Valorizzazione del patrimonio letterario della lingua italiana. Il corpus italiano e la legge 488.

M. Sassi, S. Cucurullo, P. Picchi

ILC-CNR

Pisa, 2006

Indice

Premessa

L'Europa occidentale è entrata, con gli Stati Uniti e l'Estremo Oriente, nella "Civiltà dell'informazione". Nella Unione Europea oltre due terzi della mano d'opera attiva lavorano, sui simboli dell'informazione relativa ad oggetti e a servizi con i quali non hanno contatti fisici.

Il personale che, nelle amministrazioni, nell'industria, nel commercio, utilizza strumenti informativi per il trattamento dei testi è oggi più numeroso della mano d'opera complessivamente impiegata nell'agricoltura. Le operazioni sulle informazioni sono divenute un elemento centrale dell'efficienza dei sistemi per la creazione e la distribuzione della ricchezza. Inoltre la qualità e la quantità del trasferimento dell'informazione e delle conoscenze sono cruciali nell'attuale competizione economica e industriale.

Le nuove tecnologie delle telecomunicazioni moltiplicano enormemente la capacità di trasmettere l'informazione nella sua forma parlata e scritta, e stanno trasformando il nostro pianeta in un vero e proprio villaggio globale, nel quale una vastissima comunità di utenti può entrare in contatto reciproco trascendendo barriere geografiche e politiche. Il paradigma della "Società dell'informazione" sta cambiando il modo nel quale si fanno gli affari, si esercitano i commerci e le professioni, si operano i servizi pubblici, si riceve l'educazione, si diffonde la conoscenza, si svolge il lavoro quotidiano.

La quantità di informazioni disponibili cresce costantemente. Pertanto cresce parallelamente la necessità di strumenti che consentano di operare su di esse in modo automatico, efficace, economico.

Una grande massa di informazioni e potenti connessioni telematiche non forniscono di per sé servizi utilizzabili e significativi. É necessario fornire strumenti che rendano semplici e naturali l'accesso e l'utilizzo dell'informazione. Poiché l'informazione è veicolata dalle lingue naturali, ed oggi risiede essenzialmente su supporti elettronici, è urgente poter disporre di strumenti capaci di automatizzare - almeno in parte - le operazioni linguistiche che devono essere compiute per produrre, trasmettere, archiviare, ricuperare, accedere, elaborare l'informazione. É compito delle ricerche e delle applicazioni nel settore interdisciplinare che si occupa del Trattamento Automatico del Linguaggio (TAL) fornire questi strumenti. Strettamente connesso all'obiettivo di un accesso democratico all'informazione, è il problema del multilinguismo.

Da un lato, la globalizzazione dell'economia e i nuovi servizi telematici che attraversano le frontiere pongono il problema delle interfacce tra lingue nella comunità internazionale.

Dall'altro, i profondi mutamenti e la tecnologia pervasiva crescente dei sistemi di comunicazione e informazione potrebbero avere notevoli ripercussioni sulle lingue che utilizziamo. Nel peggiore dei casi, i cittadini che non sono in grado di comunicare agevolmente nelle lingue principali in uso, potrebbero vedersi negata la piena partecipazione ad una società sempre più basata sulla informatizzazione delle risorse.

Fonti autorevoli hanno avvertito che le lingue per le quali non vengono sviluppati strumenti adeguati di trattamento automatico, rischiano di perdere gradualmente il proprio posto nella società globale dell'informazione, assieme alle culture che esse veicolano, con grave danno per uno dei patrimoni più preziosi: la diversità culturale. Per scongiurare tali pericoli è necessario garantire il supporto per l'uso dell'informazione multilingue, come è stato precisato in un vertice del G7 (Il Consiglio europeo, nella sessione di Corfù del 24 e 25 giugno 1994, ha sottolineato l'importanza degli aspetti culturali e linguistici della società dell'informazione e, nella sessione di Cannes del 26 e 27 giugno 1995, ha ribadito l'importanza per l'Unione europea della sua diversità linguistica; la Conferenza dei Ministri del G7 tenutasi a Bruxelles il 25 e 26 febbraio 1995 ha richiamato l'attenzione

sull'importanza della diversità linguistica e culturale nella società dell'informazione globale).

Da un punto di vista generale, la tecnologia della lingua apporterà un valore aggiunto significativo alle tecnologie informative e comunicative che, a loro volta, apporteranno novità importanti nell'ambito della organizzazione sociale, proprio in virtù del cambiamento del concetto di comunicazione e di interazione. La tecnologia della informazione e della comunicazione (ICT) favorisce lo sviluppo della industria dei contenuti, che produce prodotti immateriali destinati a soddisfare la domanda di informazione, formazione, cultura ed intrattenimento.

Molti siti hanno riconosciuto che è necessario un supporto nazionale all'informatizzazione del trattamento delle rispettive lingue nazionali, per proteggere la competitività delle industrie e i diritti dei cittadini. Essi sono così giunti alla conclusione che, tra le diverse infrastrutture immateriali e materiali indispensabili per uno spazio economico, è necessaria anche un'infrastruttura linguistica costituita essenzialmente da Risorse Linguistiche (RL) adeguate, mono e multilingui, per la lingua nazionale.

Gli operatori industriali del nostro paese sono condizionati, in particolare nella competizione internazionale, dalla mancanza di RL adeguate per la nostra lingua ed il programma si propone di soddisfare un bisogno primario e diffuso presso industrie che operano nei settori più diversi.

Senza queste risorse, i passi necessari per la costruzione delle applicazioni restano un'impresa estremamente dispendiosa e comunque non alla portata della maggior parte dei soggetti economicamente interessati.

La creazione di RL di questo tipo richiede la convergenza e la collaborazione di competenze specialistiche, complementari, a livello scientifico, tecnico, industriale, applicativo. Di norma queste competenze sono distribuite tra soggetti diversi, generate in comunità tradizionalmente distinte e spesso separate tra loro.

Il costo di creazione è ingente, ed è necessario assicurare il riutilizzo

di RL parziali esistenti, attraverso apposite strutture organizzative, e la loro cumulabilità attraverso la adozione di specifiche comuni.

Per la mancanza di queste RL molti operatori, in particolare PMI, pur dotati dei requisiti imprenditoriali necessari, si sono ritirati dal settore, o hanno rinunciato ad impegnarsi nello sviluppo di prodotti, anche in presenza di opportunità e condizioni molto favorevoli di mercato.

Scopo del progetto

Lo scopo del progetto è stato quello di costituire una serie di RL - dati, conoscenze, componenti software - di riferimento. Esso mira a rendere possibile, efficace e tempestiva la risposta dell'industria, della ricerca, dei servizi italiani alle sfide poste dalla Società dell'Informazione, nel contesto europeo e globale.

