Costellazioni tematiche in un corpus letterario italiano

Margareta Kastberg Sjöblom

ATST, Centre Jacques Petit, EA3187, Université de Franche-Comté

UFR SLHS, 30, rue Mégevand, 25030 Besançon cedex

Riassunto

L'obiettivo del nostro studio è la semantica lessicale applicata ai testi letterari. Gli strumenti forniti dalla statistica e dalla "textométrie", aiutano in una maniera oggettiva a ricostruire i principali temi di un corpus. Questo studio permette infatti l'identificazione di temi ricorrenti e isotopie di un corpus letterario italiano. L'analisi automatica delle collocazioni e della microdistribuzione dei termini con una estrazione automatica di "universi semantici" o "mondi lessicali" può essere effettuata con diverse metodologie. Da un lato l'estrazione di un universo tematico, che ruota attorno ad un "polo lessicale", e dall'altro il censimento dei co-occorrenti e delle sequenze dei diversi lessemi ripetute sempre uguali nel corpus. L'oggetto dell'analisi è rappresentato da un corpus costituito di opere letterarie romanzesche dell'autore ligure Anton Giulio Barrili, (1836-1908). Questo corpus sarà sottoposto al trattamento statistico di *Hyperbase* nella sua recente versione italiana, seguendo la lemmatizzazione con Tree Tagger nella sua versione per il trattamento dell'italiano, che permette anche di trarre conclusioni più formali sulla morfologia e sulla sintassi del corpus.

Abstract

This article focuses on the analysis of textual data and the extraction of lexical semantics. The techniques provided by lexical statistics tools today open the door to many avenues of research in the field of corpus linguistics, including reconstructing the major semantic themes of a textual corpus in a systematic way, thanks to a computer-assisted semantic extraction. What are the different semantic formations and constellations of a text? The automatic analysis of the semantic collocations and micro-distributions makes it possible to investigate different techniques of automatic extraction of co-occurrences and "semantic universes" revolving around a pole. The object used as a testing ground is a corpus made up by parts of the literary work of the Ligurian author Anton Giulio Barrili (1836-1908). These texts are submitted to statistical analysis, using the lexical statistical tool *Hyperbase* in its recent Italian version. The corpus is tagged and lemmatized using the Tree Tagger lemmatization tool for the Italian language, which also allows more formal conclusions to be drawn in particular concerning morphology and syntax.

Keywords: frequency, co-occurrences, thematic associations, co-occurrential microcosm, Italian literature

1. Introduzione

L'obiettivo del nostro studio è l'analisi di dati testuali e più precisamente la semantica lessicale applicata ai testi letterari. Gli strumenti forniti dalla statistica e dell'analisi chiamata oggi in Francia "textométrie" aiutano in una maniera oggettiva a ricostruire i principali temi di un corpus. Questo studio, estensione della semantica lessicale, permette infatti l'identificazione di temi ricorrenti e isotopie di un corpus letterario italiano.

Quali sono le varie costellazioni semantiche di un testo? Cercheremo di trarre vantaggio dalle recenti innovazioni in materia di analisi dei dati testuali per l'identificazione di isotopie o iso-

tropie di un corpus (Viprey, 2005). L'analisi automatica delle collocazioni e della micro-distribuzione dei termini con una estrazione automatica di "mondi lessicali" può essere effettuata con diverse metodologie. Da un lato l'estrazione di un mondo lessicale o universo tematico, che ruota attorno a un "polo lessicale", e dall'altro il censimento dei co-occorrenti e delle sequenze dei diversi lessemi ripetute sempre uguali nel corpus.

L'oggetto d'analisi è rappresentato da un corpus costituito di opere letterarie romanzesche dell'autore ligure Anton Giulio Barrili (1836-1908), che ha partecipato alla lotta per l'indipendenza italiana accanto a Garibaldi. Si tratta di un personaggio conosciuto nella storia italiana, non solo per la sua azione politica, ma soprattutto per la sua importanza letteraria. Abbiamo costruito un corpus di nove dei suoi romanzi: L'Olmo e l'edera, La Montanara, L'undicesimo comandamento, Galatea, Il ritratto del diavolo, Tra cielo e terra e La notte del Commendatore. Questo corpus sarà sottoposto al trattamento statistico di Hyperbase nella sua recente versione italiana, seguendo la lemmatizzazione con Tree Tagger nella sua versione per il trattamento dell'italiano, che permette anche di trarre conclusioni più formali sulla morfologia e sulla sintassi del corpus.

2. Il corpus: Anton Giulio Barrili

Il destino di Anton Giulio Barrili è rimasto per sempre legato all'Italia e ai movimenti politici dell'Ottocento. Nasce nel 1836 a Savona, trascorre l'infanzia a Nizza (ora in Francia, ma all'epoca appartenente al Regno di Sardegna), termina gli studi superiori a Savona, poi si laurea in Lettere e Filosofia all'Università di Genova. Barrili intraprende quindi la carriera di giornalista e diventa in seguito redattore.

