IL TRATTAMENTO AUTOMATICO DEL LINGUAGGIO NATURALE

Antonio Zampolli (*), Nicoletta Calzolari (**)

(*) CNR, Istituto di Linguistica Computazionale
($) Università di Pisa, Dipartimento di Linguistica

Abstract

La attenzione e l'interesse per il trattamento automatico del linguaggio naturale (natural language processing = NLP) fino a pochi anni fa limitati all'ambito accademico, si stanno ora estendendo ad organismi politici e industriali.

Le cause di questa estensione sono da ricercare nel fatto che le lingue naturali costituiscono il veicolo primario per la produzione, la memorizzazione, il trasferimento, il recupero dell'informazione, attività che rappresentano una parte sempre più cospicua dell'economia mondiale.

La comunicazione esaminerà i principali problemi che devono essere ancora risolti nel settore del NLP. In particolare verrà esaminata la necessità di creare delle risorse linguistiche riutilizzabili, in forma leggibile dal calcolatore (corpora di riferimento, lessici, thesauri, terminologie, ecc.). Si illustreranno le principali iniziative internazionali nel settore, in particolare europee, alle quali il CNR collabora, sottolineando i rapporti tra tali iniziative e le attività di ricerca e sviluppo svolte nell'ambito del progetto **"TRASFERIMENTO DELLE TECNOLOGIE DEI PF"**, con particolare riguardo alla creazione di basi di conoscenze lessicali mono e plurilingui, e alla creazione o sperimentazione di strumenti computazionali interattivi per l'accesso a testi e documenti.

1. Le attività dell'Istituto di Linguistica Computazionale nel quadro nel progetto "Trasferimento delle Tecnologie dei Progetti Finalizzati".

1.1 La partecipazione dell'Istituto di Linguistica Computazionale del CNR al progetto "Trasferimento delle Tecnologie dei Progetti Finalizzati" consiste principalmente nello studiare e sperimentare l'utilizzo di metodi e strumenti della Linguistica Computazionale come sussidio alle operazioni di accesso alle informazioni contenute in testi in linguaggio naturale. Metodi di questo tipo potrebbero essere applicati alle parti testuali dei documenti che costituiscono la base dati dei progetti finalizzati.

I metodi e gli strumenti in questione dovrebbero, essenzialmente, consentire di operare sui testi per due scopi distinti, anche se complementari:

- riconoscere unità linguistiche di vario livello e le loro relazioni sintagmatiche, associando ad esse conoscenze relative alle proprietà linguistiche, alle relazioni paradigmatiche, alle strutture concettuali, ecc., per metterle a disposizione degli algoritmi di ricerca e recupero delle informazioni;
- studiare, analizzare, descrivere i vari aspetti qualitativi e quantitativi dei sottolinguaggi dei documenti, per facilitare la costruzione di strumenti e l'affinamento di metodi linguistici sulla base delle caratteristiche specifiche di tali sottolinguaggi.

Tra i metodi dei quali si prevede la utilizzazione, citiamo qui:

- <u>L'analisi morfosintattica</u>

Scopo dell'analisi è quello di offrire all'utente la possibilità di "esplodere" le parole che vuole ricercare nei testi, in modo tale che il sistema ricerchi anche tutte le parole ad esse collegate morfologicamente, sia perché appartenenti allo stesso lemma (flessioni, coniugazioni), sia perché connesse da meccanismi di derivazione.
L'analizzatore dovrebbe anche risolvere, per mezzo di regole, e/o di metodi statistici fondati sull'analisi di corpora testuali, alcune classi di ambiguità grammaticale e lessicale, e proporre il trattamento di neologismi creati dai meccanismi più produttivi nelle terminologie settoriali (suffissi, affissoidi, composti, ecc.).

- <u>Analisi delle relazioni paradigmatiche</u>

Le ricerche in corso sembrano indicare che è possibile estrarre, con procedure

semi-automatiche di analisi, dalle definizioni di dizionari disponibili in "machine readable form", relazioni semantiche che strutturano e collegano variamente gli elementi del lessico. Per esempio, sembra possibile estrarre delle tassonomie di tipo concettuale che, nel loro insieme, potrebbero essere organizzate in modo da costituire, per così dire, un "thesaurus" del lessico. Altri tipi di relazioni possono essere estratte dall'analisi di corpora testuali.

Sono già stati compiuti alcuni esperimenti che mostrerebbero la utilità di adoperare relazioni semantiche così costruite nella interrogazione di testi di diverso tipo: letterari, giornalistici, politici, ecc.

In pratica, il sistema "espande" le parole che l'utente vuole cercare nei testi, ricercando nella base dati unità lessicali ad essa collegate da relazioni semantiche di vario tipo (sinonimi, iperonimi, iponimi, ecc.).
In prospettiva, le relazioni estratte da lessici e corpora potrebbero confluire in una base di conoscenza alla quale applicare metodi e strumenti specifici del settore della "knowledge representation".

Aspetti multilingui dell'interrogazione

Ci si ripromette anche di esplorare la possibilità di creare strumenti multilingui che facilitino l'accesso ai testi a utenti di altre lingue. Questa fase del progetto non è tuttavia ancora iniziata.

1.2 Poiché altre comunicazioni del Convegno illustrano e discutono l'uso di componenti linguistici nei sistemi di "information retrieval" applicati alla banca dati dei progetti finalizzati, ci limiteremo qui a descrivere brevemente alcune recenti tendenze della linguistica computazionale e alcuni progetti internazionali, ai quali l'Istituto partecipa, i quali si propongono di creare risorse linguistiche di base adeguate alle esigenze del trattamento di "real world texts", quali sono quelli che figurano nella banca dati del CNR.
La disponibilità di tali risorse appare oggi come una condizione essenziale per lo sviluppo di sistemi applicativi basati sul trattamento delle lingue naturali (NLP).

E' un fatto ben noto che una larga parte delle trasformazioni in corso, a livello politico, sociale, economico, industriale è in parte causa, e in parte effetto, del crescente flusso di informazioni multilingui nella "società dell'informazione" che caratterizza il cosiddetto "villaggio globale".

Le lingue naturali sono ancora oggi i veicoli privilegiati per produrre, codificare, trasmettere, ricuperare l'informazione. Inoltre, l'informazione oggi è, in massima parte, prodotta, memorizzata, distribuita attraverso calcolatori e reti di calcolatori. La quantità globale di informazione e il bisogno di comunicazione rapida ed efficiente sono tali che è naturale cercare di utilizzare i calcolatori per ridurre il problema a dimensioni dominabili e per ridurre i costi, elevatissimi, delle operazioni umane richieste dalla produzione ed elaborazione di testi.

Alcuni hanno sostenuto che, se il problema del trattamento del linguaggio naturale non sarà risolto, la possibilità di evoluzione della nostra società potrebbe essere ridotta o compromessa.

L'elaborazione del linguaggio naturale, potenzialmente, concerne non solo varie "professioni linguistiche" tradizionali (traduzione, documentazione, lessicografia, editoria, insegnamento delle lingue, ecc.), ma anche nuovi tipi di applicazioni rilevanti, sul piano sociale e industriale, per l'utente finale: produzione di documenti, macchine per dettare, interazione uomo-macchina,

gestione della posta elettronica, input vocale per domini ristretti, comunicazione aumentata per portatori di handicaps, ecc.

Molte di queste applicazioni, soprattutto in Europa, devono incorporare funzioni multilingui, per assistere la comunicazione tra parlanti di lingue diverse: generazione di messaggi multilingui, produzione di documenti in più lingue, accesso multilingue a basi di dati, classificazione automatica di posta elettronica, ecc.

La capacità di incorporare componenti linguistici per produrre, memorizzare, accedere alle informazioni rappresentate in lingua naturale, è critica per una larga varieta di sistemi industriali. Il mercato potenziale è - globalmente - molto rilevante. Tuttavia, sistemi ed applicazioni possono essere prodotti solo se si dispone di un "saper fare" e di una "tecnologia" linguistica adeguati.

Il carattere multilingue dell'Europa costituisce una dimensione aggiuntiva alla complessità del problema. Il multilinguismo può frapporre barriere e difficoltà allo sviluppo del mercato comune europeo. Alcuni però osservano che lo stimolo a sviluppare una tecnologia che aiuti a superare le differenze linguistiche potrebbe conferire all'Europa un know-how specifico superiore a quello degli altri due grandi blocchi economici, America e Giappone.

I problemi posti dal multilinguismo, le evidenti potenzialità di mercato dell'industria delle lingue, alcune implicazioni strategiche del trattamento del linguaggio naturale, hanno determinato, in questi ultimi anni, un crescente interesse delle principali organizzazioni internazionali (CEE, Consiglio d'Europa) e di alcuni enti nazionali (DARPA, NSF, ICOT, CNRS, DTI, etc.).

Queste organizzazioni, con la collaborazione delle principali Associazioni Scientifiche (ACL, ALLC, ACH, ecc.), hanno identificato, nella disponibilità di risorse linguistiche per il NLP, una esigenza pioritaria per lo sviluppo delle potenzialità del settore, ed hanno promosso attività e progetti per la costruzione di tali risorse.

L'Istituto di Linguistica Computazionale del CNR, in collaborazione con il Dipartimento di Linguistica dell'Università di Pisa, ha avuto un ruolo di primo piano nella ideazione, nella programmazione e nel coordinamento di queste attività in ambito internazionale.

Ci sembra opportuno sottolineare il fatto che, nell'ambito delle esplorazioni condotte sulla possibilità di avviare la cooperazione tra DARPA e varie attività di ricerca della CEE, il tema delle risorse linguistiche è stato giudicato prioritario.

Nel corso della primavera 1991 organizzeremo a Pisa un workshop su questo tema, in occasione del quale si incontreranno rappresentanti di NSF, DARPA, CEE.

Descriviamo rapidamente, qui di seguito, a titolo di esempio, alcuni progetti i cui risultati potrebbero interessare direttamente la realizzazione di sistemi di information retrieval applicabili a basi testuali di grandi dimensioni.

2.Costruzione di basi lessicali per il NLP

2.1 Lessicografia e lessicologia computazionale e il concetto di "Riusabilità"

Lo scopo prioritario oggi è la creazione di un vasto "reservoir" di conoscenza linguistica, nella forma di descrizioni linguistiche riutilizzabili,

il più complete possibili, strutturate in una vasta "Lexical Knowledge Base" (LKB) o in vari tipi di basi linguistiche interconnesse (grammaticali, lessicali, testuali, di conoscenza).

Data la richiesta crescente di sistemi di NLP di vasta scala e data la necessità che un sistema di NLP sia in grado di trattare decine di migliaia di entrate lessicali per applicazioni reali, in aggiunta al fatto che la lessicografia ha una lunga tradizione, che la creazione di un "Lexical Data Base" (LDB) di contenuto e dimensione adeguati è molto dispendioso in termini sia di tempo sia di denaro, e che la duplicazione degli sforzi è perciò da evitarsi, il termine "riusabilità" è divenuto, recentemente, una delle parole-chiave nel campo della linguistica computazionale. Questa parola va intesa in due sensi principali: uno rivolto al passato, cioè rispetto ad informazioni già esistenti, uno rivolto al futuro, cioè rispetto ad applicazioni potenziali.

Nel primo caso si tratta soprattutto di riutilizzare informazioni lessicali implicitamente o esplicitamente presenti in risorse lessicali preesistenti (ad esempio "Machine Readable Dictionaries" (MRD)[1], banche dati terminologici, corpora testuali, ecc.) come un aiuto nella costruzione di un LKB.

Per valutare la fattibilità e provare la utilità di riutilizzare i MRD, abbiamo proposto il progetto ESPRIT BRA "Acquisition of Lexical Knowledge for Natural Language Processing Systems" (ACQUILEX), che è in corso di svolgimento con la partecipazione di gruppi di ricerca di Cambridge, Amsterdam, Dublino, Barcellona e Pisa (coordinatore). Scopo principale del progetto è lo sviluppo di tecniche e metodologie per l'uso di MRD esistenti nella costruzione di componenti lessicali per sistemi di NLP. L'estrazione di informazioni lessicali viene inoltre effettuata da molteplici MRD e in un contesto multilingue, allo scopo di creare una singola LKB multilingue. "La base di conoscenza avrà le sue radici in una struttura concettuale/semantica comune che è legata a, e definisce, le accezioni individuali delle parole delle lingue trattate e che dovrebbe essere sufficientemente ricca da permettere di supportare un modello di elaborazione del linguaggio basato su un tipo di conoscenza "profonda". La base di conoscenza conterrà un consistente vocabolario di tipo generale, con associate informazioni di tipo fonologico, morfologico, sintattico e semantico/pragmatico, che possa essere impiegato nei componenti lessicali di un'ampia varietà di sistemi pratici di NLP" (Boguraev et al. 1988).

Il secondo significato del termine riusabilità è collegato a due proprietà che consideriamo essenziali in un LDB.

La prima proprietà ha essenzialmente a che fare con il punto di vista applicativo. Un LDB dovrebbe essere "multifunzionale". Dovrebbe cioè funzionare come un deposito centralizzato di dati che possono essere riutilizzati per scopi diversi e in molteplici applicazioni, attraverso interfacce appropriate, per utenti sia procedurali sia umani.

Il lessico è ovviamente un componente essenziale di qualsiasi siatema di NLP (per il parsing, la generazione, la traduzione automatica, l'information

---

[1] Il termine MRD indica, sostanzialmente, dizionari che sono stati registrati in forma leggibile dal calcolatore per produrre, mediante sistemi di fotocomposizione, l'edizione a stampa.

retrieval, il question-answering, ecc.). Il procedimento usuale consiste nel costruire un componente lessicale ad hoc per ciascun progetto di NLP, con un evidente spreco e duplicazione di sforzi. Diventa ora necessario dirigersi verso la costruzione di lessici estesi (sia per numero di entrate sia per profondità di rappresentazione), dove l'informazione sia rappresentata in modo tale da poter essere facilmente interfacciata da diverse procedure applicative, in base agli specifici bisogni di ciascuna applicazione. Ciò significa che lo stesso insieme di dati può essere condiviso dalle varie applicazioni. Ogni interfaccia proietterà sull'applicazione specifica solo quel punto di vista sui dati che è rilevante per le esigenze particolari dell'applicazione stessa. Un LDB deve essere anche facilmente estendibile, in modo che diversi ricercatori possano aggiungere le loro informazioni idiosincratiche in modo consistente rispetto al contenuto attuale dell'LDB..

La seconda proprietà di un LDB ha a che fare con il punto di vista teorico, e consiste nel suo poter essere "politeorico", cioè multifunzionale rispetto a diverse teorie linguistiche. Nell'ambito della linguistica computazionale si è fino ad ora svolta una grande quantità di lavoro su prototipi sperimentali, producendo di conseguenza solo sistemi lessicali prototipo di piccole dimensioni. Inoltre, fino ad ora si è prestata tradizionalmente più attenzione alla rappresentazione, organizzazione e uso della conoscenza linguistica in quanto incapsulata ed espressa da regole e procedure linguistiche. I dati lessicali sono stati invece spesso considerati di secondaria importanza o per lo meno facili da trattare.

