Tirrenia Stampatori

A tre voci

Seminari dell'Istituto di Filologia Moderna
Università degli Studi di Parma

La bella e la bestia

(ITALIANISTICA E INFORMATICA)

2

La linguistica computazionale

ANTONIO ZAMPOLLI

1. *Premessa storica*

Agli inizi, e cioè nel decennio che va dalla fine degli anni '40 all'inizio degli anni '60, gli utilizzi del "calcolo elettronico" per la elaborazione di dati linguistici si articolano in due filoni principali:

- gli spogli elettronici di testi (DET), a supporto della ricerca in varie discipline umanistiche (1948 inizio di *Index Thomisticus*) (R. Busa, 1968 e 1968a);
- i tentativi di Traduzione Automatica (TA; in inglese MT: Machine translation) (MIT e National Phisical Laboratory, 1948; Booth 1958).

I ricercatori che operano nei due filoni sono consapevoli di condividere problemi metodologici e tecnici, spesso derivanti dai limiti delle tecnologie di quei tempi, e interagiscono per cercare e promuovere soluzioni comuni.

I problemi discussi riguardano, per esempio, la codifica/rappresentazione della varietà di caratteri che compaiono nei testi a stampa, gli algoritmi per la manipolazione dei caratteri, lo sviluppo di metodi per il computo e l'uso delle frequenze, o per la costruzione e la memorizzazione di grandi dizionari (mono e bilingui) ("Cahiers de Lexicologie", 1961; M. Kay, 1967).

Per il primo decennio non c'è un termine generale che designi le elaborazioni di dati linguistici.

Queste ricerche vengono classificate, per esempio,[1] come:

- linguistica applicata;
- linguistica matematica (applicata);
- analisi automatica di dati linguistici/letterari.

[1] Per una descrizione più approfondita, si veda Zampolli (1974).

Nel 1965, D.G. Hays (1965 p. 1) afferma che la Linguistica Computazionale (LC) "encompasses every use to which computers can be put in the manipulation of ordinary language".

Oggi, la definizione di Hays è troppo estesa: i calcolatori – inizialmente macchine per trattare i numeri – sono ora usati soprattutto per operare su testi.

Ne segue che molte "manipolazioni" di materiali testuali fanno parte della tecnologia informatica e delle telecomunicazioni "tout-court": per es., la rappresentazione di alfabeti diversi, i sistemi di "word-processing", la ricerca di parole (o stringhe di lettere) nei testi, ecc.

Uno dei modi più efficaci e semplici per definire la LC è ancora oggi quello di considerare i principali settori nei quali si sono sviluppati metodi e tecniche, procedure specializzate, gli scopi per i quali queste sono usate, e quali contributi esse possano dare alla ricerca e alle applicazioni.

A partire dalla metà degli anni '60 i due filoni si "allontanano", e le interazioni diventano via via più rare se non assenti del tutto.[2]

Il primo filone (Humanistic Text Processing - HTP) concentra gli sforzi soprattutto sulla utilizzazione delle possibilità offerte dallo sviluppo tecnologico (rappresentazione di caratteri, uso di terminali, linguaggi di programmazione, collegamenti in rete, ecc.) e accumula rapidamente "biblioteche" di testi letterari sempre più estese. Però le unità di analisi/elaborazione restano, per lo più, i caratteri/stringhe di caratteri/"parole grafiche".

Il secondo filone (Natural Language Processing - NLP), nonostante le raccomandazioni ALPAC,[3] si concentra essenzialmente sullo studio di metodi per la analisi sintattica di frasi (monolingui) scelte "ad hoc" per studiare e validare le proprietà di teorie e modelli linguistici: pochi dati, spesso "artificiali", indipendentemente dal loro uso in contesti comunicativi reali.[4]

Si crea così una potenziale contrapposizione. Nel secondo filone (NLP) i dati vengono praticamente "ignorati"; la copertura linguistica (lessicale e

[2] Per una analisi delle cause di questo allontanamento si veda Zampolli 1997a.

[3] Nel 1965, le Agenzie del Governo Americano che avevano finanziato per quasi 10 anni grossi progetti di Traduzione Automatica, di fronte alla mancanza di risultati applicabili chiesero a un gruppo di esperti di preparare un rapporto (il famoso ALPAC Report 1966). In questo rapporto appare per la prima volta usato in una fonte ufficiale il termine *computational linguistics* per designare una nuova disciplina che, si asserisce, è nata come sviluppo della TA. Il Rapporto ritiene indispensabile che essa si occupi di raccogliere e studiare (anche contrastivamente fra lingue diverse) corpora testuali estesi, lessici di grandi dimensioni, grammatiche a larga copertura.

[4] Si veda l'"Overview" di J.J. Godfrey e A. Zampolli nel capitolo "Language Resources", in G.B. Varile, A. Zampolli (eds.) 1997, pp. 381-384.

sintattica) dei modelli e dei componenti studiati è estremamente ristretta; le descrizioni delle strutture sintattiche che i parsers sono capaci di promuovere sono spesso insufficienti per applicazioni reali; i sistemi proposti mancano della "robustezza" necessaria per le applicazioni, soprattutto se esse vogliono essere interamente automatiche, ed operano su testi "reali".

Nel primo filone (HTP), le elaborazioni si limitano prevalentemente al livello "grafemico": non si utilizzano conoscenze e metodi, sviluppati nel secondo filone, per accedere ad altri livelli (lemmatizzazione, ecc.) di analisi dei testi.

Le sole eccezioni sono rappresentate da alcuni Centri europei che, in ottemperanza al proprio mandato istituzionale, perseguono programmaticamente le ricerche in entrambi i filoni. Un esempio è costituito dalla Divisione Linguistica del CNUCE di Pisa che promuove la definizione di standards per la codifica di testi e per la annotazione della analisi lessicale/morfosintattica, e l'utilizzo di dizionari macchina e di disambiguatori morfosintattici per potenziare la ricerca testuale.

A partire dalla metà degli anni '80 si affermano, come vedremo più avanti, due nuovi paradigmi (denominati rispettivamente "industria delle lingue" (Vidal-Beneyto, 1991) e "Risorse linguistiche riutilizzabili" (Grosseto 1986): (Walker et alii, 1995).

Ciò favorisce una rinnovata collaborazione tra HTP e NLP, per sfruttare la complementarità delle loro conoscenze, stimolata dalla possibilità di accedere a finanziamenti di nuove dimensioni.

In questa breve conferenza, vorrei mostrare appunto alcuni risultati di questa interazione con qualche esempio tratto dai lavori di miei collaboratori dell'Istituto di Linguistica Computazionale del CNR di Pisa, Istituto nel quale è stata da alcuni anni trasformata la originaria Divisione Linguistica del CNUCE.[5]

Ho cercato esempi che mostrino come metodi e componenti "di analisi linguistica", sviluppati dal NLP possano arricchire e potenziare alcuni risultati che, grazie all'uso tempestivo delle possibilità offerte dalla adozione di tecnologie innovative (gamma di caratteri disponibili, browsing interattivo, accesso via Internet, ecc.), fanno ormai parte del bagaglio o, per meglio dire, del patrimonio metodologico del HTP.

[5] Per una descrizione della Sezione Linguistica del CNUCE e delle sue attività si veda A. Zampolli (1973a).

2. *Esempi di convergenza tra i due filoni*

2.1. *Accesso interattivo ai testi*

Dalla iniziale produzione di indici e concordanze,[6] gli spogli elettronici di testi[7] sono passati, dapprima, all'utilizzo di sistemi per la interrogazione interattiva, on-line, di testi residenti nella memoria, accessibile "localmente", del PC utilizzato dallo studioso,[8] all'accesso in rete che consente a molteplici utenti di interrogare – via Internet – grandi "biblioteche" di testi digitali.[9]

L'"interrogazione" interattiva di testi permette al ricercatore di richiedere che i contesti siano "tagliati" e ordinati tra loro diversamente per rispondere a diverse esigenze metodologiche, eventualmente richiedendo che i contesti relativi a certe occorrenze vengano estesi "a piacere". I programmi, che sono applicabili a lingue diverse e ad alfabeti non latini permettono di ricercare la cooccorrenza di *n* parole, indicate dallo studioso, nel testo (Figura 1), oppure di applicare la formula della cosiddetta "mutual information"[10] che segnala le parole che cooccorrono con maggior significatività nell'intorno delle parole studiate dal ricercatore (Figura 2).

Applicando procedure di lemmatizzazione, è possibile richiedere che le stesse funzionalità vengano applicate ai lemmi anziché alla forme grafiche, ottenendo così, per esempio, la segnalazione delle cooccorrenze associate ad un determinato lemma (Figura 3), oppure indici di frequenza relativi ai lemmi, o a certe categorie grammaticali (Figura 4).

La tradizione anglosassone degli spogli elettronici è stata indotta, dal fatto che le poche variazioni dovute in inglese alla morfologia influiscono scarsamente sulla consultazione degli indici e delle concordanze, a trascurare, fondamentalmente, il problema della lemmatizzazione, che invece si

[6] Si vedano gli articoli di Burton nei numeri di *Computers and the Humanities* (1981).

[7] Gli spogli di testi e, in genere, le tecniche di analisi testuale costituiscono, per così dire, un elemento "trasversale" comune alle discipline umanistiche. Si veda per es. Genet, Zampolli (eds.) 1992.

[8] Si veda E. Picchi per la descrizione del sistema DBT (Data Base Testuale), uno dei primi e certamente il più completo esempio di stazione di lavoro per l'analisi testuale.

[9] Vedi M. Tavoni (2000), per una descrizione del progetto CIBIT.

[10] Per una descrizione di questa formula e delle possibilità che essa offre si veda per es. De Marcken (1999), pp. 198 e 208.

è imposto, molto presto, alla attenzione di quanti operavano su lingue la cui flessione è molto ricca.[11]

D'altra parte, la lemmatizzazione attraverso la consultazione di un dizionario-macchina costituisce, con la cosiddetta "tokenizzazione" (cioè la segmentazione del testo nelle unità che lo compongono, e in particolare la individuazione delle unità grafiche che possono costituire occorrenze di unità lessicali), la prima operazione che i sistemi per il trattamento automatico delle lingue devono compiere.

Le informazioni lessicali ottenute con la lemmatizzazione (per es., parti del discorso, categorie morfologiche, proprietà sintattiche, ecc.) sono, per così dire, la "materia prima", costituiscono cioè i dati di input per quasi tutte le operazioni di analisi, generazione, interpretazione, automatiche di un testo o di un enunciato.

La automazione delle operazioni che costituiscono la lemmatizzazione è, pertanto, uno dei naturali punti di convergenza tra i due filoni.

Un altro argomento che interessa entrambi i filoni è lo studio delle corrispondenze tra testi paralleli, cioè tra versioni corrispondenti dello stesso testo: per es. tra le sue traduzioni, o edizioni diverse.

La prima operazione da compiere è ovviamente il cosiddetto "allineamento", cioè la individuazione di porzioni di testo corrispondenti nei testi paralleli. Nelle prime concordanze contrastive,[12] che risalgono alla fine degli anni '60, i testi paralleli venivano segmentati manualmente in unità successive numerate progressivamente: un identico numero progressivo contrassegnava ("allineava") porzioni di testo corrispondenti nelle diverse versioni.

Le ricerche del secondo filone fanno un uso crescente di estese collezioni di dati, e in particolare di corpora paralleli in lingue diverse dai quali estraggono automaticamente glossari terminologici per i traduttori, memorie di traduzione, ecc.[13]

I ricercatori di NLP hanno compiuto in anni recenti uno sforzo notevole per automatizzare le operazioni di allineamento: i sistemi da loro prodotti per lo più assumono – e sono quindi capaci di riconoscere automaticamente – come unità per l'allineamento il paragrafo, e cominciano ora ad essere usati sempre più spesso da ricercatori delle discipline umanistiche.

[11] Si veda per es. Zampolli (1976).

[12] I primi esperimenti risalgono a Bujas (v. in Bibliografia) e Zampolli (1973b) al finire degli anni '60.

[13] Si veda l'articolo di Dagan et alii (1999).

Al nostro Istituto, abbiamo studiato e stiamo perfezionando metodi che, avvalendosi anche di dizionari bilingui su supporto digitale[14] (qualora disponibili) allineano automaticamente porzioni di testo "più piccole": per es. sintagmi o singole parole (Figure 5 e 6).

È così possibile ottenere automaticamente concordanze contrastive relative a singole parole (Figura 7), o a combinazioni di parole (Figura 8).

Questi meccanismi, che forniscono una documentazione rilevante per studi testuali di tipo umanistico, possono essere usati per applicazioni innovative nel WWW, per esempio per compiti di "cross-language information retrieval": l'utente specifica un insieme di parole che, nella propria lingua, sono correlate al soggetto che gli interessa, e il sistema ricerca parole equivalenti nei testi di altre lingue, senza bisogno di ricorrere a un dizionario bilingue, ma stabilendo le equivalenze con metodi di esplorazione statistica del corpus di testi nei quali la ricerca deve essere compiuta (Figura 9).

2.2. *Basi di dati dialettali*

Una versione specializzata del sistema di interrogazione dei testi viene usata per gestire e interrogare gli archivi dialettali dell'ALT (Atlante Linguistico Toscano), basati su un modello che consente di rappresentare i dati raccolti in modo da ricuperare le informazioni ricercandole attraverso una combinazione di caratteri diversi.

Nell'archivio ogni attestazione riceve una caratterizzazione rispetto alle coordinate "località, informatori, domanda".

Diverse configurazioni di tratti corrispondono a diversi tipi di attestazione: risposte canoniche a domande del questionario; materiali integrativi; tipici contesti di uso delle attestazioni lessicali raccolte (es. fraseologia, proverbi); descrizione di usi e costumi, o altro ancora, legate ai materiali linguistici raccolti.

Alcuni strumenti linguistici vengono messi a disposizione dell'utente – in particolare non specialista – per arricchire le sue interazioni con l'ALT, dandogli per esempio la possibilità di ricercare domande su argomenti dei quali non conosce la esatta formulazione (Figura 10), attraverso l'uso di

[14] Sempre più frequentemente, oggi, i dizionari vengono fotocomposti, se non addirittura redatti direttamente in un database elettronico. Si veda Atkins, Zampolli (eds.) 1994.

algoritmi per l'analisi delle definizioni e di basi di relazioni semantiche (vedi n. 3.2), o di ricercare forme dialettali delle quali non conosce la forma fonetica esatta, ma solo una forma canonica, a partire dalle quali regole di equivalenza fonetiche "calcolano" possibili varianti locali (vd. Montemagni).

Grazie a questi accorgimenti, DBT-ALT si configura come strumento esplorativo del corpus dei materiali dialettali: il dato raccolto sul campo diventa accessibile da prospettive diverse, che l'utente può definire sulla base dei propri interessi di ricerca.

2.3. *Una stazione di lavoro per la filologia computazionale*

Un settore che sta assumendo una fisionomia specifica sempre meglio caratterizzata è quello della filologia computazionale. Come ha detto Raul Mordenti, fin dall'inizio i calcolatori sono stati usati per produrre, attraverso spogli elettronici, materiali documentari (indici, concordanze, frequenze, ecc.) a supporto delle ricerche filologiche, le quali naturalmente non possono che trarre beneficio dalla costituzione di estese biblioteche di testi digitalizzati, accessibili in rete, e dalla diffusione di metodi per la analisi linguistica che permettano di accedere alle unità linguistiche (per es. lemmi, o sequenze di categorie grammaticali) oltreché alle unità grafiche.