Nel 1859 si arruola come volontario nell'esercito piemontese, partecipando a varie campagne. Barrili combatte al fianco di Garibaldi in Trentino, nel Corpo Volontari Italiani, ed è ferito nella celebre battaglia di Mentana, quando le truppe garibaldine tentavano di liberare la capitale. Ritornato a Genova, fonda il celebre quotidiano *Il Caffaro*. Candidatosi alla Camera nelle liste della Sinistra, è eletto deputato nel 1876, gli viene conferita la cattedra di Letteratura italiana all'Università di Genova, dove sarà nominato Magnifico Rettore.

La vasta opera letteraria di Barrili è composta da numerose novelle e una cinquantina di romanzi, dei quali Benedetto Croce (Carrannante, 2009: 334) ha elogiato lo «stile limpido e scorrevole, senza stento, senza disuguaglianze e insieme accurato e corretto».

Oggi, l'opera di Barrili può sembrare un po' superata, alcune opere sono dimenticate e il suo insieme non è proprio sopravvissuta al ventesimo secolo. Tuttavia, il suo lavoro è importante, come testimone del suo tempo, superando la dimensione letteraria; è una testimonianza politica e storica di un periodo cruciale della storia italiana.

Abbiamo costruito un corpus di nove suoi romanzi. *L'Olmo e l'edera* (1869), *La Montanara* (1886), *L'undicesimo comandamento* (1891), *Galatea* (1896), *Il ritratto del diavolo* (1905), *Tra cielo e terra* (1907) e *La notte del Commendatore* (1908). Questo corpus digitalizzato di 9 romanzi contiene circa 700.000 occorrenze.

Esso è stato digitalizzato ed analizzato con *Hyperbase* nella sua ultima versione 8.0. (Brunet, 2009). *Hyperbase* è un software elaborato per l'esplorazione quantitativa dei corpora di testo di grandi dimensioni e consente una vasta gamma di trattamenti per corpus di testi predefiniti o scelti dall'utente. Consente il trattamento automatizzato di dati e statistiche che possono servire come piattaforma per vari studi. Esso permette il funzionamento e l'ottenimento di contesti docu-

mentari [1] e concordanze. La distribuzione di un elemento lessicale può essere studiato in tutti i testi che compongono il corpus di lavoro attraverso la visualizzazione grafica. L'esplorazione statistica permette varie indagini, non solo come quelle tradizionali della dimensione del lessico, dell'aumento del vocabolario, della distanza lessicale, della correlazione cronologica, ma permette anche di analizzare le specificità lessicali. *Hyperbase* può analizzare non solo gli elementi lessicali, ma anche i lemmi, i codici grammaticali, costellazioni sintattiche. Per questo lavoro, il nostro corpus è stato precedentemente lemmatizzato con TreeTagger, che procede ad una lemmatizzazione automatica dei dati.

TreeTagger è un analizzatore gratuito sviluppato presso l'Università di Stoccarda per diverse lingue, ed è un analizzatore di tipo statistico e non linguistico (Schmitt, 2010). All'uscita di ogni trattamento l'elemento lessicale è accompagnato da un massimo di due informazioni: il genere grammaticale e il lemma: -token –lemma -sgml.

Etienne Brunet, inventore e designer del software Hyperbase, lo ha integrato con la possibilità di analizzare dati in altre lingue oltre al francese esplorando programmi di questo tipo. Questi lavori l'hanno portato a creare versioni di Hyperbase in diverse lingue permettendo anche il trattamento di corpus lemmatizzati. Ora, quindi, c'è una versione italiana di Hyperbase che permette di condurre un'analisi sintattica adattata alla lingua italiana. In Fig. 1 si vede la distribuzione delle parti del discorso.