I risultati del progetto saranno armonizzati con gli standards emergenti a livello europeo e internazionale, così da poter essere immediatamente utilizzati nelle diverse applicazioni industriali del TAL. In tal modo si possono abbassare sostanzialmente i costi e i tempi di produzione, facendo quindi aumentare la competitività delle nostre industrie sul piano internazionale e la capacità di fornire servizi efficaci ai cittadini.

Questa breve relazione si propone di illustrare la struttura del lavoro ed i risultati ottenuti, relativamente alla creazione di corpora e di strumenti software realizzati con i finanziamenti della legge 488.

Le linee di lavoro seguite sono state essenzialmente due:

1. ricerca di nuovi metodi per il trattamento automatico dell'italiano sia scritto che parlato.
2. sviluppo di strumenti e risorse per l'inserimento dell'italiano in applicazioni multilingue, con particolare riguardo alle relazioni tra lingua e cultura italiana e mondo arabo.

In particolare si illustreranno i progetti, le risorse e gli strumenti sviluppati dall'Istituto di Linguistica Computazionale, che sono stati adattati anche al trattamento automatico della lingua araba.

Descrizione generale del lavoro

In conseguenza del rapido sviluppo della società dell'informazione, diventa sempre più importante disporre di risorse linguistiche capaci di facilitare la comunicazione tra culture diverse. La possibilità di disporre di queste risorse per la lingua italiana ci permetterà di essere competitivi con i paesi tecnologicamente più avanzati anche dal punto di vista della competitività delle nostre imprese sui mercati internazionali.

Le Risorse Linguistiche (RL) necessarie sono costituite da insiemi estesi di dati linguistici accompagnati o costituiti da annotazioni o rappresentazioni articolate su diversi livelli di descrizione linguistica.

Le Risorse Linguistiche

Già dalla fine degli anni '80 è stato riconosciuto che le RL sono premessa essenziale a qualsiasi ricerca nel settore del TAL.

Esse sono fondamentalmente costituite da:

- corpora di lingua parlata e scritta,
- lessici computazionali, raccolte terminologiche,
- grammatiche computazionali,
- contenuti e copertura linguistica costruiti e strutturati per l'uso nel TAL,
- metodi e componenti software di base che assicurino le funzioni fondamentali per i sistemi TAL,
- interfacce e ambienti di sviluppo per l'integrazione dei componenti predetti nelle diverse applicazioni industriali.

Per la realizzazione di queste Risorse, si è sfruttata la presenza di altre risorse parziali già disponibili, armonizzandole con ciò che si stava costruendo per le altre lingue europee e cooperando con i potenziali utilizzatori.

Strumenti per il Trattamento Automatico del Linguaggio

Le applicazioni per il TAL debbono possedere caratteristiche di robustezza tali da poter gestire la varietà di uso reale delle lingue. Strutturalmente sono costituiti da regole associate a metodi statistici e hanno una copertura lessicale e terminologica quantitativamente e qualitativamente adeguata. La copertura lessicale è costituita tipicamente da decine di migliaia di unità lessicali, corredate dalle opportune informazioni linguistiche necessarie per le elaborazioni dei dati linguistici.

Nell' ambito del Progetto 8 "**Diffusione della cultura e valorizzazione del patrimonio letterario della lingua italiana e della lingua araba attraverso una diffusione telematica di banche di dati letterarie**" sono state approntate le Risorse Linguistiche descritte nelle pagine che seguono.

Per lo sviluppo del progetto sono stati raccolti dei corpus testuali sia in lingua italiana che araba. Qui di seguito si riportano i criteri di formazione dei corpora che sono stati usati per la creazione degli strumenti software stand-alone ed on-line utilizzabili dalla sezione Software.

Il corpus dei testi italiani è stato costruito dall'Istituto di Linguistica Computazionale del CNR di Pisa. Il corpus dei testi arabi, è stato sviluppato dall'Istituto di Studi Orientali dell'Università di Napoli.

In queste note verrà descritto in particolare il corpus in italiano.

Nell'ambito dell'Obiettivo 7: "**Promozione degli scambi linguistici e culturali con il mondo arabo**" ci si propone di studiare e predisporre metodi e sistemi per la diffusione dei risultati del progetto e per la promozione e la valorizzazione della lingua e cultura italiana ed araba attraverso strumenti telematici.

Ci si propone in particolare di disegnare:

1. una biblioteca elettronica di testi letterari particolarmente rilevanti per la diffusione della cultura italiana nel mondo arabo;

2. Un sistema intelligente per l'accesso ai testi, ai lessici ed agli strumenti prodotti dai precedenti obiettivi, su rete telematica per utenti di lingua italiana e di lingua araba.

Nell'ambito del Progetto 6: "**Corpus bilingue italiano-arabo**" sono stati acquisiti e codificati i testi italiani in forma leggibile dal calcolatore.

Si è creato un corpus per mezzo di un software che convertisse i testi già disponibili su supporto elettronico dal formato originale a quello di progetto. Altri testi sono stati scelti ed elaborati tramite scanner con un buon OCR, seguito da una buona revisione. I dati, così acquisiti e codificati, hanno permesso la formazione di un corpus italiano di circa 4.000.000 di parole per un corpus parallelo con l'arabo e altre 5.000.000 di parole per la costituzione di un corpus più generico.

I testi in italiano

I testi italiani del corpus, si collocano in un periodo cronologico abbastanza limitato, che comprende l'Ottocento e che arriva fino ai primi decenni del Novecento, con qualche incursione nella letteratura contemporanea tramite una breve antologia. E' stato deciso di porre questi paletti cronologici poiché abbiamo ritenuto culturalmente più opportuno diffondere testi abbastanza recenti, validi dal punto di vista letterario e soprattutto tali da proporre attraverso la letteratura varie sfaccettature, peculiari in gran parte della società italiana contemporanea. Sono state in particolare privilegiate:

- le autrici,
- gli autori di letteratura per l'infanzia,
- le opere minori di scrittori illustri,
- gli scritti di personalità famose non per meriti letterari.
- Sono state inoltre inserite, rispettando i vincoli del diritto d'autore,
- un'antologia rappresentativa di poeti contemporanei,
- un'antologia di testi di scrittori contemporanei.

Le autrici

Da un'analisi dei siti dedicati alla letteratura italiana è stato notato che la diffusione di opere di scrittrici è piuttosto limitata. Una ricerca bibliografica attenta a questo aspetto della vita culturale italiana ha messo in evidenza personalità di scrittrici che hanno avuto un ruolo culturale importante nella società italiana dell'Ottocento e del Novecento. Grazia Deledda, Matilde Serao si possono a buon diritto affiancare Emma Perodi, Elisa Cappelli, Luisa Saredo, Flavia Steno. Abbiamo potuto inserire anche alcuni testi di una scrittrice aquilana, Laudomia Bonanni, grazie all'autorizzazione concessa al nostro Istituto dall'erede dei diritti d'autore.