E' un fatto riconosciuto che diverse teorie linguistiche e diverse organizzazioni computazionali pssono avere conseguenze rilevanti sul disegno di una grammatica. E' stata prestata invece meno attenzione alle loro conseguenze sul lessico. Abbiamo però l'intuizione che lessici disegnati per diverse teorie linguistiche possano in linea di principio contenere informazioni che da un certo punto di vista sono equivalenti, in quanto descrivono gli stessi fatti linguistici. E' necessario dimostrare la validità di questa intuizione prima di iniziare l'implementazione di un LDB con tutte le informazioni richieste dai sistemi di NLP.
E' attualmente in corso uno studio di fattibilità per valutare la possibilità di raggiungere un certo grado di consenso fra teorie diverse allo scopo di condividere lo stesso nucleo di informazioni lessicali, e, se questo è fattibile, per valutare per quali livelli di analisi linguistica si possa disegnare una rappresentazione delle proprietà linguistiche che sia neutrale o politeorica.

Abbiamo promosso a questo scopo un gruppo di lavoro di cui fanno parte rappresentanti delle più imprtanti "scuole linguistiche". Questo gruppo sta analizzando in dettaglio la possibilità di rappresentare le informazioni linguistiche usate più frequentemente nei parsers e nei generatori (ad esempio le parti del discorso, la sottocategorizzazione e complementazione, le classi di verbi, le tassonomie di sostantivi, ecc.) in modo tale da poter essere riutilizzate nell'ambito dei seguenti quadri teorici: government and binding, generalized phrase structure grammar, lexical functional grammar, relational grammar, systemic grammar, categorial grammar.

Queste ricerche, da un lato, si svolgono con la cooperazione delle Università di Standford, MIT, Berkeley, Princeton, Cambridge, Heidelberg, Pisa, e dei gruppi di ricerca di IBM USA, BBN, Bell Research Lab; dall'altro, hanno originato uno studio di fattibilità promosso dalla CEE, cui partecipiamo.

2.2 Riusabilità di dati preesistenti nella forma di MRD

Vogliamo qui mettere in rilievo in modo particolare quella che consideriamo l'evoluzione naturale del lavoro fatto in questo settore nell'ultimo decennio, cioè la possibilità di uno sfruttamento procedurale della "totalità" dell'informazione semantica contenuta implicitamente nei MRD.

In questo quadro il dizionario viene considerato come una fonte primaria di conoscenza generale di base, e oggigiorno molti progetti hanno come obiettivo principale l'acquisizione del significato delle parole da MRD, e la organizzazione della conoscenza acquisita in un LKB. Il metodo usato è induttivo e la strategia adottata è di tipo euristico: attraverso generalizzazioni progressive dagli elementi comuni trovati nelle definizioni in linguaggio naturale si tende a formalizzare la conoscenza generale che è implicitamente contenuta nelle definizioni dei dizionari, allo scopo di estrarre i concetti più generali e le relazioni semantiche di base. Ciò significa andare ben al di là dell'estrazione e organizzazione di tassonomie, la cui metodologia di acquisizione è ormai consolidata (Chodorow et al. 1985, Calzolari 1982, 1984). Per estrarre le tassonomie dobbiamo semplicemente analizzare la prima parte della definizione allo scopo di identificare il cosiddetto "genere prossimo". Questo può essere effettuato basandosi sul fatto che le definizioni sono dei Gruppi Nominali quando il definiendum è un nome, sono Gruppi Verbali per i verbi, e Gruppi Aggettivali per gli aggettivi: il programma deve perciò cercare la "testa" del Gruppo Nominale, Verbale o Aggettivale che sarà rispettivamente un nome, un verbo o un aggettivo. Questi sono i "generi prossimi" e sono connessi al definiendum da un legame di tipo IS-A.

Quando riorganizziamo un MRD in una struttura tassonomica, in cui solo gerarchie di tipo IS-A sono esplicitate, usiamo l'MRD come una fonte di conoscenza, ma in uno soltanto dei possibili modi di acquisire da esso (in modo induttivo) un concetto, cioè attraverso il legame di questo concetto a tutti i suoi esempi; tutti gli esempi appartenenti ad una stessa classe o categoria sono cioè estratti e connessi fra di loro attraverso il puntatore al loro immediato superordinato.

Nell'approccio di un progetto come ACQUILEX, il dizionario stesso è considerato come uno "strumento di classificazione" molto più potente, cioè come uno strumento empirico per evidenziare non solo concetti generali ma anche numerosi tipi di relazioni lessicali/semantiche.

L'approccio metodologico da noi seguito può essere riassunto in questi punti:

a) analisi delle definizioni nella loro struttura tipica, che consiste in "genere prossimo" e "differentia";
b) analisi della loro struttura sintattica;
c) conversione e riorganizzazione delle definizioni in formati strutturati, equivalenti dal punto di vista delle informazioni, costituiti da nodi e relazioni che collegano questi nodi.

Illustriamo ora con alcuni esempi il processo di analisi delle definizioni. Nelle figure che seguono tentiamo di simulare il processo di consultazione dell'LDB dell'italiano sviluppato a Pisa, e la navigazione all'interno del dizionario alla ricerca di parole particolari, di strutture, di "patterns" sintattici o semantici, ecc. Metteremo in evidenza in questo modo alcuni dei dati semantici che è possibile cercare e recuperare in un MRD strutturato in

modo appropriato. La figura 1 mostra parte della tassonomia per la parola *libro*, cioè l'insieme delle parole definite come "tipi di" libri (nella figura vediamo alcuni di queste parole insieme con le loro definizioni).

In un dizionario però possiamo trovare altre informazioni relative ai libri. Possiamo anche estrarre l'insieme dei verbi relativi ai libri (figura 2), o l'insieme degli aggettivi e degli altri nomi che hanno a che fare con i libri (figure 3 e 4). Nel paragafo 3.2.5 torneremo a parlare di "libri", sottolineando quei tipi di informazioni che, mancando nei dizionari, possono invece essere estratti da corpora testuali. Attualmente stiamo lavorando alla formalizzazione anche di tutti gli altri tipi di relazioni - non così semplici come le relazioni tassonomiche - che intercorrono fra parole, o fra parole e concetti, e per la cui estrazione dobbiamo analizzare e elaborare l'intera definizione e non soltanto il "genere prossimo".

Ne diamo alcuni esempi. Nella figura 5 troviamo le prime delle circa 300 parole collegate nel nostro LDB da un legame di tipo tassonomico alla parola *strumento*. La parola *attrezzo* compare in questa lista. La figura 6 mostra i primi iponimi di *attrezzo* insieme con le loro definizioni. Dall'analisi di queste definizioni è abbastanza semplice estrarre relazioni semantiche che potremmo chiamare **USED FOR, USED IN, SHAPE, MADE OF**, ecc. Queste relazioni sono estratte per mezzo di una procedura di "pattern-matching" che analizza la "differentia" delle definizioni: per mezzo di questa procedura i diversi modi in cui ciascuna relazione è di fatto lessicalizzata nelle definizioni vengono associati con il nome della relazione pertinente. La relazione **USED FOR**, per esempio, viene da patterns lessicali come: *per, usato per, atto a, che serve a, utile a*, ecc.; questi patterns lessicali acquistano questo particolare significato relazionale quando si trovano in posizioni particolari nella definizione degli iponimi della parola *strumento*. Naturalmente possono acquistare significati diversi in altri contesti. Il risultato di questa analisi del contenuto delle definizioni sarà ristrutturato in una porzione di una rete concettuale (schematizzata nella figura 7).

Altri tipi di relazioni semantiche che si possono estrarre abbastanza facilmente e direttamente dalle definizioni possono essere illustrate attraverso alcuni esempi.

Una è la relazione **SET OF**, che può a sua volta essere ulteriormente specificata per quanto riguarda i tipi di elementi dell'insieme. Troviamo esempi di parole che denotano **SET OF** *persone* (figura 8), *oggetti* (figura 9), ecc.

Altri tipi di dati interessanti per i sistemi di NLP riguardano le informazioni sulle restrizioni di selezione per verbi o per aggettivi: questi dati derivano principalmente dal pattern lessicale *detto di*, dopo il quale si trova il tipo di sostantivi di cui un aggettivo o un verbo può essere tipicamente predicato. La figura 10 mostra alcuni aggettivi e verbi che tipicamente si usano per nomi denotanti *persone*, mentre la figura 11 elenca aggettivi che cooccorrono con nomi di colori, o nomi generici di colore, o nomi specifici di colore quali *giallo, rosso*, ecc.

Un tipo interessante di dato relazionale che si può estrarre per alcuni tipi di azione riguarda l'informazione sulle parole che sono lessicalizzazioni dei tipici ruoli tematici dell'azione stessa. Possiamo chiarire quanto detto con due esempi. Nella figura 12 troviamo l'output che si ottiene quando interroghiamo l'LDB chiedendo tutte le entrate nelle cui definizioni appare la forma flessa *vende* (non in posizione di genere prossimo). Il risultato della interrogazione è il seguente: toviamo 242 lemmi di cui ben 221 sono nomi di persone che "tipicamente vendono" qualcosa, cioè di tipici **AGENTS** rispetto all'azione del

vendere. Questi lemmi rappresentano "case/role fillers" lessicalizzati nel "case-frame" di *vendere*. Ciò è ovviamente dovuto al particolare pattern usato nelle definizioni, cioè *chi vende*.

Riguardo a questo esempio si possono fare alcune osservazioni interessanti. La prima riguarda il fatto che lo stesso tipo di risultato è stato ottenuto effettuando una ricerca analoga su dizionari di inglese. Avendo visto gli esempi dei dati ottenuti per l'italiano, il gruppo di ricerca dell'IBM Yorktown ha ripetuto l'esperimento ottenendo lo stesso tipo di risultati per l'inglese (Byrd 1989). Questo dimostra che c'è in realtà una corrispondenza fra i patterns definizionali usati nella pratica lessicografica indipendentemente dalla lingua. Una tale somiglianza nelle convenzioni lessicografiche appare da molti altri esempi, e sarà sfruttata per la creazione del LKB multilingue che è l'obiettivo finale del già menzionato progetto ESPRIT.

Un'altra osservazione riguarda la cooccorrenza nelle definizioni del verbo *vendere* con verbi quali *fabbricare, fare, preparare*, ecc. Molti di questi nomi di Agente si applicano pertanto anche all'azione del "fare", e perciò appartengono a due distinte porzioni della rete concettuale che stiamo costruendo.

Si può anche notare che il Gruppo Nominale che segue il verbo denota il tipo di oggetto che è tipicamente venduto (o anche fabbicato) da questi Agenti.

E' ovviamente possibile ottenere lo stesso tipo di informazioni sui nomi di Agente per l'azione del vendere effettuando una ricerca di tutti i nomi il cui "genere prossimo" è la parola *venditore*: così recuperiamo altri 131 nomi d'Agente (alcuni di questi compaiono nella figura 13). Anche in questo caso alcuni dei nomi sono collegati anche con l'azione del "fare", mentre qui il Gruppo Preposizionale introdotto dalla preposizione *di* esprime l'oggetto che è venduto.

Questo esempio mostra come esattamente lo stesso tipo di informazione possa essere recuperato attraverso diversi modi di consultazione del dizionario, sfruttando appieno la conoscenza della sua struttura interna (in particolare la struttura interna delle definizioni). Nel LKB che si sta costuendo tutti questi dati saranno unificati in un'unica porzione di rete concettuale, indipendentemente dai diversi modi di lessicalizzazione di alcuni concetti e relazioni.

Con un tipo di ricerca leggermente diverso possiamo facilmente recuperare anche i nomi di **LOCATIONS** dove viene tipicamente eseguita l'azione del vendere. La figura 14 mostra il risultato della ricerca dei lemmi nelle cui definizioni la parola *vendono* è presente. Anche in questo caso la possibilità di trovare questi nomi di luogo è dovuta alla seguente "formula definitoria" usata dai lessicografi: *dove/in cui si vendono*. Tutti i 33 lemmi recuperati hanno in comune questo pattern: la ricerca è completamente senza "rumore".

In questi esempi possiamo osservare che i "generi prossimi" sono o il nome generico *luogo*, o quelli fra i suoi iponimi che sono nomi generici per i luoghi in cui si vende qualcosa, cioè *negozio, bottega, bancarella*. Questi sono a loro volta superordinati dei lemmi definiti. Questo tipo di informazione gerarchica è già codificato formalmente nelle tassonomie dell'LDB sviluppato a Pisa.

Ciò che qui ci interessa far notare è la possibilità di formalizzare e implementare nel LKB gli altri tipi di relazioni semantiche, quali **LOCATION**, **AGENT** e **THEME**, rispetto alle azioni del "vendere" e del "fare". La relazione **THEME**, cioè gli oggetti che sono tipicamente venduti nei posti definiti, è di nuovo espressa dal Gruppo Nominale oggetto del verbo.

Anche in questo caso, naturalmente, dati simili sono recuperati attraverso

la ricerca nel dizionario degli iponimi di *negozio*, *bottega*, ecc. Noi ci proponiamo di formalizzare tutti questi tipi di informazioni in una rete semantica, illustrata parzialmente nella figura 15.

Gli esempi citati mostrano che un LDB può essere utilmente sfruttato per analizzare e per estrarre dati linguistici che dovranno poi essere ristrutturati e rappresentati nel LKB. Questi tipi di concetti e di relazioni, e le interdipendenze fra le varie accezioni dei lemmi, saranno esplicitati completamente nel LKB. Quando andiamo al di là delle semplici tassonomie verso la costruzione di un LKB, riusciamo a stabilire molteplici tipi di associazioni semantiche che possono essere rappresentate in una rete concettuale, e quando ci muoviamo da un ambiente "monolingue" verso un ambiente "multilingue" stabiliamo anche associazioni fra lingue diverse. Tali associazioni sono ottenute (per quelle parti delle lingue che possono essere ridotte ad un insieme comune di concetti e relazioni) attraverso la rete concettuale comune costruita lavorando su vari dizionari di lingue diverse ma all'interno di uno stesso "quadro di ricerca", cioè tentando di formalizzare nella rete semantica:

- la "stessa" conoscenza "del mondo",
- per gli "stessi" obiettivi (NLP, Text Processing),
- con la "stessa" metodologia,
- dagli "stessi" tipi di risorse (MRD),
- in uno "stesso" tipo di rappresentazione.

La rete semantica comune diventerà perciò il punto di convergenza dei risultati delle strategie di acquisizione della conoscenza applicate a un certo numero di risorse diverse ma omogenee fra di loro, e l'ambiente multilingue costituirà un valido "testbed" per la valutazione della strategia seguita nel disegno e implementazione della LKB.

2.3 Riusabilità dei dizionari bilingui

Non solo i dizionari monolingui in MRF, ma anche i dizionari bilingui possono essere utilmente adoperati come fonti di informazioni lessicali per la creazione di LDB e LKB. Questi dizionari possono essere analizzati ed elaborati con un duplice obiettivo: da un lato sono anch'essi una fonte di interessanti informazioni "monolingui"; dall'altro si possono ovviamente sfruttare come fonte di legami fra due LDB monolingui.