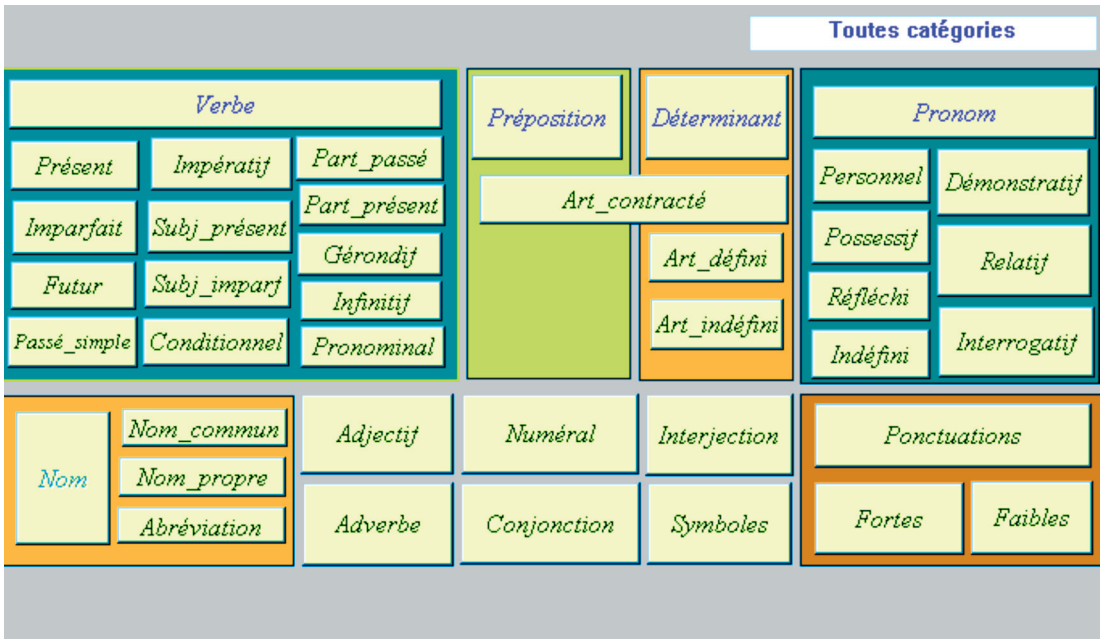

Figura 1: Le categorie grammaticali di Tree Tagger

2. Eplorazione statistica delle analisi delle co-occorrenze

Il trattamento statistico dei dati con Hyperbase consente oggi diversi approcci relazionali e ricostruzione di "reti" tematiche di un corpus. In primo luogo, permette l'estrazione di co-occor-

[1] Si intende il paragrafo attorno ad un polo lessicale.

renti e mette in relazione anche tutti i termini ubicati in un ambiente immediato di una data occorrenza.

Oltre a recuperare gli elementi più frequenti dal corpus, in questo studio cerchiamo di trovare la correlazione tra due elementi. Questo trattamento di co-occorrenze è stata a lungo una delle principali sfide della statistica lessicale. Inoltre, nella ricerca letteraria è diventato uno strumento importante, in particolare per l'estrazione dei sistemi isotopici.

Si tratta di un metodo basato sulla correlazione, per analogia con le procedure offerte dal software Alceste (Reinert, 2008). Invece di essere basato su una pre-segmentazione istituita da un thesaurus precedentemente costituito, qui le occorrenze o i lemmi possono riconoscere la loro propria "famiglia tematica", semplicemente grazie alla loro vicinanza negli stessi contesti.

Il programma di "correlazione" comincia con una lista di occorrenze (sostantivi o aggettivi che non sono né molto rari né molto frequenti) e registra tutti i loro incontri, occasionali o persistenti, nella stessa pagina [2]. Un nesso è stabilito tra due occorrenze, quando esse tendono ad incontrarsi. La tendenza rispecchia il numero di co-occorrenze, per la quale un registro è tenuto in una matrice quadrata in cui gli stessi elementi sono indicati sulle righe e colonne.

La scelta della pagina permette di evitare certe difficoltà sintattiche che imporrebbero la scelta di unità linguistiche più brevi (frase o paragrafo). L'eliminazione degli elementi lessicali funzionali consente inoltre di concentrarsi sulle relazioni semantiche o tematiche, piuttosto che sulle relazioni sintattiche. La convivenza a lunga distanza in un unico testo non conta. Conta solo nelle le co-occorrenze immediate o ravvicinate sulla stessa pagina, dove sono più probabili da trovare le isotopie.

La scelta delle occorrenze è automatica. Il programma di selezione registra tra i 200 e i 400 elementi, tutti sostantivi o aggettivi. Poi segue una fase di esplorazione sequenziale del corpus. In ogni pagina si cerca la presenza o l'assenza di elementi nella lista, rilevando le co-occorrenze più importanti.

50.86 montagna re	15.84 cielo terra
46.06 francesco signor	15.53 convento frate
26.35 commendatore signor	15.14 camera letto
24.29 labbro sorriso	14.56 altro uno
24.02 conta signor	14.18 miracolo san
22.16 capo cenno	13.65 acqua lago
18.23 convento san	13.37 marito moglie
17.62 capegli oro	13.31 conte marito
17.28 casa padrone	13.25 arte compagno
17.07 sera teatro	13.24 carta visita
15.89 francesco ospite	13.10 piazza santo

Tabella 1: Co-occorrenze

Le associazioni (Tab. 1) non sono affatto sorprendenti, troviamo al primo posto *montagnare*, seguito da *Francesco-signor*, *commendatore-signor* e *labbro-sorriso*. Chi conosce la scrittura di Barrili non è sorpreso nel notare l'associazione importante e simbolica del re e della

[2] Si intende la pagina del libro, identificata con *Hyperbase.*

montagna che ci riporta all'idea e all'ideale anche di una Italia unita, con il re nella capitale Torino e l'elevazione quasi religiosa della cattedrale Superga, la montagna del re.

L'analisi fattoriale delle corrispondenze conduce ad una mappa sintetica delle co-occorrenze. Il grafico (Fig. 2) con il primo piano fattoriale ottenuto per il nostro corpus può sembrare molto caotico, ma l'interpretazione è abbastanza chiara. Infatti, da sinistra a destra si passa dal concreto all'astratto: gli elementi costruiti dall'uomo (casa, salotto, strada, chiesa), le parti del corpo umano (viso, labbro, occhio, bocca), infine i sentimenti e il pensiero (amore, dolore, sentimento), l'arte e la creazione. Il secondo fattore sembra separare ciò che appartiene alla società e ciò che appartiene al singolo individuo.