Presso il nostro Istituto è da tempo attiva una sezione che si occupa in particolare di adattare i metodi, gli strumenti informatici, e le risorse linguistiche alle esigenze specifiche della filologia, e in particolare della filologia classica, per es. attraverso la costituzione di dizionari di macchina, e dei corrispondenti analizzatori morfologici, per il greco e il latino.[15]

Qui voglio mostrare alcuni esempi di risultati che si possono ottenere associando alle tecniche, per così dire, tradizionali di spoglio, metodi dell'intelligenza artificiale, della linguistica computazionale, dell'"image processing"; per esempio:

– restaurare automaticamente le informazioni testuali contenute nelle immagini digitali di antichi testi a stampa (Figure 11 e 12);

[15] Io ho personalmente sostenuto (nello *steering committee* della TEI) la necessità di assicurare che le *guidelines* della TEI coprissero il settore della filologia, e in particolare delle edizioni critiche, attraverso la costituzione di un gruppo di lavoro dedicato.

- riconoscere automaticamente i caratteri di questo tipo di documenti (usando metodi quantitativi, del tipo Hidden Markov Models (Figura 13));[16]
- produrre una stazione di lavoro[17] dotata di funzioni automatiche che consentano, per esempio, di visualizzare contemporaneamente i contesti della parola studiata, sia nella trascrizione digitalizzata del testo, sia nella riproduzione fotografica – anch'essa memorizzata – del testo originale. È così possibile generare delle concordanze testo-immagine, nelle quali, in due finestre parallele dello schermo, appaiono da un lato la parola ricercata (con i suoi contesti) evidenziata nella trascrizione, mentre nell'altra finestra è evidenziata la immagine originale della parola nel suo contesto-immagine. Poiché il software stabilisce automaticamente il legame tra la trascrizione di una parola e la sua immagine, la ricerca può iniziare indifferentemente selezionando una parola nella trascrizione o nell'immagine (Figura 14). Come mostra la Figura 15, queste tecniche possono essere utilizzate anche su testi manoscritti;
- usare le informazioni fornite dall'apparato critico, che registrano le varianti trasmesse dalle diverse fonti, per proporre una rappresentazione/visualizzazione tridimensionale delle relazioni tra le fonti che il filologo ha sottoposto a collazione. L'algoritmo opera su classificazioni delle varianti e della loro rilevanza indicate dallo studioso (Figura 16).

Il compito di "leggere" manoscritti corrotti è facilitato non solo dai componenti già citati (per es. per la restaurazione dei caratteri e il loro riconoscimento automatico) che, come si è detto, usano tecnologie proprie dell'AI e dell'"immage processing", ma anche da moduli linguistico-statistici che integrano tali tecnologie, per esempio proponendo possibili interpretazioni di parole incomplete sulla base della consultazione di un *thesaurus* esteso (oltre un milione di forme) o di tabelle che riportano la frequenza di lettere, diagrammi, trigrammi, sia assolute sia distribuite nella varie posizioni (prima, seconda, ennesima) di parola, calcolate in corpora omogenei al tipo di lingua e al dominio del testo da interpretare (Bozzi, 2000 c).

[16] Per una descrizione di questi modelli, la cui adozione è stata determinante per il riconoscimento del parlato, si veda per es. R. De Mari, F. Brugnano, in G.B. Varile, A. Zampolli (eds.) 1997, pp. 21-29 e Bedini et alii, in Bozzi (ed. 2000).

[17] Si veda il volume di A. Bozzi (ed. 2000).

3. *Gli strumenti del NLP*

Uno dei compiti principali del NLP è quello di associare a un testo una descrizione linguistica formalizzata, specificando, per esempio, quali unità lo compongono (fonemi, parole, sintagmi, frasi, ecc.); quali sono le loro proprietà; quali siano le relazioni tra loro; le funzioni delle relazioni; quali strutture costituiscano; ecc.

La teoria linguistica di solito fornisce il formalismo di rappresentazione, le categorie di descrizione e classificazione, la forma delle relazioni.

Le informazioni di base vengono fornite da risorse linguistiche (v. n. 3.4) associate ai sistemi di analisi, che, in genere, adottano un principio comune: unità meno estese si combinano tra loro per formare unità più estese.

Scopo di molte teorie linguistiche è modellare questa combinazione.

Il processo avviene attraverso passi successivi, ciascuno dei quali produce una descrizione (della frase) che serve da input per il passo successivo.

A questi passi possono corrispondere componenti distinti, spesso corrispondenti a livelli di descrizione linguistica: lessico, morfologia, sintassi, semantica, ecc.

I risultati prodotti dai sistemi di analisi consentono, per esempio, di classificare alcune/tutte le parole di un testo (per esempio, contrassegnare i nomi propri di persona, luogo, Compagnia, ecc.; le parti del discorso; tutte le parole di un dominio semantico), o di "rappresentare" il "significato" di una frase (per es., nell'interazione uomo-macchina, identificare la natura della frase dell'interlocutore – domanda, risposta, comando – o l'argomento di cui si chiede la spiegazione).

3.1. *Lemmatizzazione e analisi morfologica*

Prendiamo come esempio le operazioni che si devono effettuare per lemmatizzare un testo, compito che, come abbiamo visto, ha sempre costituito motivo di convergenza tra i due filoni.

La lemmatizzazione consiste essenzialmente in due fasi principali.

Innanzitutto si devono ricercare, per ciascuna occorrenza grafica del testo, tutti i lemmi possibili nella lingua considerata. Un metodo molto semplice consiste nel memorizzare un *lessico computazionale*, nel quale ad ogni lemma sono associati una sua parte invariabile, e cioè comune alle sue diverse forme e flessioni (spesso il tema, per cui si parla di solito impropriamente di "dizio-

nari di temi"), e l'indicazione del paradigma di flessione. Una serie di tabelle fornisce poi, per ciascun paradigma, l'elenco delle desinenze possibili.

Un "analizzatore morfologico" scompone le parole da lemmatizzare in parte invariabile + desinenza; l'esistenza del "tema", o dei "temi" proposti viene verificata cercandoli nel lessico computazionale; nello stesso tempo si verifica, nella tabella, la compatibilità della desinenza (o desinenze) col tema (temi) proposto/i.[18] Si veda per esempio la Figura 17.

Oggi esistono in commercio analizzatori molto efficienti ed economici per molte lingue. Essi adottano strategie diverse sia per trattare le "flessioni irregolari", sia per ovviare ad alcuni problemi ancora non completamente risolti: per es. il riconoscimento o la lemmatizzazione di parole non "coperte" dal lessico computazionale (es. neologismi, sigle, nomi propri, e, in italiano, almeno per alcuni analizzatori, gli alterati non lessicalizzati).

La consultazione del lessico computazionale può proporre, per una certa forma, come risultato una sola analisi (forme univoche) o più di una analisi (forme omografe).

L'omografia si può presentare per ragioni diverse: per es. tra lemmi semanticamente diversi, o tra lemmi appartenenti a parti del discorso diverse.

Quest'ultimo caso è molto frequente.

Esistono oggi, per alcune lingue, dei componenti capaci di disambiguare automaticamente con una soglia di errore dell'ordine del (o inferiore) al 3%, i casi di forme che possono appartenere a parti del discorso diverse.

Questi "taggers" possono funzionare sulla base di regole che specificano sequenze possibili o impossibili di categorie grammaticali (vedi Figura 18), o sulla base di probabilità di sequenze di categorie apprese da un corpus annotato (training corpus).

3.2. *Analizzatori sintattici (parsers)*

Evidentemente, un analizzatore capace di riconoscere la struttura sintattica dell'intera frase (parser), ha anche l'effetto di disambiguare l'omografia di livello morfosintattico, mentre la disambiguazione dell'omografia lessicale richiederebbe un discorso a parte.[19]

[18] Per un panorama dello stato dell'arte degli analizzatori morfologici si veda Battista, Pirrelli.

[19] Non ho qui il tempo di accennare alle ricerche nel settore della "wordsense disambi-

Un analizzatore sintattico è, come si è detto, un processo – basato su conoscenze linguistiche codificate nella grammatica e nel lessico – capace di riconoscere e rappresentare le relazioni tra i componenti (es. soggetto, predicato, oggetto) di una frase, e determinarne la natura.

Le regole della grammatica specificano come un costituente sia formato da costituenti più piccoli e come le informazioni da associare a un costituente derivino dalle informazioni associate alle sue parti.

Le proprietà dei costituenti elementari ("parole" del testo) sono fornite, di solito, da un lessico computazionale, come risultato del processo di "lemmatizzazione".

Un semplicissimo esempio di regole di grammatica (un sottoinsieme "ridicolmente" ristretto di regole per l'analisi dell'italiano), e del risultato della loro applicazione, è riportato nella Figura 19.

Spesso a una data sequenza di costituenti può essere applicata più di una regola.

Si consideri per es., l'ambiguità, che concerne il "punto di attacco" del gruppo preposizionale (che potrebbe dipendere dal gruppo nominale o dal verbo) nelle strutture del tipo *vedo X con Y*. Per esempio:

- osservo le ragazze nel parco colle statue;
- osservo le ragazze nel parco col binocolo.

Spesso le conoscenze necessarie per evitare che il parsers "overgeneri", cioè per fargli scegliere la analisi appropriata, sono di natura semantica, o pragmatica, o contestuale.

Allo stato dell'arte queste conoscenze sono rappresentabili solo per domini particolari (per es. previsioni meteorologiche o orari di un aeroporto).

Si consideri per es. la frase della Figura 20.

Per risolvere l'ambiguità si dovrebbe trovare, nel lessico, l'informazione che

– per mangiare si può utilizzare, come strumento, la forchetta;
– i cibi possono essere conditi da altri cibi, o accompagnati da altri cibi, ecc.

Parsers capaci di trattare in modo soddisfacente testi "reali" non esistono, allo stato attuale dell'arte, per nessuna lingua.[20] Essi si "arrestano", spes-

guation": rinvio alla rivista "Special Issue on Word Sense Disambiguation: The State of the Art", a cura di Ide N., Véronis J. (1998) n. 24 (1), pp. 1-40, Numero Speciale di Computational Linguistics.

[20] Per uno stato dell'arte nel settore del parsing, si vedano per es. Allen (1987), Hausser (1999) e le varie sezioni del cap. 12 di "Language Analysis and Understanding" (A. Zaenen chapter ed.), in G.B. Varile, A. Zampolli (eds.) 1997.

so dopo poche frasi, perché le grammatiche ad essi associate non coprono l'intera varietà di strutture sintattiche di fatto usate nei testi o, perché i lessici computazionali disponibili sono incompleti (per es., non contengono o non trattano in modo adeguato le cosiddette "multiwords").

Uno degli ostacoli maggiori è costituito dalla presenza di casi di ambiguità, che il parser non "sa" risolvere. In altre parole, il parser non ha le conoscenze sufficienti per scegliere tra analisi alternative: per es., a livello lessicale, tra significati diversi di una stessa forma grafica, o, a livello sintattico, tra strutture alternative che possono essere assegnate a una frase. Una classificazione delle conoscenze che sarebbero necessarie di fatto non esiste: esse possono essere esprimibili a livello di descrizione semantica delle unità lessicali e delle relazioni possibili tra loro, oppure sono di natura pragmatica o extralinguistica. In questi casi sarebbe necessario che il parsers fosse dotato di meccanismi capaci di compiere inferenze basate sulla conoscenza della realtà.[21]

Oggi siamo capaci di integrare questo tipo di conoscenza in sistemi di NLP solo circoscrivendole a un dominio molto ristretto della realtà, e con riferimento ad applicazioni specifiche per tale dominio.

Molte delle ricerche in corso nel NLP sono rivolte a risolvere, in qualche modo, questi problemi.

3.3. *Shallow Parsers*

Un modo radicale è quello di rinunciare a costruire analisi complete per l'intera frase, e limitarsi a riconoscere alcuni costituenti senza indicare necessariamente quali relazioni esistano tra loro.

Si parla in questo caso di "shallow parsers", che restituiscono per una frase una analisi sintattica 'piatta', sottospecificata, a costituenti non ricorsivi. Essi operano essenzialmente attraverso algoritmi di "chunking", che identificano segmenti sintattici la cui analisi è incontrovertibile, minimizzando o evitando problemi quali la generazione di analisi multiple, o il fallimento dell'analisi dovuto ad informazione lessicale mancante.

Per esempio, nell'esempio precedente, uno "shallow parsers" si limite-

[21] Si pensi alla impossibilità di tradurre in inglese o in francese la semplice frase italiana *arriva oggi*, se non si conosce il genere del soggetto, sottointeso in italiano ma obbligatoriamente espresso da un pronome personale nelle altre due lingue.

rebbe a segmentare la frase in gruppo nominale, verbo, gruppo nominale, gruppo preposizionale, senza voler decidere "dove attaccare" il gruppo preposizionale (v. Figura 21).

Essenzialmente, gli "shallow parsers" operano solamente con regole che adoperano categorie morfosintattiche, ed evitano problemi la cui soluzione richiederebbe conoscenze semantiche, fornite dal lessico, o pragmatiche.

È tuttavia già accertato che analisi sottospecificate di questo tipo possono essere usate utilmente in molte applicazioni: per es. per la acquisizione automatica di informazione lessicale da corpora; il recupero e filtraggio di informazioni da testi; l'estrazione della terminologia del dominio; l'acquisizione di espressioni polilessicali ("multi-word expressions").

L'informazione acquisita può essere usata come punto di partenza per algoritmi di analisi sintattiche a base lessicale (es. analisi funzionale), necessarie per applicazioni che richiedono una "comprensione" completa del testo, quali sommarizzazione e traduzione automatica.

Poiché gli "shallow parsers" sono in grado di "funzionare" su testi reali, mi sembra che sarebbe importante promuoverne l'uso anche nell'ambito delle ricerche dell'HTP. Ciò richiede naturalmente una serie di interventi, quali l'adattamento dei lessici di base e delle regole di analisi (se necessario) allo stato di lingua usata nei testi sui quali si debbono svolgere le ricerche.

3.4. *Risorse linguistiche*

Per tutti i compiti discussi finora c'è la necessità di disporre di risorse linguistiche (RL) adeguate. Il termine RL si riferisce ad insiemi, di solito molto estesi, di dati e descrizioni linguistiche su supporto leggibile dal calcolatore, raccolte per essere usate nel costruire, migliorare, o valutare algoritmi e sistemi di linguistica computazionale. Esempi di RL sono corpora parlati e scritti, database lessicali, grammatiche e terminologie. Il termine viene di solito esteso fino a includere gli strumenti di base di acquisizione, preparazione, raccolta, gestione, uso.

Lo sviluppo di sistemi robusti ed efficienti di elaborazione della lingua dipende in modo cruciale dalla disponibilità di RL di vario tipo.

3.4.1. *Risorse lessicali*

Abbiamo fatto riferimento più volte ai lessici computazionali, come fonti di conoscenze indispensabili per moltissime elaborazioni. Essi forni-

scono per ogni entrata lessicale informazioni formalizzate che ne descrivono le proprietà e il comportamento a diversi livelli: morfologico, sintattico, semantico, ecc. In molte applicazioni, i sistemi devono poter "conoscere" decine se non centinaia di migliaia di entrate lessicali.

