia il limite entro il quale è conveniente codificare in un lessico la rappresentazione di informazioni (anche in questo caso soprattutto di tipo semantico) senza avere previamente la evidenza della loro utilizzabilità in un sistema concreto;
- quali siano le distinzioni tra 'sensi' che, riportate dai normali dizionari di consultazione, 'vale la pena' di mantenere nella codifica delle entrate di una base di conoscenza lessicale.

Particolarmente istruttivo è a questo riguardo l'esercizio di SENSEVAL, un'iniziativa internazionale per la valutazione di sistemi di disambiguazione semantica (effettuata attraverso corpora annotati prevalentemente sulla base di reti semantiche alla WordNet), cui partecipiamo anche come co-organizzatori.

b) Acquisizione "dinamica" di informazioni da fonti testuali

Un aspetto essenziale, con riflessi sia teorici sia pratici, dello sviluppo e dell'arricchimento della BCL è costituito dall'interazione di modi di estensione manuali e automatici o semi-automatici. Questi due modi di arricchimento sono legati ai concetti complementari di risorse lessicali "statiche" e "dinamiche". Da un punto di vista teorico è impossibile che una risorsa lessicale "statica", comunque estesa, abbia copertura adeguata rispetto a qualsiasi corpus e/o esigenza applicativa. Da ciò consegue la necessità di associare a nuclei "statici" di risorse lessicali "di base" strumenti sempre più raffinati di acquisizione "dinamica" di informazioni lessicali (ai diversi livelli di descrizione linguistica) a partire da diversi tipi di corpora - in modo da riflettere l'uso della lingua in contesti comunicativi reali - e/o per diverse esigenze applicative. L'acquisizione deve consistere nell'arricchimento o nell'adattamento delle risorse di base a diversi tipi di testi e/o applicazioni. Tale processo di acquisizione comporta un ciclo di: *i*) analisi/annotazione di corpora, *ii*) acquisizione di informazioni a un determinato livello linguistico, *iii*) valutazione dei risultati, *iv*) ritorno a *i*) per una analisi/annotazione o più completa o a un livello linguistico più 'alto'.

Corpora testuali, orali, e multimodali

I corpora testuali, collezioni di testi in formato elettronico, costituiscono:

- la fonte naturale di conoscenza per studiare e descrivere le caratteristiche dell'uso delle lingue in diversi contesti comunicativi, sia in ambito monolingue, sia in ambito contrastivo plurilingue, da utilizzare nel programmare i componenti dei sistemi computazionali che devono trattare automaticamente testi o enunciati prodotti in tali contesti;
- il riferimento rispetto al quale valutare le *performances* di metodi e sistemi dell'ingegneria linguistica.

Seguendo le specifiche formulate da EAGLES, stiamo attuando *i)* l'estensione continua di un corpus di riferimento dell'Italiano, inserito in un piano di creazione di un *monitor corpus*, e cioè un corpus di riferimento aggiornato periodicamente per riflettere le caratteristiche evolutive della lingua nel tempo, finalizzato all'impiego in applicazioni di mercato che non possono prescindere dall'utilizzo di sempre maggiori quantità di dati testuali opportunamente codificati e annotati a livello linguistico, *ii*) la creazione di un primo nucleo di corpora paralleli (costituiti da testi tradotti in lingue diverse), la cui richiesta sta crescendo da parte degli operatori industriali, *iii*) la annotazione di corpora di scritto e di parlato a diversi livelli di descrizione linguistica, *iv*) la creazione di componenti software per l'annotazione automatica multi-livello in XML di corpora testuali e orali, *v)* la produzione, a partire dal corpus di riferimento, di un nuovo Lessico di Frequenza della Lingua Italiana, *vi)* la raccolta e annotazione di corpora multimodali.

Per poter sfruttare adeguatamente i corpora che vengono raccogliendosi per molte lingue, è necessario:

- definire criteri oggettivi per la scelta della composizione di corpora di diverso tipo, così che possano essere considerati campioni "adeguati" di un dato universo linguistico;

- approntare metodi per automatizzare al massimo la annotazione dei corpora a livelli linguistici e extralinguistici di complessità crescente, in particolare rispetto a classi di domini e compiti applicativi specifici;
- mettere a punto metodi e annotazioni che permettano la estrazione (semi-)automatica di conoscenze linguistiche dai corpora;
- accrescere la varietà dei corpora relativi a domini e scopi applicativi diversi (per es., fra le lingue di specialità, in particolare corpora di linguaggio infantile sono necessari per studiare l'evoluzione della competenza linguistica, anche allo scopo di fornire informazioni per lo sviluppo di strumenti didattici e di componenti per il settore *consumers*.

Annotazione linguistica multi-livello di corpora dialogici

Grandi quantità di dati orali annotati sono necessarie per modellare gli effetti delle diverse sorgenti di variabilità (ambiente e contesto acustico, modi e canale di comunicazione, stato emotivo e sociale del parlante, dialetti, ecc.) su unità linguistiche quali fonemi, pronuncia di parole e sequenze di parole, ecc. La maggioranza di strumenti e tecnologie per il trattamento automatico del linguaggio attualmente disponibili sono state progettate per applicarsi al linguaggio scritto e monologico. La conseguenza più evidente è che gli strumenti applicativi basati sulle tecnologie esistenti mal si adattano ad essere impiegati per il trattamento dell'interazione dialogica, lasciando scoperta un'area di grande potenziale sviluppo nella società dell'informazione, considerato il ritmo corrente di crescita e diffusione tecnologica nel settore delle telecomunicazioni. Questa ricerca si propone di *a*) validare le specifiche linguistiche per l'annotazione del dialogo ai livelli morfo-sintattico, sintattico e pragmatico su un campione rappresentativo di dialoghi, *b*) estendere il corpus (200 dialoghi uomo-macchina e 200 uomo-uomo) raccolti, nel quadro del progetto nazionale TAL (v. infra) in collaborazione con i principali operatori industriali del settore, e annotare in modo incrementale un corpus di mille dialoghi, utilizzando strumenti di pre-

annotazione automatica conformi alle specifiche e di un editor per l'annotazione manuale.

Corpora multimodali

Per modellare il dialogo interattivo, sia uomo-uomo che uomo-macchina, e operare adeguatamente nella recente tendenza verso un uso *human-centered* dei calcolatori, è necessario disporre di risorse multimodali (parlato, scritto, gesti, ecc.). Si comunica non solo attraverso le parole, ma anche attraverso l'intonazione, lo sguardo, le mani, i gesti, l'espressione facciale, ecc. Nella comunicazione, queste modalità si integrano e si complementano per fornire diversi tipi di informazione. C'è oggi un impeto crescente all'interno dell'HLT a considerare tutti i diversi canali comunicativi. Utilizzando i risultati di un esperimento svolto all'interno di ELSNET, assieme alle raccomandazioni del gruppo NIMM di ISLE, l'ILC sta iniziando la creazione di corpora multimodali e la loro annotazione, ed è iniziato il disegno di strumenti software per la gestione e l'annotazione di dati multimodali, in collaborazione al progetto HTL-NITE.