Gli autori di letteratura per l'infanzia

Poca conosciuta è altresì la produzione letteraria dedicata all'infanzia anche di scrittori quali Massimo D'Azeglio, Guido Gozzano, Luigi Capuana, che proponiamo insieme alle opere di Luigi Bertelli, alias Vamba, e di Carlo Lorenzini, alias Carlo Collodi. Cospicua è poi la produzione da parte di scrittrici: Emma Perodi, ad esempio, nata a Cerreto Guidi nel 1850, ha pubblicato volumi di racconti e romanzi per l'infanzia. Sotto la sua direzione, Il Giornale dei Bambini, fu letto da grandi e piccini per la prima volta il Pinocchio di Collodi apparve sulla rivista con il titolo "Storia di un burattino".

Le opere minori di scrittori illustri

Federico De Roberto, Emilio De Marchi, Luigi Capuana, Antonio Fogazzaro sono scrittori di opere già ampiamente presenti nelle biblioteche elettroniche disponibili via Internet. Nel nostro data base abbiamo privilegiato le loro opere minori, che sono spesso interessanti come documenti del costume, della società del tempo.

Gli scritti di personalità famose non per meriti letterari

Una sezione del database è stata dedicata a scritti di personalità che hanno contribuito alla nascita e all'evolversi della società italiana nell'Ottocento. Per questo motivo sono in linea pagine di Giuseppe Mazzini, Giuseppe Garibaldi, Carlo Cattaneo, Paolo Mantegazza, di

Pellegrino Artusi la cui opera *La scienza in cucina e l'arte di mangiar bene* è tutt'oggi il più importante trattato per conoscere la cucina tradizionale italiana anche attraverso aneddoti e divagazioni.

Un'antologia rappresentativa di poeti contemporanei

Carlo Michelstaedter, Guido Gozzano, Dino Campana, Emilio Praga, Sergio Corazzini, Cesare Pavese, Arturo Onofri.

Un'antologia di testi di scrittori contemporanei

Dacia Maraini, Andrea Camilleri, Niccolò Ammaniti, Antonio Tabucchi.

Creazione del sottocorpus italiano del corpus bilingue.

I testi che estendono il componente (sottocorpus) italiano del corpus bilingue sono stati scelti la rappresentatività di usi linguistici che caratterizzano le attività socio-economiche che nella odierna società dell'informazione si servono dei prodotti e dei servizi dell'ingegneria linguistica.

I testi scelti sono stati acquisisti con le tecnologie opportune in base al tipo di supporto e codificati secondo gli standards definiti nel progetto 6, standards compatibili con quelli emergenti in ambito internazionale (TEI, CES PAROLE/EAGLES). Si sta infatti imponendo all'attenzione degli operatori del settore la necessità di utilizzare i network globali di informazione per mettere a disposizione dei cittadini contenuti di rilevanza culturale fruibili con l'aiuto di strumenti "intelligenti". Si attuerà così una grande biblioteca elettronica distribuita nei vari paesi della UE, consultabile dallo utente finale con metodi e strumenti comuni. I testi prodotti saranno controllati manualmente per verificare la correttezza formale della codifica e la corrispondenza al testo sorgente.

L'ingegneria linguistica è chiamata a contribuire attraverso la creazione di interfacce "user-friendly" e di basi di conoscenza che fungono da riferimento per potenziare le capacità di interazione dell'utente finale. Con la realizzazione dell'obiettivo, il progetto si

propone di collocare i risultati del piano nella rete globale di contenuti culturali "attivi" che costituisce - secondo le previsioni di numerosi esperti - una delle maggiori prospettive di sviluppo e crescita dei networks globali di informazione.

Corpora testuali di grandi dimensioni, opportunamente articolati in sottoinsiemi tipologici, orientati alle applicazioni dell'ingegneria linguistica, sono di per sé una risorsa di indiscutibile utilità. Questa utilità è accresciuta notevolmente dalla annotazione linguistica, cioè dalla categorizzazione delle unità linguistiche a un determinato livello di descrizione linguistica.

L'annotazione consiste nel segmentare il testo nelle unità costitutive del livello prescelto, e nell'associare a ciascuna di esse le categorie che la classificano nel sistema di analisi linguistica adottato. Sistemi d'annotazione standard, consentono di riutilizzare software d'elaborazione comuni e di scambiare, confrontare, cumulare i risultati ottenuti in gruppi di ricerca diversi su corpora diversi.

L'annotazione linguistica può essere utilizzata sia per arricchire e rendere più efficace l'accesso dell'utente finale ai testi (per esempio, per individuare e/o ricercare, schemi grammaticali, per selezionare usi specifici di forme omografe), sia come base per lo sviluppo di metodi statistici fondati sulle frequenze di uso delle categorie linguistiche e delle loro sequenze. Metodi statistici di questo tipo trovano applicazioni sempre più numerose ed efficaci sia sul versante dello scritto che su quello parlato, in compiti che vanno dalla acquisizione (semi)automatica di conoscenze, alla estrazione e recupero d'informazioni, alla comprensione del parlato, alla "summarization", al "data mining", ecc.

Il sistema PiSystem e Internet

Il PiSystem è un sistema integrato per il trattamento di materiali testuali e lessicali che trae origine dal famoso DBT, Brevetto CNR del dr. Eugenio Picchi, dirigente di ricerca dell'ILC-CNR di Pisa (http://www.ilc.cnr.it/pisystem/index.html).

Lo scopo del software per l'accesso remoto ai testi è quello di

facilitare e rendere più efficace la interazione dell'utente finale con le risorse testuali e lessicali memorizzate. A tal fine, è stata posta particolare attenzione allo sviluppo di opportune funzioni di ricerca basandosi sulle funzioni e le interfacce che costituiscono il PiSystem. Un sistema di ricerca e consultazione interattiva che è utilizzato da centinaia di utenti dell'ILC e di basi testuali da questi prodotti.

Il PiSystem, che era disponibile solo in versione locale, è stato trascritto, con tecnologie e linguaggi multipiattaforma (indipendenti dall'ambiente di sviluppo) in modo da renderlo accessibile con qualsiasi "browser" disponibile in Internet.