Le risorse lessicali di maggior uso possono essere raggruppate in due tipi principali:

- i lessici computazionali propriamente detti, nei quali ciascuna entrata lessicale è accompagnata da informazioni, morfologiche, sintattiche (v. Figura 22) o semantiche (v. Figura 23);[22]
- i networks lessicali, nei quali le entrate lessicali non sono accompagnate da questo tipo di informazioni, ma sono collegate tra loro da relazioni di vario tipo (iperonimia/iponimia, sinonimia, ecc.) (v. Figura 24, nella quale sono elencati gli iponimi del lemma *tessuto*).[23]

Stiamo già sperimentando l'utilizzo di networks lessicali di questo tipo per assistere il ricercatore di discipline umanistiche nelle consultazioni interattive dei testi, descritte prima (n. 2.1). Per esempio, un ricercatore che era interessato a studiare gli influssi della moda francese in Italia nella prima metà dell'800, ha esplorato una raccolta del giornale "Il Corriere delle Dame" ricercando in essa tutti i nomi di tipo di tessuto, aiutandosi appunto con l'elenco degli iponimi di *tessuto* fornitigli dalla nostra risorsa lessicale.

3.4.2. *Corpora parlati e scritti*

Nel campo dello *speech processing*, i sistemi sono costruiti con tecnologie che sono strettamente basate su dati quantitativi ricavati da corpora di dati che rappresentano il dominio delle applicazioni desiderate.

Nell'elaborazione di testi scritti, i corpora testuali sono riconosciuti come la fonte primaria delle informazioni che sono necessarie per descrivere – ai fini computazionali – gli "usi reali" delle lingue in diversi contesti comunicativi.

[22] Gli esempi in figura sono presi dal lessico italiano costruito nell'ambito dei progetti comunitari PAROLE e SIMPLE, che avevano il compito di creare i nuclei iniziali armonizzati di lessici computazionali codificati, rispettivamente, a livello sintattico e semantico, per tutte le lingue dei paesi dell'Unione Europea. Si veda Ruimy et alii.

[23] Gli esempi in figura sono presi dal lessico italiano costruito nel progetto comunitario EuroWordNet, che ha costruito strutture lessicali di questo tipo per italiano, olandese, spagnolo. Cfr. per es. P. Vossen (ed. 1998) "Special Issue on EuroWordNet. Computers and the Humanities", vol. 32, n. 2-3.

Le ricerche in corso cercano innanzitutto di risolvere i problemi metodologici posti dalla composizione dei corpora, che devono essere costituiti in modo da rappresentare al meglio il tipo di lingua che si deve studiare.

Se un corpus deve consentire la estrazione di informazioni sull'uso della lingua a un dato livello di descrizione (per es. morfosintattico), è necessario che esso sia *annotato*, cioè che vengano individuate in esso le unità di analisi proprie di tale livello (per es. le parti del discorso) ed eventualmente ulteriori proprietà (per restare nell'esempio in oggetto, le categorie morfologiche: genere, numero, ecc.).

Spesso la annotazione interessa più livelli, le cui unità di analisi devono essere correlate tra loro, e con le occorrenze grafiche. Nella Figura 25 sono rappresentate le relazioni tra vari livelli di annotazione usati per i dialoghi raccolti nei progetti MATE, TAL, ecc.[24]

Corpora annotati sono usati per "addestrare" analizzatori che operano su base statistica, per es. disambiguatori morfologici che prendono come base le frequenze delle parti del discorso, o delle sequenze di due, tre, *n* parti del discorso. Le frequenze rilevate in un "training corpus", annotato a mano, vengono considerate come probabilità che il "tagger" utilizza per disambiguare casi di occorrenze che possono appartenere a due o più parti del discorso in altri corpora, preferibilmente omogenei, per i tipi di lingua e dominio, al primo. Lo stesso dicasi per allineatori statistici di corpora paralleli.[25]

Questi metodi si prestano naturalmente ad essere utilizzati nell'ambito dell'HTP.

Di particolare rilevanza sono i metodi di autoapprendimento (automatic learning), che associano tecniche dell'AI (quali le reti neuronali) ed elaborazione statistiche, per "estrarre" da corpora (annotati o meno) conoscenze di vario tipo.

Per es., poiché il costo per "scrivere" delle grammatiche formali o completare dei lemmi è molto elevato, si stanno studiando metodi per acquisire automaticamente – da estesi corpora rappresentativi – nuove conoscenze grammaticali da inserire nelle grammatiche e nel lessico, quali per es. i

[24] Per lo stato dell'arte e l'importanza degli studi per il trattamento automatico dei dialoghi si veda il cap. 6 "Discourse and Dialogue" a cura di H. Uszkoreit, in G.B. Varile, A. Zampolli (eds.) 1997. Per una descrizione del progetto MATE, cui la figura si riferisce, si veda Pirrelli, Soria.

[25] Si vedano gli articoli pertinenti in S. Armstrong et alii (eds.) 1999.

quadri di sottocategorizzazione, le restrizioni di selezione (o preferenze) sugli argomenti, collocazioni lessicali tipiche, ecc.

Segnalo qui una ricerca innovativa, che peraltro sta già ottenendo buoni risultati, per ricavare automaticamente da corpora non annotati "classi" di lemmi omogenei rispetto ad alcune proprietà semantiche, e le relazioni semantiche che intercorrono tra di esse: per es., tra una classe di verbi e una classe di sostantivi che possono fungere da loro soggetto o oggetto.[26]

Mi sembra urgente riutilizzare metodi di questo tipo per estrarre da testi letterari, o comunque di interesse culturale, "classi" di parole concettualmente rilevanti, e scoprire quali relazioni intercorrano tra di loro.

3.4.3. *Tipi di azione da svolgere nel settore delle RL*

Per comune consenso, le operazioni da svolgere si articolano in tre tipi o sottosettori.

Specifiche comuni

Si deve cercare il consenso tra le diverse prospettive teoriche e i diversi approcci al disegno dei sistemi, così da produrre *de facto standards* per la creazione delle RL che ne assicurino la riutilizzabilità in applicazioni diverse e la armonizzazione a livello internazionale e multilingue.

La "Text Encoding Initiative" (TEI) ha prodotto una serie di *guidelines* per la codifica di testi.

Il progetto comunitario EAGLES (Expert Advisory Group on Linguistic Engineering Standards) raccoglie gli sforzi europei di ricercatori e industrie per la creazione di standards, consensuali, per l'orale e per lo scritto, per corpora, lessici, tecnologie del parlato, risorse multimodali, formalismi e valutazione dei sistemi.

Costruzione delle Risorse

Si tratta di: (i) assicurare che le risorse monolingui delle varie lingue siano compatibili tra loro; ciò, fra l'altro, faciliterebbe il loro collegamento per formare risorse multilingui; (ii) assicurare la ripartizione degli sforzi, l'utilizzazione delle diverse competenze necessarie, e il riutilizzo delle risorse parziali disponibili.

[26] Si veda Allegrini et alii.

Ciò richiede il coordinamento di progetti nazionali e la applicazione del principio di sussidiarietà tra attività comunitarie e nazionali.

Un esempio è il progetto comunitario PAROLE-SIMPLE, che ha lo scopo di produrre, per tutte le lingue dell'Unione Europea, un insieme armonizzato di corpora e di lessici i quali, pur essendo di dimensioni limitate, dovrebbero costituire il nucleo attraverso il quale organizzare la creazione di risorse armonizzate più estese, di copertura e dimensioni adeguate ai bisogni delle singole lingue, con il supporto di risorse finanziarie dei rispettivi paesi. (Estensione che sta già avvenendo per l'italiano, il greco, il danese, ed altre lingue dell'UE).

Distribuzione delle risorse

È necessario creare delle infrastrutture cooperative per raccogliere, mantenere, disseminare e mettere a disposizione le RL a beneficio dell'intera comunità di ricerca e sviluppo del TAL (Trattamento Automatico delle Lingue).

In USA, questo compito è svolto dal "Linguistic Data Consortium" (LDC), localizzato presso l'Università di Pensilvania. Esso dipende da finanziamenti pubblici (ARPA e NSF) e dalle sottoscrizioni di enti utilizzatori. Nei primi 3 anni di attività LDC ha rilasciato 250 CD-Rom di dati per uso pubblico.

In Europa la Commissione della UE ha sponsorizzato la fondazione di ELRA (European Language Resource Association), che, entrata nel secondo anno di attività, ha nel proprio catalogo, con prezzi differenziati per enti di ricerca e industrie e per soci e non soci, alcune centinaia di risorse, di vario tipo, e per diverse lingue.

Mi sembra che il campo delle RL sia uno dei settori nei quali la collaborazione tra NLP e MTP è più urgente e necessaria, per molte ragioni, ad alcune delle quali accenno brevemente.

Gli operatori dell'HTP possono offrire il know-how che hanno sviluppato nel corso degli anni nel settore della costruzione e analisi di corpora.

D'altro lato, le loro analisi possono essere facilitate e rese più efficaci da metodi di annotazione automatica, dai metodi per la estrazione automatica di conoscenze, dalle procedure di scoperta di classi concettuali e delle loro relazioni. Si pensi per esempio alle sinergie possibili tra le competenze dei due filoni nello studio dei fenomeni legati allo stile, o al genere, ecc., sulla base della analisi quantitativa di corpora.

Per quanto possa non apparire evidente a prima vista, variazioni linguistiche dovute, per esempio, a fenomeni di "stile" legati a un genere di testi,

debbono essere prese in considerazione da sistemi di analisi del linguaggio naturale, e possono in molti casi semplificare il loro compito, riducendo le ambiguità: si pensi, per esempio, all'espressione del tipo "il titolo ha chiuso a .....", in testi finanziari, dove l'"attachment" del gruppo preposizionale introdotto da *a* e la relazione espressa dalla preposizione sono determinabili automaticamente dato il tipo di linguaggio.

4. *Le applicazioni della LC*

Alle applicazioni della LC è stato riconosciuto, da Organizzazioni Internazionali quali la Unione Europea, il Consiglio d'Europa, l'ONU, un ruolo strategico di grande rilievo per lo sviluppo socioeconomico della cosiddetta Società dell'Informazione. Questa società è resa possibile dalla convergenza di informatica e telecomunicazioni. Essa apre nuove possibilità e modi di lavoro e nuove forme di organizzazione della vita sociale, nella quale i networks telematici mettono in comunicazione cittadini di paesi e lingue diverse. La necessità di automatizzare, almeno in parte, le operazioni linguistiche che servono per produrre, gestire, accedere all'informazione, in ambito sia monolingue sia plurilingue, stimola la ricerca di metodi "robusti" per il trattamento delle lingue naturali parlate e scritte, e lo sviluppo di sistemi di comunicazione (uomo-macchina e uomo-uomo) e informazione capaci di utilizzarli per tutta una serie di applicazioni di alto valore economico e di grande impatto sociale.

Elenchiamo, a titolo puramente indicativo, alcune tra le molte tipologie di attività, in cui il TAL ha un ruolo strategico:

- l'insegnamento delle lingue;
- l'insegnamento in generale;
- le applicazioni giudiziarie, ad esempio le trascrizioni, il riconoscimento del parlatore, la verbalizzazione;
- le applicazioni avioniche (guida di elicotteri, cabina d'aereo, simulatori di volo);
- navigazione mono e multilingue su Internet e WWW;
- industria dei contenuti (valorizzazione del patrimonio culturale e artistico italiano);
- servizi stampa;
- localizzazione di prodotti e del software;
- aiuto alla traduzione;

- servizi turistici;
- collaborazione e interazione tra diversi comparti e funzioni della Pubblica Amministrazione;
- accesso dei cittadini ai servizi della Pubblica Amministrazione (agenzie di collocamento e imprese, per offerte e domande di lavoro on-line, servizi di sicurezza sociale, per l'accesso a informazioni personali distribuite, ecc.);
- classificazione automatica di documenti;
- supporto a processi decisionali (per es. basati su normative);
- servizi di documentazione mono e multilingue;
- supporto alla produzione di documenti mono e multilingui;
- stazioni di lavoro personalizzate per l'accesso a biblioteche elettroniche;
- flusso e scambio di notizie tra amministrazioni diverse;
- servizi multilingui di polizia, di sicurezza, di emergenza;
- estrazione di informazioni strategiche da testi e documenti;
- ecc.

Nel settore del trattamento del parlato (*speech processing*), negli anni più recenti sono state sviluppate tecnologie che permettono di analizzare il segnale acustico e applicare conoscenze sulle lingue per compiti quali:

- identificazione del parlatore (verifica della identità: accesso a servizi, locali, ecc.);
- identificazione della lingua (indirizzare un utente a operatori che parlano la sua lingua);
- riconoscimento del parlato (convertire le onde sonore, catturate da un microfono, in una serie di "parole scritte" che il sistema potrà utilizzare per preparare documenti – macchine per dettare –, interrogare banche dati, ecc.);
- output vocale: dall'input ortografico, alla trascrizione fonetica con marche prosodiche, alla sintesi della voce.

La tecnologia ha già oltrepassato la soglia della concreta applicabilità commerciale. Le applicazioni possibili sono di grande impatto economico. Si pensi, per esempio, ad applicazioni quali i servizi automatizzati come le segreterie telefoniche, le consultazioni telefoniche di basi di dati commerciali (pagine gialle, orari ferroviari ed aerei, teleacquisti); i servizi di CTI (Interfacce Telefoniche Computerizzate); le applicazioni militari (comando vocale di mezzi e di armi, protezione biometrica di accessi, etc.); i giochi, ed in generale l'intrattenimento; la robotica (autovetture, elettrodo-

mestici, etc.); le applicazioni di aiuto ai disabili (lettura dei testi, comandi vocali); interfacce vocali con il calcolatore.

Il passo successivo, sul quale si concentrano attualmente gli sforzi di centri pubblici e privati di ricerca, concerne (i) la comprensione del parlato: le parole *riconosciute* diventano l'input di un *processo linguistico* che ha lo scopo di produrre una qualche rappresentazione formale del significato degli enunciati emessi dall'utilizzatore; (ii) *la gestione del dialogo*: per es., sistemi che forniscono istruzioni a un operatore assistendolo attraverso un dialogo nello svolgimento di un compito (es. manutenzione); oppure sistemi di informazione medica che forniscono aggiornamenti in tempo reale durante una operazione chirurgica.

Gli sforzi attuali sono diretti a ridurre progressivamente il grado di dipendenza dei sistemi da un dominio semantico ristretto.

Alcune innovazioni tecnologiche (per es. la tecnologia di "Internet mobile", cioè l'accesso vocale a Internet per mezzo della telefonia portatile) e alcuni recenti successi tecnologici (per es. la possibilità di trascrivere automaticamente il parlato da video sonori, e di applicare ai testi così trascritti sistemi automatici di indicizzazione e "retrieval" dell'informazione) hanno attirato la attenzione sulle necessità di prendere in considerazione i settori della multimedialità e delle multimodalità.

Nella interazione uomo-uomo (mediata dalla tecnologia), uomo-macchina, uomo-informazione, è importante, per una comunicazione facile ed efficiente, poter utilizzare i diversi canali disponibili: vista, udito, tatto, ecc.[27] Le tecnologie della lingua, scritta e parlata, della visione, del "tatto" sono ancora imperfette, ma stanno già iniziando gli esperimenti di integrazione per rendere la comunicazione uomo-macchina più efficiente. Domini di applicazione molto ristretti presentano le opportunità più indicate per tentare queste sinergie. Sono fattori critici: le analisi dei fattori umani, lo studio di corpora di interazioni, la costituzione di "lessici" interpretativi di segni, gesti, intonazioni, ecc.