Architettura modulare di componenti software per l'annotazione automatica multi-livello in XML di corpora di lingua scritta e parlata

È in fase di studio un'architettura software per l'annotazione di testi a diversi livelli di analisi linguistica, laddove ciascun livello è concepito come concettualmente autonomo e dichiarativo, ancorato a un documento comune contenente il testo di base (grezzo o "emendato"), ed eventualmente interfacciato con gli altri livelli di annotazione attraverso puntatori XML. La modularità dell'architettura dev'essere tale da rendere naturale l'innesto di ulteriori livelli di annotazione sul tronco fondamentale rappresentato dai seguenti livelli: morfologico, morfosintattico (tagging), sintattico a costituenti immediati (chunking), funzionale, delle relazioni anaforiche e pragmatico. Chiaramente, ciascuno dei moduli sviluppati per l'annotazione "per livello" può essere

utilizzato come componente funzionalmente indipendente, da integrare in sistemi software destinati ad applicativi specifici.

4.1.2. *Metodi e strumenti per la ricerca umanistica*

Lo studio, la sperimentazione di prototipi, l'utilizzo di metodi computazionali a supporto della ricerca di diverse discipline umanistiche, in particolare per quanto riguarda le discipline linguistiche, letterarie, filologiche e lessicografiche, hanno costituito, in un certo senso, la piattaforma sulla quale si è venuto costituendo il nostro gruppo. A livello internazionale, l'attività dell'ILC si ispira ai seguenti principi:

- promuovere l'interscambio di conoscenze, metodi, risorse, strumenti con gli altri filoni della LC;
- incorporare nelle procedure studiate per gli umanisti le possibilità offerte dai più recenti sviluppi tecnologici;
- basare il disegno dei componenti e degli strumenti computazionali sull'analisi del *modus operandi* delle diverse discipline umanistiche, così da costruire delle stazioni di lavoro dedicate modulari e flessibili;
- promuovere la consapevolezza del ruolo dello *Humanistic Text Processing* nella Società dell'Informazione, e incoraggiare gli umanisti ad assumere il ruolo che ad essi compete attraverso l'utilizzo di metodi innovativi e proposte di progetti nazionali e internazionali.

Si deve notare che l'evoluzione tecnologica sta gradualmente arricchendo il panorama dei possibili utilizzi del calcolatore nelle scienze umane: accanto alla dimensione, per così dire, inizialmente dominante del *text processing*, i nuovi media oggi disponibili (suono, immagini, colori) permettono nuovi sviluppi metodologici nell'utilizzo del calcolatore nelle discipline umanistiche. Questo apre nuove prospettive di dialogo tra partner socio-economici e le scienze umane, fornitrici privilegiate di contenuto linguistico,

letterario, culturale per i nuovi "media" digitali.[8] Ma è da notare come il trattamento delle immagini sia divenuto un prezioso sussidio anche per le applicazioni di analisi testuale e filologica.[9] Le possibilità offerte al ricercatore variano a seconda dei suoi interessi disciplinari specifici: dall'ausilio negli studi stilometrici, all'utilizzo di meccanismi ipertestuali e di *hyperlinking*, all'uso di tecniche di *data-mining* su dati storici, documenti, testi letterari, allo studio delle proprietà formali delle strutture dei dati e degli algoritmi che assistono il ricercatore nei suoi compiti istituzionali, all'utilizzo di RL, metodi e tecniche di analisi linguistica (semi)automatica per la individuazione di informazioni, strutture, *features* linguistiche in corpora di lingua, antica o contemporanea, di ragazzi o di adulti, di diverse comunità sociolinguistiche, e per lo studio della loro evoluzione. Le ricerche si sono sviluppate in due direzioni, distinte ma complementari.

Metodi e tecnologie per basi di dati testuali e linguistici multifunzionali

Nel primo filone si sono, per così dire, tradotti, in forma interattiva (interamente utilizzabili su Internet) i programmi utilizzati per decenni dagli utenti del CNUCE per gli spogli elettronici dei testi. Si sono progressivamente aggiunte funzionalità diverse: tagger morfosintattici a base statistica; estrema flessibilità dell'algoritmo di contestualizzazione; visualizzazione di dati e computi statistici in tempo reale; capacità di operare su più alfabeti e lingue

[8] Ciò apre nuove opportunità di impiego per operatori culturali capaci di utilizzare le tecniche informatiche: non è un caso che le richieste di corsi al nostro Istituto vengano in particolare in questo settore.

[9] Mentre alcune ricerche del nostro Istituto utilizzano la multimedialità a scopi didattici o di apprendimento, ed altre si interessano alla creazione di un ambiente software per l'etichettatura semiautomatica multimediale di video (che intendiamo combinare a scopi applicativi con l'uso di ontologie nelle quali ai concetti sono collegati - ove possibile - immagini o *frames* prototipali), le ricerche del settore filologico-testuale utilizzano l'associazione testo-immagine, sia nel contesto - per così dire applicativo - dei servizi culturali ai cittadini (per es. nel settore delle *digital libraries*), sia nel contesto delle ricerche più propriamente linguistico-filologiche (per es. per trattare l'immagine del testo e per allinearla alle sue trascrizioni).

contemporaneamente; interrogazione del testo attraverso gruppi (famiglie) di parole, formulati dall'utente, o derivati automaticamente da basi di conoscenze lessicali associate (parole legate tra loro da relazioni semantiche di vario tipo - sinonimia, iponimia, iperonimia, ecc.-, per es., tratte da ItalWordNet); sistemi per l'allineamento automatico di testi paralleli; ricerca automatica di equivalenti di traduzione in corpora paralleli comparabili; ecc. Queste funzionalità si sono rivelate di grande utilità per gli studi linguistici, letterari, di storia del pensiero, ecc. Tutti questi moduli sono integrati in una vera e propria stazione lessicografica, che include componenti atti ad assistere il lessicografo nei suoi compiti istituzionali (disegno della struttura dell'entrata, scelta e classificazione dei contesti; ecc.)

Filologia computazionale

Nel secondo filone, ci si è concentrati soprattutto sul trattamento congiunto di immagine e trascrizione del testo, in particolare al servizio dei filologi. L'obiettivo generale è quello di creare un insieme di metodi, strumenti software, dati sperimentali, risorse linguistiche integrate in una stazione di lavoro che consenta ai ricercatori umanisti, in particolare filologi, l'uso agevole ed efficace di diverse tecnologie, allo stato dell'arte della LC, della intelligenza artificiale, del trattamento delle immagini nello studio dei testi.