La banca-dati descritta in questo lavoro è raggiungibile tramite il seguente link, che si apre con qualsiasi browser e si presenta con la maschera successiva:

http://serverdbt/Sito488/startDBT_ITA.html

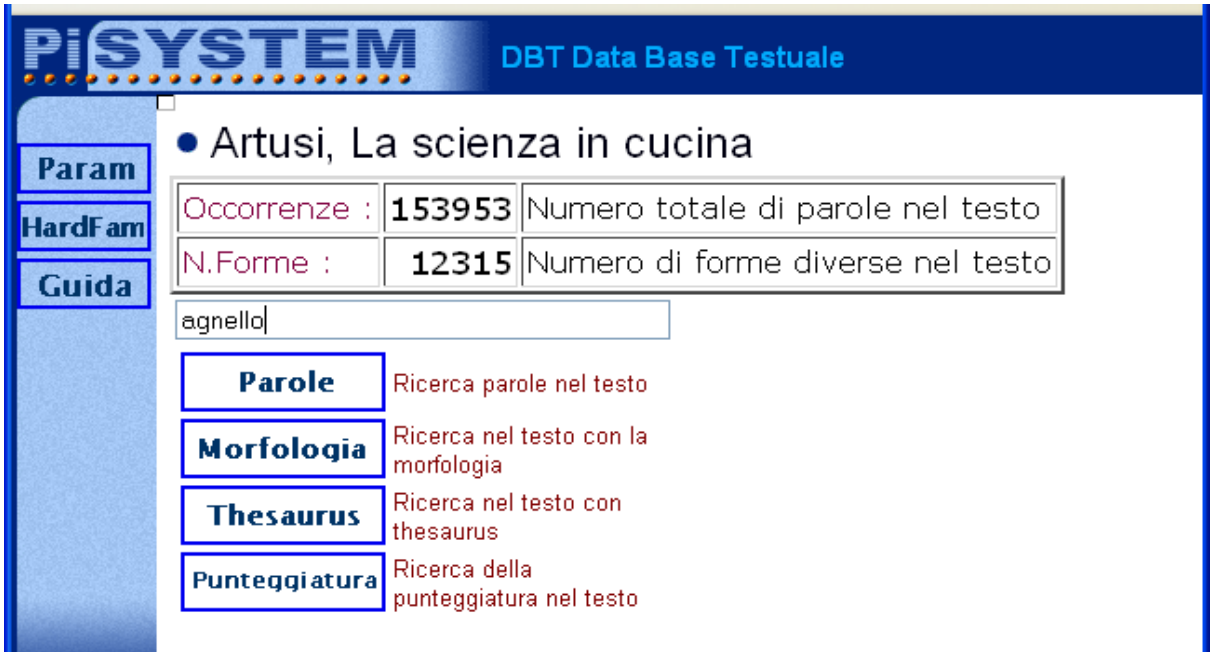
Descrizione dell'interfaccia

Nelle pagine seguenti vengono descritte le funzionalità del sistema di interrogazione via web, con una esemplificazione pratica delle ricerche che si possono effettuare.

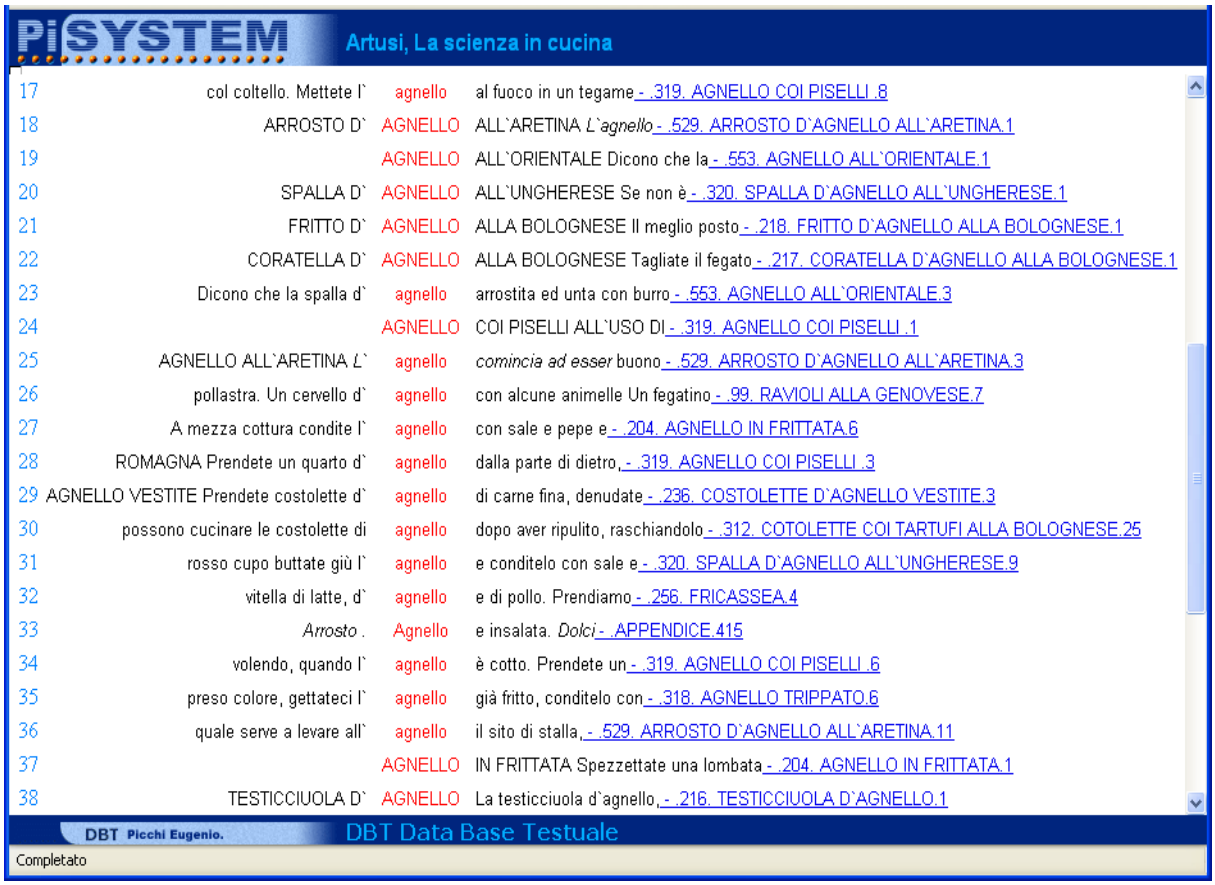
Sono possibili due tipi di consultazioni: per ogni singolo testo oppure globalmente per il corpus completo.

In entrambi i casi l'interfaccia è molto simile; nel caso del corpus, appariranno anche i riepiloghi per ogni testo.

Ad esempio si può procedere alla scelta del testo singolo, in questo caso l'Artusi, e alla prima richiesta, dopo aver digitato la parola nello spazio apposito, si può usare il pulsante **Parole**, o **Morfologia** o **Thesaurus**.