In particolare si vanno sviluppando sistemi di rappresentazione *multicodale*, usate per codificare strutture di elementi atomici, sintattici, semantici, pragmatici associati con media e modalità.

Si parla molto oggi di multimedialità, soprattutto per la produzione e la distribuzione del cosiddetto 'contenuto digitale'.[28]

[27] Si veda il capitolo 8 "Multilinguality" a cura di A. Zaenen, in G.B. Varile, A. Zampolli (eds.) 1997.

[28] Questa espressione include, nell'uso della CEE, intrattenimento, educazione, pub-

La Società dell'Informazione è necessariamente multilingue. I produttori e distributori di contenuto digitale dovrebbero ricorrere alle tecnologie della lingua sia per produrre, sia per rendere accessibile il contenuto in molte lingue.

Qui mi limito a portare l'esempio di uno strumento multimediale (ADDIZIONARIO) (vd. Turrini), sviluppato nel mio Istituto come laboratorio linguistico ipermediale, che vuole facilitare bambini e ragazzi ad usare in modo piacevole e poco faticoso i dizionari, e nell'apprendimento delle lingue. Nell'aula del laboratorio multimediale sono disponibili due strumenti: un dizionario per bambini, fatto da bambini; un quaderno che il bambino può usare per costruire un proprio dizionario correlandolo di pronuncia, definizioni, esempi, disegni e suoni.

ADDIZIONARIO, pur nella sua semplicità di concezione, sta avendo grande successo e risonanza nelle scuole, nei giornali, alla televisione, ecc.

Voglio solo sottolineare come la memorizzazione strutturata delle definizioni e degli esempi, prodotti da bambini diversi, offra un materiale di studio inedito per quanti studiano e desiderano conoscere le capacità linguistiche, concettuali, creative dei bambini, come psicolinguisti, psicologi, terapisti, insegnanti curriculari e di sostegno.

Inoltre, la applicazione di metodologie della linguistica computazionale permettono di elaborare ulteriormente i materiali linguistici raccolti: per esempio, applicando tecniche di analisi note, ricostruire relazioni concettuali (ipernominie, ecc.) che strutturano le conoscenze lessicali dei bambini.

Traduzione Automatica (TA)

Un altro esempio di applicazione che sta richiamando l'interesse dei media e delle industrie legate ad Internet e al WWW è dato dal progetto UNL (Universal Networking Language).

Come è noto, la storia della Traduzione Automatica è fatta di grandi aspettative da parte dei finanziatori e del pubblico, spesso deluse, alimentate da promesse dei gestori dei progetti di TA formulate allo scopo di ottenere dei fondi. Ancora oggi la valutazione dei sistemi di TA è un problema irrisolto. Sembra comunque certo che la TA "di qualità" sia per ora riservata a testi il cui dominio è molto ristretto.

blicità, "e-commerce" digitali o qualsiasi loro combinazione, indipendentemente dal mezzo o formato specifico.

Il termine TA include ogni processo che, utilizzando il calcolatore, trasporta (o aiuta l'utilizzatore a trasportare) un testo da una lingua naturale ad un'altra.

Si distinguono più precisamente 3 tipi di traduzione: (i) la traduzione interamente automatica, eseguita senza intervento umano durante tutto il processo; (ii) la traduzione assistita dall'uomo: il calcolatore compie il processo chiedendo aiuto all'uomo (mono o bilingue) in casi "difficili"; (iii) la traduzione (umana) assistita dal calcolatore: il traduttore compie il processo di traduzione usando come "assistenti" sistemi computazionali (per es. dizionari, spelling chekers, memorie di traduzione, ecc.).

Le tecniche impiegate possono essere raggruppato in tre categorie principali: (i) approccio basato su "regole": i sistemi commerciali tipici usano dizionari tra 250 mila-500.000 parole e 500-1000 regole grammaticali per lingua; (ii) metodi statistici: tecniche statistiche basate su corpora paralleli (e comparabili) per produrre le regole di corrispondenza tra lingue; (iii) approccio basato su esempi, sull'analogia, sulla memoria di traduzione. Il sistema ha una serie di coppie di "espressioni" bilingui memorizzate in un data base di esempio.

Il sistema cerca nel DB il "matching" migliore per l'"espressione" da tradurre.

Sembra auspicabile che in un prossimo futuro ci si avvalga dei 3 tipi di tecnica assieme.

Altri criteri di classificazione prendono in considerazione il canale di cui si avvale la comunicazione (interazione *scritta*, per es. via posta elettronica, o *orale*, con un sistema di traduzione che media tra i parlanti), e la direzione della comunicazione. Si distingue tra assimilazione – quando un individuo o una organizzazione vogliono convertire nella propria lingua materiali scritti da altri in altre lingue, e disseminazione, se, invece, essi vogliono diffondere propri materiali – scritti in una lingua – in una varietà di lingue del mondo.

La qualità della traduzione richiesta è diversa, per esempio, se il testo tradotto viene solo scorso ("lettura diagonale") per identificarne gli argomenti principali e per identificare eventuali parti interessanti, o se la traduzione (per es. di un manuale tecnico di manutenzione) deve essere letta accuratamente.

Due sono gli approcci tradizionali alla TA.

Transfer: la struttura formalizzata (di tipo sintattico-semantico) assegnata dal componente di analisi al testo da tradurre, propria della lingua

sorgente, viene trasformata, da un componente di transfer[29] nella rappresentazione strutturale propria delle lingue di arrivo, a partire dalla quale il componente di generazione produce il testo tradotto in tale lingua.

In questi sistemi è necessario sviluppare tanti componenti di transfer quante sono le coppie di lingue tra cui tradurre, moltiplicate per due (il transfer dalla lingua *a* alla lingua *b* richiede di solito regole diverse dal transfer inverso, da *b* verso *a*).

Interlingua: gli analizzatori forniscono una rappresentazione che è la stessa per tutte le lingue (interlingua), dalla quale partono le generazioni del testo nelle diverse lingue di arrivo.

Evidentemente, se concretamente fattibili, i sistemi ad interlingua presentano il vantaggio di eliminare i componenti di transfers, il cui numero cresce con il crescere delle coppie di lingue da tradurre.

Il sistema UNL (si veda lo schema di Figura 26; Prodanoff) è di questo secondo tipo.

Lo sforzo per ora si è concentrato (i) sul disegno di una interfaccia, costituita da un catalogo di concetti universali e di relazioni, le quali generalizzano le relazioni espresse, con mezzi diversi (casi, preposizioni, funzioni grammaticali, ecc.), dalle varie lingue e (ii) nella implementazione di generatori che, a partire da questa "rappresentazione universale", generano i testi nelle diverse lingue. Gli esperimenti condotti fino ad ora hanno dato risultati promettenti. Si tratta di verificare ora se lo stato dell'arte è maturo per lo sviluppo di sistemi di analisi capaci di produrre automaticamente le strutture di interfaccia. Questo interrogativo è ancora oggetto di discussione. Alcuni sostengono però che, anche in mancanza di analizzatori adeguati, lo stesso autore del testo (cioè della parte in lingua naturale di un sito WWW) potrebbe curare la conversione del proprio testo nel linguaggio universale di rappresentazione (UNL), interamente a mano o interagendo con un analizzatore capace di riconoscere i casi difficili (per es. di ambiguità) e di chiedere il suo aiuto.

L'autore potrebbe così rendere il proprio testo accessibile e comprensibile a chiunque abbia accesso a un generatore che fa da interlingua verso la propria lingua.

[29] La CEE dovrà probabilmente affrontare i problemi posti dal probabile accesso di nuovi paesi dell'Est alla Unione.

5. *Il supporto pubblico alla linguistica computazionale*

Come si è detto, il potenziale strategico sul piano economico e sociale viene sempre più di frequente riconosciuto da autorità nazionali e sopranazionali.

Per esempio, nei successivi Programmi Quadri di Ricerca, la CEE ha manifestato il proprio crescente interesse per le applicazioni della cosiddetta "industria delle lingue", destinando al TAL 40 MECU nel 3°, 80 MECU nel 4°, 200 MECU nel 5° Programma Quadro di Ricerca (PQR), con l'intento di promuovere la competitività dell'industria europea, la qualità dei servizi ai cittadini, il multilinguismo dei servizi, il supporto alle categorie deboli (es. disabili), la valorizzazione del patrimonio culturale, ecc.

In particolare, gli obiettivi assegnati al settore nel 5° PQR possono essere così riassunti: (i) aggiungere multilingualità a tutti gli stadi dei sistemi di informazione; (ii) assicurare la interattività naturale (dialoghi multimodali, ecc.); (iii) assimilazione attiva del contenuto digitale (sommari automatici, ecc.).

Inoltre, è interessante notare che questo è stato il primo settore del programma IST (Information Science and Technology) a implementare l'accordo transatlantico di cooperazione tecnico-scientifica (CEE/NSF, ARPA).

Se le autorità di organismi sopranazionali sono motivate soprattutto dalla necessità di gestire le diverse lingue dei paesi che le costituiscono, le autorità dei singoli paesi trovano una forte motivazione,[30] oltreché nel valore strategico (industriale e commerciale) delle applicazioni della LC, anche nel valore sociale (facilitare un accesso "democratico" all'informazione a tutti i cittadini)[31] e soprattutto nella consapevolezza che lingua e cultura so-

[30] All'epoca della sua vicepresidenza, Al Gore ha affermato che la vera sfida tecnologica a cavallo dei due millenni consiste nel far sì che i cittadini di tutti i paesi possano accedere ai networks globali di informazione nella propria lingua, usando la voce e non una tastiera.

[31] Si vedano Maccanico (1997) e Zampolli (1997) in Ridolfi, Piraino, eds. (1997).

Nella dichiarazione comune dei Ministri di 29 Paesi rilasciata al termine della Conferenza Ministeriale Europea sui GIN (Global Information networks) (Bonn, 6-8 luglio 1997), si leggono le seguenti raccomandazioni:

- è necessario che le informazioni e il contenuto dei GIN siano prodotti, e resi accessibili, nelle diverse lingue;
- è essenziale che non si crei una divisione tra quelli che possono accedere all'informazione e quelli che non possono;
- si deve promuovere la "democrazia elettronica", attraverso lo sviluppo, per ciascuna lin-

no indissolubili, e che (come ha detto il Presidente Mitterand in un celebre discorso alla Accademia di Francia) "le lingue che non sapranno informatizzarsi rischiano di perdere il ruolo veicolare nella Società dell'Informazione multilingue".

In Italia,[32] in occasione della preparazione della partecipazione italiana ad alcuni progetti comunitari, si è costituito per iniziativa del Ministero PTT e dell'ISPT, un gruppo di lavoro, i cui membri provengono da diversi Ministeri, dalla Pubblica Amministrazione, dall'Università, dal CNR e altri Enti di ricerca pubblici e privati, dall'industria, dal commercio, dalla scuola, dal settore degli erogatori di servizi.

Il gruppo di lavoro ha ritenuto opportuno intraprendere due iniziative:

- sulla base della analisi dello stato dell'arte del TAL in Italia e dei bisogni prioritari del nostro paese, ha delineato un primo piano di lavoro, da proporre su scala nazionale;
- ha organizzato una Conferenza sul tema "Trattamento automatico della lingua nella Società dell'Informazione" (Roma - gennaio 1997 - PTT ISPT) che è stata aperta dal Ministro PTT, dai Sottosegretari alla Ricerca e all'Industria, da altre importanti Autorità. Hanno presenziato oltre 300 partecipanti, per la maggior parte dei settori dell'industria e dei servizi.

La conferenza ha segnato un importante riconoscimento del TAL come settore disciplinare autonomo, del suo potenziale per lo sviluppo socioeconomico e culturale del nostro paese nella società globale multilingue, e della necessità di promuovere la lingua italiana in questo contesto, ed ha espresso il consenso della comunità sugli obiettivi identificati nel piano iniziale di attività proposto dal Gruppo di Lavoro.

Alcuni obiettivi del piano potranno essere realizzati grazie a due progetti di interesse nazionale, complementari, da poco avviati dal MURST: "Risorse Linguistiche Infrastrutturali per il trattamento automatico delle lingue" (1999-2001), e "Linguistica Computazionale, Ricerche monolingui e multilingui" (2000-2002).

Gli obiettivi dei due progetti possono essere così riassunti:

gua, di strumenti TAL che facilitino la interazione dei cittadini con i diversi servizi offerti nei GIN (per es.: accesso al patrimonio culturale; grandi archivi di informazione; commercio elettronico; ecc.).

[32] Per una breve storia dello sviluppo della LC in Italia si veda Zampolli (2000).

A. Creazione di alcune risorse linguistiche di base per lo sviluppo del Trattamento automatico della Lingua italiana, e del software relativo: corpora di parlato, sia di carattere generale sia per applicazioni specifiche; corpora paralleli contrastivi (lingua scritta); annotazione dei corpora parlati e scritti a diversi livelli linguistici (ivi compreso il livello semantico); lessici computazionali, con informazioni fonologiche, morfologiche, sintattiche, semantiche; basi di conoscenze lessicali strutturate di relazioni semantiche; sistemi per la estrazione e la acquisizione dinamica di conoscenze linguistiche da corpora.
B. Ambienti di sviluppo per facilitare l'utilizzo di strumenti, risorse, tecnologie linguistiche, da parte di integratori che intendano creare applicazioni per lo speech o il NL Processing: (i) ambienti per lo sviluppo e/o l'integrazione in sistemi applicativi di grammatiche, lessici, strutture semantiche, (ii) ambiente per l'integrazione di tecnologie vocali esistenti in prodotti e servizi di vario tipo.
C. Valutazione, utilizzazione/dimostrazione di sistemi, strumenti e risorse creati in A e in B nella realizzazione o nel miglioramento di prototipi applicativi: traduzione; estrazione, filtraggio, recupero dell'informazione; supporto all'"authoring" e all'uso di testi tecnici; navigazione su Internet; varie applicazioni vocali; ecc.
D. Creazione di un network di operatori italiani: cooperazione e interazione tra ricerca, sviluppo, applicazioni, servizi, pubbliche amministrazioni, utenti; diffusione dei risultati; ecc.

Questo network verrà esteso e consolidato, anche con l'apporto dei progetti citati, con lo scopo di contribuire ai seguenti obiettivi:

- promuovere e intensificare la cooperazione e gli scambi tra ricerca pubblica, ricerca privata, industrie, integratori di applicazioni, fornitori di servizi, utenti attuali e potenziali;
- assicurare la diffusione dei risultati dei progetti già in corso d'opera, collaborando alla loro valutazione;
- potenziare il settore delle tecnologie della lingua, presentando una comunità articolata ma cooperante, e creando un framework per discutere dei bisogni e delle priorità del paese e per diffonderne la consapevolezza;
- promuovere lo sviluppo di obiettivi/aree "urgenti", già identificate nel piano ma non ancora incluse nei progetti finanziati;
- definire, adeguandoli agli sviluppi dello stato dell'arte, della società, e delle esigenze del mercato, gli obiettivi di auspicabili follow-up di interesse nazionale, e promuoverne la attuazione;

– favorire la collaborazione e la partecipazione italiana alle attività internazionali, e in particolare a quelle promosse dalla Commissione;
– promuovere a tutti i livelli la formazione universitaria e professionale nel settore della tecnologia delle lingue, come settore disciplinare autonomo;
– promuovere la organizzazione periodica di Conferenze nazionali.