È da notare come, con lo sviluppo della tecnologia digitale per le biblioteche e per gli archivi, si siano aperte nuove frontiere per la ricerca filologica e linguistica, oltre che per la conservazione e la fruizione dei beni librari. La ricerca, nella quale integriamo prodotti ideati dall'Istituto[10] con prodotti industriali specializzati e con prodotti disponibili sul mercato, offre una serie di funzionalità,

[10] Sia per il trattamento dei testi, sia per il trattamento delle immagini, i prodotti commerciali disponibili non sono assolutamente sufficienti: essi non possiedono la necessaria specializzazione richiesta dalla particolarità delle applicazioni, che trova origine nella struttura dei dati formali specifici della disciplina e nella funzionalità degli algoritmi che utilizzano questa struttura, e devono essere basati sulla analisi metodologica dell'operare del ricercatore.

che possono essere combinate in vario modo a seconda della applicazione in corso, e che possono essere così riassunte:

- *funzionalità di data base management*: funzionalità coordinate di *information retrieval*, gestione dei diversi livelli informativi di annotazione, supporto al lavoro collaborativo (gestione di apporti multipli a un unico documento), funzionalità di input/output, standard e gestione dei formati descrittivi conformi a diversi sistemi di metadata in vigore;
- *funzionalità di image processing*: supporto alla segmentazione dell'immagine, al matching dinamico testo-immagine, *enhancement* e restauro elettronico, scansione automatica delle immagini, riconoscimento automatico di caratteri a stampa antichi assistito da un componente linguistico-computazionale, protezione dell'integrità ed autenticità dei documenti, gestione dei formati multirisoluzione;
- *funzionalità per il lavoro filologico:* supporto alla analisi del testo e alla gestione della annotazione attraverso il ricorso a fonti di conoscenza linguistica (indici statistici, thesauri, lemmatizzatori, ecc.), al collegamento ipermediale di più documenti tra loro, nonché alla gestione del tagging normalizzato (SGML, HyTime e XML);
- *funzioni di supporto e di utilità*: funzioni per l'annotazione a più livelli, separata dal documento, funzioni per seguire un proprio percorso di segmentazione, indici verbali e visivi delle parole e dei luoghi, ecc.

In particolare, allo scopo di potenziare questo filone, ho promosso o sto promuovendo le seguenti iniziative a livello internazionale:

- una serie di incontri tra specialisti per disegnare una 'roadmap' del settore, per uscire dall'impasse attuale, che è caratterizzata da un continuo affinamento delle tecnologie cui non sempre corrispondono innovazioni metodologiche sostanziali sul piano della ricerca scientifica;
- una serie di tavole rotonde, in seno per es. alle Conferenze dell'ALLC-ACH, nelle quali rappresentanti di Agenzie come NEH, ESF, ecc. tipicamente interessate alle scienze umane,

rappresentanti della Società dell'Informazione (NSF, CE, ecc.), ricercatori ed eventuali utilizzatori commerciali discutano priorità, strategie, possibilità di collaborazione, e i metodi più efficaci per inserire il contributo delle discipline umanistiche nella società dell'informazione;
- un incontro tra lessicografi computazionali, lessicografi "accademici", lessicografi "commerciali", per fare il punto della situazione, a distanza di oltre 20 anni dal Convegno patrocinato dalla ESF a Pisa nel 1981, per esplorare e valutare quali sinergie i nuovi sviluppi tecnico-scientifici consentano di stabilire.

È da notare che, appunto per le potenziali connessioni con il contenuto culturale (che hanno per esempio indotto la CE a lanciare un programma - al di fuori del Programma Quadro - chiamato eContent) le ricerche di questo settore stanno trovando sempre più spesso complementi e supporti in varie iniziative[11].

Nel settore poi dell'*electronic publishing*, si moltiplicano le imprese editoriali che fanno uso del nostro software e dei nostri metodi, per pubblicazioni testuali e multimediali, su CD, HDVD, ecc. Sono da segnalare in particolare le imprese lessicografiche, sia per la produzione di dizionari "tradizionali" su CD-ROM, sia per la pubblicazione di dizionari "innovativi", connessi da un lato a descrizioni di vari livelli linguistici, dall'altro a testi di riferimento annotati.

4.1.3. *Modelli e metodi per il trattamento delle lingue naturali, e prototipi applicativi mono e multilingui*

Come detto in precedenza, uno dei risultati delle prime edizioni della Scuola Estiva Internazionale è stato quello di far conoscere fin nei dettagli operativi i metodi per l'analisi e la generazione di frasi in linguaggio naturale (per es. di tipo ATN e/o CHART) che, basati su regole formali e su meccanismi di inferenza o deduzione

[11] Per es. progetti speciali (BIBLOS) e strategici del CNR (Beni Culturali), progetti MURST (ex 40%, CIBIT; Programmi Parnaso: Bibliofilo), progetti del Ministero del Lavoro (ADAPT: TECLA), progetti CE (BAMBI, MEMORIA) e CE-NSF (CHLT).

da basi di conoscenza, hanno caratterizzato la LC nei primi decenni. Buona parte delle attività dell'istituto sono state dedicate ad approfondire questi aspetti; per es., abbiamo scritto frammenti sempre più estesi di grammatiche formali dell'italiano, destinate a parser di vario tipo. Quando, in particolare con l'affermazione del paradigma dell'industria delle lingue, la robustezza, la copertura linguistica e la riduzione del numero di strutture "overgenerate" sono diventati requisiti fondamentali per i componenti di analisi dei testi di input, si è dovuto ricorrere, come in molti altri paesi all'utilizzo di modelli statistici e a metodi *shallow*.

a) Modelli statistici per l'induzione di modelli computazionali della lingua

La necessità di confrontarsi con la varietà e complessità di linguaggi tecnici e settoriali, nonché con la frammentarietà e dipendenza contestuale del linguaggio in uso, ad esempio nell'interazione dialogica per il conseguimento di un risultato (transazione economica, scambio di informazioni ecc.), richiede l'utilizzo di metodi induttivi che, attraverso l'applicazione di algoritmi di apprendimento automatico per il reperimento, la organizzazione e la gestione delle conoscenze linguistiche contribuiscano a conferire maggiore robustezza e copertura agli strumenti sviluppati deduttivamente. In questo settore, l'istituto è attivamente impegnato nella messa a punto di una serie di tecniche per l'*induzione* di modelli computazionali del linguaggio scritto e parlato a partire da evidenza linguistica attestata, eventualmente annotata a uno o più livelli di analisi. In particolare, la ricerca si articola intorno a quattro livelli di analisi linguistica:

- morfologia: tecniche di reperimento e riconoscimento automatico di classi di unità lessicali di importanza cruciale per il *processing* robusto di testi reali: derivati, composti, unità polirematiche, ecc.
- sintassi: sviluppo di modelli di correlazione tra i) etichette morfosintattiche, ii) costituenti sintattici non ricorsivi (o *chunks*), iii) etichette funzionali, ai fini dello sviluppo di

tecniche markoviane e analogiche per l'analisi di sequenze di dette unità;
- semantica-lessicale: sviluppo di tecniche per l'induzione automatica di classi lessico-semantiche e tassonomie lessicali sulla base di domini tecnici specifici;
- pragmatica: sviluppo di tecniche per l'acquisizione di misure di correlazione tra strutture linguistiche e atti comunicativi.

Il metodo di ricerca prevede due fasi:

- Fase A: indagine esplorativa delle cause statisticamente critiche di fallimento dell'analisi automatica e studio preliminare delle correlazioni tra unità linguistiche pertinenti per ciascuno dei suddetti livelli di analisi.
- Fase B: sviluppo di componenti software per l'induzione di modelli computazionali per i suddetti livelli.

b) Shallow parsing

Si sono recentemente disegnati e si stanno implementando analizzatori il cui obiettivo è non solo quello di analizzare in maniera robusta testi di vario tipo, ma anche di individuare nuclei e relazioni sintattiche rilevanti per successive elaborazioni che mirino ad identificare (a vari livelli di granularità) aspetti più propriamente "semantici" (classificazione semantica, links a thesauri e ontologie, ecc.). Il sistema di analisi sintattica si basa su due strumenti, il "Chugger" e l'Analizzatore Funzionale.