Questa stessa struttura si trova anche in altre monografie, come quelle di Proust, Flaubert, Le Clézio e Gracq. Senza dubbio vi si deve leggere non l'utilizzo di uno stesso tema, ma un riflesso della composizione propria del romanzo, che prevede nella stessa opera delle pagine con scrittura narrativa, descrittiva, riflessiva o dei dialoghi, caratteristica formale del genere letterario del romanzo.

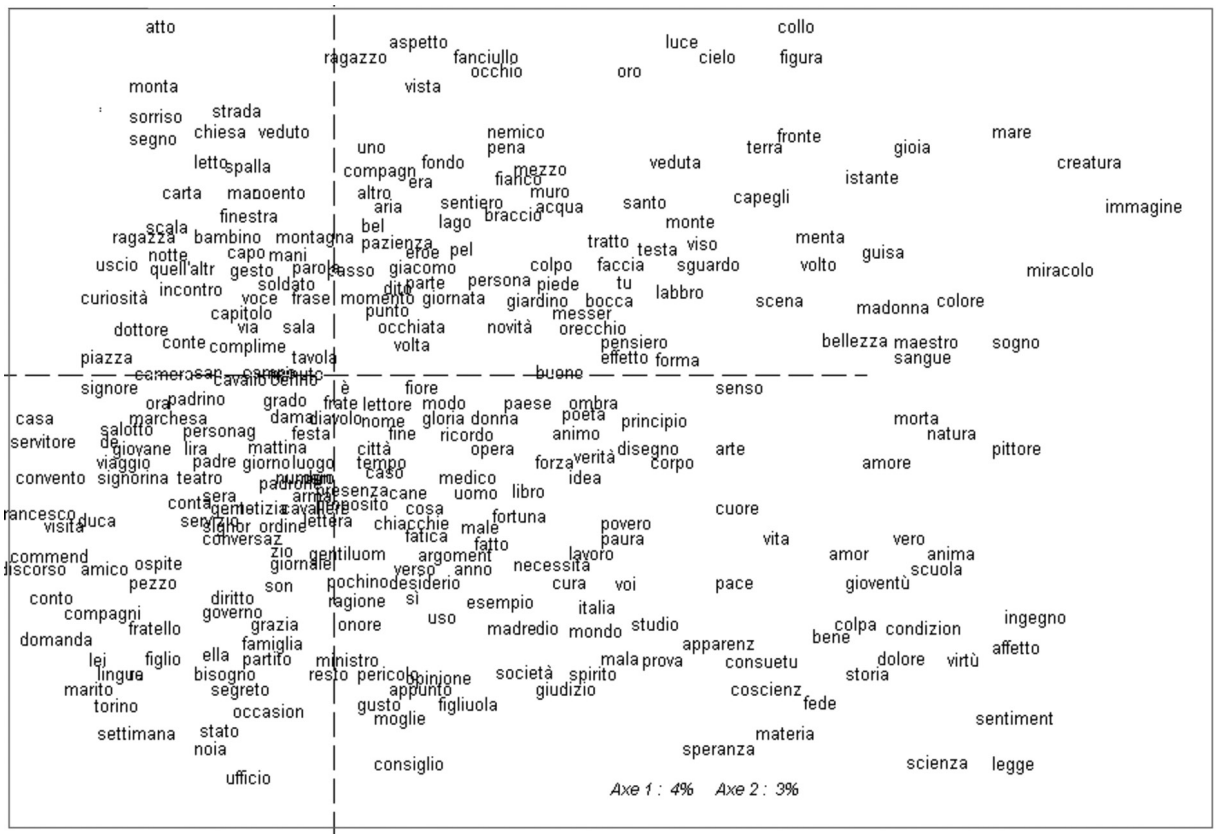

Figura 2: Analisi fattoriale delle corrispondenze delle co-occorrenze

Possiamo anche considerare gli isotopi specifici, non solo analizzando la frequenza, ma con un approccio diverso: quello delle sequenze. Questa analisi ci permette di considerare le associazioni delle occorrenze nella loro posizione adiacente, ignorando la divisione dei testi.

3. Associazioni tematiche

L'estrazione di co-occorrenze con Hyperbase, come si è detto, permette di creare una rete di elementi lessicali che costituiscono la base per i calcoli successivi.

Nel corpus è stata riscontrata forte presenza del lemma Italia. Lo stato italiano nasce nel 1861 dopo il lungo periodo del Risorgimento. Quel periodo della storia d'Italia, in cui l'affermarsi di una coscienza nazionale porta all'unità politica e all'indipendenza della nazione italiana, è anche il periodo della gioventù di Barrili. Quelli della nascita del Regno d'Italia sotto la dinastia di Casa Savoia sono decenni sempre presenti nell'immaginario dello scrittore.

Il lemma Italia è anche molto presente nella nostra lista di co-occorrenze. L'Italia è in co-occorrenza diretta con i lemmi seguenti [3]:

Italia – Torino (9.29), *Italia – soldato* (6.27), *Italia – fortuna* (5.87), *Italia – poeta* (5.67), *Italia – nome* (5.50), *Italia – città* (5.17), *Italia – idea* (4.82), *Italia – colore* (4.37), *Italia – volta* (3.03), *Italia – marchesa* (2.99), *Italia – occhio* (2.68).