Bibliografia

Allegrini P., Montemagni S., Pirrelli V. (2000): *Example-based automatic Induction of semantic Classes from Text Corpora through entropic Scores*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).

Allen J. (1987): *Natural Language Understanding*, The Benjamin/Cummings Publishing Company.

ALPAC Report (1966): *Language and Machine: Computers in Translation and Linguistics*, Washington, DC, National Research Council. Automatic Language Processing Advisory Committee.

Armstrong S., Church K., Isabelle P., Manzi S., Troukermann E., Yarowsky D. (eds. 1999): *Natural Language Processing Using Very Large Corpora*, Kluwer.

Atkins B.T.S., Zampolli A. (eds. 1994): *Computational Approaches to the Lexicon*, OUP.

Battista M., Pirrelli V. (2000): *Syntagmatic and paradigmatic Issue in computational Morphology*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).

Bedini L., Minutoli S., Tonazzini A. (2000): *Blind Restoration of degraded Texts based on Wiener Filtering*, in Bozzi A. (ed. 2000), pp. 97-120.

Biagini L., Bindi R., Goggi S., Marinelli R., Orsolini P. (2000): *Criteria and Methods for building the Italian PAROLE Corpus*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).

Booth A.D., Cleave J.P., Brandwood B.A. (1958): *Mechanical Resolution of Linguistic Problems*, London, Butterworths Scientific Publications.

Bozzi A. (ed. 2000): *Computer-aided Recovery and Analysis of damaged Text Documents*, Bologna, CLUEB.

Bozzi A. (2000a): *Image and Text: the Workstation Prototype*, in Bozzi A. (ed. 2000), pp. 201-211.

Bozzi A. (2000b): *The Linguistic Module*, in Bozzi A. (ed. 2000), pp. 187-200.

Bozzi A. (2000c): *Verso una filologia computazionale: metodi e procedure per lo studio e l'edizione dei Beni Librari*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).

Bujas Z., *Computers in Serbocroat-English Contrastive Analysis Project*, ICCL.

Burton D.M. (1981): *Automated Concordances and Word Indexes: The Early sities and the Early centers*, in "Computers and the Humanities", n. 15, pp. 83-100.

Busa R. (1951): *Sancti Thomae Aquinatis Hymnorum ritualium varia specimina concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*, Milano, Bocca.

Busa R. (1968): *De lexico electronico latino*, in *Acta omnia Gentium ac Nationum Conventus Latinis Litteris*, Roma.

Busa R. (1968a): *Traitement de mots d'une fréquence extrême*, in *Les machines dans la linguistique*, Praga.

"Cahiers de Lexicologie. Revue Internationale de Lexicologie et Lexicographie", Paris, Didier Larousse.

Cappelli A., Caligaris C., Catarsi M.N., Moretti L. (2000): *Tecnologie per l'elaborazione del contenuto: alcune applicazioni*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).

Cappelli A., Moretti L. (2000): *Strumenti linguistici basati sulla conoscenza*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).

Cignoni L., Coffey S. (2000): *The Notion of lexical Frequency as applied to phraseological Units*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).

Dagan I., Church K., Gale W. (1999): *Robust Bilingual Word Alignment for Machine Aided Translation*, in Armstrong S., Church K., Isabelle P., Manzi S., Troukermann E., Yarowsky D. (eds.).

De Marcken C. (1999): *On the Unsupervised Induction of Phrase - Structure Grammars*, in Armstrong S. et alii (eds.), pp. 191-208.

Genet J.P., Zampolli A. (eds. 1992): *Computers and Humanities*, Dartmouth, European Science Foundation.

Grishman R. (1999): *Iterative Alignment of Syntactic Structures for a Bilingual Corpus*, in Armstrong S., Church K., Isabelle P., Manzi S., Troukermann E., Yarowsky D. (eds.).

Hays D.G. (1965): *Readings in Automatic Language Processing*, American Elsevier.

Hays D.G. (1967): *Introduction to Computational Linguistics*, New York, American Elsevier.

Hussler R. (1999): *Foundations of Computational Linguistics, Man-Machine Communication in Natural Language*, Springer.

Ide N., Véronis J. (eds. 1998): *Special Issue on Word Sense Disambiguation: The State of the Art*, in "Computational Linguistics", n. 24 (1), pp. 1-40.

Kay M. (1967): *Standards for Encoding Data in a Natural Language*, in "Computers and the Humanities", v. I, n. 5, pp. 170-177.

Lepschy G. (1966): *La linguistica strutturale*, Torino, Einaudi.

Les Machines dans la Linguistique (1968), Praga.

Maccanico A. (1997): *Discorso di apertura della Conferenza*, in Ridolfi R. Piraino (eds.), pp. 17-19.

Montemagni S., Pirrelli V. (2000): *Shallow parsing of Italian*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).

Picchi E. (2000): *Linguistica Computazionale - Analisi testuale e lessicale* (eds. 2000).
Picchi E. (2000a): *PiSystem: strumenti integrati per l'analisi testuale*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).
Picchi E. (2000b): *Corpora multilingui paralleli, comparabili, cross-language, information retrieval*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).
Picchi E., Montemagni S., Biagini S., *DBT-ALT: a System for Storying and Querying the Data of the Atlante Lessicale Toscano (ALT)*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).
Pirrelli V., Soria C. (2000): *MATE: a multilevel annotation scheme for dialogue annotation*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).
Prodanoff I. (2000): *UNL*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).
Ridolfi R. Piraino (1997): *Trattamento automatico della lingua nella Società dell'informazione*, Atti del Convegno, Roma.
Ruimy N., Corazzari O., Gola E., Spanu A., Calzolari N., Zampolli A. (2000): *The PAROLE Model and the Italian Syntactic Lexicon*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).
Tavoni M. (ed. 2000): *Biblioteca Italiana telematica - Risultati e prospettive di una ricerca*, Pisa, CIBIT.
Turrini G. (2000): *Addizionario*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).
Van der Eijk P. (1999): *Comparative Discourse Analysis of Parallel Texts*, in Armstrong S., Church K., Isabelle P., Manzi S., Troukermann E., Yarowsky D. (eds.).
Varile G.B., Zampolli A. (eds. 1997): *Survey of the State of the Art in Humane Language Technology*, Pisa-Cambridge, Giardini-Cambridge University Press.
Vidal-Beneyto J. (ed. 1991): *Las industrias de la lengue*, Madrid, Ediciones Piramide.
Vossen P. (ed. 1998): *Special Issue on EuroWordNet*, in "Computers and the Humanities", vol. 32, n. 2-3.
Walker D., Zampolli A., Calzolari N. (eds. 1995): *On Automating the Lexicon. Research and Practice in a Multilingual Environment*, Proceedings of a Workshop held in Grosseto, OUP.
Wu D. (1999): *Trainable Coarse Bilingual Grammars for Parallel Text Bracketing*, in Armstrong S., Church K., Isabelle P., Manzi S., Troukermann E., Yarowsky D. (eds.).
Zampolli A. (1969): *Due conversazioni sullo stato attuale della Linguistica Computazionale*, Pisa.
Zampolli A. (1970): *Progetti e metodi della Sezione Linguistica del CNUCE*, in "Revue", Organisation Internationale pour l'Étude des Langues Anciennes par ordinateur, III, pp. 39-83.
Zampolli A. (1973): *Linguistica Matematica e Calcolatori. Atti del Convegno della Prima Scuola Internazionale (Pisa, 16 agosto-6 settembre 1970)*, Firenze, Olschki.

Zampolli A. (1973a): *La Section Linguistique du CNUCE*, in Zampolli A. (1973), pp. 133-199.

Zampolli A. (1973b): *Humanities Computing in Italy*, in "Computers and the Humanities", vol. VII, n. 6, pp. 343-360.

Zampolli A. (1974): *Problemi di Linguistica applicata computazionale*, Pisa, CNUCE - Consiglio Nazionale delle Ricerche.

Zampolli A. (1976): *Les dépouillements électroniques: quelques problèmes de méthode et d'organisation*, in Fattori M., Bianchi M. (eds. 1976), *I Colloquio Internazionale del Lessico Intellettuale Europeo*, Roma, Edizioni dell'Ateneo, pp. 173-178.

Zampolli A. (1997): *Introduzione*, in Ridolfi R. Piraino (eds.), pp. 7-14.

Zampolli A. (1997a): *The PAROLE Project in the General Context of the European Actions for Language Resources*, in Marcinkeviciene R., Volz N. (eds.), *TELRI Proceedings*, Kaunas (Lithuania), pp. 185-210.

Zampolli A. (2000): *Introduzione*, in Zampolli A., Calzolari N., Cignoni L. (eds. 2000).

Zampolli A., Calzolari N., Cignoni L. (eds. 2000): *L'Istituto di Linguistica Computazionale del CNR: attività di ricerca*, Pisa, Giardini (in corso di stampa).

[CntFam] - Famiglia n. 1

N.Famiglie definite = 1 [seq] [5,12]
primo & (amor | amore)
N. Contesti associati : 5

1) fecemi la divina podestate,/ la somma sapienza e 'l primo amore./ Dinanzi a me non fuor cose create/ - **Inf., can.3.6**
2) Cesare fui e son Iustiniano,/ che, per voler del primo amor ch'i' sento,/ d'entro le leggi trassi il - **Par., can.6.11**
3) Ben parve messo e famigliar di Cristo:/ che 'l primo amor che 'n lui fu manifesto,/ fu al primo consiglio - **Par., can.12.74**
4) vero a l'intelletto mio sterne/ colui che mi dimostra il primo amore/ di tutte le sustanze sempiterne./ Sternel la - **Par., can.26.38**
5) panno fa la gonna;/ e drizzeremo li occhi al primo amore,/ sì che, guardando verso lui, penètri/ - **Par., can.32.142**

Figura 1. Ricerca di cooccorenze di parole nella "Divina Commedia"

Co-Occorrenze statistiche

1)	4	13	7.620	1.500	ACCENDE
2)	2	8	7.320	1.500	ACCESO
3)	2	9	7.150	1.000	ACCESE

[Cnt] - [Co_occ]: FOCO

cima di giudicio non s'awalla/ perché foco d'amor compia in un punto/ ciò che de' sodisfar chi
di lieve si comprende/ quanto in femmina foco d'amor dura,/ se l'occhio o 'l tatto spesso - *Pu*
raggiò nel monte Citerea,/ che di foco d'amor par sempre ardente,/ giovane e bella in - *Purg*
mi rispuose tanto lieta,/ ch'arder parea d'amor nel primo foco:/ «Frate, la - *Par., can.3.6*
e credendo s'accese in tanto foco/ di vero amor, ch'a la morte seconda/ fu degna di - *Par.,*

12)	5	67	5.576	2.400	FOCO
13)	3	54	5.150	1.333	VERO
14)	2	39	5.035	2.000	MOSSE
15)	2	42	4.928	2.500	ANIME
16)	2	44	4.861	2.000	QUINCI

Figura 2. Mutual information: cooccorrenze statistiche di una o più parole

Co-Occorrenze statistiche

1)	14	*22589*	2.160	1.000	IN
2)	9	*24370*	1.413	1.000	A
3)	6	*25000*	0.792	1.000	DI

[Cnt] - [Co_occ]: IN

1) della non-Europa>. Chi non crede nello scenario virtuoso proposto da C
2) sono preoccupazioni relative perché a Bonn si crede in un prossimo raf
3) \Tit\... di rilancio del premier Miyazawa Tokio non crede nella ripresa G
4) è, come l'opposizione voleva far credere in campagna elettorale, l'affoss
5) mentre quei pochi che continuano a credere nell'azionario non potevano
6) <l'esito del voto permette di credere nella capacità del Governo Suchoc
7) alle speranze di Sapin e di non credere in una riduzione dei tassi tedesc
8) hotel in Italia. Il mercato sembra credere in una rapida soluzione delle t
9) risollevare le sorti del Ristretto? A credere nei poteri ricostituenti di q
10) non credo di esserlo. Io credo nell'ordine della società. Un libertario –
11) \Tit\... difficile <Non tutti i tedeschi credono in Dio - si lamenta il presid
12) Cee, Jacques Delors - ma tutti credono nella Bundesbank>. Ma in che
13) movimento referendario e di tutti quelli che credono nelle riforme>. In

Figura 3. *Mutual information*: ricerca automatica delle preposizioni associate ad un determinato lemma

Data Base Testuale

Testi lemmatizzati

Divina Commedia

Metti in quadro | **Fai famiglia** | **Contesti lemma** | **Quadro**

Nr	Forma	Lemma	CatGr	Freq.
1	va	andare	vi*1ips2	9
2	vassi	andare	vi*ips3	1
3	anderemo	andare	vi+1ifp1	1
4	andrai	andare	vi+1ifs2	1
5	andavamo	andare	vi+1iip1	1
6	andava	andare	vi+1iis1	5
7	andava	andare	vi+1iis3	1
8	van	andare	vi+1ipp3	1
9	vo	andare	vi+1ips1	1
10	va	andare	vi+1ips2	18
11	vai	andare	vi+1ips2	1
12	vassi	andare	vi+1ips3	3
13	andaro	andare	vi+1irp3	1
14	vanno	andare	vi+ipp3	5
15	andasse	andare	vi1cis3	3
16	andiam	andare	vi1cpp1	1
17	andiamo	andare	vi1cpp1	2
18	andianci	andare	vi1cpp1	1