Il Chugger implementa tecnologie di parsing a stati finiti e realizza contemporaneamente l'etichettatura morfo-sintattattica delle parole (identificazione della categoria sintattica con cui una forma occorre in un dato contesto linguistico) e la segmentazione del testo in costituenti sintagmatici non ricorsivi (chunks). Esso unisce dunque le funzionalità tipiche di un tagger e quelle di un chunker. Il Chugger è il risultato dell'evoluzione di un modulo pre-esistente, *Chunk-it*, che realizza la segmentazione in chunks di testi precedentemente taggati e disambiguati.

L'analizzatore funzionale prende in input l'output del Chugger e, basandosi su una grammatica a dipendenze a stati finiti, riconosce le principali relazioni grammaticali tra gli elementi nella frase: es. identificazione del soggetto, dei complementi, dei nuclei nominali complessi, ecc. Rappresenta il componente fondamentale per l'estrazione di informazione semantica. Esso comprende una interfaccia con un database per l'utilizzazione di informazione lessicale, sia sintattica che semantica, estratta dai lessici computazionali disponibili.

L'insieme di questi componenti, più altri strumenti di analisi di base, sembra ora costituire una procedura pratica utilizzabile in diverse applicazioni che richiedono una analisi "robusta" dei testi di input[12]. Inoltre questi componenti hanno destato l'interesse di ditte che operano sul Web, le quali si propongono di incorporare nei propri prodotti le analisi da essi effettuate, per migliorare la 'performance' dei propri sistemi. Queste ditte danno alla ricerca un importante contributo, fornendo un ambiente applicativo già funzionante e degli utenti già attivi.

Ciò non esclude la continuazione delle ricerche secondo l'approccio - per così dire - tradizionale, o 'classico', degli anni '70 e '80, che prosegue con l'obiettivo di creare un ciclo di analisi teorica, progettazione, sperimentazione e metodologia per le principali applicazioni del NLP. Per esempio, la richiesta di programmi di gestione dell'informazione linguistica e testuale, sempre crescente con l'estendersi dei servizi distribuiti e di rete, impone l'attuazione di modelli e prototipi sempre più avanzati sia nel settore del reperimento e gestione di documenti on-line, sia per quanto concerne interfacce flessibili e modulari. Gli scenari di applicazione sono molti, ma riconducibili a moduli e prototipi ben precisi, come l'interpretazione di *queries* espresse nella forma più naturale, il riconoscimento di *pattern* informativi nei documenti resi disponibili in rete, i sistemi di presentazione dell'informazione, la generazione di frasi in linguaggio naturale. La nostra ricerca:

- svolge lavoro sperimentale relativamente alle tecniche di interpretazione del linguaggio naturale, di formalizzazione e

[12] Per es. in due progetti recentemente approvati dalla CE: MLIS-MUSI e POESIA.

reperimento dell'informazione, di generazione, anche multimediale, di risposte;
- produce programmi di analisi e gestione dell'informazione linguistica indipendenti dal tipo di applicazione e dal deposito di conoscenze (sintattiche, lessicali, semantiche, etc.) eventualmente disponibile, programmi che possono essere combinati modularmente per formare prototipi di applicazioni;
- fornisce prototipi preindustriali di sistemi di gestione dell'informazione;
- offre metodologie di analisi di problemi applicativi e di costruzione di sistemi di gestione e presentazione dell'informazione.

Si esplorano ed impiegano sia modelli basati su regole formali, sia metodi per l'induzione di modelli a partire dall'analisi dei dati, sia modelli basati sulla rappresentazione della conoscenza, sia metodi per l'utilizzo congiunto dei tre approcci citati. La ricerca, la costruzione e la verifica di modelli computazionali delle diverse operazioni linguistiche sono indispensabili sia per lo studio delle competenze linguistico-cognitive coinvolte nell'uso delle lingue attraverso implementazioni e simulazioni, sia per la realizzazione di prototipi e di prodotti adeguati alle esigenze di applicazioni innovative che si basano sul, o includono il, trattamento automatico delle lingue naturali. Un chiaro esempio è la "computazione" (in qualche forma) del significato, il cui trattamento è indispensabile per la realizzazione di nuovi strumenti che rendano più efficaci i processi di comunicazione, di recupero dell'informazione, di comprensione dei testi, ecc.

Gli obiettivi applicativi che sono attualmente perseguiti, sono i seguenti:

<u>Modelli di gestione dell'informazione testuale per servizi di pubblica utilità</u>

I testi non presentano la stessa "densità di informazione": alcune porzioni possono essere considerate come irrilevanti rispetto alla classificazione del loro contenuto. Lo sviluppo di tecniche di preprocessing ed eliminazione di parti irrilevanti, nonché di analisi locale (ad isole) delle parti rilevanti è un momento essenziale

dell'adattamento delle tecniche di analisi proprie del NLP a specifiche applicazioni.

Domande e offerte di lavoro: servizio automatico su Internet

Lo scopo è di fare incontrare efficacemente domanda e offerta da una parte, e di programmare corsi di formazione professionale tenendo conto della richiesta del mercato del lavoro dall'altra. Il servizio è previsto per Internet. L'attività prevista consiste nella sperimentazione e valutazione di tecniche e strumenti di analisi per l'estrazione dal testo di informazione a vari livelli (lessicale, concettuale, sintattico) e nel disegno di un modello capace di "ragionare" sulle informazioni estratte dal testo per "costruire" la descrizione delle competenze che caratterizzano il profilo dell'offerta.

Interrogazione in linguaggio naturale di una base di conoscenza di origine manualistica

La formulazione di domande di chiarificazione non richiede un'interpretazione "profonda" e completa; è sufficiente, per lo più, identificare gli elementi significativi della domanda, che, attraverso un meccanismo di mapping, permettono di identificare l'insieme dei fatti che possono costituire una risposta. Anche in questo caso, obiettivo della ricerca è identificare e sviluppare le tecniche per l' eliminazione delle porzioni "inutili" di ogni domanda e per l'analisi sintattico-semantica delle parti utili, allo scopo di ottenere risposte pertinenti da una base di conoscenze.

Multilinguismo e servizi di rete

I servizi di gestione dell'informazione testuale acquisiscono maggior rilevanza se inseriti nel contesto di ricerca multilingue in rete. La sutura tra le tecniche di multilinguismo e quelle di trattamento dei documenti non è né immediata né ovvia; tuttavia è un passaggio obbligato verso la messa in opera di un reale servizio internazionale. Il progetto UNL promosso dalla UNU/IAS (*United Nations University/Institute for Advanced Studies*), che coordino

per l'Italia, come Direttore dell' *Italian Language Center*, sta studiando la fattibilità di costruire un servizio permanente disponibile in rete, attraverso il quale i documenti espressi in un linguaggio naturale (*source language*) vengono tradotti in un formato "universale" (l'*Universal Networking Language* - UNL -, basato su un insieme di relazioni logiche molto generali) per poi essere generati nelle varie lingue *target*. Il progetto UNL sta costituendo una rete di "centri nazionali", ciascuno dei quali ha il compito di fornire i programmi necessari per tradurre testi scritti nella rispettiva lingua nazionale nell'UNL, e viceversa. Le lingue finora incluse nel network sono: italiano, tedesco, francese, spagnolo, portoghese-brasiliano, russo, lituano, arabo, indù, indonesiano, mongolo, cinese, giapponese e swahili.