Uno dei nostri principali obiettivi consiste nell'integrare i diversi tipi di informazioni che sono tradizionalmente contenuti nei dizionari monolingui e bilingui, in modo da espandere il contenuto informativo dei singoli componenti nel nuovo sistema integrato. I dizionari bilingui contengono ad esempio, di solito, maggiori informazioni, rispetto ai dizionari monolingui, per quanto riguarda gli esempi d'uso, le espressioni fisse o gli idioms. Questo tipo di dati linguistici può ovviamente essere integrato nei dizionari monolingui, e può anche essere reso facilmente accessibile.

Le entrate lessicali monolingui di partenza, in un sistema integrato, possono essere aumentate con informazioni che provengono dalla corrispondente entrata bilingue: differenti discriminazioni di senso, altri esempi, informazioni sintattiche, collocazioni, idioms, ecc. Possiamo anche rovesciare la prospettiva e guardare alle entrate bilingui come fornite delle informazioni tradizionalmente contenute nelle entrate monolingui: per lo più le definizioni. Entrambi i punti di vista sul sistema lessicale, sono virtualmente presenti nel sistema bilingue integrato. Noi tendiamo a mantenere in un'unica struttura sia le caratteristiche

indipendenti dei dizionari monolingui e bilingui di partenza, sia l'integrazione dei due tipi di dizionari, ovviamente con diversi punti di vista e modi di accesso ai dati.

Uno schema generale del sistema di LDB bilingue che stiamo implementando è rappresentato nella figura 16. Anche per quanto riguarda i dizionari bilingui, il nostro metodo consiste nel riutilizzare dati già disponibili in MRF analizzando e trasformando con procedure computazionali le informazioni contenute nei dizionari tradizionali. La procedura di analisi e elaborazione degli MRD bilingui è abbastanza simile a quella accennata sopra per i dizionari monolingui (cioè parsing dell'entrata lessicale, disegno di una nuova struttura, riorganizzazione computazionale, ecc.). Effettuata questa parte preliminare, anche in questo caso diventa immediatamente evidente l'utilità della consultazione di un dizionario bilingue strutturato come banca dati, sfruttando gli elementi strutturali già formalizzati nel LDB, allo scopo di scoprire proprietà e strutture non immediatamente visibili e accessibili nel dizionario stampato, ma utili per un ulteriore uso del dizionario computerizzato.

Dopo le prime fasi di elaborazione dei dati del dizionario bilingue, non farà più alcuna differenza quale delle due lingue verrà presa come punto di partenza nell'interrogazione. In un certo senso, non avremo più una lingua "source" e una lingua "target", poiché le procedure di consultazione e di accesso saranno indipendenti e neutrali rispetto alla direzione (il sistema lessicale diventa bidirezionale). Saranno anche automaticamente generate "cross-references" bidirezionali per quanto riguarda le informazioni contenute a livello di ciascuna accezione come indicatori semantici: sinonimi, iperonimi o indicatori contestuali.

Un'altra possibilità che è di particolare interesse nel contesto di questo Progetto Finalizzato, è l'uso del data base lessicale monolingue come strumento per espandere l'informazione, data nel bilingue sotto forma di una singola parola, sostituendo a tale parola l'intera "famiglia" di parole alle quali questa in realtà fa riferimento. Consideriamo, per esempio, l'entrata *vivido*, per la quale il dizionario riporta diverse traduzioni, da usare a seconda degli indicatori contestuali che si riferiscono al nome di cui l'aggettivo si predica (in parentesi):

vivido ..... ***(colori)*** bright, vivid

In casi come questo, le restrizioni semantiche generiche sulla traduzione si possono considerare come un "tratto semantico", e possono essere espanse automaticamente - dal Thesaurus monolingue - a tutti i possibili iponimi (questo si può fare on-line al momento dell'interrogazione). In tal modo può essere selezionata la traduzione appropriata in qualunque contesto in cui compaia un nome specifico di *colore*: questo diventa possibile nel disegno del nostro sistema integrato. L'informazione che può essere formalizzata a livello semantico in un dizionario monolingue - che serve a discriminare fra le diverse accezioni di un'entrata - dovrebbe in linea di principio essere dello stesso tipo di quella data nei dizionari bilingui sotto forma di "indicatori semantici" o "condizioni di selezione" per restringere la scelta a una particolare traduzione.

Dopo la riorganizzazione del MRD bilingue in un ben strutturato LDB, dobbiamo affrontare il difficile compito di usare i dati in esso contenuti per costruire dei legami fra i due LDB monolingui. La difficoltà ovviamente deriva dalla polisemia delle parole usate, sia a livello di entrata, sia come traduzioni. Non è esplicitamente detto, infatti, quale accezione è quella usata

in una situazione specifica. Ci proponiamo di risolvere questo problema, per quanto è possibile, nel già menzionato progetto ESPRIT, per lo più cercando di sfruttare gli indicatori semantici del bilingue e le tassonomie e altre informazioni concettuali dei monolingui.

Lo stabilire corrispondenze fra accezioni dei dizionari monolingui e traduzioni nei dizionari bilingui è uno dei problemi più interessanti per quanto riguarda la connessione di questi diversi tipi di dizionari. Dal momento che uno dei problemi principali in traduzione è la scelta corretta fra i diversi significati di parole ambigue dal punto di vista lessicale, noi pensiamo che sia necessario anche per un sistema di traduzione automatica o di traduzione assistita dal calcolatore di essere collegato a un data base linguistico integrato, cioè a una fonte di informazioni lessicali organizzate sotto forma di un Thesaurus attraverso tassonomie multidimensionali, in cui la possibilità di disambiguare fra diverse accezioni sia almeno in parte semi-automatizzata.

Sistemi di questo tipo dovrebbero essere molto utili come aiuto per i traduttori. Il risultato finale può in realtà essere visto come una "translator workstation", che fornisce l'accesso a molti tipi di dizionari e ad altre risorse lessicali, in cui vengono sfruttate al massimo le potenzialità e le funzioni delle banche dati lessicali e di quelle testuali.

Altri scopi di un sistema bilingue del tipo di quello appena disegnato possono essere i seguenti:

- uno strumento per i lessicografi;
- uno strumento per studi lessicologico-contrastivi;
- un mezzo per migliorare LDB monolingui;
- un aiuto nella costruzione di dizionari per la traduzione automatica;
- uno strumento per l'insegnamento di lingue straniere;
- un dizionario computerizzato per utenti finali.

A nostro avviso, uno dei principali vantaggi di un LDB bilingue è il tipo completamente diverso di "navigazione" al suo interno che è reso possibile sia dall'accesso multiplo ai suoi dati sia dai legami con il LDB monolingue. In particolare, non solo è possibile creare dei legami fra coppie di parole di L1 e L2, come nel dizionario stampato, ma principalmente fra "gruppi o famiglie" di parole semanticamente connesse, cosa che riteniamo sia una caratteristica essenziale per un vero dizionario bilingue e per tutti gli obiettivi che abbiamo appena elencato.

3. I corpora testuali di riferimento

3.1 Corpora testuali e sistema di NLP

La motivazione generale per costruire ed analizzare un corpus testuale può essere riassunta, semplificando, come segue.

E' impossibile, per descrivere una lingua, accedere a tutti i testi prodotti in un dato periodo di tempo ("popolazione"). Pertanto si analizza un insieme di testi opportunamente scelti (corpus), considerandolo come un "campione rappresentativo", nell'aspettativa di ritrovare, in altri testi della stessa popolazione, gli stessi fatti, comportamenti, distribuzioni, osservati nei testi campione.

Un sistema di NLP, che si proponga applicazioni concrete sui testi scritti in una lingua, deve essere fondato sulla evidenza di come tale lingua è di fatto usata in testi reali.

L'analisi di corpora "rappresentativi" è un mezzo insostituibile per

ottenere tale evidenza.

In particolare, lavori sperimentali recenti, nei settori della comprensione del parlato, dell'information retrieval, della classificazione di messaggi strategici, ecc., hanno mostrato che è molto utile - e forse necessario - trarre vantaggio dalle caratteristiche e proprietà specifiche dei vari "sottolinguaggi": usi diversi della stessa lingua in contesti comunicativi diversi, con diversi canali di comunicazione, per convogliare informazione in domini specifici, ecc.

Le differenze, nella varietà e distribuzione di alcune classi di fenomeni linguistici, tra sottolinguaggi diversi, possono essere sfruttate per ridurre il numero di situazioni linguistiche "non trattabili" automaticamente, migliorando il rendimento dei sistemi, aumentando la accettabilità da parte degli utenti, e quindi ampliando il ventaglio delle applicazioni fattibili. E' possibile, per esempio, restringere l'elenco degli usi sintattici, o la molteplicità delle restrizioni di selezione negli argomenti dei verbi, così da ridurre il numero delle ambiguità da risolvere nei testi.

L'analisi di corpora adeguati è il solo mezzo conosciuto per descrivere i diversi sottolinguaggi. I corpora testuali sono utilissimi per la descrizione contrastiva di lingue diverse, e per definire metodi e costruire strumenti per la valutazione di componenti e sistemi per il NLP.

Alcuni dei sistemi di NLP più recenti di maggior successo sono fondati essenzialmente sull'impiego di metodi statistici: per esempio, su modelli markoviani costruiti incorporando probabilità derivate dallo studio delle frequenze in corpora testuali.

Alcuni esempi di estrazione di conoscenze di corpora testuali saranno brevemente discussi più sotto, al punto 3.2.4.

Alcune organizzazioni internazionali (CEE, Consiglio d'Europa) e nazionali (DTI, DARPA, ERDI, ecc.), riconoscendo che la creazione di corpora testuali è una esigenza di massima priorità per lo sviluppo delle applicazioni del NLP, hanno avviato progetti di vario tipo.

Descriviamo brevemente, qui di seguito, una iniziativa della CEE alla quale partecipiamo.

3.2 Il Progetto per un network europeo di corpora di riferimento

Accogliendo una nostra proposta, la CEE ha promosso uno studio che ha lo scopo di definire le modalità di costituzione di un network europeo di corpora di riferimento (NERC). Lo svolgimento dello studio è affidato a 6 Istituti europei, con il coordinamento scientifico e organizzativo dell'Università di Pisa e del nostro Istituto.

Il programma di lavoro è articolato negli obiettivi seguenti.

3.2.1 Tipologia e composizione dei corpora

Nel raccogliere i primi corpora, ancora prima dell'uso del calcolatore, gli autori hanno sempre cercato di definire un insieme di parametri e di condizioni per la scelta dei materiali testuali da includere nel corpus campione, allo scopo di assicurgli la massima rappresentatività, nei confronti della "popolazione", inaccessibile nella sua totalità. Venivano discussi, in particolare:

la stratificazione del corpus: quanti e quali sottoinsiemi (sottolinguaggi, tipi di testo, argomenti, modalità di comunicazione, ecc.) riconoscere nella

popolazione e quali è in che proporzione includere nel corpus;
la dimensione del corpus: quale sia la dimensione minima del corpus e dei suoi sottoinsiemi che può assicurare una rappresentatività adeguata;
i criteri di campionamento: quanti testi includere per ogni sottoinsieme e con quali criteri sceglierli; quali possano essere le unità testuali minime: frasi, paragrafi, capitoli, testi interi, ecc.

Il nostro studio si propone di rispondere a questi problemi, per assicurare, da un lato, la omogeneità e comparabilità dei corpora costruiti per le diverse lingue europee, dall'altro la loro adeguatezza alle necessità del NLP.

A tale scopo si esploreranno e confronteranno alternative diverse. Per esempio:
- Mettere in atto una struttura organizzativa per la raccolta continua di testi in MRF, costruendo così archivi nazionali continuamente aggiornati. I testi sarebbero raccolti essenzialmente in base alla loro disponibilità. L'archivio sarebbe a disposizione dei ricercatori che ne estrarrebbero di volta in volta dei testi per le proprie ricerche.
- Definire una tipologia di sottolinguaggi; valutarne la priorità nei termini di possibili applicazioni specifiche di sistemi di NLP; promuovere, di conseguenza, la costruzione di corpora specializzati indipendenti.
- Stabilire un disegno, fondato su criteri scientifici, per un corpus generale, multifunzionale, bilanciato, per ogni lingua.

Lo sviluppo del corpus potrebbe iniziare con i sottoinsiemi di più alta priorità, all'interno del disegno generale, oppure costruendo inizialmente un nucleo multifunzionale bilanciato di dimensioni adeguate. Esso potrebbe servire come riferimento per valutare progressivamente i criteri di composizione e guidare la estensione del corpus. Potrebbe essere usato per testare metodi, procedure, programmi. Potrebbe fornire un primo insieme significativo di materiali testuali per una prima risposta alle esigenze di diverse categorie di utenti.

3.2.2 Armonizzazione della annotazione linguistica dei testi

La maggior parte dei corpora disponibili o in via di costituzione non contiene alcuna analisi linguistica ("raw texts"). Pochi corpora contengono la indicazione della classificazione morfosintattica delle parole (parti del discorso e categorie flessionali: "tagged texts"). Praticamente assenti i corpora annotati a livello sintattico e semantico ("parsed texts").

Questa situazione non è di certo dovuta alla mancanza di utenti potenziali, che invece abbondano. Essa è dovuta invece:
- al costo proibitivo di una analisi completamente manuale;
- alla inadeguatezza dei parsers finora costruiti, che non sono sufficientemente "robusti" per trattare corpora reali;
- alla mancanza di schemi di analisi comuni accettati dai diversi tipi di utilizzatori possibili.

Il nostro studio si propone innanzitutto di definire la fattibilità di uno schema di annotazione comune.

Nella comunità scientifica, esistono due posizioni chiaramente distinte. Alcuni ricercatori ritengono che uno schema comune di annotazione sarebbe incapace di soddisfare i bisogni di utenti diversi, e che, fondamentalmente, uno

schema neutrale rispetto alle diverse teorie sarebbe impossibile. Conseguentemente, essi suggeriscono che ci si dovrebbe concentrare sulla creazione di procedure flessibili, semiautomatiche, di analisi, lasciando al singolo ricercatore di decidere un proprio specifico schema di analisi, attraverso la definizione delle regole del parser.

Altri ricercatori ritengono invece che sia necessario cercare di definire uno schema di annotazione comune, e, attraverso procedure volte ad ottimizzare gli inevitabili interventi manuali, applicarlo alla annotazione di corpora opportunamente scelti.

Il nostro studio, per fornire degli elementi oggettivi di risposta, cercherà di esaminare in dettaglio se, fino a che punto, e per quali livelli linguistici, sia possibile disegnare uno schema multifunzionale di analisi, tale cioè che diverse categorie di utilizzatori possano derivare, almeno in parte, attraverso interfacce appropriate, dalla annotazione comune le informazioni linguistiche di cui abbisognano.

Ciò richiede un esame comparativo degli schemi attualmente usati, sia nei corpora sia nei diversi sistemi di NLP, e degli schemi proposti, più o meno esplicitamente, dalle diverse teorie linguistiche. Richiede inoltre la identificazione delle esigenze specifiche di diverse categorie di potenziali utilizzatori.