Figura 4. *DBT per testi lemmatizzati: ricerca lemma "andare"*

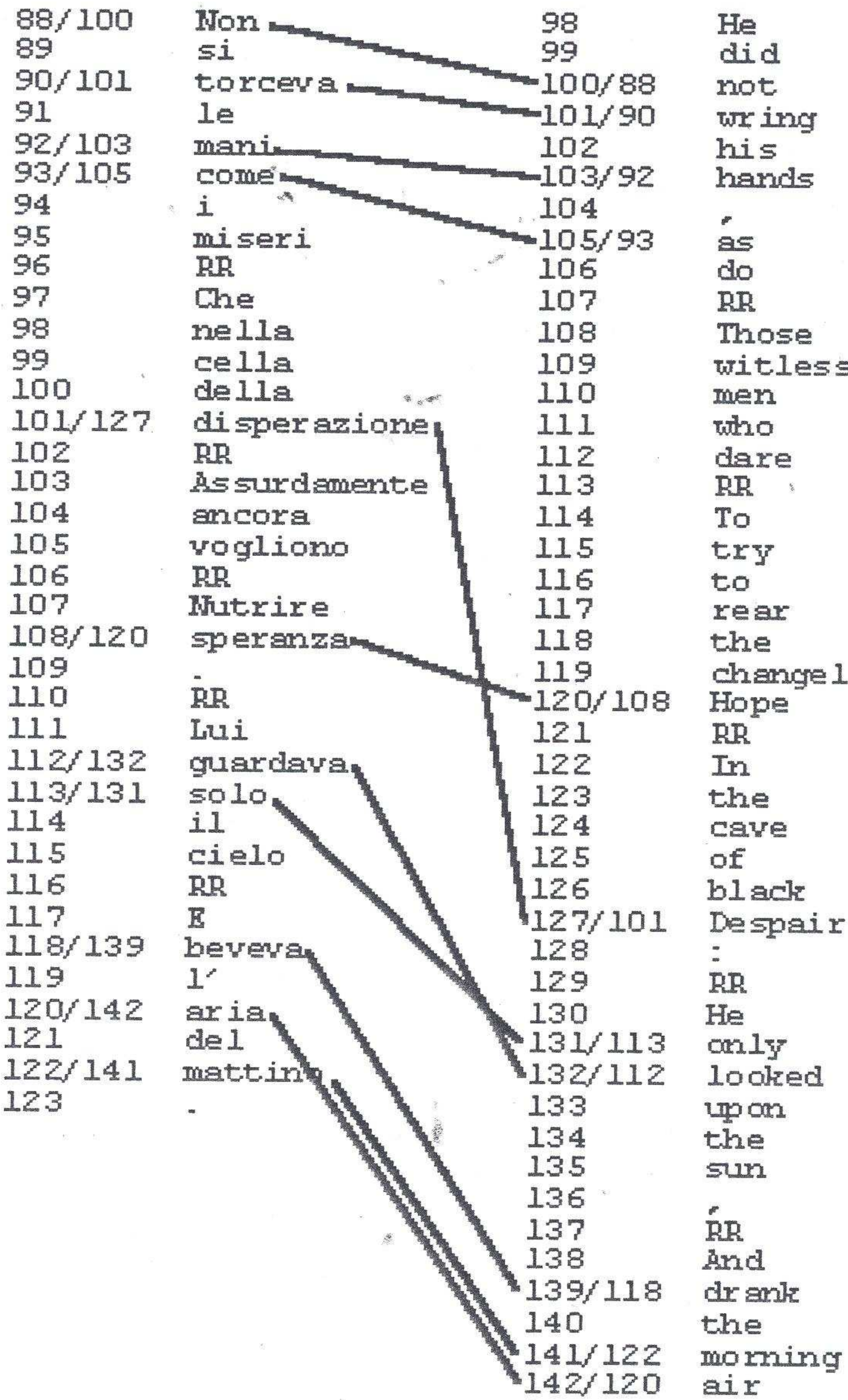


Figura 6. Creazione automatica della rete di links tra testi paralleli

[Cnt] - {E}colour

1) {E} She sounded so sorry for us when she said it, that I couldn't help crying." Another chick, lighter in colour and a few hours younger than the first, so that he could still remember with gratitude the cosy atmosphere which had surrounded - **E-madre.41**
1) {I} tapini perché ci manca la madre. Lo disse con un accento di sì forte compassione ch'io dovetti piangere." Un altro più bianco e di qualche ora più giovane del primo, per cui ricordava ancora con gratitudine l'atmosfera dolce da cui era - **I-madre.38**

Figura 7. La rete di collegamenti "cross language" per la sincronizzazione di testi paralleli

[Cnt] - {E}colour

1) {E} She sounded so sorry for us when she said it, that I couldn't help crying." Another chick, lighter in colour and a few hours younger than the first, so that he could still remember with gratitude the cosy atmosphere which had surrounded - **E-madre.41**
1) {I} tapini perché ci manca la madre. Lo disse con un accento di sì forte compassione ch'io dovetti piangere." Un altro più bianco e di qualche ora più giovane del primo, per cui ricordava ancora con gratitudine l'atmosfera dolce da cui era - **I-madre.38**
2) {E} But which bird? It was alone no longer; the whole sky over there was a beating of wings of every colour and shape. Leaning out from the edge of our world I watched the continent floating nearer. "It's going to - **E-bird.138**
2) {I} sorvolando. Ma quale uccello? Non era più solo, tutto il cielo là sopra era uno sbattere d'ali d'ogni colore e d'ogni forma. Sporgendomi dall'orlo della nostra terra io guardavo avvicinarsi il continente alla deriva. "Ci viene - **I-bird.135**
3) {E} The system should also allow for noise cleaning, edge enhancement, contrast modification and pseudo colour display, which are used to modify the aspect of displayed images in order to facilitate visual interpretation. 3. - **E-spray.81**
3) {I} sistema deve consentire l'esecuzione delle operazioni di pulitura del rumore, esaltazione di contorni, modifica del contrasto e rappresentazione con pseudocolori, che permettono di modificare l'aspetto dell'immagine visualizzata al fine di facilitarne l'interpretazione visiva. 3. DESCRIZIONE DELLA - **I-spray.75**

Figura 8. Contesti contrastivi italiano-inglesi della parola inglese "colour"

[CntFam] - Family n. 1

N.Famiglie definite = 1
(10,25)colour & change
N. Contesti associati : 3

1) {E} The sun filled late September days continue, followed by those splendid October days marked by golden sunset and skies which change colour from green to gold, as in the Venetian paintings of Cima da Conegliano and Titian. Those who live in - **E-impressi.63**
1) {I} altro succede. Continuano le giornate di sole della fine di settembre, e le splendide ottobrate dai tramonti dorati, dai cieli che trascolorano dal verde all' oro, come quelli che si trovano nella pittura veneta da Cima di Conegliano a Tiziano. - **I-impressi.66**
2) {E} we begin to notice the melancholy signs that the season is hurrying to an end. The corn begins to change colour in the fields. The leaves are partly green and partly the dry dirty gray colour of straw and stubble. The - **E-impressi.134**
2) {I} a lanciare i suoi segnali melanconici di stagione che corre verso la fine. Le piante del granoturco cominciano a cambiare colore nel campo. Le foglie sono mezzo verdi e mezzo secche, di un grigio sporco in cui c' è anche un ricordo - **I-impressi.132**
3) {E} rapidly advanced, that winter is now just around the corner. The leaves on all the trees have started to change colour. Everything begins to turn yellow or red. Though it seems like a time of celebration and of the triumph of colour - **E-impressi.167**
3) {I} grandi passi e che l' inverno è ormai dietro la porta. Le foglie di tutti gli alberi hanno cominciato a mutare colore. Tutto inizia a ingiallire o a rosseggiare, Pare un momento di festa e di trionfo del colore, mentre è un - **I-impressi.162**

Figura 9. Contesti contrastivi dell'espressione "change colour"

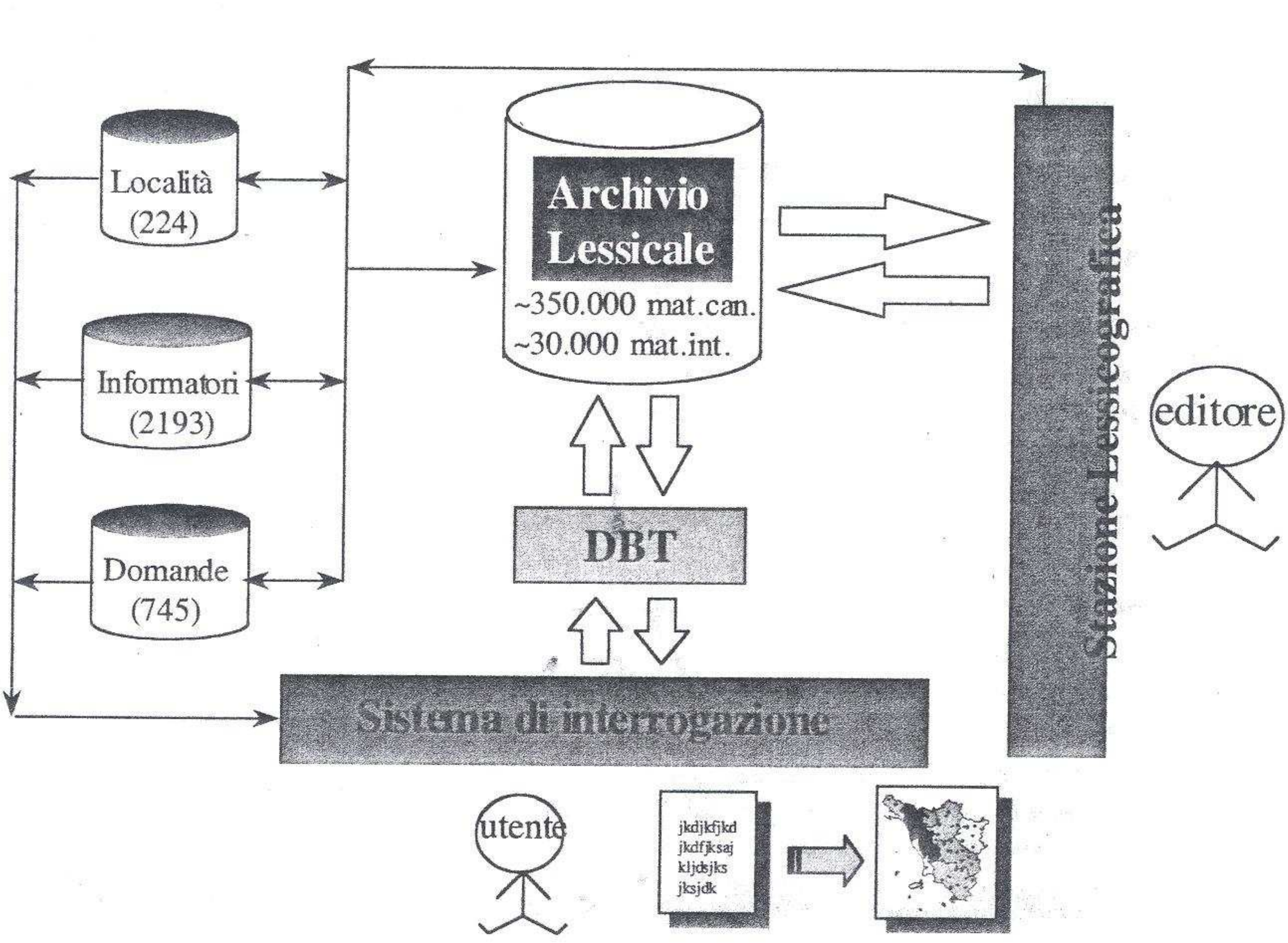


Figura 10. DBT-ALT: architettura del sistema

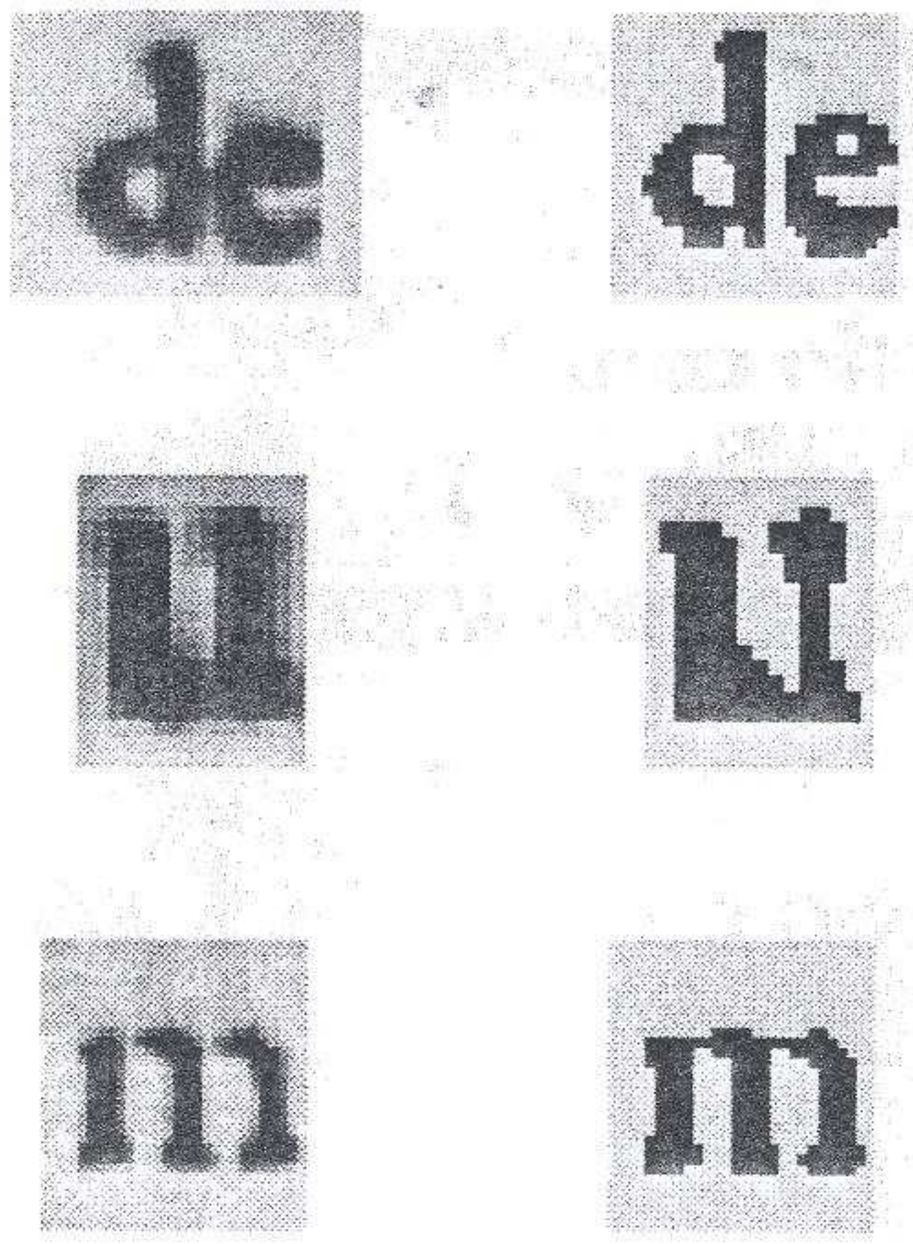


Figura 11. Esempi di restauro di caratteri

– alla sinistra, caratteri da sottoporre a restauro
– alla destra, gli stessi caratteri restaurati