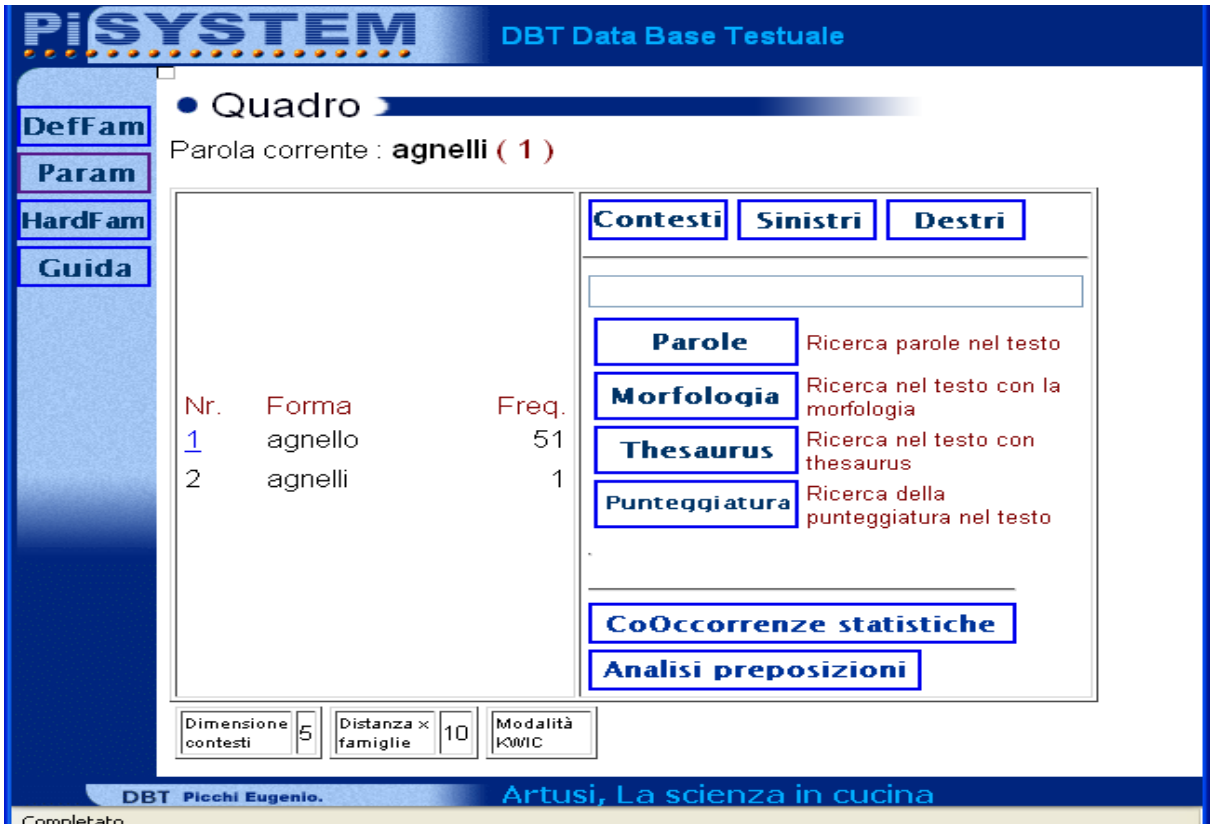


Nel primo caso si otterrà il risultato seguente, in cui sono evidenziati i contesti della parola *agnello*, ordinati alfabeticamente, secondo la parte destra del testo che segue la parola richiesta.

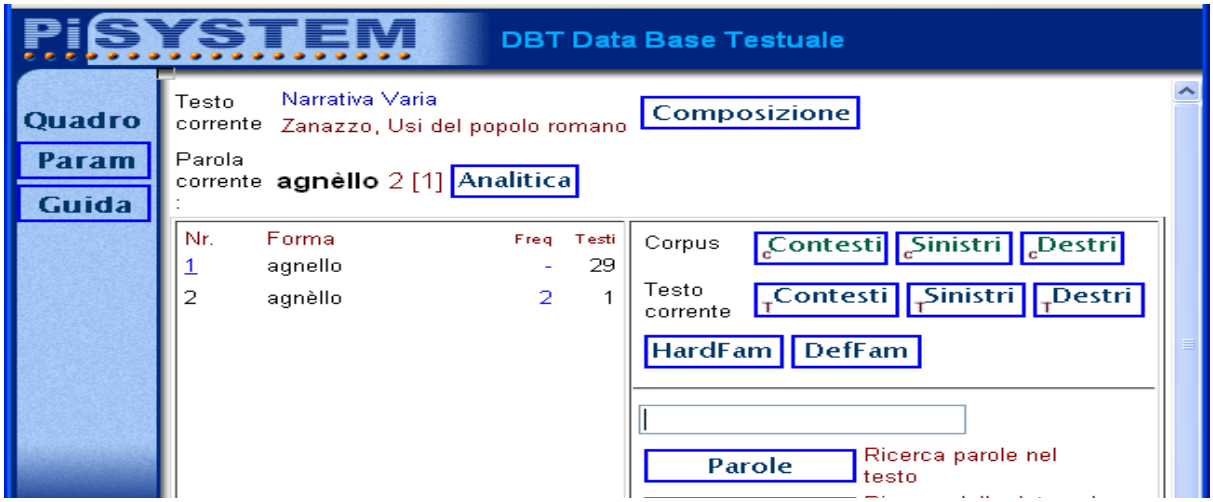


Ogni riga di contesto breve fornisce il link di accesso al contesto largo, che riproduce una porzione di testo più ampia.

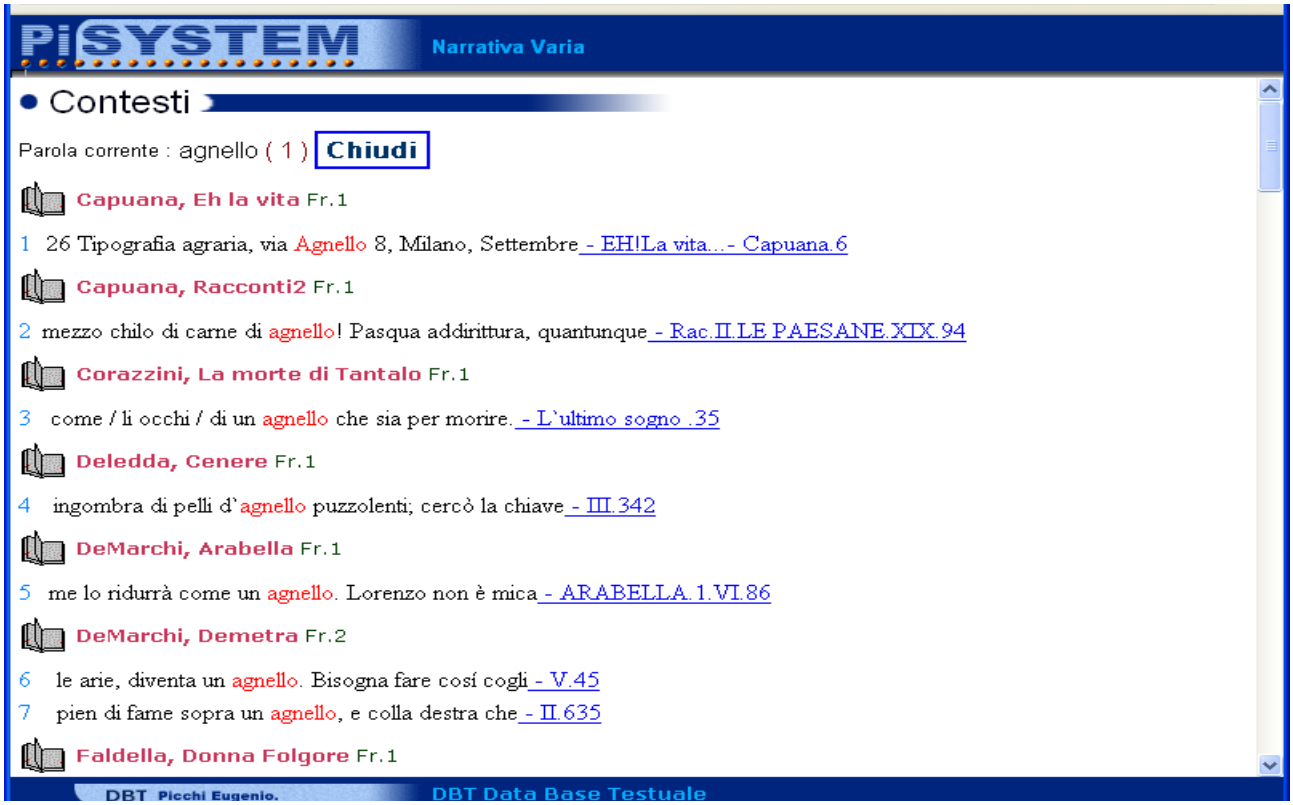
Digitando la stessa parola, ma usando il pulsante Morfologia si otterrà come risultato il recupero di tutte le forme del lemma richiesto (in questo caso il sostantivo anche al plurale, nel caso di un verbo si otterrano tutte le sue forme flesse trovate nel testo).