3.2.3 Disegno di software e procedure comuni

La possibilità di ripartire tra diversi Istituti la costruzione del software necessario per la creazione, la gestione, l'accesso, l'analisi, la elaborazione, la distribuzione dei corpora, è importante in considerazione del fatto che, mentre, per alcune procedure, si tratta essenzialmente di ottimizzare e standardizzare metodi conosciuti, per altre funzioni è necessario un notevole sforzo di ricerca. E' il caso, per esempio, della costruzione di parsers "robusti", capaci di analizzare la varietà di fenomeni che occorre nei testi reali, utilizzando lessici di grandi dimensioni e grammatiche che coprano sottoinsiemi linguistici molto estesi. Tali parsers inoltre dovrebbero essere capaci di:

- "failing gracefully", cioè continuare ad operare anche nei casi nei quali non riescono a conseguire il livello di analisi desiderato, fornendo in ogni caso dei risultati a livelli inferiori, e ricorrendo all'aiuto umano interattivo ove necessario;
- sfruttare le specificità dei diversi sottolinguaggi: limitazioni del vocabolario, riduzione, in complessità ed estensione, della sintassi, uso di patterns di selezione di restrizione specifici, ecc.
- utilizzare, ed eventualmente combinare, tecniche tradizionali di parsing fondate su regole gammaticali, e sistemi probabilistici, basati su frequenze di transizione tra categorie o strutture adiacenti.

Un compito particolarmente importante è quello di ideare, costruire, sperimentare metodi per la estrazione, dai corpora, di conoscenze di vario tipo.

3.2.4 Estrazione di conoscenza da corpora testuali e integrazione in un LKB

Abbiamo già visto nella sezione relativa ai lessici che i MRD sono delle fonti molto ricche di informazioni lessicali e semantiche. Sfortunatamente, essi non contengono tutto ciò che è necessario conoscere sul lessico di una lingua. Ci sono alcune informazioni molto rilevanti che nei MRD sono o completamente

mancanti o incomplete o semplicemente non sono molto buone o affidabili o facilmente recuperabili.

Alcuni tipi di dati si possono utilmente estrarre, più o meno direttamente, attraverso l'elaborazione di corpora molto estesi di dati testuali. I risultati di queste elaborazioni dovranno poi essere analizzati e valutati dal linguista e/o dal lessicografo. E' importante rendersi conto che per certi tipi di fenomeni linguistici l'analisi di un corpus testuale è pressoché indispensabile: esempi tipici sono le collocazioni e le "frasi fisse" (si veda Calzolari, Bindi, 1990). Un elenco provvisorio e non esaustivo di informazioni lessicali per le quali possiamo trovare dei dati nei corpora testuali, con vari gradi di difficoltà e a vari livelli di completezza, è il seguente:

- dati di frequenza (a livello di entrata, forma flessa, accezione, associazioni fra parole, ecc.);
- sottocategorizzazione;
- collocazioni, frasi fisse, idioms;
- ruoli tematici, valenze;
- restrizioni semantiche sugli argomenti;
- soggetto tipico, oggetto, modificatore, ecc.;
- informazione aspettuale;
- nomi propri.

Consideriamo di nuovo in questo contesto la parola *libro* per vedere un esempio di informazioni ottenibili da testi. Se consideriamo i verbi relativi ai libri nel dizionario dell'italiano da noi elaborati possiamo notare che non figurano né *leggere*, né *scrivere*, *pubblicare*, ecc. Anche in questo caso la stessa osservazione è stata fatta per quanto riguarda i dizionari inglesi (si veda Boguraev et al., 1989), il che non succede certo per caso, ma è di nuovo una chiara indicazione della somiglianza che sussiste fra dizionari di lingue diverse.

Se osserviamo le definizioni di questi verbi, troviamo di solito parole più generiche relative a "cose stampate", quali *scrittura*, *parole*, *segni*, *lettere*, *scritto*, *opera*, *volume*, *giornale*. La parola *libro* appare, invece, solo in alcuni esempi. Il legame potebbe dunque essere stabilito solo in modo indiretto, dato che la parola *libro* è definita a sua volta attraverso parole quali *volume*, *opera*, *scritti*, *stampati*, ..., le stesse parole cioè che appaiono nelle definizioni dei verbi sopra citati.

Questi verbi sono invece associati in modo diretto con *libro* nel corpus testuale. Qui, infatti, su 3.222 occorrenze del lemma *libro*, troviamo che i verbi citati appaiono insieme con *libro* rispettivamente:

leggere 187 volte; *scrivere* 196; *pubblicare* 107.

E' dunque l'analisi di grandi corpora testuali che rende possibile il recupero di questo tipo di informazioni sulle collocazioni. Abbiamo implementato degli strumenti statistici che permettono l'estrazione semi-automatica di questi e altri tipi di dati dal nostro corpus (si veda Calzolari, Bindi, 1990).

Nell'analizzare un corpus di molti milioni di parole, siamo in un certo senso "costretti" a scoprire e a descrivere:

- usi che non sono descritti nei dizionari tradizionali;
- frequenze relative delle diverse accezioni, e dei diversi patterns sintattici;

- soprattutto, i segnali sintattico/grammaticali attraverso i quali si può almeno parzialmente ottenere una disambiguazione semantica, dato il fatto che, in presenza di una diversa sottocategorizzazione sintattica spesso cambia il significato di una parola, mentre, viceversa, non necessariamente troviamo una sola accezione con lo stesso quadro sintattico.

Nel raccogliere questo tipo di dati per un certo numero di parole, spesso ci rendiamo conto che tali dati dovrebbero essere riorganizzati in modo diverso rispetto a come si trovano ora nei dizionari tradizionali, se vogliamo che questi siano conformi all'uso attuale della lingua.

Allo scopo di automatizzare il recupero di questo tipo di informazioni direttamente dal corpus, dovremmo prima essere in grado di annotare il corpus per quanto riguarda almeno le parti del discorso. Per questo scopo esistono già alcuni sistemi funzionanti per l'inglese (si vedano ad esempio Hindle 1989, Webster, Marcus 1989). Questi sistemi verranno analizzati in dettaglio nel corso del progetto, essendo legati anche al problema, di estremo interesse per i sistemi di NLP, dell'estrazione con metodi statistici di conoscenze morfosintattiche e sintattiche da corpora testuali.

Questa stessa strategia di guardare a segnali sintattici (e collocazionali) per la disambiguazione semantica (da usarsi per scegliere fra diverse traduzioni di una stessa parola) è ora valutata in un progetto pilota - supportato dal *Consiglio d'Europa* - che stiamo sviluppando in un contesto multilingue.

3.2.5 Aspetti legali e organizzativi

Per preparare la costruzione di un network europeo, il nostro progetto si propone anche di chiarire gli aspetti seguenti:

- problemi legali connessi alla inclusione nel corpus di testi protetti da copyright;
- protezione dei diritti derivanti del valore "aggiunto" ai testi attraverso le operazioni di memorizzazione, strutturazione, analisi, ecc.;
- valutazione dei costi richiesti dalle diverse alternative e dalle diverse fasi di lavoro;
- identificazione di possibili sorgenti di materiali testuali in machine readable form, e valutazione della loro utilizzabilità sul piano tecnico e giuridico;
- identificazione di possibili partners e di condizioni per il loro coinvolgimento (autorità nazionali, industrie, agenzie di ricerca, ecc.);
- descrizione e classificazione dei potenziali utenti;
- scenari organizzativi per l'aggiornamento periodico dei corpora, e per la gestione dei servizi.

4 Iniziative internazionali di supporto alla creazione di risorse linguistiche per il NLP

4.1 Survey delle risorse linguistiche in machine readable form

E' evidente che qualsiasi progetto inteso a studiare la fattibilità di costruire risorse linguistiche adeguate per le varie lingue, deve prendere in considerazione le risorse già esistenti e la possibilità di riutilizzarle per incorporarle, in tutto o in parte.

Riconoscendo questa necessità, A.Zampolli e D.Walker hanno promosso un

survey (distribuito, con l'aiuto della CEE, ai membri di oltre 20 Associazioni scientifiche e professionali e alle industrie del settore), il quale ha lo scopo di creare un database delle risorse esistenti, o in corso di costruzione, per quanto attiene:

- collezioni e corpora di testi
- dizionari leggibili dal calcolatore
- basi di conoscenze lessicali per il NLP
- banche terminologiche
- basi di dati orali per il trattamento del parlato.

Le informazioni cercate riguardano, essenzialmente:

- la natura, la composizione, la provenienza dei dati
- il sistema di rappresentazione
- il sistema di acquisizione
- eventuali interventi in preeedizione
- tipo di uso e di utenti ai quali i dati sono destinati
- livelli di analisi e sistema di annotazione
- software per la gestione, l'accesso, la elaborazione
- condizioni e modalità per l'utilizzo dei dati da parte di ricercatori e/o industrie.

4.2 Verso uno standard internazionale: la "Text Encoding Initiative"

La possibilità di scambiare i corpora tra i vari centri del network europeo, di distribuire testi alla varietà di utenti esterni al network, di ripartire il costo delle ricerche e della implementazione di software specializzato per l'accesso e la elaborazione dei corpora, richiede, come condizione necessaria, che i materiali testuali siano rappresentati in "machine readable form" secondo uno schema comune di codificazione. Per rispondere a questa esigenza, il progetto NERC ha deciso di utilizzare lo standard proposto dalla cosiddetta "Text Encoding Initiative".

L'uso di calcolatori per lo studio di testi si è diffuso in varie discipline umanistiche (filologia, storia della letteratura, lessicografia, filosofia, antropologia, storia, ecc.) a partire dai primi anni '50. In tutti questi anni la comunità scientifica non è riuscita a sviluppare degli schemi comuni per il "mark-up" di "machine readable texts", e la situazione è stata descritta come un "virtual chaos".

Lo scambio di testi e la loro elaborazione per mezzo di software comuni sono molto difficoltosi, mentre i recenti sviluppi tecnologici e la diffusione capillare di mezzi di calcolo promettono di aumentare di un ordine di grandezza la quantità di testi disponibili in MRF.

In questa situazione, le tre maggiori Associazioni scientifiche del settore (ACH, ACL, ALLC) hanno promosso la "Text Encoding Initiative", un progetto internazionale che si propone di formulare e diffondere delle "Guidelines" per la codificazione e lo scambio di testi in MRF.

Il progetto è promosso e diretto da uno Steering Committee formato da due rappresentanti di ciascuna delle Associazioni promotrici. Quattro comitati, composti in parti eguali da ricercatori europei e nordamericani, hanno il compito, rispettivamente, di:

- Definire il metalinguaggio da utilizzare nel mark-up dei testi.

E' stato scelto a tale scopo lo SGML, anche in considerazione della analoga scelta dei maggiori organismi internazionali e nazionali, tra i quali la American Publisher Association, che coopera alla TEI.

- Studiare e definire gli standards relativi alla documentazione di testi in MRF.
Questa documentazione comprende una varietà di informazioni, che vanno dalle tradizionali indicazioni bibliografiche, alla specificazione degli interventi apportati ai testi in preedizione, alla scelta degli elementi testuali codificati, ecc.
- Identificare gli elementi testuali che possono comparire nei diversi tipi di testi nelle varie lingue e sono rappresentati nella tradizione tipografica, descriverli nelle loro strutture e funzioni, e proporre un insieme di "tags" standardizzati.
- prendere in esame i tipi più frequenti di analisi eseguite per arricchire i testi con annotazioni di vario tipo (linguistica, letteraria, ecc.), identificare le categorie descrittive utilizzate, e proporre un sistema comune di rappresentazione che tenga conto delle strutture e dei livelli concorrenti di descrizione.

Trenta tra le maggiori associazioni scientifiche e professionali internazionali, riunite in un advisory board, si sono impegnate a promuovere tra i propri membri le Guidelines prodotte dalla TEI.

Il progetto è finanziato, per la partecipazione degli americani, dal National Endowment for the Humanities, mentre la partecipazione degli europei è finanziata dalla CEE attraverso il coordinamento dell'Università di Pisa e del nostro Istituto.

La prima versione delle Guidelines è apparsa nel giugno 1990.
La versione definitiva è prevista per il luglio 1992.

PASSIONARIO	1SM	ANTICO LIBRO LITURGICO CATTOLICO	3		
OMILIARIO	1SM	ANTICO LIBRO LITURGICO CONTENENTE OMELIE	1		
EPISTOLARIO	1SM	LIBRO CHE CONTENEVA BRANI DI EPISTOLE E VANGELO	3		
ORA	1SF	LIBRO CHE CONTENEVA LE OPERAZIONI PROPRIE DELLE VARIE ORE	9		
SALTERIO	2SM	LIBRO CHE CONTIENE I SALMI	3		
RITUALE	2SM	LIBRO CHE CONTIENE LE NORME CHE REGOLANO UN RITO	3		
UFFICIOLO	1SM	LIBRO CHE CONTIENE LE PREGHIERE IN ONORE DELLA VERGINE	3		
UFIZIOLO	1SM	LIBRO CHE CONTIENE LE PREGHIERE IN ONORE DELLA VERGINE	3		
CANTORINO	1SM	LIBRO CHE CONTIENE LE REGOLE DEL CANTO FERMO	3		
PORTULANO	1SM	1LIBRO CHE DESCRIVE MINUTAMENTE LA COSTA	34Z		
GUIDA	1SF	LIBRO CHE INSEGNA PRIMI ELEMENTI DI ARTE O TECNICA	3		
GRADUALE	2SM	LIBRO CHE RACCOGLIE I GRADUALI DELL'ANNO LITURGICO	3		
GIORNALMASTRO	1SM	LIBRO CHE RIUNISCE IL GIORNALE E IL MASTRO,PER CONTABILITA'	3		
ANNUARIO	1SM	LIBRO CHE SI PUBBLICA ANNUALMENTE	3		
....					
EFEMERIDE	1SF	LIBRO IN CUI ERANO ANNOTATI I FATTI CHE ACCADEVANO OGNI GIOR	3		
EFFEMERIDE	1SF	LIBRO IN CUI ERANO ANNOTATI I FATTI CHE ACCADEVANO OGNI GIOR	3		
COPIAFATTURE	1SM	LIBRO IN CUI SI COPIANO LE FATTURE	3		
SALDACONTI	1SM	LIBRO IN CUI SONO REGISTRATI I CREDITI E I DEBITI	3		
TASCABILE	2SM	LIBRO IN EDIZIONE ECONOMICA E PICCOLO FORMATO	3		
PERGAMENO	1SM	1LIBRO IN PERGAMENA	3	1	E
BENEDIZIONALE	1SM	LIBRO LITURGICO	3		
MESSALE	1SM	LIBRO LITURGICO CATTOLICO	3		
LEZIONARIO	1SM	LIBRO LITURGICO CON LE#LEZIONI(LEZIONE)DI UFFICI DIVINI	3		
CORALE	2SM	LIBRO LITURGICO CONTENENTE GLI UFFICI DEL#CORO()	1		
EVANGELIARIO	1SM	LIBRO LITURGICO CONTENENTE PASSI DELL' EVANGELO	1		
INNARIO	1SM	LIBRO LITURGICO,NEL CATTOLICESIMO E NELLE CHIESE ORIENTALI	3		
....					
CORANO	1SM	LIBRO SACRO DEI MUSSULMANI	3		
AVESTA	1SM	LIBRO SACRO DELLA RELIGIONE ZOROASTRIANA	3		
GENESI	1SF	PRIMO LIBRO DEL PENTATEUCO NELLA BIBBIA	3		
ALBO	2SM	SPECIE DI LIBRO CONTENENTE FOTOGRAFIE,DISCHI,FRANCOBOLLI	3		
LEVITICO	2SM	TERZO LIBRO BIBLICO DEL PENTATEUCO	9		
SAPIENZA	1SF	UNO DEI LIBRI DELL'ANTICO TESTAMENTO	3		
SAPIENZIA	1SF	1UNO DEI LIBRI DELL'ANTICO TESTAMENTO	3		

Fig. 1. Alcuni degli iponimi di *libro*.