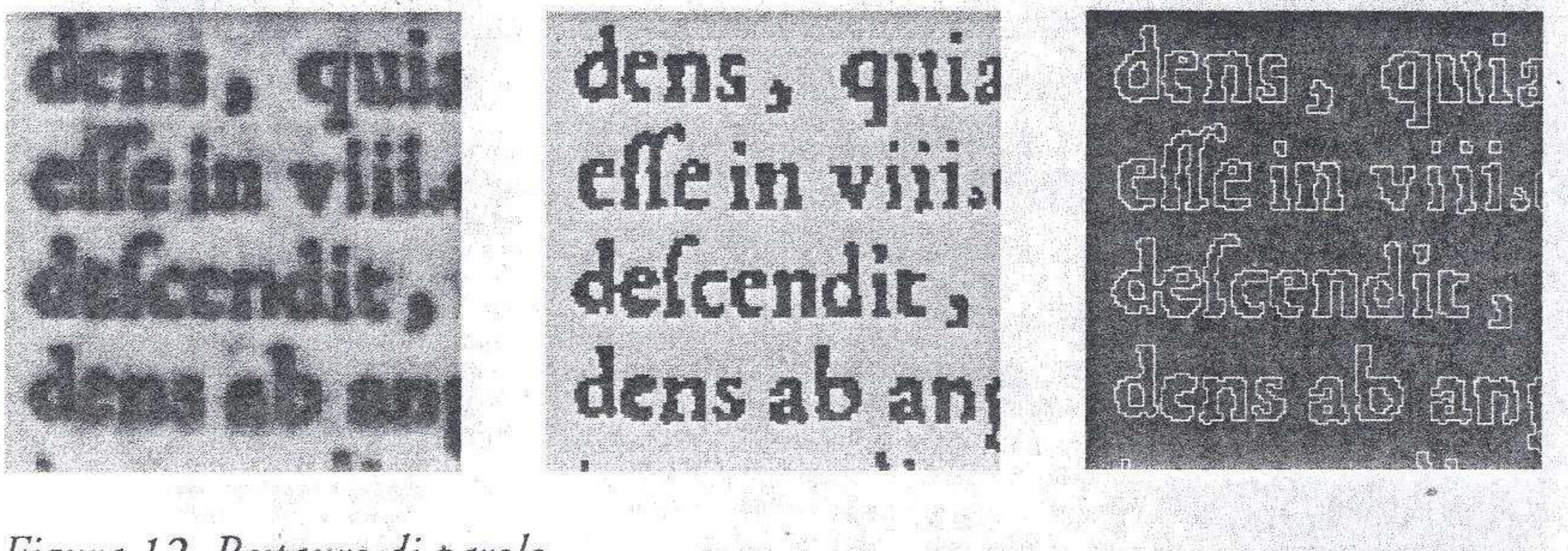


Figura 12. Restauro di parole

Figura 13. Segmentazione e riconoscimento del testo

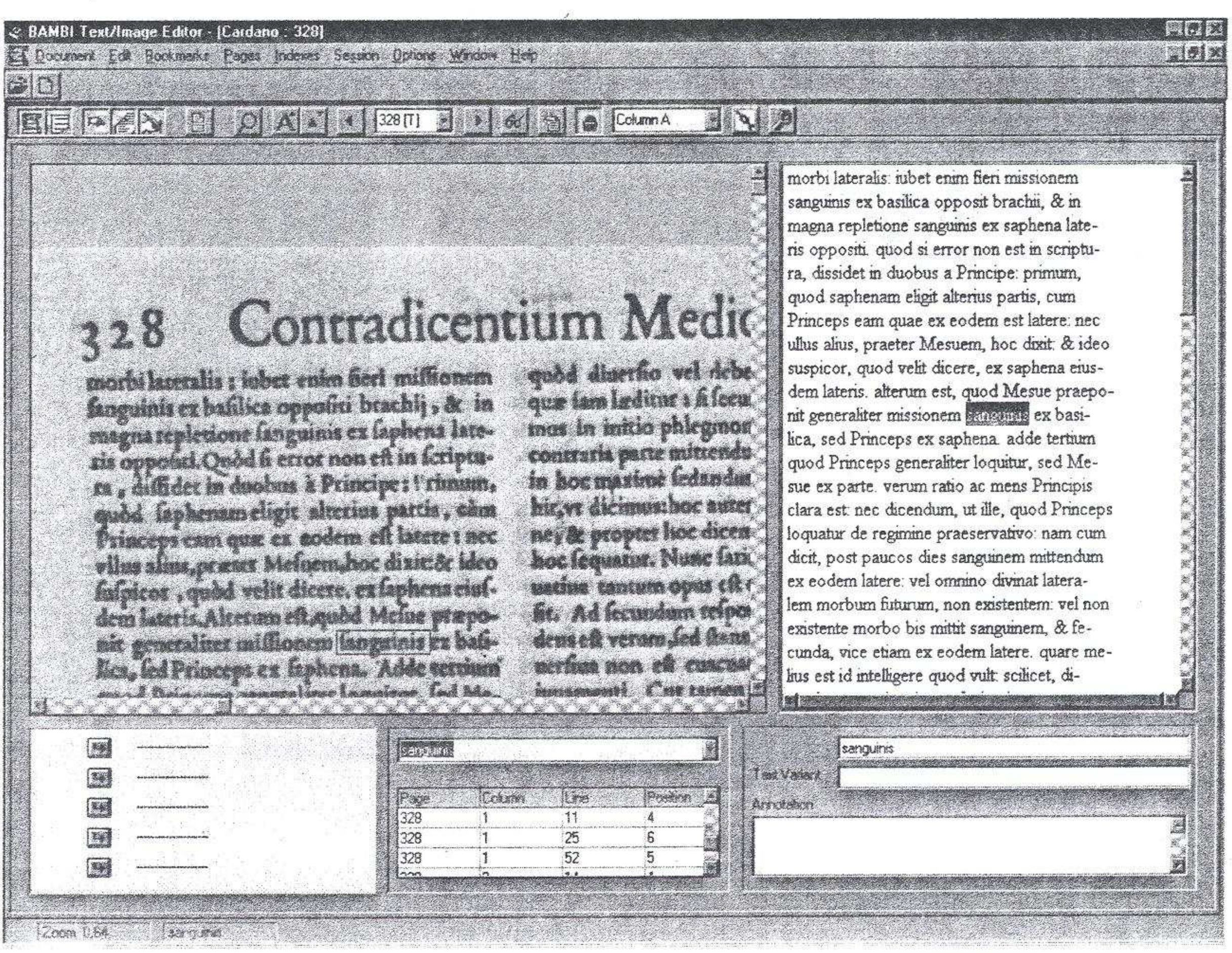


Figura 14. Corrispondenza immagine-trascrizione

a destra: la parola nella trascrizione
a sinistra: l'immagine corrispondente

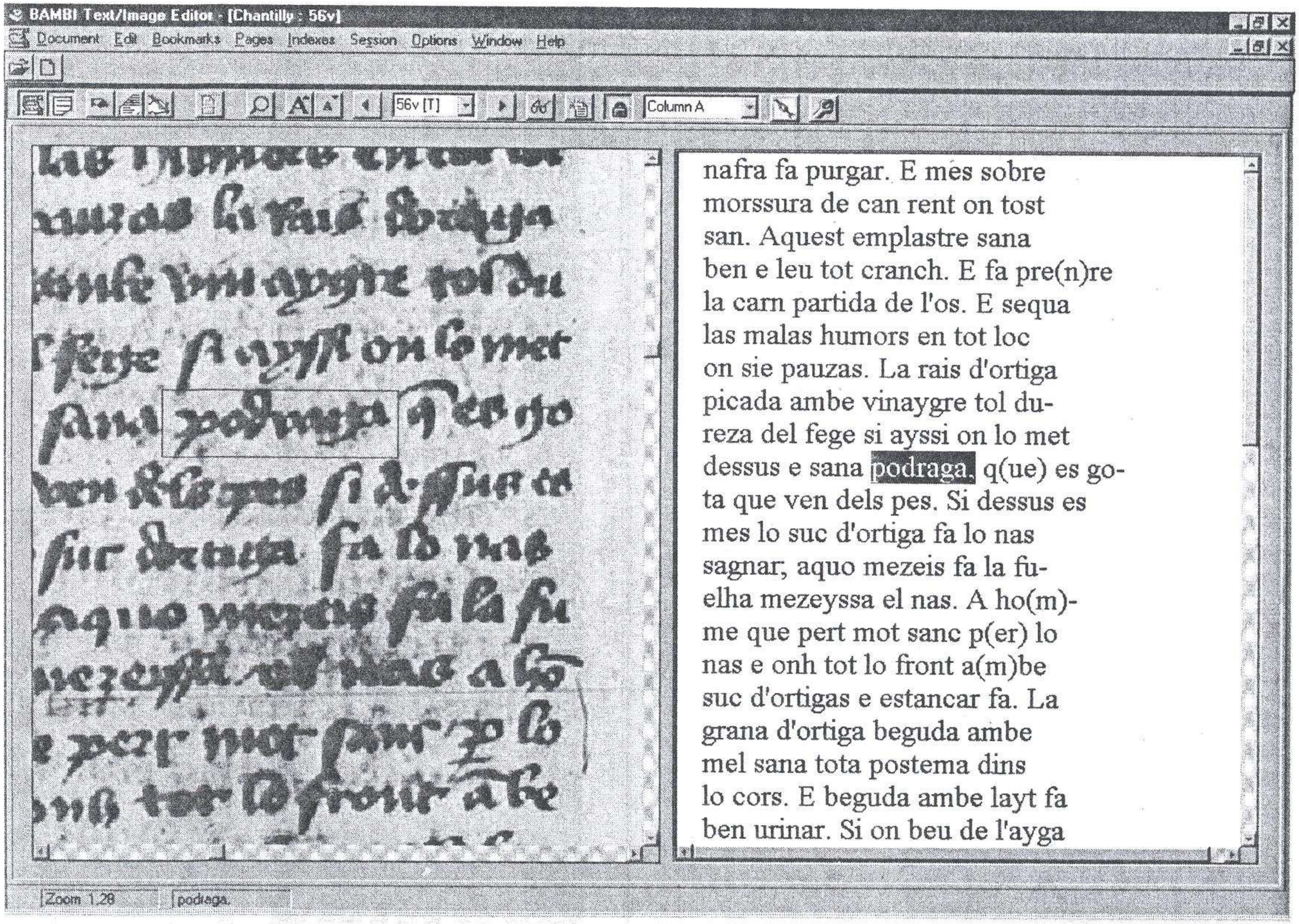
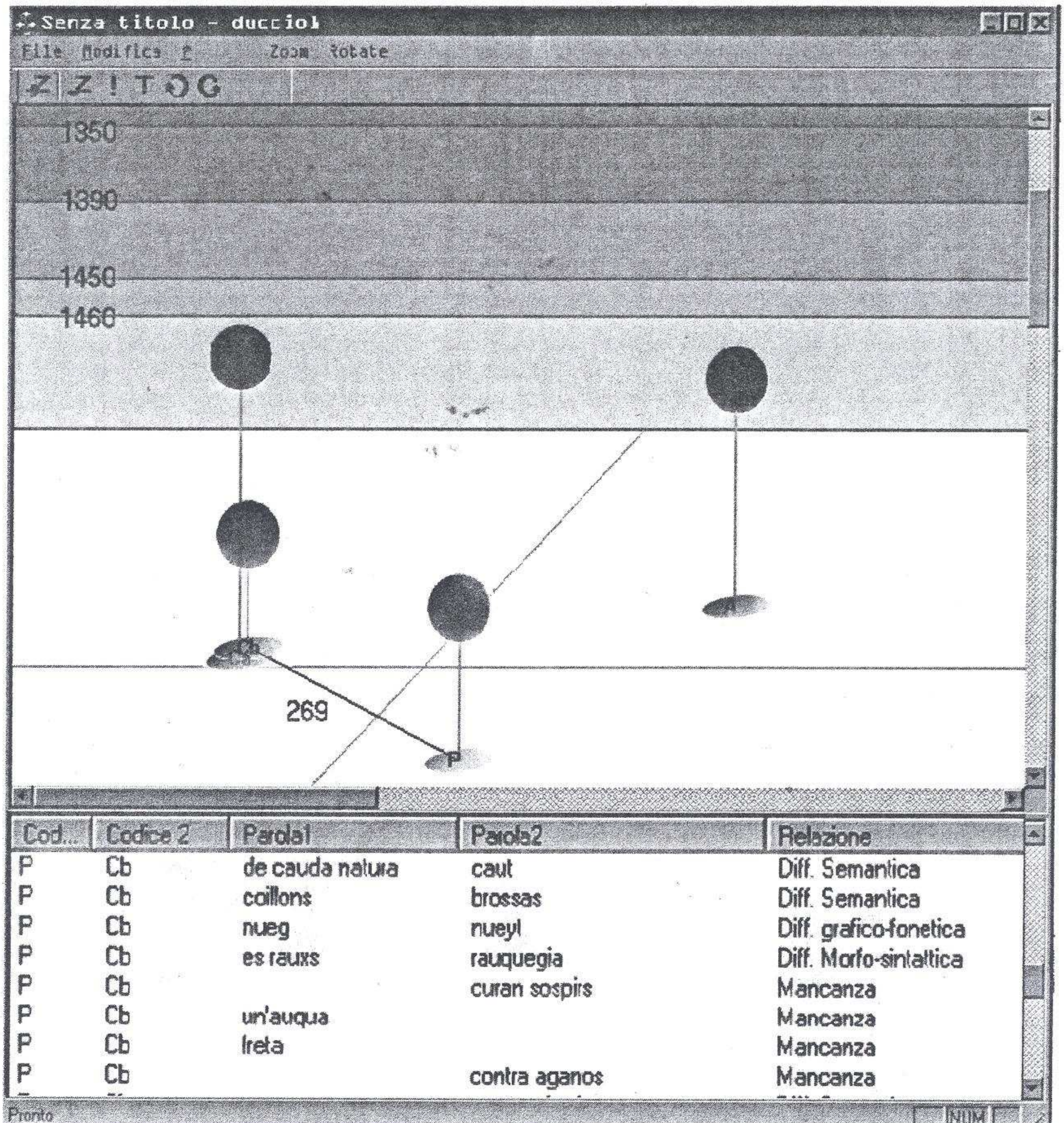


Figura 15. Corrispondenza immagine (di manoscritto)-trascrizione



Cod...	Codice 2	Parola1	Parola2	Relazione
P	Cb	de cauda natura	caut	Diff. Semantica
P	Cb	caillons	brossas	Diff. Semantica
P	Cb	nueg	nueyl	Diff. grafico-fonetica
P	Cb	es rauxs	rauquegia	Diff. Morfo-sintattica
P	Cb		curan sospirs	Mancanza
P	Cb	un'auqua		Mancanza
P	Cb	lreta		Mancanza
P	Cb		contra aganos	Mancanza

Pronto NUM

Figura 16. Visualizzazione delle relazioni tra le fonti sottoposte a collazione

(Esempio semplificato)

Forme da analizzare

AMARE

AMO

Lessico Computazionale

Parte Invariante	Codice di Paradigma Flessionale
AM-	S, M, 0
AM-	VB,1
AMAR	AG,4

Tabelle di desinenze

Codice di Paradigma Flessionale	Desinenze
S, M, 0	[0(S); I(P)
VB,1-	[0(Ind, Pres, S,1)-I(Ind, Pres, S, 2) – (….ARE (Inf, Pres)]
AG,4-	[0 (M,S); A(F,S) – I(Ms, P) – E(F, P)]

Figura 17. Analisi morfologica di Forme

Forme da analizzare	io non LA AMO più	
analizzatore morfologico	(V. figura precedente)	
analisi multiple* 1]	LA: [Art, F, S] LA: [PN, F, S]	AMO: [N, M, S] AMO: [VB, Ind, Pres, S,
TAGGERS		
a) regole di coocorrenza	1) – ART + VB	NO SI (solo se genere e numero uguali)
	2) – ART + N	SI NO
	3) – PN + VB	
	4) – PN + N	
b) statistico (probabilità di sequenze di categ. Morfologiche derivante da frequenze in un 'training corpus')		
Analisi proposta automaticamente	- LA: - AMO:	[PN, F S] [VB, I, P, S, 1]
Eventuale controllo manuale		
Analisi Finale	V. Sopra	

*Trascuriamo, nell'esempio, l'analisi LA: sost.