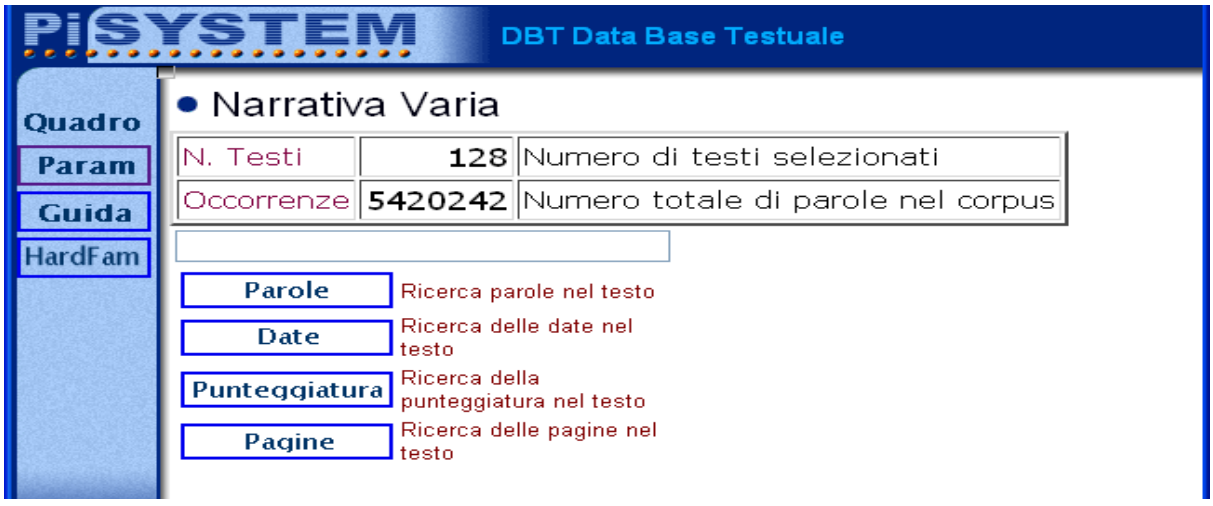
La stessa ricerca operata sul Corpus dà il risultato seguente, dove si può vedere il riepilogo che la parola, comprensiva della variante ortografica accentata, si trova in 29 testi diversi.

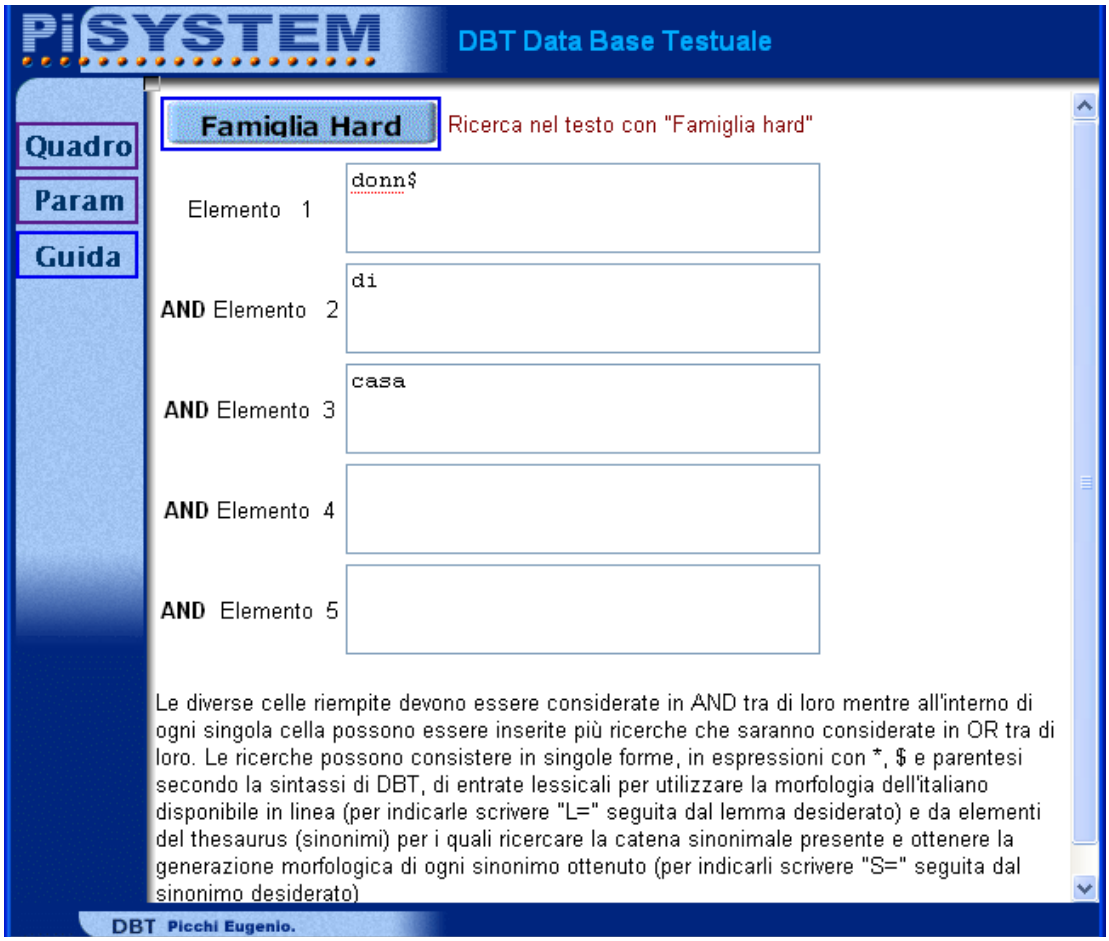


Con il pulsante *Contesti* si possono visualizzare tutti i contesti, in ordine alfabetico di autore, con l'indicazione dell'opera a cui appartengono e con l'indicazione della suddivisione interna (capitoli o paragrafi, qualora siano presenti).