ALLIBRARE	1VT	REGISTRARE SU UN LIBRO DI CONTI	1		
CARTOLINARE	1VT	RILEGARE UN LIBRO ALLA RUSTICA	3		
CIRCOLARE	1VIT	PASSARE DALL'UNA ALL'ALTRA PERSONA,DI DANARO,LIBRI	3		E
DISTRIBUIRE	1VT	DIFFONDERE TRA TUTTI I RIVENDITORI LIBRI,GIORNALI	3		
DIVOLGARE	1VTP	1RENDERE FINANZIARIAMENTE DISPONIBILI LIBRI,SAGGI	3		E
DIVULGARE	1VTP	RENDERE FINANZIARIAMENTE DISPONIBILI LIBRI,SAGGI	3		E
INTERFOGLIARE	1VT	INTERPORRE,CUCIRE TRA I FOGLI DI UN LIBRO FOGLI BIANCHI	3		
INTESTARE	1VTP	FORNIRE DI INTESTAZIONE O TITOLO UN LIBRO	1		
RITONDARE	1VT	1PAREGGIARE,TAGLIANDO LE SPORGENZE,DETTO DI LIBRI,TESSUTI	3	1	
SCARTABELLARE	1VT	SCORRERE IN FRETTA E DISORDINATAMENTE LE PAGINE D'UN LIBRO	3		
SCOMPAGINARE	1VTP	DISFARE,ROVINARE LA LEGATURA DI LIBRI	3		
SCRITTURARE	1VT	ANNOTARE,REGISTRARE SU LIBRI O SCRITTURE CONTABILI	3		
SFASCICOLARE	1VT	SCOMPORRE UN LIBRO,UN QUADERNO NEI FASCICOLI DI CUI E' FATTO	3		
SFOGLIARE	2VTP	SCORRERE UN LIBRO RAPIDAMENTE	3		
SFOGLIARE	2VTP	TAGLIARE LE PAGINE DI UN LIBRO	3	3	
SQUADERNARE	1VTP	3VOLTARE E RIVOLTARE PAGINE DI LIBRI,QUADERNI	3	3	
TOSARE	1VT	PAREGGIARE I FOGLI DEI LIBRI NEL RILEGARLI	3	3	E

Fig. 2. Verbi relativi a *libri*.

ADESPOTA	1A	3ANONIMO/DETTO DI LIBRO,CODICE,MANOSCRITTO DI AUTORE IGNOTO	5		
ADESPOTO	1A	ANONIMO/DETTO DI LIBRO,CODICE,MANOSCRITTO DI AUTORE IGNOTO	5		
APOCRIFO	1A	DETTO DI LIBRO NON RICONOSCIUTO COME CANONICO	3		
CARTOLIBRARIO	1A	DI COMMERCIO DI LIBRI E OGGETTI DA CANCELLERIA	3		
CIRCOLANTE	1A	CHE DA' LIBRI A PRESTITO AGLI ABBONATI A TURNO	9		
COMMERCIALE	1A	DETTO DI LIBRO,FILM CHE MIRA SOLO A OTTENERE BUONI INCASSI	3		F
COPERTINATO	1A	DETTO DI LIBRO O FASCICOLO CON COPERTINA	1		
DEUTEROCANONICO	1A	DEI LIBRI DELL'ANTICO TESTAMENTO RESPINTI COME APOCRIFI	3		
EDITORE	1A	CHI PUBBLICA LIBRI,RIVISTE	3		
ERUDITO	1A	LIBRO ERUDITO		T	
INTESTATO	1A	FORNITO DI TITOLO O INTESTAZIONE,DETTO DI LIBRO,LETTERA	3		
INTONSO	1A	3DI LIBRO CUI NON SONO ANCORA STATE TAGLIATE LE PAGINE	3		F
LIBERIANO	3A	CHE RIGUARDA IL LIBRO	36K		
LIBRARIO	1A	DI,RELATIVO A LIBRO	1		
LIBRESCO	1A	CHE DERIVA DAI LIBRI E NON DALLA VIVA ESPERIENZA	1		P
MASTRO	2A	LIBRO MASTRO		L	
MOSAICO	2A	RELATIVO AI LIBRI BIBLICI	3		
PAGA	4A	LIBRO PAGA		L	
POSTUMO	1A	DI LIBRO PUBBLICATO DOPO LA MORTE DELL'AUTORE	3		
PROTOCANONICO	1A	DETTO DI CIASCUN LIBRO BIBLICO INSERITO PER PRIMO NEL CANONE	3		
SAPIENZIALE	1A	CHE SI RIFERISCE AI LIBRI SAPIENZIALI	3		E

Fig. 3. Aggettivi relativi a *libri*.

RISVOLTO	1SM	ALETTA/ PARTE DELLA SOPRACOPERTA DI LIBRO RIPIEGATA	5		
BIBLIOFILO	1SG	AMATORE,RICERCATORE,COLLEZIONISTA DI LIBRI	3		
BIBLIOFILIA	1SF	AMORE PER I LIBRI	3		
REGGILIBRI	1SM	ARNESE PIEGATO AD ANGOLO RETTO PER REGGERE IN PIEDI LIBRI	3		
BIBLIOIATRICA	1SF	3ARTE DEL RESTAURO DEI LIBRI	3	3	
ERMENEUTICA	1SF	ARTE DI INTERPRETARE MONUMENTI,LIBRI ANTICHI	3		
SFOGLIATA	2SF	ATTO DELLO SCORRERE UN LIBRO E SIMILI	1		
PUBBLICAZIONE	1SF	ATTO EFFETTO DEL RENDERE PUBBLICO O DEL PUBBLICARE LIBRI	1		
BANCHEROZZO	1SM	1BANCARELLA DI LIBRI ALL' APERTO	3	1	
ZAZZERA	1SF	BARBA,RICCIO/ PARTE RUVIDA INTONSA DEI LIBRI	5		
PORTACARTE	1SM	BORSA PER METTERVI CARTE,DOCUMENTI,LIBRI	3		
BOTTELLO	1SM	3CARTELLINO CHE SI METTE SU LIBRI E BOTTIGLIE	3	3	
CARTOLIBRERIA	1SF	CARTOLERIA AUTORIZZATA ALLA VENDITA DI LIBRI	3		
CANONE	1SM	CATALOGO DEI LIBRI SACRI RICONOSCIUTI AUTENTICI	3		
REDATTORE	1SN	CHI CURA FASI PER PUBBLICAZIONE DI LIBRI IN CASE EDITRICI	3		
CARRETTINISTA	1SM	CHI ESPONE O VENDE LIBRI SU UN CARRETTINO	1		
BIBLIOTECA	1SF	COLLEZIONE DI LIBRI SIMILI PER FORMATO ARGOMENTO EDITORE	3		
LIBRATA	1SF	COLPO DATO CON UN LIBRO	1		
....					
BIBLIOTECA	1SF	EDIFICIO CON RACCOLTE DI LIBRI A DISPOSIZIONE DEL PUBBLICO	3		
BIBLIOGRAFIA	1SF	ELENCO DI LIBRI CONSULTATI PER COMPILAZIONE DI OPERE	3		
INDICE	1SM	ELENCO ORDINATO DI CAPITOLI O PARTI DI LIBRO	3		
BIBLIOLATRIA	1SF	FEDE CIECA NEI LIBRI STAMPATI	3		
....			39Q		
LIBRERIA	1SF	LUOGO O MOBILE IN CUI SONO ACCOLTI E CUSTODITI I LIBRI	3		C
BIBLIOTECA	1SF	LUOGO OVE SONO RACCOLTI E CONSERVATI LIBRI	3		
BIBLIOMANIA	1SF	MANIA DI RICERCARE E COLLEZIONARE LIBRI	3		
BIBLIOTECA	1SF	MOBILE A MURO CON SCAFFALI PER LIBRI	3		
CLASSIFICATORE	1SN	MOBILE PER CONTENERE LIBRI DOCUMENTI	3		
LIBRERIA	1SF	NEGOZIO O EMPORIO DI LIBRI			
FRONTISPIZIO	1SM	PAGINA ALL' INIZIO DI UN LIBRO CON TITOLO NOTE TIPOGRAFICHE	3		
ANTIPORTA	1SF	PAGINA CON TITOLO PRECEDENTE FRONTESPIZIO DI LIBRO	3		
TAVOLA	1SF	PAGINA FOGLIO DI LIBRO CON ILLUSTRAZIONI	3		
INTERFOGLIO	1SM	PAGINA INTERPOSTA TRA I FOGLI DI UN LIBRO	3		
LIBRERIA	1SF	RACCOLTA DI LIBRI LIBRO	1		
BIBLIOLOGIA	1SF	SCIENZA DEI LIBRI	3		
LIBRAIO	1SN	VENDITORE DI LIBRI	1		
LIBRARO	1SN	1VENDITORE DI LIBRI			
VERSO	3SM	VERSETTO/SUDDIVISIONE IN FRASI DELLE PARTI DI LIBRI SACRI	5		E

Fig. 4. Alcuni dei nomi relativi a *libri*.

```
STRUMENTO          ---->>ABBASSALINGUA                1SM    00
                         ABERROMETRO                  1SM    00
                         ACCELEROGRAFO                1SM    00
                         ACCELEROMETRO                1SM    00
                         ACCHIAPPAMOSCHE              1SN    00
                         ACCIAINO                     1SM    00
                         AEROFONO                     1SM    00
                         AEROMETRO                    1SM    00
                         AEROSCOPIO                   1SM    00
                         AFFILATOIO                   1SM    00
                         AGGUAGLIATOIO                1SM    00
                         AGO                          1SM    0A
                         ALCOOLIMETRO                 1SM    00
                         ALGESIMETRO                  1SM    00
                         AMMOSTATOIO                  1SM    00
                         AMPEROMETRO                  1SM    00
                         ANALIZZATORE                 1SN    00
                         ANCORA                       1SF    10
                         ANEMOMETRO                   1SM    00
                         ANEMOSCOPIO                  1SM    00
                         ANGELICA                     1SF    00
                         APRIBOCCA                    1SM    00
                         APRICASSE                    1SM    00
                         ARCHIPENDOLO                 1SM    00
                         ARMA                         1SF    00
                         ARMONICA                     1SF    00
                         ARMONIO                      1SM    00
                         ARMONIUM                     1SM    00
                         ARPA                         1SF    10
                         ARPEGGIONE                   1SM    00
                         ARRIDATOIO                   1SM    00
                         ASPERSORIO                   1SM    00
                         ASPIRATORE                   1SM    00
                         ASSIOMETRO                   1SM    00
                         ASTIGMOMETRO                 1SM    00
                         ASTROFOTOMETRO               1SM    00
                         ASTROGRAFO                   1SM    00
                         ASTROLABIO                   1SM    00
                         ATTINOMETRO                  1SM    00
                         ATTREZZO                     1SM    0A
                         AUDIOMETRO                   1SM    00
                         AULOS                        1SM    00
                         AVENA                        1SF    00
                         BADILE                       1SM    00
```

Fig. 5. I primi iponimi di *strumento*.

```
AFFOSSATORE        1SN    ATTREZZO AGRICOLO PER SCAVARE FOSSI                                  3
ALLARGATESE        1SM    ATTREZZO USATO PER ALLARGARE LE TESE DEI CAPPELLI                    3
ALLISCIATOIO       1SM    ATTREZZO USATO IN FONDERIA PER PREPARARE LE FORME                    3
ANELLO             1SM    ATTREZZO GEMELLARE IN GINNASTICA                                     3
APISCAMPO          1SM    ATTREZZO PER IMPEDIRE L' ASCESA DELLE API AL MELARIO                 3
APPOGGIO           1SM    ATTREZZO GINNICO FORMATO DA BLOCCHETTI RETTANGOLARI DI LEGNO         3
ARATRO             1SM    ATTREZZO AGRICOLO ATTO A ROMPERE,DISSODARE IL TERRENO                3
ARNESE             1SM    ATTREZZO DA LAVORO                                                   3
ASPO               1SM    ASPA,ANNASPO,NASPO/ ATTREZZO CHE SERVE AD ESEGUIRE L'ASPATURA       54E
ASTA               1SF    ATTREZZO DI FORMA TUBOLARE NELL' ATLETICA                            3
BACCHETTA          1SF    ATTREZZO PER ESERCIZI GINNICI COLLETTIVI                             3
BARRAMINA          1SF    ATTREZZO PER LA PERFORAZIONE DELLE ROCCE                             3
BASTONCINO         1SM    ATTREZZO DEGLI SCIATORI CON RACCHETTA CIRCOLARE                      3
BASTONE            1SM    MAZZA/ ATTREZZO SPORTIVO                                             5
CACCIAVITE         1SM    ATTREZZO PER STRINGERE O ALLENTARE LE VITI                           3
CAVALLINA          1SF    ATTREZZO PER ESERCIZI DI VOLTEGGIO NELLA GINNASTICA                  3
CAVALLO            1SD    ATTREZZO PER ESERCIZI DI VOLTEGGIO NELLA GINNASTICA                  3    5
CERCHIO            1SM    ATTREZZO STRUTTURA FIGURA A FORMA DI CERCHIO                         3
CESTA              1SF    CHISTERA/ ATTREZZO DI VIMINI USATO NELLA PELOTA BASCA                5
CHIAVE             1SF    ATTREZZO METALLICO PER PROVOCARE CONTATTI                            3
CHIAVE             1SF    ATTREZZO METALLICO PER METTERE IN MOTO MECCANISMI                    3
CHIAVE             1SF    ATTREZZO METALLICO PER ALLENTARE E STRINGERE VITI O DADI             3
CHIODO             1SM    ATTREZZO IN METALLO DEGLI ALPINISTI                                  3
CHIOVO             1SM    1ATTREZZO IN METALLO DEGLI ALPINISTI                                 3    1
CILINDRO           1SM    ATTREZZO CILINDRICO NELLA GINNASTICA                                 3
CLAVA              1SF    ATTREZZO IN LEGNO USATO PER ESERCIZI GINNICI                         3
COLTIVATORE        2SN    ATTREZZO PER SMUOVERE E SMINUZZARE LA SUPERFICIE DEL TERRENO         3
CORDA              1SF    ATTREZZO DA ALPINISMO O GINNASTICA                                   39L
CUCCHIAIA          1SF    ATTREZZO PER ESTRARRE DETRITI DI ROCCIA                              3
CUCITRICE          2SF    ATTREZZO USATO NEGLI UFFICI PER UNIRE FOGLI                          3
DISCO              1SM    ATTREZZO CIRCOLARE CHE SI LANCIA IN GARE SPORTIVE                    3
ERPICE             1SM    ATTREZZO DI FERRO PER LAVORARE IL TERRENO                            3
ESTENSORE          2SI    ATTREZZO GINNICO                                                     3
ESTIRPATORE        3SM    ATTREZZO PER SMUOVERE O LIBERARE IL TERRENO DA ERBACCE               3
FALCE              1SF    ATTREZZO PER TAGLIARE A MANO CEREALI ED ERBE                         3
FIOCINA            1SF    ATTREZZO CON TRE O PIU' DENTI FISSI PER CATTURARE PESCI              3
....
UTENSILE           2SM    OGNI ATTREZZO PER LAVORARE LEGNO,PIETRE,MATERIALI                    3
VANGHETTA          1SF    ATTREZZO LEGGERO DI SOLDATO PER PICCOLI LAVORI DI STERRO             3
VOGADORE           1SI    1ATTREZZO GINNICO PER MOVIMENTO DA REMATORE                          3
VOGATORE           1SN    ATTREZZO GINNICO PER MOVIMENTO DA REMATORE                           3
VOLTARISO          1SM    ATTREZZO PER RIVOLTARE SULL'AIA MODESTE QUANTITA' DI RISO            3
ZAPPA              1SF    ATTREZZO MANUALE PER LAVORARE IL TERRENO                             3
```

Fig. 6. Alcuni degli iponimi di *attrezzo* con le loro definizioni.