Figura 18. Procedura di Disambiguazione Morfosintattica ("Tagging")

- FRASE DA ANALIZZARE Lo zio mangia la pasta
- uscita dalla consultazione del lessico computazionale

LO	: - ART, M, S	LA	: - ART, F, S
	: - PN, M, S		: - PN, F, S
ZIO	: - N, M, S	PASTA:	- N, F, S
MANGIA	: - VB, Ind, Pres, S, 3		

- REGOLE: ART (=) + N(=) → GN VB + GN → GV

  PN+VB → GV GN+GV → F

- ANALISI:

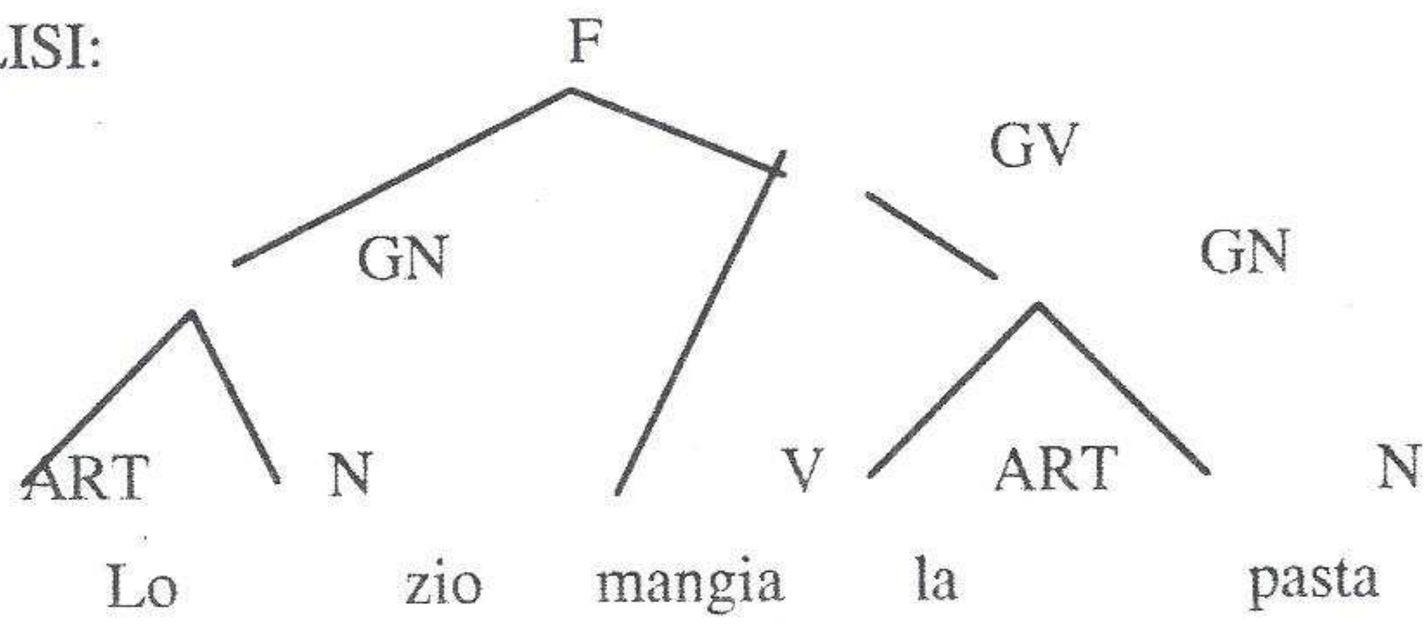


Figura 19. Frammento di grammatica e analisi di una frase

•FRASI DA ANALIZZARE: Lo zio mangia la pasta con il pomodoro

Lo zio mangia la pasta con la forchetta

•uscite aggiuntive della consultazione con := PZ
la :- ART,F,S

dal lessico:

il := ART,M,S la :-
PN,F,S

pomodoro := N,M,S

forchetta :- N,F,S

•regole aggiuntive: PZ + GN = GPZ

GN+V+GN+GPZ = F

GN + GPZ = GN

•ANALISI

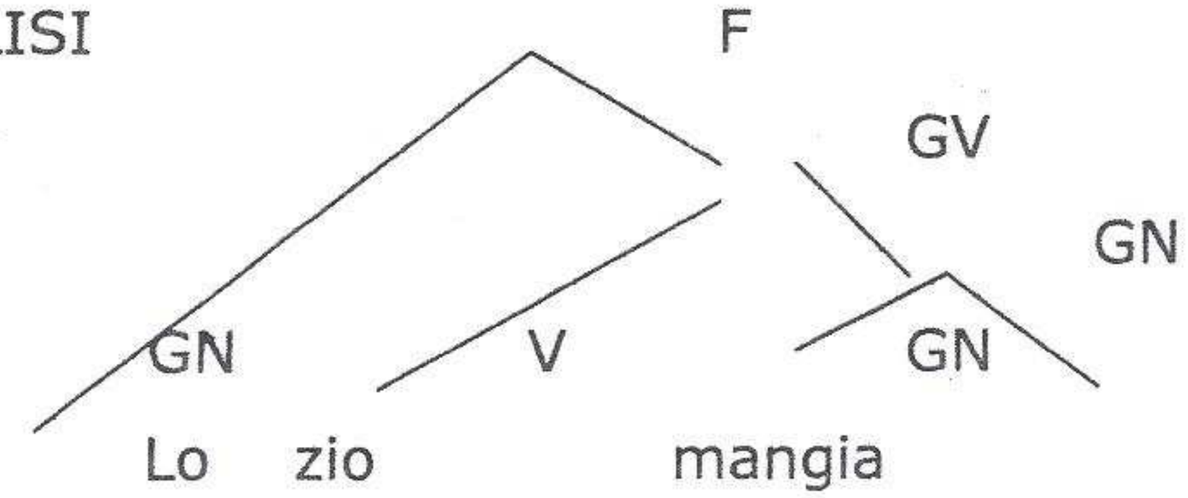


Figura 20. "Attachment" di un gruppo preposizionale

Analisi sintattica completa
a base lessicale conforme allo standard di EAGLES

```
[S
[NP-subj Antonio NP-subj]
[VERB vede VERB]
[NP-obj le ragazze
        [PP nel parco PP] NP-obj]
[PP col telescopio PP]
. S]
```

CHUNKING: analisi sintattica sottospecificata a costituenti non ricorsivi

[N_C Antonio]
[FV_C vede]
[N_C le ragazze]
[P_C nel parco]
[P_C col telescopio]
[PUNC_C .]

Figura 21.

ENTRATA SINTATTICA
FORMATO MACRO

SynU: ***chiarire***
Description :

Example : *che tu faccia ciò/sentire questo/questo fatto mi chiarisce un dubbio/che volevo partire/di aver sbagliato*

[Construction :

Syntlabel: **Clause**
P0 [opt:**yes**]: [function: **subject**]
[cat: **np**]
[cat: **cl**] [synsubcat: **infcl**] [introd: **0**]
[cat: **cl**] [synsubcat: **thatcl**] [mood:**sub**]
P1 [opt:**no**]: [function: **object**]
[cat: **np**]
[cat: **cl**] [synsubcat :**thatcl**] [mood: **ind**]
[cat: **cl**] [synsubcat: **infcl**][introd: **di**]
P2 [opt:**yes**]: [function: **indirectobject**]
[cat: **pp**] [introd: **a**]]

[Self: [Intervconst: **V**[func: **head**][morphsubcat: **main**] [aux:**avere**]]
]]]

Figura 22. Lessico di "PAROLE"

ENTRATA SEMANTICA

'binocolo'

naming = " **binocolo** "
freedefinition = "Strumento costituito da due cannocchiali gemelli,
osservare oggetti lontani"
weightvalsemfeaturel =
"Semantic class: INSTRUMENT
Domain: OPTICS
Template INSTRUMENT
UnificationPath Concreteentity-ArtifactAgentive-Telic"
Relations:
semr = " Isa"
weight = " PROTOTYPICAL"
target = " **strumento** " (Artifact)

semr = " Createdby"
weight = " PROTOTYPICAL"
target = " **fabbricare** " (Physical_Creation)

semr = " Usedfor"
weight = " PROTOTYPICAL"
target = " **osservare** " (Perception)

Figura 23. Lessico di "SIMPLE"

Figura 24. Iponimi di "tessuto" in EUROWORDNET italiano

Un esempio: MATE

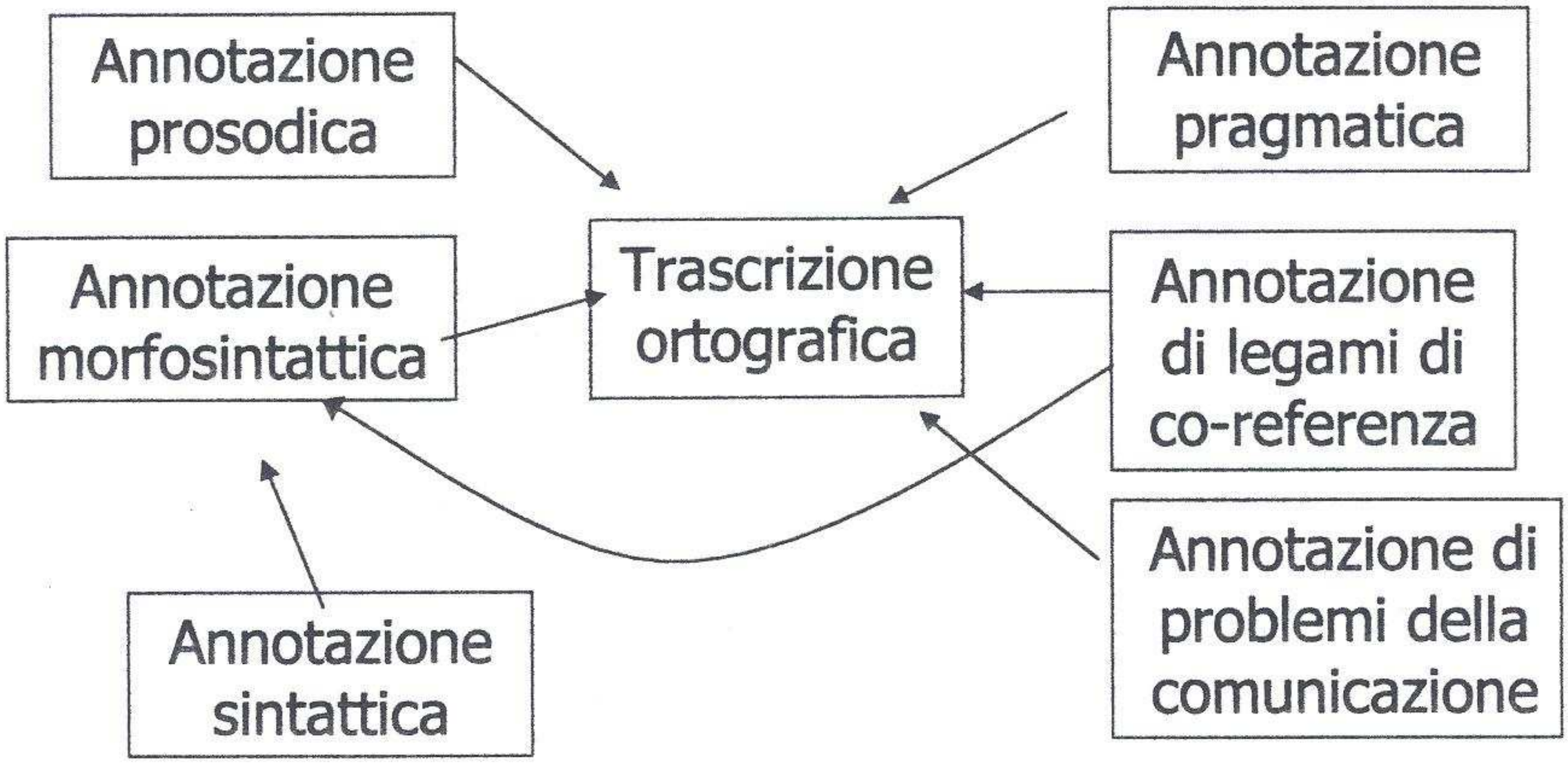
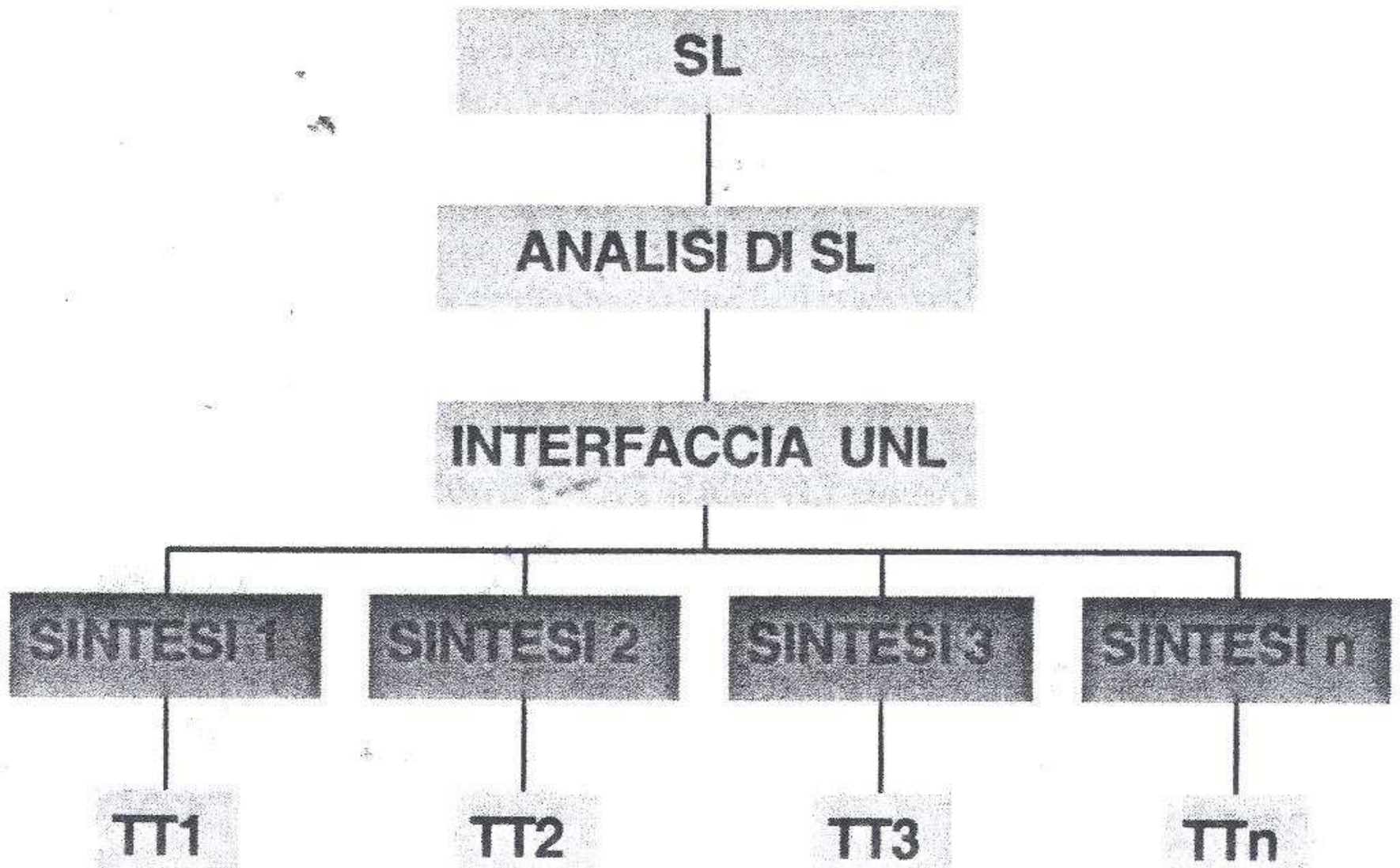


Figura 25. Relazioni tra livelli di dialoghi

Universal Language: un data base di concetti
una lista di relazioni tra concetti



Lingue correnti: italiano, spagnolo, portoghese, francese, russo, lituano, cinese, mongolo, giapponese, arabo, swaili, indu, indonesiano

SL: Lingua sorgente; TT: lingua target, di arrivo.

Figura 26. Esempio UNL (universal language)