Queste sono le co-occorrenze binarie più forti. È da notare la forte associazione dei lemmi Italia e Torino, la capitale dell'Italia dal 1861 fino al 1865, solo quattro anni. Il lemma Italia può essere considerato come un polo e Hyperbase permette d'estrarre le 25 co-occorrenze più vicine.

argomento, *cenno*, *città*, *colore*, *cura*, *disegno*, *fortuna*, *idea*, *lago*, *madre*, *maestro*, *marchesa*, *nome*, *occhio*, *poeta*, *povero*, *re*, *sala*, *sangue*, *segreto*, *soldato*, *Torino*, *verso*, *volta.*

La figura 3 fornisce una rappresentazione grafica dei collegamenti preferenziali che tessono una rete intorno all'occorrenza scelta per il polo, in questo caso *Italia*.

Il calcolo della struttura grafica, i nodi e gli archi, è garantita dal software libero GraphViz. I dati forniti dal programma e i risultati sono riportati da Hyperbase con una rappresentazione grafica che tiene conto non solo delle posizioni, ma anche dei pesi relativi degli incontri degli elementi lessicali. Le occorrenze in grigio (in rosso nella versione colore) corrispondono ai nodi di frequenza elevata, e quelli in nero ai nodi meno frequenti, non avendo un contatto diretto con il polo, Italia. Le linee in grassetto corrispondono alla co-occorrenza diretta con il polo e le linee più sottili alla co-occorrenza indiretta, cioè "gli amici degli amici" [4].

Così, vediamo non solo "la prossemica", la relazione degli elementi, ma anche il loro potere di prossimità. I legami più forti dell'Italia sono *Torino*, *soldato*, *nome*, *poeta* e *cenno*. Inoltre, il grafico permette di identificare i legami che mantengono queste occorrenze con altri elementi lessicali, Torino è legata alla città, alla marchesa, all'argomento, al nome, alla cura e all'idea.

Ma possiamo anche considerare altrimenti l'associazione di lessemi, questo microcosmo co-occorrenziale. Prendiamo lo stesso elemento lessicale, *Italia*, uno dei lemmi più frequenti del corpus. Questa volta, abbiamo estratto il contesto immediato che circonda tutte occorrenze del lemma *Italia*.

---

[3] Il programma calcola e ordina in coppia tutti gli indici che misura la distanza tra le parole della lista. La selezione del filtro è impostato secondo un valore appropriato, date le dimensioni del corpus.

[4] La misura della co-occorrenza è quella di Dunning (1993). Questo indice si basa su quattro parametri
- a: numero di co-occorrenze di due occorrenze nel campo esplorato (*ivi* il paragrafo);
- b: numero di occorrenze della prima occorrenza, in assenza della seconda;
- c: numero di occorrenze della seconda in assenza della prima occorrenza;
- d: numero di occorrenze, altre co-occorrenze RV = -21ogL = 2(sl-s2).

Per s1 = a log a + b log b + c log c + d log d + (a+b+c+d)log(a+b+c+d)
s2 = (a+c)log(a+c) + (b+d)log(b+d) + (a+b)log(a+b) + (c+d)log(c+d)
A partire dal valore 4, l'indice di Dunning è considerato come sfuggevole (soglia del 5%). Dal valore 4, l'indice di Dunning è considerato non affidabile.