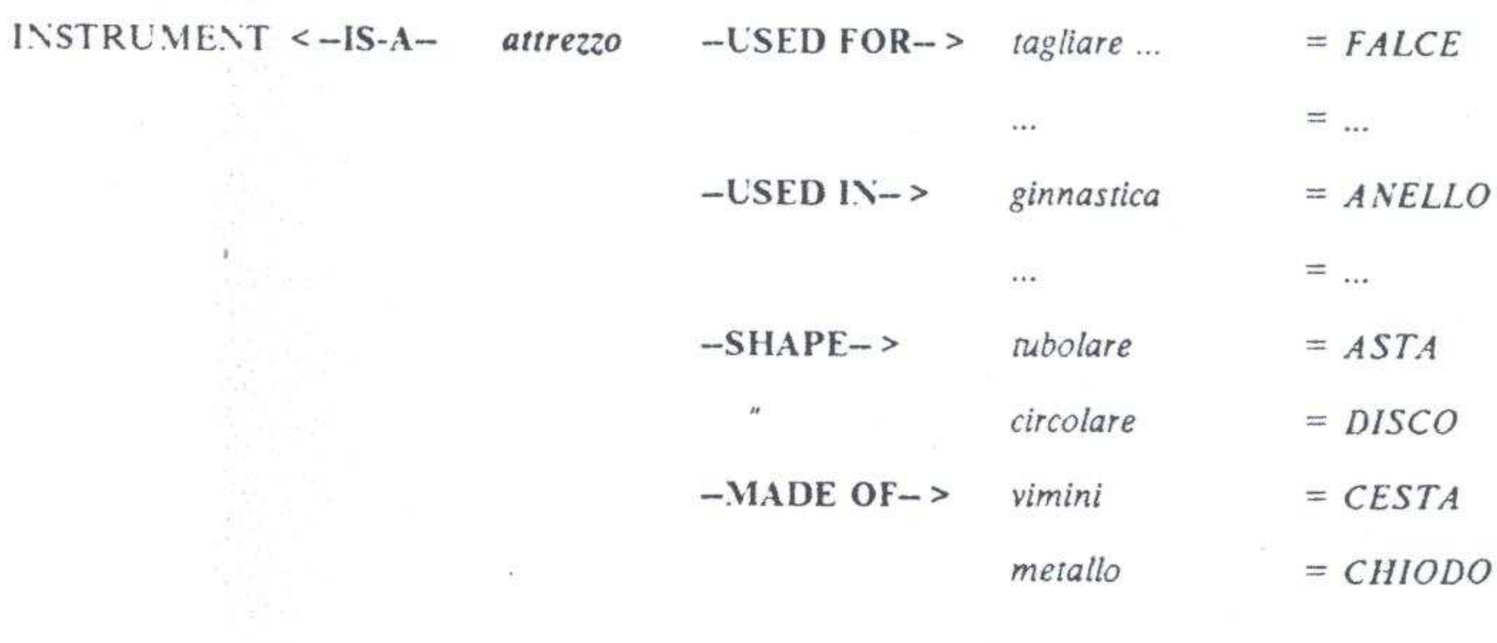

Fig. 7. Diagramma di una porzione di rete per *attrezzo*.

FORMICAIO	SM	MOLTITUDINE DI	PERSONE
GREGGE	SN	MOLTITUDINE DI	PERSONE
STORMO	SM	MOLTITUDINE DI	PERSONE
MANO	SF	GRUPPO DI	PERSONE
ROSA	SF	CERCHIA/ GRUPPO INSIEME DI	PERSONE
BRANCO	SM	INSIEME DI	PERSONE
CIRCOLO	SM	CENACOLO,SODALIZIO/INSIEME DI	PERSONE
COMMISSIONE	SF	GRUPPO DI	PERSONE A CUI E' AFFIDATO UN UNCARICO PUBBLICO
POPOLAZIONE	SF	INSIEME DELLE	PERSONE ABITANTI IN UN LUOGO
ORGANICO	SM	COMPLESSO DI	PERSONE ADDETTE A CERTE ATTIVITA'
SEGRETERIA	SF	INSIEME DELLE	PERSONE ADDETTE A UNA SEGRETERIA
SQUADRA	SF	COMPLESSO DI	PERSONE ADDETTE A UNO STESSO LAVORO
CIURMA	SF	INSIEME DELLE	PERSONE ADDETTE AI LAVORI DELLA TONNARA
NAZIONE	SF	INSIEME DI	PERSONE APPARTENENTI A STESSA STIRPE
FAMIGLIA	SF	COMPLESSO DI	PERSONE AVENTI UN ASCENDENTE DIRETTO COMUNE
VICINATO	SM	INSIEME DI	PERSONE CHE ABITANO UNA STESSA CASA
CORTE	SF	GRUPPO DI	PERSONE CHE ACCOMPAGNA UN PERSONAGGIO IMPORTANTE
LEGA	SF	INSIEME DI	PERSONE CHE AGISCONO PER UTILE PROPRIO
AUDITORIO	SM	UDITORIO/COMPLESSO DI	PERSONE CHE ASCOLTANO
UDIENZA	SF	UDITORIO/INSIEME DI	PERSONE CHE ASCOLTANO
CAROVANA	SF	GRUPPO DI	PERSONE CHE ATTRAVERSANO CON CARRI LUOGHI DESERTI
CORO	SM	GRUPPO DI	PERSONE CHE CANTANO INSIEME
MALAVITA	SF	L'INSIEME DELLE	PERSONE CHE CONDUCONO VITA DISSOLUTA
CROCCHIO	SM	GRUPPO DI	PERSONE CHE CONVERSANO
CORO	SM	GRUPPO DI	PERSONE CHE DICONO,GRIDANO Q.C. CONTEMPORANEAMENTE
CONCISTORO	SM	GRUPPO DI	PERSONE CHE DISCUTONO
FINANZA	SF	COMPLESSO DI	PERSONE CHE ESPLICANO ATTIVITA' BANCARIA
....			
FRONTE	SN	COMPLESSO DI	PERSONE OMOGENEO PER FINALITA' CONSUETUDINI
ARISTOCRAZIA	SF	COMPLESSO DI	PERSONE PIU' QUALIFICATE PER UNA ATTIVITA'
CHIESA	SF	INSIEME DI	PERSONE PROFESSANTI LA MEDESIMA DOTTRINA
DRAPPELLO	SM	GRUPPO DI	PERSONE RACCOLTE INSIEME
COMPAGNIA	SF	COMPLESSO DI	PERSONE RIUNITE INSIEME PER ATTIVITA' COMUNI
GRUPPO	SM	INSIEME DI	PERSONE UNITE DA VINCOLI NATURALI O DI INTERESSE

Fig. 8. Alcuni dei nomi denotanti **SET OF** *persone*.

ARCIPELAGO	SM	GRUPPO INSIEME DI	OGGETTI
ANTIQUARIATO	SM	COMMERCIO O RACCOLTA DI	OGGETTI ANTICHI
SERVIZIO	SM	INSIEME DI	OGGETTI CHE SERVONO A UN DETERMINATO SCOPO
TROFEO	SM	INSIEME DI	OGGETTI CHE TESTIMONIANO SUCCESSI E VITTORIE
AFFARDELLAMENTO	SM	COMPLESSO DEGLI	OGGETTI CONTENUTI NELLO ZAINO DEL SOLDATO
ARGENTERIA	SF	COMPLESSO DI	OGGETTI D'ARGENTO
ORERIA	SF	COMPLESSO DI	OGGETTI D'ORO
COLLEZIONE	SF	RACCOLTA DI	OGGETTI DELLA STESSA SPECIE
CRISTALLERIA	SF	INSIEME DEGLI	OGGETTI DI CRISTALLO DA TAVOLA
CIANFRUSAGLIA	SF	CHINCAGLIERIA/INSIEME DI	OGGETTI DI POCO PREGIO
CIANFRUSCAGLIA	SF	CHINCAGLIERIA/INSIEME DI	OGGETTI DI POCO PREGIO
ASSORTIMENTO	SM	INSIEME DI	OGGETTI DI STESSO GENERE DIVERSI NEI PARTICOLARI
ARSENALE	SM	INSIEME DI	OGGETTI DIVERSI
SUPPELLETTILE	SF	OGGETTO O INSIEME DI	OGGETTI IN UNA SCUOLA CHIESA E SIMILI
INTRECCIO	SM	COMPLESSO DI	OGGETTI INTRECCIATI
ATTREZZERIA	SF	INSIEME DI	OGGETTI NECESSARI PER UNA SCENA TEATRALE
SUPPELLETTILE	SF	OGGETTO O INSIEME DI	OGGETTI NELL'ARREDAMENTO DELLA CASA
ARREDO	SM	OGGETTO O COMPLESSO DI	OGGETTI PER GUARNIRE AMBIENTI
COMPLETO	SM	INSIEME DI	OGGETTI PER UN USO DETERMINATO
BAROCCUME	SM	INSIEME DI	OGGETTI PRETENZIOSI E DI CATTIVO GUSTO
GIOIELLERIA	SF	INSIEME DI	OGGETTI PREZIOSI
SUPPELLETTILE	SF	OGGETTO O INSIEME DI	OGGETTI RINVENUTI IN UNO SCAVO

Fig. 9. Nomi denotanti **SET OF** *oggetti.*

ASSESTATO	A	ASSENNATO,AVVEDUTO,DETTO DI	PERSONA
BARLACCIO	A	MALATICCIO,DEBOLE,DETTO DI	PERSONA
INSENSATO	A	STUPIDO,DEMENTE,DETTO DI	PERSONA
PRIMITIVO	A	C=INCIVILITO/SEMPLICE,ROZZO,CREDULONE,DETTO DI	PERSONA
PROVETTO	A	MATURO,DETTO DI	PERSONA
RIMESSO	A	LANGUIDO,LENTO,FIACCO,DETTO DI	PERSONA
RINCRESCIOSO	A	CHE SENTE RINCRESCIMENTO,DETTO DI	PERSONA
RIPOSANTE	A	CALMO,TRANQUILLO DETTO DI	PERSONA
RISPETTOSO	A	CHE HA,E' PIENO DI*RISPETTO(),DETTO DI	PERSONA
ROBUSTO	A	FORTE/CHE POSSIEDE FORZA,ENERGIA,DETTO DI	PERSONA
ROCO	A	RAUCO,DETTO DI	PERSONA
ROGNOSO	A	MISERO,MESCHINO,NOIOSO,DETTO DI	PERSONA
RUDE	A	ROZZO,GROSSOLANO,DETTO DI	PERSONA
RUGIADOSO	A	SANO,FLORIDO,DETTO DI	PERSONA
RUSTICO	A	NON MOLTO SOCIEVOLE NE' RAFFINATO,DETTO DI	PERSONA
RUVIDO	A	DI MANIERE ROZZE,DI CARATTERE ASPRO,DETTO DI	PERSONA
....			
ADOMBRARE	VTE	INSOSPETTIRSI,TURBARSI,DETTO DI	PERSONA
ARRABBIARE	VIE	ESSERE PRESO DALL'IRA,DALLA COLLERA DETTO DI	PERSONA
CORVETTARE	VI	SALTARE,BALZARE,DETTO SPEC. DI	PERSONA
CUCCIARE	VET	GIACERSI/STARE A LETTO,DETTO DI	PERSONA
IMBIZZARRIRE	VET	INCOLLERIRE O DIVENTARE IRREQUIETO DETTO DI	PERSONA
IMPROSCIUTTIRE	VI	DIVENTARE ASCIUTTO COME UN PROSCIUTTO,DETTO DI	PERSONA
RABBRUSCARE	VEY	ADOMBRARSI/OFFUSCARSI IN VOLTO,DETTO DI	PERSONA
RICEVERE	VT	AMMETTERE,DETTO DI	PERSONA
RIDURRE	VT P	METTERE IN CONDIZIONI PEGGIORI,DETTO DI	PERSONA
RIMETTERE	VT PI	RISTABILIRSI,DETTO DI	PERSONA
RINFIERIRE	VI	INFIERIRE DI NUOVO O DI PIU',DETTO DI	PERSONA
RINSECCHIRE	VIT	DIVENTARE MAGRO,ASCIUTTO,DETTO DI	PERSONA
RINVENIRE	VI	RIANIMARSI,RIAVERSI/RICUPERARE I SENSI DETTO DI	PERSONA
RISALTARE	VNI	EMERGERE,DISTINGUERSI,DETTO DI	PERSONA
RISORGERE	VI T	SOLLEVARSI,RIAVERSI DETTO DI	PERSONA
RISPUNTARE	VIT	RIAPPARIRE,RICOMPARIRE,DETTO DI	PERSONA
RISURGERE	VI T	SOLLEVARSI,RIAVERSI,DETTO DI	PERSONA
RIUSCIRE	VI	RAGGIUNGERE IL FINE,LO SCOPO,DETTO DI	PERSONA
ROTOLARE	VTIR	GIRARSI SU DI SE',VOLTOLARSI,DETTO DI	PERSONA
ROVINARE	VITR	CADERE IN BASSO,DETTO DI	PERSONA
....			
CORDIALE	A	DETTO DI	PERSONA AFFABILE,GENTILE,APERTA
LONGO	A	CHE SI ESTENDE IN ALTEZZA,DETTO DI	PERSONA ALTA E MAGRA
LUNGO	A	CHE SI ESTENDE IN ALTEZZA,DETTO DI	PERSONA ALTA E MAGRA
PRODIGIO	A	DETTO DI	PERSONA CHE E' ECCEZIONALE
SUPINO	A	C=PRONO/DETTO DI	PERSONA CHE GIACE SUL DORSO
LACERO	A	CENCIOSO/DETTO DI	PERSONA CHE INDOSSA VESTITI LOGORI
SCIVOLOSO	A	DETTO DI	PERSONA CHE NASCONDE LE SUE VERE INTENZIONI
IMPREGIUDICATO	A	DETTO DI	PERSONA CHE NON HA AVUTO CONDANNE PENALI
IMPETTITO	A	DETTO DI	PERSONA CHE STA ERETTA E COL PETTO IN FUORI
ASOCIALE	A	DETTO DI	PERSONA CHIUSA INTROVERSA
....			
NAUFRAGARE	VI	ESSERE SUL BASTIMENTO CHE ROMPE IN MARE,DETTO DI	PERSONE
RICONGIUNGERE	VT D	CONGIUNGERSI DI NUOVO,RIUNIRSI,DETTO DI	PERSONE
RIMESCOLARE	VTP	INTROMETTERSI,MISCHIARSI A UN GRUPPO,DETTO DI	PERSONE
ROVESCIARE	VTP	ABBANDONARSI,DETTO DI	PERSONE
SBOCCARE	VIT	ARRIVARE IN UN DATO LUOGO,DETTO DI	PERSONE
SCHIAMAZZARE	VI	VOCIARE,STREPITARE,DETTO DI	PERSONE
SPELLICCIARE	VTB	PICCHIARSI,AZZUFFARSI RABBIOSAMENTE,DETTO DI	PERSONE
ULULARE	VI	EMETTERE PROLUNGATI,CUPI LAMENTI,DETTO DI	PERSONE