4.1.4. *Tecnologia della lingua per la didattica e la disabilità*

Una particolare applicazione degli strumenti e delle tecniche della LC è la loro utilizzazione nella didattica e nella didattica speciale[13]. È stato realizzato, in collaborazione con il Dipartimento di Informatica di Torino, *Addizionario*, un laboratorio linguistico ipermediale per lo studio dell'italiano sia come lingua materna, sia come lingua straniera. Si tratta di un ambiente didattico interattivo che offre un contesto motivante ad attività di solito ritenute abbastanza difficili e noiose, quali la consultazione di un dizionario e l'arricchimento del lessico, e fa emergere le conoscenze e i vissuti che il bambino associa alle parole per poi servirsene come aggancio per fissare le conoscenze nuove. Nel laboratorio linguistico sono disponibili due strumenti multimediali strettamente interrelati: *Addizionario*, un dizionario per bambini, scritto e illustrato dai bambini e *Quaderno attivo*, lo strumento creativo per mezzo del quale i bambini possono costruire il loro dizionario personale, modificando e adattando alle loro esigenze l'informazione importata da Addizionario, o aggiungendo, alla lista delle 1000 parole del dizionario di base, parole nuove, corredate di

[13] Il progetto speciale del CNR "Uso del calcolatore nell'insegnamento nelle lingue di specialità" aveva permesso un primo studio di fattibilità.

una vasta gamma di informazioni quali la definizione, gli esempi d'uso, i disegni, i suoni, ecc.
Il prodotto è destinato a diverse categorie di utenti: gli alunni della scuola dell'obbligo che, per arricchire il loro lessico, avranno a disposizione uno strumento gradevole e coinvolgente; gli insegnanti che potranno utilizzarlo per preparare unità didattiche disegnate specificatamente per soddisfare le necessità dei loro studenti; gli psicologi e terapisti che se ne potranno servire per la diagnosi dei disturbi dello sviluppo e dell'apprendimento del bambino; e infine i redattori dei dizionari per giovanissimi che potranno accedere a una grande quantità di materiali da cui prendere spunto per realizzare dizionari per bambini sempre più "attraenti", facili da usare e rispettosi delle capacità, dei gusti e degli interessi dei loro utenti.

La linea di ricerca si propone ora una larga sperimentazione di utilizzo di Addizionario in diversi tipi di contesto sociale del nostro e di altri paesi[14], sia con soggetti normali che portatori di handicap. L'analisi di un corpus di linguaggio infantile dovrebbe fornire informazioni per il perfezionamento degli strumenti di ausilio didattico.

4.1.5. *Coordinamento di attività nazionali e internazionali*

Ho già descritto sopra alcuni degli scopi di questa attività, consistente nell'ideare, proporre, organizzare, coordinare iniziative che mirino a promuovere lo stato dell'arte attraverso collaborazioni che assicurino la convergenza degli sforzi (tra diversi tipi di attori italiani e di altri paesi, pubblici e privati) verso l'innovazione in settori che le nostre analisi indicano come cruciali - sotto il profilo scientifico-tecnico e organizzativo - per uno sviluppo della disciplina che risponda alle esigenze della Società globale, e in particolare ai bisogni prioritari del nostro paese.

[14] Sperimentazioni sono in corso in Galles, in Mexico e in Spagna.

Riassumendo:

A livello nazionale

Tra le iniziative di carattere nazionale che hanno visto protagonista l'Istituto di Linguistica Computazionale a livello di promozione e coordinamento sono da menzionare in particolare le seguenti.
Un network nazionale che riunirà progressivamente i diversi tipi di attori interessati al trattamento automatico delle lingue nel nostro paese, allo scopo di identificare bisogni e priorità, definire programmi di lavoro, monitorare i progressi, evitare la duplicazione degli sforzi, creare convergenze e promuovere la condivisione di risorse linguistiche, facilitare il trasferimento delle tecnologie, integrare diversi tipi di competenze, sviluppare nuovi percorsi formativi, ecc.

Programmi a carattere nazionale che - utilizzando ed estendendo conoscenze e risultati prodotti dall'Istituto - rispondono ad esigenze riconosciute come prioritarie per lo sviluppo della comunità nazionale (si vedano in particolare i due progetti descritti nella sezione seguente). Pressoché tutti i temi nei quali si articolano questi programmi corrispondono a ricerche e studi dell'Istituto. L'Istituto può così rafforzare le proprie attività di ricerca attraverso sinergie e collaborazioni istituzionalizzate con i più accreditati attori nazionali, svolgendo appieno quella funzione di coordinamento nazionale per il settore della LC che è tra i compiti istituzionali previsti dal suo Statuto, e i suoi ricercatori possono contare su risorse umane e finanziarie addizionali.

A livello internazionale

Anche in questo ambito mi sono impegnato a perseguire, con l'aiuto dei miei colleghi dell'ILC, il compito assegnato dallo Statuto dell'ILC (stabilire gli opportuni collegamenti tra attività italiana e attività internazionale), attraverso iniziative rivolte a potenziare le sinergie con i centri di ricerca dei paesi più avanzati - europei ed extraeuropei, pubblici e privati - e promuovere lo sviluppo internazionale dello stato dell'arte in settori strategici, operando nel contempo affinché l'italiano sia incluso nella rete multilingue della

società globale. Queste iniziative prendono varie forme organizzative. Per esempio:

- Accordi bilaterali (attualmente con Istituti di Giappone, USA, Bulgaria, Spagna, Cuba, Messico, Argentina, Francia).
- Partecipazione alla attività di Associazioni, spesso con responsabilità di coordinamento attraverso la presidenza di dette Associazioni (per es. ELRA, PAROLE, ALLC, EURALEX, ecc.), o comunque con la partecipazione ai loro comitati direttivi (ICCL, AILA, ACL, ACH, SIGLEX, ecc.)
- Organizzazione di eventi internazionali: per esempio la *International Conference on Language Resources and Evaluation*; diversi workshop su vari temi della LC, con partecipanti europei, asiatici, nordamericani; tavole rotonde in diversi convegni (COLING, LREC, ALLC, ecc.).
- Partecipazione a network internazionali, quali il *Network of Excellence for Natural Language and Speech* (ELSNET), nel cui Management Board rappresento l'Italia, il gruppo di coordinamento dei progetti nazionali dei governi europei (ENABLER), il Comitato Internazionale per le Risorse Linguistiche, OntoWeb *Ontology-based information exchange for Knowledge Management and Electronic Commerce*.
- Promozione della partecipazione dell'Istituto a progetti internazionali (abbiamo partecipato a più di 30 progetti europei negli ultimi anni).
- Promozione della collaborazione tra l'Unione Europea, gli Stati Uniti, ed altri paesi tecnologicamente avanzati, nel settore del trattamento automatico della lingua (TAL).