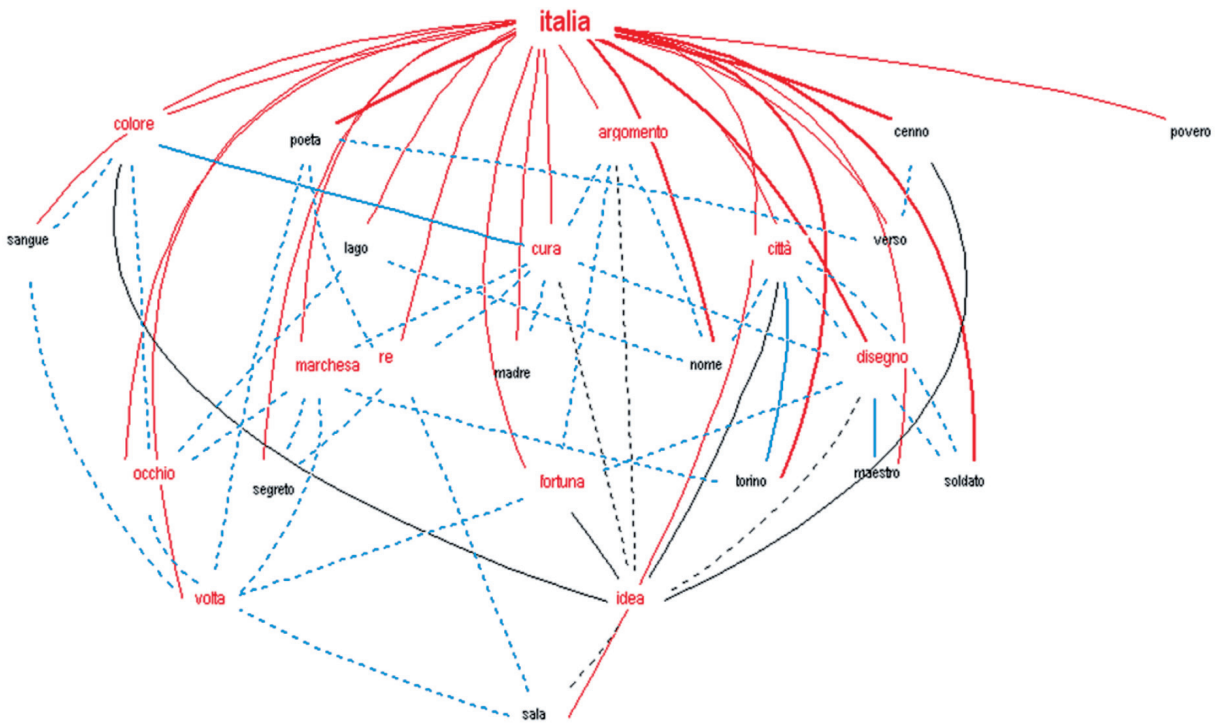

Figura 3: Associazioni tematiche del lemma Italia

L'estrazione automatica del contesto di un elemento lessicale (che è il paragrafo) permette la creazione di un sub-corpus che è oggetto di un calcolo di specificità, dal momento che non cerchiamo più un rapporto tra un'occorrenza ed un testo, ma un rapporto tra le occorrenze stesse, che misura anche il calcolo di correlazione, quando due serie sono giustapposte in uno stesso grafico. Dunque, questa procedura non è ridotta a misurare la prossimità di due occorrenze, ma ogni infinita relazione di tutte le co-occorrenze che possono esistere intorno ad un'occorrenza definita come polo, come *Italia* [5].

Noi (si ha da mettere in sodo) non abbiamo in ITALIA il cuore, il capo, e il ventre della nazione in una sola città.

_____ Olmo Page: 4 c _____

Firenze, Milano, Napoli, Palermo, Genova, Venezia, Roma, Bologna e Torino (Ella vede che le pongo a mazzo) sono esse tutta l'ITALIA, e ognuna di esse, insieme coi caratteri comuni a tutte, porta i suoi caratteri particolari.

_____ Olmo Page: 5 a _____

Magnifica dimora è il cimitero di Staglieno, e quando sarà finito, nessun'altra città d'ITALIA potrà vantare il somigliante per ricchezza di marmi e di disegno.

_____ Olmo Page: 211 d _____

Insomma, il cannone fu preso, ma Sandro era ferito a morte, ed ebbe appena il tempo di alzare il cappello sulla punta della baionetta, facendo cenno ai compagni di affrettare il passo, e gridare: viva l'ITALIA !

_____ Olmo Page: 251 a _____

L'ITALIA mi ha preso il mio Sandro, ma ora nessuno verrà più a metterci i piedi sul collo.

_____ Olmo Page: 251 b _____

- - Sì, di là dall'ITALIA, e ci andremo sull'Amerigo Vespucci -.

_____ Olmo Page: 372 a _____

[5] Estratto dell'elenco che contiene 95 contesti.

Nell'aprile del 1849, prostrate sui campi di Novara le fortune d'ITALIA, Francesco V era ritornato tra i frementi suoi sudditi, ma con le baionette austriache al fianco, e primo a muovergli incontro, per dargli il benvenuto nei suoi felicissimi Stati, fu il conte Jacopo Malatesti.

Montanara Page: 400 b

Nata intorno al 1820, era del 1857 una bellezza matura e stupenda, citata a Piacenza come a Bologna, a Torino come a Firenze, nota insomma a tutta l'ITALIA superiore per la sua alta galanteria, per il suo matto spendere, per le teste che aveva fatte girare.

Montanara Page: 403 b

Distanza	*Corpus*	*Estratti*	*Parola*	*Distanza*	*Corpus*	*Estratti*	*Parola*
37.58	95	96	italia	2.57	245	5	(
7.54	2247	45	d'	2.57	245	5	)
6.22	19	6	piemonte	6.37	8	5	«
4.76	67	6	patria	3.45	122	5	»
4.43	8	3	fortune	2.15	53	2	acque
4.40	388	11	nome	2.19	50	2	antichi
4.29	4509	47	l'	2.17	51	2	atti
4.18	32	4	paolo	3.51	2071	24	aveva
4.07	13	3	begli	3.42	7	2	banchetto
3.95	41	4	soldato	2.53	31	2	barba
3.79	19	3	poeti	2.69	24	2	barca
3.79	19	3	circondario	4.07	13	3	begli
3.71	98	5	torino	2.91	17	2	bologna
3.52	27	3	corona	3.02	14	2	burchiello
3.51	2071	24	aveva	2.67	25	2	capi
3.51	6	2	schiatta	2.91	17	2	caratteri
3.42	7	2	banchetto	3.39	32	3	case
3.39	32	3	case	2.32	42	2	causa
3.34	8	2	collare	2.06	141	3	ce
3.32	3121	31	della	3.79	19	3	circondario
3.23	39	3	nobili	3.00	176	5	città
3.17	42	3	unico	3.34	8	2	collare
3.17	42	3	cuori	3.52	27	3	corona
3.16	11	2	n'erano	3.17	42	3	cuori
3.07	13	2	regioni	7.54	2247	45	d'
3.02	14	2	burchiello	2.55	30	2	debbono
3.01	3501	32	i	2.47	1609	16	dei
3.00	176	5	città	3.32	3121	31	della
2.94	55	3	viva	2.39	1112	12	delle
2.91	17	2	felicissimo	2.22	120	3	dovuto
2.91	17	2	caratteri	2.18	14418	93	e
2.91	17	2	bologna	2.15	53	2	eccellenza
2.84	19	2	sandro	2.20	49	2	entrato
2.80	65	3	firenze	2.35	40	2	esse
2.75	22	2	popoli	2.67	25	2	europa
2.73	70	3	musica	2.91	17	2	felicissimo
2.69	24	2	roma	2.80	65	3	firenze