Nella finestra iniziale del Corpus è presente anche il pulsante *HardFam* che permette di fare ricerche più complesse.





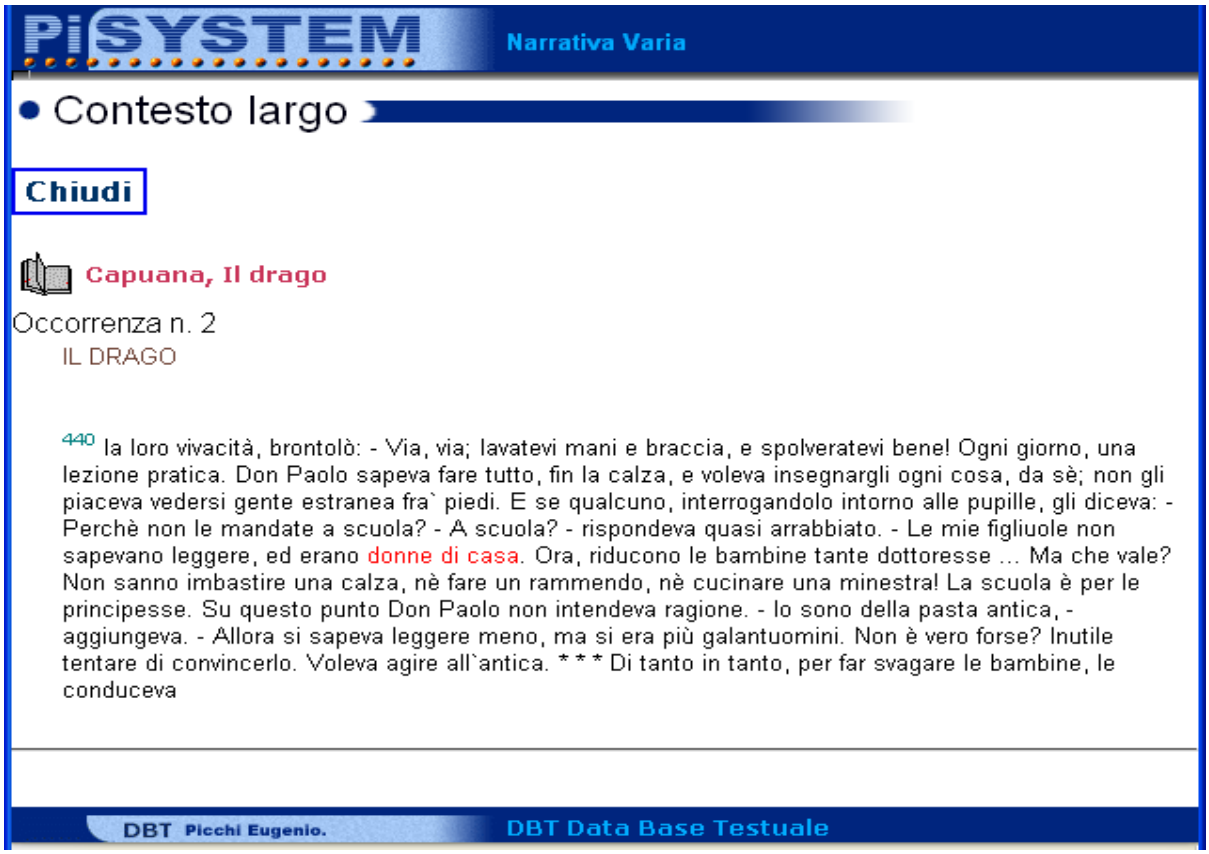
Viene proposta una tabella nelle cui celle indicare le parole richieste: in questo esempio se ne propongono tre, una per cella. La prima parola viene mascherata con il dollaro, che permette di recuperarne anche la forma plurale.

Se necessario, è possibile restringere l'ambito della ricerca in modo da individuare solo i contesti che rispettino il parametro di vicinanza delle parole; questo parametro viene impostato tramite il pulsante *Param*, presente nella maschera iniziale, che apre la seguente finestra di inserimento:

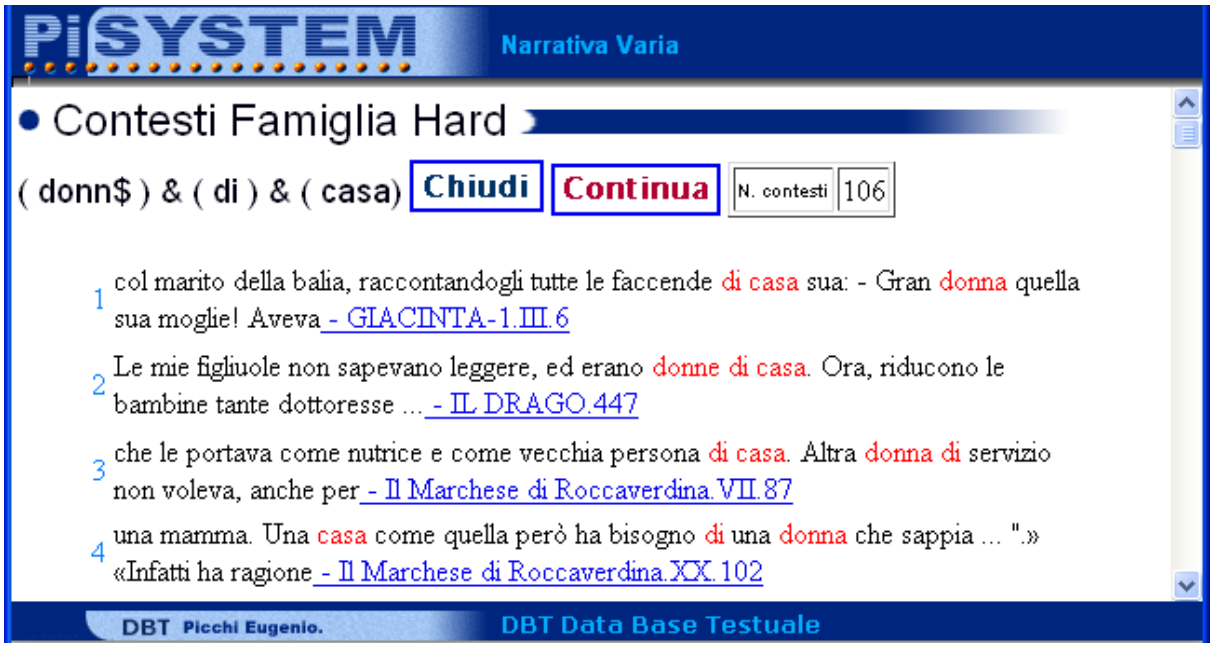
Nell'esempio riportato, la famiglia di parole viene recuperata solo se

queste si trovano ad una distanza inferiore a 2, che indica che vi può essere soltanto un elemento fra le parole richieste.