5. PROGRAMMI DI INTERESSE NAZIONALE

Due programmi di interesse nazionale[15] sono stati recentemente definiti e proposti dall'ILC sulla base delle esperienze, conoscenze, prospettive acquisite attraverso l'attività internazionale

[15] ITAL, progetto speciale del CNR, aveva consentito di compiere un primo, sia pure ridotto, studio di fattibilità.

precedentemente illustrata e la direzione delle ricerche dell'ILC. I due programmi di interesse nazionale sono:

a) TAL ("Infrastruttura nazionale per le risorse linguistiche nel settore del trattamento automatico della lingua naturale parlata e scritta"), del costo complessivo di circa 5 miliardi, finanziato dal MURST per un totale di circa 3,5 miliardi nell'ambito della legge 46/82 art.10. Il progetto, affidato ad un gruppo di 13 enti privati[16], come previsto dalla legge costitutiva, si è concluso nell'autunno 2001.
b) Il Piano "Linguistica Computazionale: ricerche monolingui e multilingui", del costo complessivo di circa 9 miliardi, finanziato dal MURST, nell'ambito della legge n.488 del 19/12/1992 (Cluster 18), con circa 6 miliardi, è articolato in 8 progetti, ciascuno affidato a un soggetto esecutore,[17] il quale si avvale peraltro di numerose collaborazioni articolate secondo svariate forme giuridiche.

5.1. PROPOSTA INIZIALE E OBIETTIVI DEI DUE PROGRAMMI NAZIONALI

Entrambi i programmi sono originati da una proposta iniziale, che è il frutto delle discussioni e delle conclusioni raggiunte all'interno di un Gruppo di lavoro di circa 30 persone, costituito nel 1996 presso il Ministero della PTT, nel quale erano inclusi rappresentanti di vari Ministeri, di Enti di ricerca pubblici e privati, Università, Industria, Fornitori di Servizi, Pubbliche Amministrazioni,

[16] I partner sono: CPR - Consorzio Pisa Ricerche; ITC - Istituto Trentino di Cultura; CSELT - Centro Studi e Laboratori Telecomunicazioni; SYNTHEMA; CVR - Consorzio Venezia Ricerche; CERTIA - Centro per la Ricerca, Sviluppo, Formazione nelle Tecnologie e Applicazioni Informatiche; QUINARY; ALCEO; COMPUTER SHARING; DELCO; GST - Gruppo Soluzioni Tecnologiche; INTERACTIVE MEDIA; NECSY - Network Control Systems.

[17] I soggetti esecutori sono rispettivamente: CPR, Pisa; CIRASS, Napoli; THAMUS, Salerno; ILC-CNR, Pisa; SYNTHEMA, Pisa; Istituto Universitario Orientale, Napoli; Dipartimento di Scienze Storiche del Mondo Antico, Università di Pisa; Sportello per la Cooperazione Scientifica e Tecnologica con i Paesi del Mediterraneo (SMED) del CNR, Napoli.

Categorie ed Associazioni Professionali di diverso tipo. Il gruppo mi incaricava di redigere un documento che individuasse gli obiettivi e le linee di lavoro prioritarie che avrebbero dovuto costituire la proposta di un programma nazionale inteso ad ovviare alle necessità più urgenti della comunità nazionale di R e S, soprattutto per migliorarne la competitività sul piano internazionale.
La suddivisione del lavoro proposto in due programmi separati è dovuta a ragioni burocratiche. Ritengo perciò utile elencare gli obiettivi principali senza mantenere una divisione tra i due programmi, ma ordinandoli piuttosto come nella proposta unica iniziale.

5.1.1. *ItalWordNet*

ItalWordNet è una risorsa tipo «rete semantica» per l'Italiano a copertura relativamente ampia (circa 50.000 entrate, in parte appartenenti al dominio finanziario/economico). ItalWordNet è strutturato secondo il modello di WordNet (costruito a Princeton), così come è stato arricchito ed esteso ad altre lingue nell'ambito dei progetti europei EuroWordNet. I sensi sono raggruppati in «synset» (gruppi di sensi sinonimi tra loro, nella terminologia di WordNet), i quali sono collegati da relazioni semantiche di vario tipo, quali iperonimia/iponimia, meronimia, ecc. La risorsa viene validata da un partner industriale attraverso l'uso in sistemi di Information Extraction.

5.1.2. *Basi di conoscenze lessicali per il trattamento computazionale dell'italiano*

La risorsa lessicale contiene informazioni codificate a diversi livelli descrittivi (fonologico, morfologico e sintattico per 55.000 entrate lessicali, e semantico per 55.000 sensi), formulate secondo il modello PAROLE-SIMPLE, opportunamente rivisitato. Viene anche implementato un software per la gestione delle risorse.

5.1.3. *Corpus di italiano parlato*

L'obiettivo precipuo è quello di avviare anche per l'italiano, al pari di quanto avviene per le lingue degli altri paesi industrializzati, la costituzione di un archivio fonico dell'italiano basato su un corpus di parlato articolato su più livelli stilistici, atto a garantire una sufficiente rappresentatività della variabilità diatopica, utile per le principali attività sia di ricerca linguistica di base che per i molteplici indirizzi applicativi individuabili nel settore industriale del trattamento automatico dei segnali vocali, per rendere più "naturale" il dialogo con il servizio. Il corpus è costituito da 100 ore di parlato articolate nei seguenti insiemi:

a) 10 ore di materiale radiotelevisivo (notiziari, interviste, *talk show*),
b) 60 ore di materiale raccolto sul campo (tramite la tecnica detta "map task"),
c) 5 ore di materiale letto in laboratorio da più parlanti (testi atti a garantire la copertura lessicale di base - LIP/VdB),
d) 10 ore di parlato telefonico (con copertura diatopica come in a e in b),
e) 10 ore di parlato acquisito per finalità applicative in domini specifici (lessico economico finanziario, informazioni ferroviarie etc., sia telefonico sia ortofonico).

5.1.4. *Treebank sintattico-semantica dell'italiano*

La Treebank è costituita da un corpus bilanciato annotato a livello sintattico (80.000 occorrenze annotate con struttura a costituenti; 360.000 occorrenze annotate a livello funzionale) e semantico (80.000 occorrenze annotate con i sensi di ItalWordNet). La disponibilità di una Treebank per l'Italiano con informazioni utili all'individuazione di modelli linguistici e alla valutazione dei risultati di diversi sistemi e tecniche è diventata una esigenza prioritaria per diverse applicazioni industriali.

5.1.5. *Acquisizione di informazioni linguistiche da corpora*

Il progetto prevede il disegno e la realizzazione di un sistema di acquisizione lessicale da corpora testuali, per arricchire dinamicamente risorse lessicali "statiche", con informazioni acquisite (semi)automaticamente da corpora di specialità. Il sistema deve essere capace di apprendere da testi liberi alcuni degli aspetti di conoscenza lessicale (relativi alle parole) che sono essenziali per le più svariate applicazioni di ingegneria linguistica, quali quadri di sottocategorizzazione e restrizioni o preferenze lessicali e semantiche sulle posizioni argomentali. Il sistema di acquisizione è costruito in modo da essere parametrizzabile per diversi tipi di variabili ed è usabile per diversi tipi di testi, rendendo così possibile incorporare dinamicamente nel lessico statico di base una conoscenza linguistica/lessicale per particolari domini, sottolinguaggi o contesti comunicativi.