Tabella 2: Elenco del microcosmo tematico del lemma Italia

Confrontando il lemma Italia col suo entourage – come si diceva, il paragrafo –, si ottiene un sub-corpus che consiste in lessemi che ruotano attorno al polo. Resta da confrontare questo sottoinsieme al grande corpus Barrili, che è 8 volte maggiore. Tutte le parole trovate sono sottoposte ai calcoli di frequenza, e la lista che risulta dei calcoli è confrontata con il dizionario. È,

infatti, il confronto classico di due corpora. I calcoli producono, alla fine del'ultima analisi, due elenchi di co-occorrenze (Tab. 2), gerarchico a sinistra e alfabetico a destra [6].

L'elenco che si ottiene da questa analisi è un po' diverso da quello che avevamo ottenuto in precedenza. In primo luogo, troviamo ovviamente la preposizione *di*. Ma più interessante è poter identificare, nel cuore del testo, all'interno dei paragrafi, un microcosmo tematico molto caratteristico di un'epoca e di un autore.

Tra i lessemi più vicini possiamo identificare *Piemonte*, *Torino*, *corona*, *nobile*, *città*, *nobiltà*, *principe*, *regno* e *re*. L'Italia di Barrili, nella sua poduzione letteraria è veramente l'Italia del re Vittorio Emanuele e della nobiltà piemontese.

Possiamo eseguire lo stesso calcolo di prima, quello relativo alle co-occorrenze, ma questa volta prendendo in considerazione come base per il calcolo, solo il nostro sub-corpus dell'entourage tematico del polo Italia:

Hyperbase permette ancora d'estrarre le co-occorrenze più vicine:

patria, *Piemonte*, *fortune*, *Paolo*, *begli*, *soldato*, *poeti*, *circondario*, *Torino*, *corona*, *schiatta*, *della*, *banchetto*, *case*, *collare*, *nobili*, *unico*, *regioni*, *città*, *burchiello*, *viva*, *felicissimo*, *caratteri*, *Bologna*, *Sandro*, *Firenze*, *popoli*, *Roma*, *Europa*, *sempre*, *segreto*, *nobiltà.*

Qui il calcolo viene eseguito come nel calcolo precedente, ma prendiamo in considerazione l'universo tematico, vale a dire, il punto che circonda il polo *Italia*. Il grafico che segue riflette le diverse relazioni semantiche, la loro forza e le loro costellazioni.