Fig. 10. Alcuni degli aggettivi e dei verbi che si possono predicare di *persone.*

ACCESO	A	VIVO,INTENSO,DETTO DI	COLORE
CHIARO	A	C=SCURO/PALLIDO,TENUE,POCO INTENSO DETTO DI	COLORE
CUPO	A	DI TONALITA' SCURA DETTO DI	COLORE
SERPATO	A	CHE E' SCREZIATO,COME LA PELLE DEL SERPENTE,DETTO DI	COLORE
SQUILLANTE	A	VIVACE,INTENSO,DETTO DI	COLORE
STABILE	A	CHE NON SBIADISCE,DETTO DI	COLORE
TENUE	A	PALLIDO/NON MOLTO VIVO DETTO DI	COLORE
RISCHIARARE	VTE	FARSI CHIARO,LUMINOSO,DETTO DI	COLORE
SCARICARE	VTRIP	PERDERE VIVACITA',SBIADIRE,DETTO DI	COLORE
BERRETTINO	A	DETTO DI	COLORE AZZURRO CINEREO SU VASI DI MAIOLICA
CALCE	A	DETTO DI	COLORE BIANCO INTENSO
GIGLIACEO	A	DETTO DI	COLORE CHE RICORDA QUELLO DEL GIGLIO
SCURO	A	C=CHIARO/DETTO DI	COLORE CHE TENDE AL NERO
BRUNO	A	DETTO DEL	COLORE DEL MANTELLO DEI BOVINI
ALBICOCCA	A	DETTO DI	COLORE GIALLO ARANCIATO
ZAFFERANO	A	DETTO DI	COLORE GIALLO INTENSO
ISABELLA	A	DETTO DI	COLORE GIALLO TIPICO DI MANTELLO EQUINO
PERLA	A	DETTO DI	COLORE LATTIGINOSO E OPALESCENTE
TERRA	A	DETTO DI	COLORE MARRONE CHIARO SFUMATO AL GRIGIO
SUDICIO	A	DETTO DI	COLORE NON BRILLANTE,NON VIVO
DISUGUAGLIATO	A	DETTO DI	COLORE NON UNIFORME DI UNA TINTURA
NEGRO	A	DETTO DEL	COLORE PIU' SCURO
NERO	A	DETTO DEL	COLORE PIU' SCURO
GIACINTINO	A	DETTO DEL	COLORE ROSSASTRO,TIPICO DEL GIACINTO
TANGO	A	DETTO DI	COLORE ROSSO ASSAI BRILLANTE
GRANATA	A	DETTO DI	COLORE ROSSO SCURO
PULCE	A	DETTO DI	COLORE TRA GRIGIO E VERDE
RUGGINE	A	DETTO DI	COLORE TRA IL MARRONE E IL ROSSO SCURO
LILLA'	A	GRIDELLINO/DETTO DI	COLORE TRA ROSA E VIOLA
GIADA	A	DETTO DI	COLORE VERDAZZURRO CHIARO
SBIADATO	A	SBIADITO,TENUE,PALLIDO,DETTO DI	COLORI
ADDOLCIRE	VTP	AMMORBIDIRE,DETTO DI	COLORI
DISCORDARE	VE	STONARE/NON ARMONIZZARE,DETTO DI	COLORI
SBIADIRE	VET	SCOLORIRE,STINGERE/DIVENTARE PALLIDO,SMORTO,DETTO DI	COLORI
SGARGIARE	VI	ESSERE ECCESSIVAMENTE VIVACE E VISTOSO,DETTO DI	COLORI
SMONTARE	VTIP	SCHIARIRE,SCOLORIRE,STINGERE,DETTO DI	COLORI
TRIONFARE	VIT	RISALTARE/FARE SPICCO,DETTO DI	COLORI
USCIRE	VIT	RISALTARE DETTO DI	COLORI
SMORTO	A	CHE E' PRIVO DI SPLENDORE E VIVACITA' DETTO DI	COLORI E SIM.
ALLEGRO	A	VIVACE,BRIOSO DETTO DI	COLORI SUONI E SIMILI
RISALTARE	VNI	SPICCARE NITIDAMENTE,DETTO DI	COLORI,DISEGNI,PITTURE
TENDERE	VT IP	AVVICINARSI AD UNA GRADAZIONE DETTO DI	COLORI,SAPORI,ODORI

Fig. 11. Alcuni degli aggettivi e dei verbi che si possono tipicamente predicare di *colori*.

VENDE	---->>AGNELLAIO	1SI	CHI MACELLA O VENDE AGNELLI	1	
	AGORAIO	1SM	CHI FA O VENDE AGHI		
	ALABASTRAIO	1SI	CHI VENDE OGGETTI DI ALABASTRO		
	ARAZZIERE	1SI	CHI TESSE E VENDE ARAZZI	1	
	ARGENTIERE	1SI	CHI VENDE OGGETTI D'ARGENTO		
	ARMAIOLO	1SI	CHI FABBRICA VENDE RIPARA ARMI		
	ASTUCCIAIO	1SI	CHI FABBRICA O VENDE ASTUCCI	1	
	BABBUCCIAIO	1SI	CHI FA O VENDE BABBUCCE	1	
	BADILAIO	1SI	CHI FA O VENDE BADILI	1	
	BERRETTAIO	1SN	CHI FABBRICA O VENDE BERRETTI	1	
	BICCHIERAIO	1SI	CHI FABBRICA O VENDE BICCHIERI	1	
	BIGLIETTAIO	1SN	CHI VENDE I BIGLIETTI PER IL VIAGGIO	1	
	BILANCIAIO	1SI	STADERAIO/CHI FABBRICA E VENDE BILANCE	4	
	BILIARDAIO	1SI	CHI FABBRICA O VENDE BILIARDI	1	
	BIRRAIO	1SI	CHI FABBRICA O VENDE BIRRA	1	
	BOCCALAIO	1SI	CHI FABBRICA O VENDE BOCCALI	1	
	BORSAIO	1SG	CHI FABBRICA O VENDE BORSE	1	
	BOTTAIO	1SI	CHI FABBRICA,RIPARA O VENDE BOTTI	1	
	BOTTONAIO	1SN	CHI FABBRICA O VENDE BOTTONI	1	
	BUSTAIA	1SF	DONNA CHE CONFEZIONA O VENDE BUSTI	1	
	CALZETTAIO	1SN	CHI VENDE O FABBRICA CALZE	1	
	CANESTRAIO	1SI	CHI FA O VENDE CANESTRI	1	
	CARBONAIO	1SM	CHI VENDE CARBONE	1	
	....				
	OROLOGIAIO	1SI	CHI FABBRICA,RIPARA O VENDE OROLOGI	1	
	ORTOPEDICO	2SI	CHI FABBRICA O VENDE APPARECCHI ORTOPEDICI	3	
	OTTICO	2SI	CHI CONFEZIONA E VENDE OCCHIALI E LENTI	3	
	PADELLAIO	1SI	CHI FA O VENDE PADELLE	1	
	PANETTIERE	1SN	FORNAIO/CHI FA O VENDE PANE		
	PANIERAIO	1SG	CHI FA O VENDE PANIERI		
	PANTOFOLAIO	1SN	CHI CONFEZIONA O VENDE PANTOFOLE	1	
	PASTAIO	1SN	CHI FABBRICA O VENDE PASTE ALIMENTARI	1	
	PASTICCERE	1SN	CHI FA O VENDE DOLCIUMI		
	PASTICCIERE	1SN	CHI FA O VENDE DOLCIUMI		
	PATACCARO	1SI	2CHI VENDE MONETE OD OGGETTI FALSI		
	PELLETTIERE	1SG	CHI PRODUCE O VENDE OGGETTI DI PELLETTERIA	1	
	PELLICCIAIO	1SN	CHI LAVORA O VENDE PELLICCE	1	
	....				
	VENDITORE	2SI	CHI V'NDE	1	
	VETRAIO	1SI	CHI VENDE TAGLIA APPLICA LASTRE DI VETRO		
	VINATTIERE	1SM	1CHE VENDE O COMMERCIA VINO	1	5
	VIOLINAIO	1SI	LIUTAIO/CHI FABBRICA O VENDE VIOLINI	4	
	ZOCCOLAIO	1SI	CHI FA O VENDE ZOCCOLI	1	

Fig. 12. Nomi di **AGENTS** per l'azione del "vendere".

VENDITORE	---->>ABBACCHIARO	1SI	2VENDITORE DI ABBACCHI	1	2	
	ACQUAVITAIO	1SI	VENDITORE DI ACQUAVITE	1		
	ARCHIBUGIERE	1SM	FABBRICANTE O VENDITORE DI ARMI	3	1	
	....					
	BIBITARO	1SI	2VENDITORE DI BIBITE	1	2	
	BORSETTAIO	1SG	FABBRICANTE O VENDITORE DI BORSE E BORSETTE	1		
	BRONZISTA	1SN	VENDITORE DI OGGETTI ARTISTICI IN BRONZO			
	BURATTINAIO	1SI	FABBRICANTE O VENDITORE DI BURATTINI			
	CALCOGRAFO	1SI	VENDITORE DI INCISIONI	3		
	CALDARROSTAIO	1SN	VENDITORE DI CALDARROSTE	1		
	CAMICIAIO	1SD	FABBRICANTE O VENDITORE DI CAMICIE	1		
	CAPPELLAIO	1SN	FABBRICANTE O VENDITORE DI CAPPELLI DA UOMO	3		
	CARAMELLAIO	1SN	FABBRICANTE O VENDITORE DI CARAMELLE	1		
	....					
	FRUTTIVENDOLO	1SN	VENDITORE DI FRUTTA E ORTAGGI	3		
	LATTAIO	1SN	VENDITORE DI LATTE	1		
	LIBRAIO	1SN	VENDITORE DI LIBRI			
	MACELLAIO	1SN	VENDITORE DI CARNE MACELLATA	3		
	....					
	PROFUMIERE	1SN	FABBRICANTE O VENDITORE DI PROFUMI E COSMETICI	1		
	SALUMIERE	1SN	VENDITORE DI SALUMI	1		
	SPEZIALE	2SI	VENDITORE DI SPEZIE	1	1	
	STRILLONE	1SN	VENDITORE AMBULANTE DI GIORNALI	3		
	VALIGIAIO	1SN	FABBRICANTE O VENDITORE DI VALIGIE BAULI,BORSE	1		
	VINAIO	1SN	VENDITORE FORNITORE DI VINO	1		

Fig. 13. Nomi di **AGENTS** per l'azione del "vendere".

VENDONO	---->>APPALTO	1SM	LUOGO DOVE SI VENDONO PRODOTTI DI MONOPOLIO DELLO STATO	3	2	
	BANCO	1SM	LOCALE DOVE SI VENDONO O SCAMBIANO BENI SERVIZI	3		
	BIGIOTTERIA	1SF	NEGOZIO DOVE SI VENDONO OGGETTI DECORATIVI NON PREZIOSI	3		E
	BIGLIETTERIA	1SF	LUOGO IN CUI SI VENDONO BIGLIETTI	1		
	BISCOTTERIA	1SF	NEGOZIO DOVE SI VENDONO I BISCOTTI			
	BOTTIGLIERIA	1SF	NEGOZIO DOVE SI VENDONO VINO LIQUORI IN BOTTIGLIA	3		
	BRICABRAC	1	NEGOZIO,BANCARELLA OVE SI VENDONO TALI ANTICAGLIE	3		E
	CALZETTERIA	1SF	NEGOZIO IN CUI SI VENDONO CALZE			
	CALZOLERIA	1SF	BOTTEGA IN CUI SI FABBRICANO O VENDONO SCARPE			
	CAMICERIA	1SF	NEGOZIO IN CUI SI VENDONO CAMICIE			
	CAPPELLERIA	1SF	NEGOZIO DOVE SI VENDONO CAPPELLI MASCHILI	1		
	CERERIA	1SF	LUOGO DOVE SI FABBRICANO E VENDONO CANDELE	3		
	CHINCAGLIERIA	1SF	NEGOZIO IN CUI SI VENDONO CHINCAGLIE			
	CONFETTURERIA	1SF	LUOGO OVE SI PREPARANO,VENDONO CONFETTURE	1		
	CREMERIA	1SF	2LATTERIA IN CUI SI VENDONO ANCHE GELATI DOLCI E SIM.	3		
	DIACCIATINO	2SN	2BOTTEGA DOVE SI VENDONO SORBETTI	3	1	
	DROGHERIA	1SF	BOTTEGA DOVE SI VENDONO DROGHE	1		
	FERRAMENTA	1SF	NEGOZIO IN CUI SI VENDONO OGGETTI DI FERRO	3		
	GELATERIA	1SF	SORBETTERIA/NEGOZIO OVE SI FANNO O VENDONO GELATI	4		
	MAGLIERIA	1SF	BOTTEGA NEGOZIO IN CUI VENDONO INDUMENTI DI MAGLIA			
	MESCITA	1SF	BOTTEGA IN CUI SI VENDONO VINO LIQUORI	3	2	
	MESTICHERIA	1SF	2BOTTEGA IN CUI SI VENDONO COLORI MESTICATI	3	2	
	NEGOZIO	1SM	BOTTEGA/ LOCALE DOVE SI ESPONGONO E VENDONO MERCI	5		
	NORCINERIA	1SF	2BOTTEGA IN CUI SI VENDONO SOLO CARNI DI MAIALE	3	2	
	OCCHIALERIA	1SF	NEGOZIO IN CUI SI VENDONO O SI RIPARANO OCCHIALI			
	OROLOGERIA	1SF	NEGOZIO DOVE SI VENDONO OROLOGI	3		
	PANTOFOLERIA	1SF	LUOGO IN CUI SI VENDONO PANTOFOLE			
	PELLETTERIA	1SF	NEGOZIO IN CUI SI VENDONO OGGETTI DI PELLE LAVORATA	3		
	PIATTERIA	1SF	BOTTEGA DOVE SI VENDONO I PIATTI	3		
	ROSTICCERIA	1SF	BOTTEGA DOVE SI PREPARANO O VENDONO ARROSTI	3		
	SALUMERIA	1SF	BOTTEGA,NEGOZIO,IN CUI SI VENDONO I SALUMI	3		
	SPACCIO	1SM	LOCALE DELLE CASERME DOVE SI VENDONO GENERI ALIMENTARI VARI	3		
	UTENSILERIA	1SF	BOTTEGA IN CUI SI VENDONO UTENSILI			

Fig. 14. Nomi di **PLACES** relativi all'azione del "vendere".