5.1.6. *Dialoghi annotati per applicazioni di interfacce vocali avanzate*

Viene registrato, trascritto, annotato a diversi livelli di analisi (morfologica, sintattica, semantica, pragmatica, prosodica) un corpus di 450 dialoghi, metà uomo-uomo, metà uomo-macchina, nel dominio del turismo, per acquisire informazioni sui modi di accedere a servizi informativi mediante messaggio vocale.

5.1.7. *Risorse Grammaticali e Sistema integrato di Supporto allo Sviluppo di Applicazioni (SiSSA)*

SiSSA si propone di creare un sistema, integrante strumenti software e risorse linguistiche, che assista l'utente nella costruzione e validazione di applicazioni che comportano l'utilizzo di conoscenze grammaticali. SiSSA vuole inserirsi immediatamente dopo la fase iniziale di analisi del problema applicativo e di stesura delle specifiche, con l'obiettivo di permettere la prototipazione rapida di soluzioni a problemi applicativi concreti, accorciando i tempi di sviluppo, e facilitando la valutazione dei risultati conseguiti.

5.1.8. *Strumenti e ambienti di sviluppo software per interfacce vocali avanzate*

L'obiettivo è di mettere a disposizione dell'utente un sistema integrato di strumenti software, motori di riconoscimento/sintesi vocale ed interfacce di comunicazione verso le basi dati ed il mondo esterno, che lo supporti nello sviluppo di applicazioni vocali interattive di tipo complesso, in particolare basate sul parlato continuo e sul dialogo uomo-macchina in linguaggio naturale parlato. Un esempio di applicazioni sono i sistemi di accesso all'informazione via telefono in cui l'utilizzatore del servizio possa formulare le proprie richieste senza eccessivi vincoli, ed il sistema sia in grado di sostenere dei dialoghi su un dominio applicativo specifico. L'utilizzo di questo ambiente software integrato consente allo sviluppatore di progettare e testare un'applicazione vocale, anche complessa, mascherando la complessità delle varie tecnologie utilizzate (riconoscimento, sintesi, comprensione e gestione del dialogo, ecc.) e dei linguaggi di programmazione specifici, concentrando viceversa le sue competenze sulle funzionalità applicative e sugli aspetti ergonomici e di usabilità dell'interfaccia utente, con un significativo risparmio economico e di tempo.

5.1.9. *Supporti alla gestione di conoscenza e testi normativi*

Il risultato specifico cui si mira è la realizzazione di applicazioni dimostrative nelle aree di *authoring* e di sistemi di supporto decisionale normativo, con particolare riferimento alle normative territoriali. Scopo generale è studiare e valutare metodi capaci di identificare, in corpora di testi relativi a un dominio applicativo ben definito, caratteristiche linguistiche specifiche, che consentano, opportunamente utilizzate ed ingegnerizzate, di restringere la complessità e in particolare la varietà e il grado di ambiguità dei fenomeni linguistici da trattare, così da poter conferire ai sistemi applicativi il grado di robustezza e il livello di affidabilità richiesti per l'utilizzazione pratica da parte di utenti reali.

5.1.10. *Corpus bilingue italiano-arabo*

Si è costruito un corpus bilingue italiano arabo: 10 milioni di occorrenze per ciascuna lingua, 4 milioni delle quali provenienti da testi paralleli allineati. Inoltre 1 milione e mezzo di occorrenze di ciascuna lingua è annotato semi-automaticamente a livello morfosintattico.

5.1.11. *Promozione degli scambi linguistici e culturali con il mondo arabo*

Lo scopo è mettere i corpora così prodotti a disposizione delle diverse categorie di utenti potenzialmente interessati, attraverso metodi intelligenti di accesso su Internet, allo scopo di facilitarne l'uso, promuovendo nel contempo gli scambi culturali tra l'Italia e il mondo arabo.

5.1.12. *Network di operatori italiani della Linguistica Computazionale*

Il Network riunisce operatori di industrie, enti di ricerca, Pubbliche Amministrazioni, Associazioni Professionali, Università. Uno dei compiti del Network è la organizzazione di infrastrutture cooperative per raccogliere, mantenere, disseminare e mettere a disposizione le RL e gli strumenti di base a beneficio dell'intera comunità di ricerca e sviluppo della LC, innanzitutto nel nostro paese, ma anche a livello internazionale. Una delle dimensioni caratterizzanti il nostro settore è infatti il multilinguismo. Il network italiano è naturalmente connesso al network europeo ENABLER.

6. L'ILC NELLA NUOVA RETE CNR (2002 →)

Dal 1 gennaio 2002 l'ILC è divenuto un istituto della nuova rete CNR. Ciò è avvenuto grazie alla consolidata posizione di centro di eccellenza[18], a livello nazionale e internazionale, conseguito dall'ILC, e anche grazie alla costante attenzione a mantenere la sua identità e a integrare fra di loro i diversi aspetti della LC o TAL. Testimoniano il riconosciuto ruolo di *leadership* dell'ILC, fra l'altro, la capacità di: *i)* attrarre cospicui finanziamenti esterni

[18] Il Ministro Maccanico, nell'aprire il Convegno sul TAL nel 1997, riconobbe esplicitamente la nostra leadership, come pure il sottosegretario di Stato al MURST Tognon nel discorso introduttivo alla cerimonia di apertura di LREC nel 1998 (Granada).

(internazionali e nazionali) in canali di natura estremamente competitiva, *ii)* influenzare le grandi visioni strategiche implementate da Enti nazionali e internazionali, *iii)* essere presente in ruoli direttivi nei maggiori organismi internazionali del settore, e recentemente *iv)* partecipare a numerose *Expressions of Interest* per *Networks of Excellence* e *Integrated Projects* per il 6° PQ (2002-2006), e *v)* essere promotore di una *Expression of Interest* che raggruppa i maggiori centri di LC in Europa e nel mondo.

Nell'atto istitutivo dell'Istituto figurano i tre grandi settori tematici del nuovo organo, che presentano una totale continuità rispetto alle linee di ricerca appena descritte, che in essi confluiscono naturalmente. All'interno di questi grandi settori si raggruppa sia l'attività istituzionale dell'ILC sia l'attività di progetti finalizzati e finanziati dall'esterno. I tre settori sono:

- Disegno di standard e costruzione di risorse linguistiche computazionali.
- Modelli e metodi per il trattamento delle lingue naturali, e prototipi applicativi mono e multilingui.
- Metodi e strumenti computazionali per la ricerca umanistica, con particolare riguardo alle discipline linguistiche, letterarie, filologiche e alla lessicografia.

Il nuovo ILC intende continuare a svolgere un ruolo determinante per promuovere la consapevolezza della necessità di sostenere il TAL, e per definire e stimolare un insieme di azioni coordinate che rispondano ai bisogni prioritari del nostro paese nel settore: dalla promozione di strategie e programmi di interesse nazionale finanziati dal MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca) alla proposta di curricula autonomi di formazione universitaria (master, dottorati), al collegamento tra comunità internazionale e nazionale, alla proposta e coordinamento di iniziative e progetti comunitari e internazionali, all'incentivazione del collegamento e del trasferimento tecnologico verso l'industria.