Nella finestra del risultato si possono quindi vedere i contesti (26) che rispettano tali condizioni. Nel riquadro successivo si può vedere il contesto largo e il riferimento al testo che lo contiene.



Non ponendo questo limite di vicinanza, il risultato della richiesta sarebbe il seguente (106).



Elenco completo delle opere disponibili nel Corpus

Nella seguente tabella sono elencati tutti gli autori e i testi che fanno parte del Corpus. In particolare viene evidenziato nella seconda colonna a sinistra il numero di parole che compongono ciascun testo.

1	2278	Campana, La Verna
2	395	Campanile, Se la luna
3	22449	Capuana, Cardello
4	45420	Capuana, Eh la vita
5	50255	Capuana, Giacinta
6	32691	Capuana, Il benefattore
7	17810	Capuana, Il drago
8	60100	Capuana, Profumo
9	75872	Capuana, Marchese Roccaverdina
10	127947	Capuana, Racconti1
11	113615	Capuana, Racconti2
12	82598	Capuana, Racconti3
13	29046	Capuana, Scurpiddu
14	9845	Capuana, Un vampiro
15	33249	Capuana, Ceraunavolta
16	38380	Capuana, Raccontafiabe
17	48050	Capuana, Chi vuol fiabe
18	23938	Capuana, Si conta
19	38921	Capuana, Ultimi racconti

20	6690	Casanova, Avventure
21	2328	Casati, Disperatamente Giulia
22	19541	Cattaneo, Psicologia menti associate
23	36879	Collodi, Storie allegre
24	1545	Corazzini, La morte di Tantalo
25	33015	Corio, Abissi Plebei
26	67595	Cuoco, Saggio storico
27	32106	D'Azeglio, Racconti Leggende
28	69414	Deledda, Cenere
29	106863	DeMarchi, Arabella
30	57394	DeMarchi, Cappello del prete
31	101190	DeMarchi, Demetra
32	13358	DeMarchi, DueMarianne
33	95433	DeMarchi, Giacomo
34	27479	DeMarchi, Signor dottorino
35	22395	DeMarchi, Vecchie storie
36	20997	DeRoberto, Morte amore
37	16051	Dossi, Colonia felice
38	65337	Faldella, Donna Folgore
39	38904	Faldella, Serenata ai morti
40	20470	Faldella, Viaggio a Roma
41	16701	Fogazzaro, Il fiasco
42	98145	Fogazzaro, Il santo
43	140281	Fogazzaro, Malombra

44	113488	Fogazzaro, Piccolo mondo antico
45	3547	Galateo, Milano visione
46	76634	Garibaldi, Clelia
47	73981	Ghislanzoni, Abrakadabra
48	5566	Giacosa, Partita a scacchi
49	5105	Ginzburg, Mai devi domandarmi
50	35523	Gozzano, Altare del passato
51	19387	Gozzano, La danza degli gnomi
52	139519	Grossi, Marco Visconti
53	15069	Leopardi, Discorso
54	19738	Loria, Caff, e altro
55	3398	Loy, Sogni d'inverno
56	3398	Loy, Sogni d'inverno
57	43005	Mantegazza, Anno3000
58	44869	Mantegazza, Madera
59	1539	Maraini, Passato e futuro
60	52552	Marchesa Colombi, Gente per bene
61	29056	Marchesa Colombi, In risaia
62	26450	Marchesa Colombi, Matrimonio in provincia
63	45264	Marchesa Colombi, Tramonto ideale
64	36552	Mazzini, Doveri
65	1023	Azzucco, Vita da vita
66	8713	Michelstaedter, Poesie
67	35773	Mioni, Sogno dell'anarchico

68	2388	Neera, Donne Milanesi
69	53791	Neera, Teresa
70	118	Onofri, poesia
71	66579	Oriani, Oro incenso mirra
72	41398	Oriani, Vortice
73	387	Panzini, Cura del moto
74	631	Pavese, Poesie
75	94452	Perodi, Caino Abele
76	36855	Perodi, Fate d'oro
77	206467	Perodi, Novelle della nonna
78	33213	Perodi, Tempo
79	33213	Perodi, Tempo
80	12286	Praga, Fiabe
81	70391	Praga, Memorie Presbiterio
82	17179	Praga, Penombre
83	10317	Praga, Tavolozza
84	11574	Praga, Trasparenze
85	6551	Rajna, Dialetto Milanese
86	9302	Sacchetti, Vita Letteraria
87	108629	Salgari, Alla conquista di un impero
88	51288	Salgari, Attravero l'Atlantico in pallone
89	50963	Salgari, I corsari delle Bermude
90	104993	Salgari, I figli dell'aria
91	69300	Salgari, I misteri della giungla nera

92	68068	Salgari, I pescatori di balene
93	59282	Salgari, I pirati della Malesia
94	81448	Salgari, I predoni del Sahara
95	109848	Salgari, Il figlio del corsaro rosso
96	95491	Salgari, Il re del mare
97	14754	Serao, La virtù
98	26912	Serao, Leggende napoletane
99	11337	Serao, O Giovannino o morte
100	142496	Serao, Paese della cuccagna
101	41733	Serao, Storia 2 anime
102	35624	Serao, Ventre di Napoli
103	24100	Steno, Contro il fato
104	995	Tabucchi, Vari
105	28024	Tarchetti, Racconti fantastici
106	32760	Tarchetti, Racconti umoristici
107	5931	Torelli, Stampa Politica
108	15623	Tozzi, Bestie
109	74031	Vamba, Gianburrasca
110	34969	Venosta, Omnibus
111	28414	Verga, Eva
112	30007	Vertua, Vita intima
113	78149	Zanazzo, Usi del popolo romano
114	32601	Zuccoli, Maleficio occulto
115	3586	Campana, Canti Orfici

116	179	Campana, Chimera
117	662	Ammaniti: Io non ho paura
118	75570	Arrighi, Nana
119	3080	Arrighi, Pungolo
120	153953	Artusi, La scienza in cucina
121	36675	Bazzero, Ugo
122	33154	Boito, Il Maestro
123	51643	Boito, Senso
124	32642	Bonanni, Il Fosso
125	56959	Bonanni, Vietato ai minori
126	390	Borghi, Ambrosiano
127	1663	Borghi, Madamine
128	3105	Camilleri, Passato futuro
	5420242	**Numero totale di parole nel corpus**