OROLOGERIA = <–**LOC**– ***"vendere"*** –**THEME**–> orologi –**IS-A**–> OBJECT

OROLOGIAIO = <–**AGENT**– " " " " "

Fig. 15. Diagramma di una porzione di rete per l'azione del ***"vendere"***.

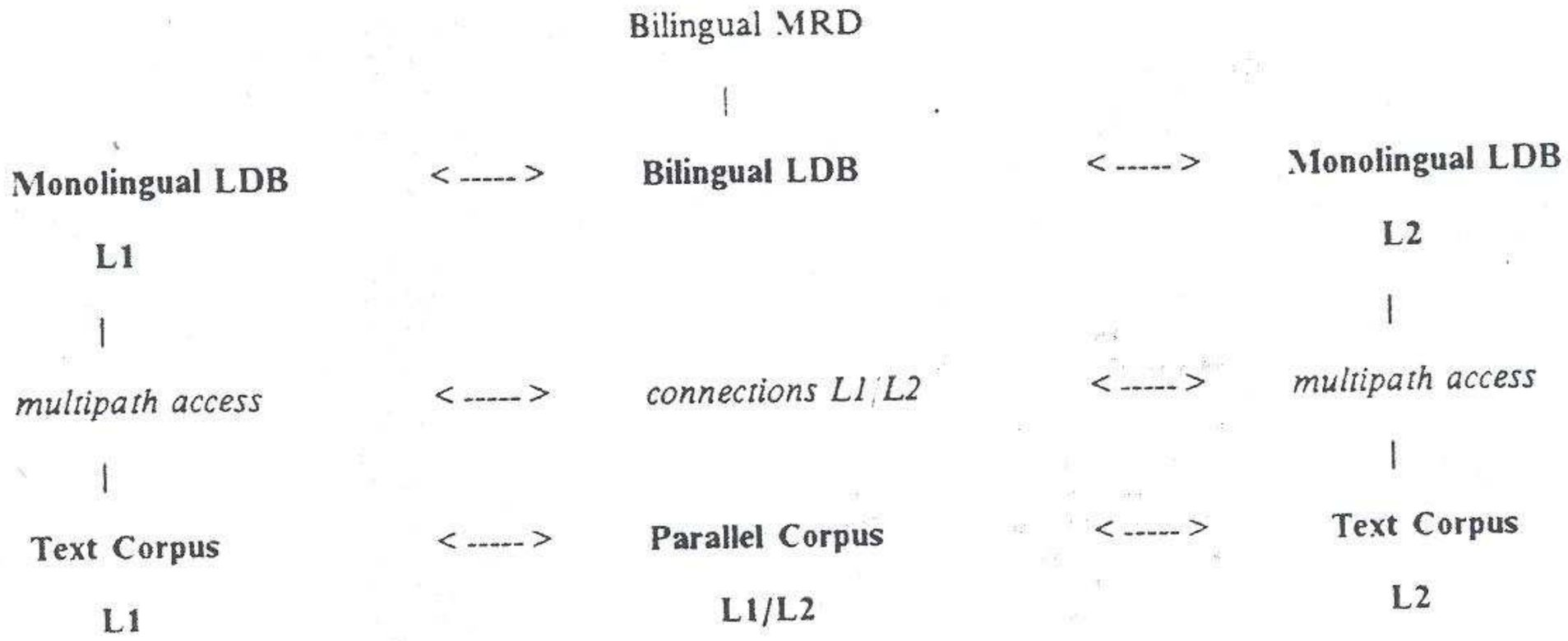

Fig. 16. Un modello per un LDB Bilingue.

Bibliografia

Ahlswede, T., Evens, M., Parsing vs. Text Processing in the Analysis of Dictionary Definitions, *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, Buffalo, New York, 1988, 217-224.

Alshawi, H., Analyzing the Dictionary Definitions, in B. Boguraev, E. Briscoe (eds.), 1989, 153-170.

Amsler, R. A., A Taxonomy for English Nouns and Verbs, *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics*, Stanford, California, 1981, 133-138.

Atkins B.T., The Uses of Large Text Databases, Semantic ID Tags: Corpus Evidence for Dictionary Senses, *Third Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, Waterloo, Canada, 1987, 17-36.

Atkins, B.T., Kegl, J., Levin, B., Explicit and Implicit Information in Dictionaries, in *Proceedings of the Conference on Advances in Lexicology*, Waterloo, 1986.

Bindi, R., Calzolari, N., Statistical analysis of a large textual Italian Corpus in search of lexical information, presented for *EURALEX 1990*, Malaga, forthcoming.

Boguraev, B., Briscoe E.J. (eds.), *Computational Lexicography for Natural Language Processing*, Longman, London, 1989.

Boguraev, B., Briscoe, E.J., Calzolari, N., Cater, A., Meijs, W., Zampolli, A., Acquisition of Lexical Knowledge for Natural Language Processing Systems, (AQUILEX), Technical Annex, ESPRIT Basic Research Action No. 3030, Cambridge, 1988.

Boguraev, B., Byrd. R., Klavans, J., Neff, M., From structural analysis of lexical resources to semantics in a Lexical Knowledge Base, in *Proceedings of the First International Lexical Acquisition Workshop*. Detroit (Michigan), 1989.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., Roossin, P., A Statistical Approach to Language Translation, *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 1988.

Byrd, R.J., Discovering Relationships among Word Senses, *Dictionaries in the Electronic Age*, Fifth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary, Oxford, 1989.

Byrd, R.J., Calzolari, N., Chodorow, M., Klavans, J., Neff, M., Rizk, O., Tools and Methods for Computational Lexicology, *Computational Linguistics*, 1987, vol. 13(3-4), 219-240.

Calzolari, N., Towards the organization of lexical definitions on a data base structure, *COLING82*, ed. by E. Hajicova, Prague, Charles University, 1982, pp.61-64.

Calzolari, N., Bindi, R., Acquisition of Lexical Information from a Large Textual Italian Corpus, *COLING90*, ed. by H. Karlgren, Helsinki University, 1990, pp.54-59.

Calzolari, N., Detecting Patterns in a Lexical Database, *Proceedings of the 10th International Conference on Computational Linguistics*, Stanford, California, 1984, 170-173.

Calzolari,N., The dictionary and the thesaurus can be combined, in *Relational Models of the Lexicon* , (Studies in Natural Language Processing series), ed. by M.Evens, Cambridge (Mass.), Cambridge University Press, 1988, 75-96.

Calzolari,N., Lexical Databases and Text Corpora: perspectives of integration for a Lexical Knowledge Base, in *Proceedings of the First International Lexical Acquisition Workshop.* Detroit (Michigan), 1989a, n.28.

Calzolari, N., Computer-aided lexicography: dictionaries and word databases, *Computational Linguistics*, edited by I.S. Batori, W. Lenders, W. Putschke, Berlin: Walter de Gruyter, 1989b, 510-519.

Calzolari, N., Structure and Access in an automated Lexicon and Related Issues, in D. Walker, A.Zampolli, N.Calzolari (eds.), forthcoming.

Calzolari, N., Picchi, E., A Project for a Bilingual Lexical Database System, *Advances in Lexicology, Second Annual Conference of the UW Centre for the New Oxford English Dictionary*, Waterloo, Ontario, 1986, 79-92.

Calzolari, N., Picchi, E., Acquisition of Semantic Information from an On-Line Dictionary, *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 1988, 87-92.

Calzolari,N., E.Picchi, A.Zampolli, The use of computers in lexicography and lexicology, in *The Dictionary and the Language Learner*, ed. by A.Cowie, Lexicographica Series Maior 17, Tubingen, Niemayer, 1987, 55-77.

Chodorow, M.S., Byrd, R.J., Heidorn, G.E., Extracting Semantic Hierarchies from a Large On-line Dictionary, *Proceedings of the Association for Computational Linguistics*, Chicago, Illinois, 1985, 299-304.

Church, K.W., A Stochastic parts program and noun phrase parser for unrestricted text, *ACL, Second Conference on Applied Natural Language Processing*, 1988, 136-143.

Church, K., Hanks, P., Word Association Norms, Mutual Information and Lexicography, *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, 1989, 76-83.

Cumming, S., The Lexicon in Text Generation, in D. Walker, A.Zampolli, N.Calzolari (eds.), forthcoming.

Fox, E., Nutter, T., Ahlswede, T., Evens, M., Markowitz, J., Building a Large Thesaurus for Information Retrieval, *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas, 1988, 101-108.

Goetschalckx, J., Rolling, L. (eds.), *Lexicography in the Electronic Age*, Amsterdam, North-Holland, 1982.

Gruppo di Pisa, Il Dizionario di Macchina dell'Italiano, in *Linguaggi e Formalizzazioni*, ed. by Gambarara, D., Lo Piparo, F., Ruggiero, G., Roma, Bulzoni, 1979, pp.683-707.

Hays, D.G., *Computational Linguistics: Introduction*, in Meetham and Hudson (eds.), 1969, 49-51.

Hindle, D., Acquiring Disambiguation Rules from Text, *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Morristown (NJ), 1988, 118-125.

Ingria, R., Lexical Information for parsing Systems: Points of Convergence and Divergence, in D. Walker, A.Zampolli, N.Calzolari (eds.), forthcoming.

Kay, M., The Dictionary of the Future and the Future of the Dictionary, in Zampolli, Cappelli (eds.), 1983, pp.161-174.

Japanese Electronic Dictionary Research Institute, *Electronic Dictionary Project*, Tokyo, 1988.

Locke, W.N., Booth, A.D., *Machine Translation of Languages*, MIT Press, 1955.

Katz, B., Levin, B., Exploiting Lexical Regularities in Designing Natural Language Systems, *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 1988, 316-323.

Klavans, J.L., Building a Computational Lexicon using Machine Readable Dictionaries, paper presented at the Third Congress of the European Association for Lexicography, Budapest, 1988.

Kucera, H., Francis, W.N., *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island, 1967.

Nagao, M., Nakamura, J., Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation, *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 1988, 459-464.

Neff, M., Boguraev, B., Dictionaries, Dictionary Grammars and Dictionary Entry Parsing, *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, 1989, 91-101.

Papp, F. Szepe, G. (eds.), *Papers in Computational Linguistics, Proceedings of the 3rd International Meeting on Computational Linguistics*, 1976.

Picchi,E., N.Calzolari, Textual perspectives through an automatized lexicon, in *Methodes quantitatives et informatiques dans l'etude des textes.* Geneve: Slatkine, 1986, 705-715.

Picchi,E., C.Peters, N.Calzolari, A tool for the second language learner: organizing bilingual dictionary data in an interactive workstation, in *Proceedings of the XX ALLC Conference*, Jerusalem, 1988, forthcoming.

Pustejovsky, J., Current Issues in Computational Lexical Semantics, Invited Lecture, *Proceedings of the Fourth Conference of the European Chapter of the ACL*, Manchester, England, 1989, xvii-xxv.

Smadja, F., Macrocoding the Lexicon with Co-occurrence Knowledge, paper presented at the First Lexical Acquisition Workshop, Detroit, 1989.

Smith, J., Ideals versus Practicalities in Linguistic Data Processing, in A. Zampolli, N. Calzolari (eds.), 1973, 895-8.

Talmy, L., Lexicalization Patterns: Semantic Structure in Lexical Forms, in T. Shopen (ed.), *Language Typology and Syntactic Description: Grammatical Categories and the Lexicon*, Cambridge University Press, Cambridge, 1985.

Van der Steen, G.J., A Treatment of Queries in Large Text Corpora, in S. Johansson (ed.), *Computer Corpora in English Language Research*, Norwegian Computing Centre for the Humanities, Bergen, 1982, 49-65.

Vossen, P., Meijs, W., den Broeder, M., Meaning and Structure in Dictionary Definitions, in B. Boguraev and E. Briscoe (eds.), 1989, 171-192.

Walker, D., Zampolli, A., Foreword, in B. Boguraev, T. Briscoe (eds.), 1989, xiii-xiv.

Walker,D., A.Zampolli, N.Calzolari (eds.), *Towards a polytheoretical lexical database.* Pisa: ILC, 1987.

Walker, D., A.Zampolli, N.Calzolari (eds.), Special Issue of the *Journal of Computational Linguistics*, 13(1987)3-4, 193.

Walker, D., Zampolli, A., Calzolari, N. (eds.), *Automating the Lexicon: Research and Practice in a Multilingual Environment*, OUP, forthcoming.

Webster, M., M. Marcus, Automatic acquisition of the lexical semantics of verbs from sentence frames, in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, 1989, 177-184.

Whitelock, P., Wood, M., Somers, H., Johnson, R., Bennett, P. (eds.), *Linguistic Theory and Computer Applications*, Academic Press, New York, 1987.

Wilks, Y., Fass, D, Guo, C.-M., McDonald J., Plate, T., Slator, B., A Tractable Machine Dictionary as a Resource for Computational Semantics, in B. Boguraev and E. Briscoe (eds.), 1989, 193-228.

Zampolli, A., Projet pour un lexique electronique de l'italien, *Calcolo*, V (1968), Suppl. al II volume, 109-26.

Zampolli, A., Lexicological and Lexicographical Activities at the Istituto di Linguistica Computazionale, in Zampolli, Cappelli (eds.), 1983, pp.237-278.

Zampolli, A., Multifunctional Lexical Databases, *Encrages*, 16(1986), 56-65.

Zampolli, A., Progetto Strategico "Metodi e strumenti per l'industria delle lingue nella cooperazione internazionale", Pisa, 1987.

Zampolli, A., Progetto Speciale "Aquisizione di una base di conoscenze lessicali per il trattamento automatico dell'Italiano: obiettivi nazionali e cooperazione internazionale", Pisa, 1989.

Zampolli, A., Calzolari, N., (eds.), *Computational and Mathematical Linguistics, Proceedings of the International Conference on Computational Linguistics 1973*, 2 Volumes, Firenze, 1973 and 1977.

Zampolli, A., Calzolari, N., Computational Lexicography and Lexicology, *AILA Bulletin*, 1985, 59-78.

Zampolli, A., Cappelli, A., (eds.), The Possibilities and Limits of the Computer in producing and publishing Dictionaries, *Linguistica Computazionale*, Pisa, III, 1983.

Zampolli, A., Cignoni, L., Rossi, S., Problems of Textual Corpora, ILC-9-2, Pisa, 1985.

Zampolli, A., Cignoni L., Peters C., Computational Lexicology and Lexicography, *Linguistica Computazionale*, Special issue dedicated to Bernard Quemada, VI (1991).