APPENDICE 1

PROGETTI INTERNAZIONALI

Elenco qui alcuni progetti comunitari cui il gruppo di Pisa ha partecipato o partecipa, spesso come coordinatore.

ESPRIT BRA-ACQUILEX-1 e 2 (coordinatori)
Tutte le applicazioni della LC richiedono conoscenze sulle parole. La prima operazione di base che ogni sistema di NLP deve compiere è "riconoscere" le parole del testo di input e le loro proprietà linguistiche. Dopo molti anni di lavoro manuale si è tentato di rendere la acquisizione di informazioni lessicali semi-automatica, attraverso l'analisi di dizionari di macchina (MRD), in particolare attraverso l'analisi automatica delle definizioni.

ET-7
Ho convinto la CEE a promuovere questo progetto per confermare il risultato positivo degli esperimenti volti a dimostrare la fattibilità di standard comuni per le risorse linguistiche: il progetto ha compiuto un survey dettagliato delle conoscenze linguistiche necessarie per i diversi sistemi di NLP.

LRE-DELIS
Obiettivo del progetto era di disegnare entrate lessicali combinando due approcci: uno teorico - la *frame semantics* di Fillmore, uno empirico, l'analisi dell'uso della lingua nei corpora.

ET-10
Scopo principale del progetto era quello di estrarre informazioni dal dizionario COBUILD.

MULTEXT
Scopo principale del progetto era quello di costruire strumenti per la creazione di corpora paralleli allineati e annotati.

MULTILEX
Scopo del progetto era quello di costruire campioni di lessici tecnici multilingui.

LRE e LE EAGLES (coordinatori)
I progetti si proponevano lo sviluppo di standard nel settore del trattamento automatico della lingua, per lessici computazionali, corpora, formalismi e grammatiche, trattamento del parlato, e valutazione di sistemi.

NERC
Il progetto si proponeva il disegno di un piano per la creazione di grandi corpora comparabili e armonizzati per le lingue europee, e delle relative specifiche.

MEMORIA
Obiettivo del progetto era la creazione di una stazione di lavoro personalizzata per l'accesso ai beni librari.

LS-GRAM
Il progetto si proponeva di estendere la copertura delle grammatiche di EUROTRA per alcune lingue.

COLSIT
Il progetto si proponeva il consolidamento delle grammatiche di EUROTRA.

RENOS
Obiettivo del progetto era lo sviluppo di un sistema di classificazione e di information retrieval su documenti attraverso la gestione di strutture lessicali.

TAMIC
Il progetto si proponeva l'uso di strumenti software per il trattamento del linguaggio naturale per facilitare l'accesso dei cittadini ai dati pensionistici.

IDEAL
Obiettivo del progetto era lo sviluppo di un sistema per la gestione automatica di dialoghi uomo-macchina.

CRISTAL
Il progetto si proponeva di sviluppare un sistema di information retrieval intelligente basato sulla rappresentazione della conoscenza lessicale.

TELRI
Scopo del progetto era la creazione di un network di risorse linguistiche per paesi dell'est europeo.

ONOMASTICA
Il progetto si proponeva di creare elenchi di nomi propri di luogo e di persona in trascrizione fonologica per uso nei servizi di telefonia.

EUROSEARCH
Obiettivo del progetto era lo sviluppo di tecniche per la navigazione multilingue su WWW.

RELATOR (coordinatori)
Il progetto ha disegnato un modello per la distribuzione di risorse linguistiche in Europa.

PROART
Il progetto si proponeva la creazione di interfacce robuste in un ambiente di comunicazione uomo-macchina.

MLAP - e LE-PAROLE (coordinatori)
I progetti hanno disegnato il modello e le specifiche, e successivamente creato, lessici con informazioni morfologica e sintattica di sottocategorizzazione per 12 lingue europee, e corpora per 14 lingue.

ELAN
Il progetto aveva lo scopo di creare una struttura per l'accesso in linea ai corpora di PAROLE e TELRI.

SPARKLE (coordinatori)
Il progetto si proponeva di far avanzare lo stato dell'arte nello sviluppo di un nuovo spettro di risorse lessicali, da acquisire automaticamente dai testi analizzati attraverso tecniche di *shallow parsing,* per applicazioni specifiche.

LE-SIMPLE (coordinatori)
Il progetto ha disegnato il modello e le specifiche, e successivamente creato, il livello di informazioni semantiche per i 12 lessici PAROLE.

EUROWORDNET
Scopo del progetto era quello di costruire delle reti semantiche, sul modello di WordNet, per l'italiano, l'olandese, lo spagnolo, l'inglese (cui

si aggiunsero poi il francese, il ceco, l'estone, il tedesco) e di collegarli tra loro attraverso riferimenti ai synset di WordNet inglese, usato come lingua pivot.

ELSE
Il progetto ha condotto uno studio di fattibilità riguardo all'organizzazione di una struttura e di una procedura di valutazione comparativa delle tecnologie.

EUROMAP/HOPE
Il progetto si proponeva la promozione, coordinata, dei programmi di LC comunitari nei diversi paesi europei.

MATE
Il progetto ha sviluppato una piattaforma software integrata per la annotazione di dialoghi a molteplici livelli di analisi linguistica.

NITE
Il progetto si propone di sviluppare una piattaforma software integrata per l'annotazione di risorse multimodali.

MLIS-MUSI (MUltilingual Summarisation Tool for the Internet) (coordinatori)
Il progetto si propone lo sviluppo di un prototipo per un sistema di sommarizzazione multilingue, attraverso un livello di Rappresentazione Concettuale Interna (Irep) indipendente dalla lingua.

NSF XMELLT
Il progetto si propone lo studio del trattamento di *multiword expressions* in lessici computazionali multilingui.

EC-NSF ISLE (coordinatori)
Il progetto Europeo e Americano si propone la crezione di standard per lessici computazionali multilingui, per corpora multimodali, e per la valutazione di sistemi di traduzione automatica.

POESIA (Public Open Source Environment for Safe Internet Access) (coordinatori)
Il progetto si propone di creare un sistema software flessibile e parametrizzabile in grado di filtrare contenuti non adatti ai giovani (per es. siti pornografici) su Web, attraverso l'uso di tecnologie di trattamento delle immagini e di NLP.

CHLT
Il progetto ha sviluppato strumenti linguistici per il recupero delle informazioni nei beni culturali.

ENABLER (coordinatori)
Il progetto si propone la creazione di un Network di progetti nazionali europei (e non) dedicati al TAL.

UNL
Il progetto si propone di produrre servizi multilingue sul web.

INTERA
Obiettivo di INTERA è lo sviluppo di un'area europea integrata per le risorse linguistiche, attraverso l'interconnessione tra dati messi a disposizione dai centri nazionali e attraverso la loro descrizione unificata mediante metadati.

CFID
Obiettivo di questo progetto era l'individuazione e la soluzione di fallimenti comunicativi nell'interazione dialogica.