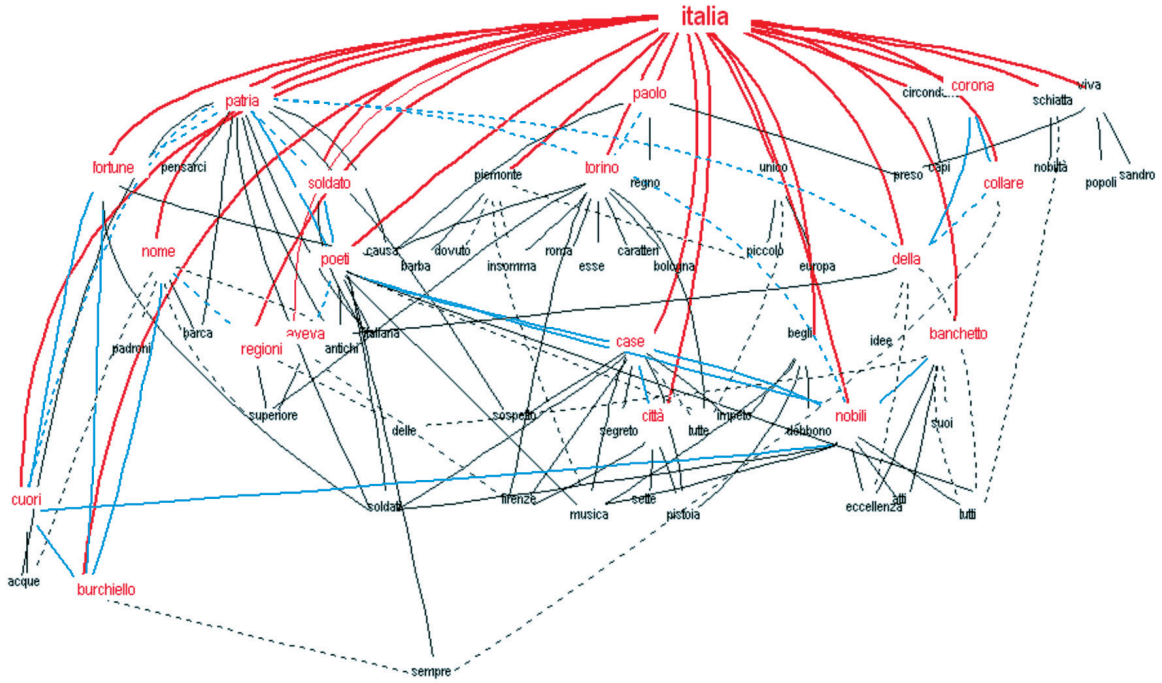

Figura 4: Associazioni tematiche del lemma Italia nel suo contesto

Il grafico in Fig. 4 è simile a quello in Fig. 3, ma con dei cambiamenti abbastanza significativi, soprattutto se si guarda alla forza dei legami e alle relazioni non dirette. *Torino*, la co-occorrenza più vicina, ha qui, a sua volta, dei legami piuttosto complessi con altri elementi lessicali in questo sub-corpus: *Roma*, *Bologna*, *Firenze*, *causa*, *superiore*.

[6] Estratto degli elenchi.

5. Conclusione

Questi diversi modi di considerare le associazioni tematiche e i microcosmi lessicali intorno ad un polo, all'interno di un corpus, permette non solo di confrontare diversi calcoli statistici, ma anche di fornire elementi di prova in merito alla robustezza di queste analisi. I risultati sono spesso molto simili e la completezza dei singoli studi fornisce una solida base e una complementarità importante per lo studio delle isotopie e collocazioni in un testo.

La statistica lessicale ha fatto molti progressi e ora si tende a considerare il testo come una struttura ordinata o come uno spazio organizzato. Il trattamento dei dati quantitativi e la classificazione automatica tengono conto oggi anche della dimensione sintagmatica del testo, della struttura lineare delle dinamiche interne e della progressione del testo. Per ragioni di spazio non possiamo interessarci a tutti questi aspetti che formano l'analisi coerente di un corpus digitalizzato.

Riguardo al corpus di Barrili le diversi analisi hanno permesso di indagare su un tema specifico in un momento particolare dell'Italia, in una prospettiva letteraria e storica. L'immagine dell'Italia che si riflette attraverso il corpus di Barrili non è forse quella che ci aspettevamo da un politico di sinistra, giornalista e patriota entusiasta. Abbiamo anche notato che dei lessemi come *repubblica*, *parlamento*, *democrazia* sono praticamente assenti nel corpus. Nell'immaginario di Barrili, almeno in questo corpus, che va dal 1869 al 1907, non c'è nessun presentimento della fine del regno di Casa Savoia, nemmeno la sola idea di una repubblica italiana.

Tuttavia, se il trattamento statistico di un corpus può fare una descrizione formale, le analisi della statistica co-occorrenziale e della "prossemica" di elementi lessicali permette di giungere ad un nuovo livello, importante per passare alla semantica e alla fraseologia. Ecco perché questa tecnica si sta rivelando una risorsa preziosa per l'analisi semantica del testo.

Riferimenti

Adam J.-M. and Heidmann U. (editors) (2005). *Sciences des textes et analyse du discours: enjeu d'une interdisciplinarité*. Genève : Slatkine Erudition.

Brunet É. (2009). *Hyperbase, Manuel de référence, version 8.00.*, Nice, CNRS-ILF, "Bases, corpus et langage" (UMR 6039).

Brunet E. (2008). Fréquenes et séquences. Mise en oeuvre dans Hyperbase. In Mellet S. and Salem A., editors, *Lexicometrica, Topographie et topologie textuelles*, http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9.htm.

Carrannante A. (2009). Per il centenario di Anton Giulio Barrili (1836-1908). *Nuova Antologia*, gennaio-marzo: 333-337.

Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. In *Computational Linguistics*, vol. 19-1: 61-74.

Kastberg Sjöblom M. (2006). *J.M.G. Le Clézio – Des mots aux themes*. Paris: Honoré Champion, collection Lettres numériques.

Raster F. (1996). *Sémantique interpretative*. Paris: PUF.

Reinert M. (2002). *Alceste, Manuel de référence*. Université de Saint-Quentin-en Yvelines: CNRS.

Reinert M. (2008). Mondes lexicaux stabilisés et analyse statistique de discours. In Heiden, S. and Pincemin, B., editors, *JADT 2008*, Lyon 12-14 marzo, pp. 981-993.

Schmitt H. (2010). Tree Tagger, Università di Stoccarda, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

Viprey, J.-M. (2005). Philologie numérique et herméneutique integrative. In Adam, J.-M and Heidmann U., editors, pp. 51-67.