Quaderni di Statistica
Vol. 7, 2005

Statistica testuale e *text mining*: alcuni paradigmi applicativi

Sergio Bolasco
Università degli Studi di Roma "La Sapienza"
E-mail: sergio.bolasco@uniroma1.it

Summary: In this paper, after reconstructing some essential phases in the evolution of automatic analysis of texts, the steps of an "ideal" strategy for the statistical analysis of textual data are defined. The characteristics of lexical and textual analysis are described, as well as some techniques of information extraction, that employ resources which are both endogenous and exogenous with respect to the texts to be examined. In order to show the potential of today's textual statistics and of the most recent Text Mining applications, some relevant case studies concerning statistical survey and document analysis are illustrated.

Keywords: Textual statistics, Text mining, Automatic analysis of texts, Lexical analysis, Textual analysis, Information extraction.

1. Nascita ed evoluzione della statistica testuale

Gli studi statistici su dati espressi in linguaggio naturale, o *dati testuali*, a partire dagli anni 1960-1970 hanno subìto forti cambiamenti strettamente legati all'evoluzione dell'informatica, fino a produrre l'analisi automatica dei testi e la *statistica testuale* (Lebart, Salem 1994). Più recentemente, la crescente disponibilità di risorse linguistiche informatizzate (Zampolli, Calzolari 1995) e la crescente diffusione dei testi consultabili on-line, quindi direttamente analizzabili, ha ulteriormente rivoluzionato criteri e tecniche in quest'ambito. Le soluzioni trovate non si fondano più soltanto su strumenti statistici, ma scaturiscono da una stretta multidisciplinarità che associa a questi, con

uguale importanza, strumenti informatici e linguistici, soprattutto nell'area, nota oggi in letteratura, con il termine di *Text Mining* (TM) (Sullivan 2001, Zanasi 2005, Bolasco *et al.* 2005).

Di fatto, nel corso del tempo, gli studi quantitativi intorno alla lingua[1] hanno cambiato progressivamente il loro obiettivo, spostandolo da una logica di tipo *linguistico*[2] (sviluppata fino agli anni sessanta del Novecento) ad una di tipo *lessicale*[3] (intorno agli anni Settanta del secolo scorso), per approdare negli anni Ottanta e Novanta ad analisi di tipo *testuale*[4] o infine *lessico-testuale*[5].

Parallelamente a questo progressivo cambiamento d'obiettivo, sono mutati tecniche e unità di analisi. Vengono sviluppati strumenti *software* e vengono proposte vere e proprie filiere per l'elaborazione dei dati testuali. Mutano anche i soggetti stessi, protagonisti di questi studi: non è infrequente infatti veder nascere contributi originali non solo in università o centri di ricerca, ma anche in aziende. Queste ultime, dovendo interagire con enormi masse di materiali testuali spesso disponibili in rete (l'80% delle informazioni in azienda, si dice che sia in forma di testi e solo il 20% in dati numerici), hanno il problema di selezionare, all'interno di fonti smisurate, i dati di loro interesse, per estrarne *informazione* capace di produrre valore. Si tratta di soluzioni di *Text Mining* orientate alla gestione della conoscenza e alla cosiddetta *Business Intelligence*.

---

1 Contributi significativi si trovano in riviste quali, fra le altre, *Cahiers de Lexicologie, Computers and Humanities, ACM Computing Surveys, Journal of Quantitative Linguistics, Linguisticae Investigationes, Literary and Linguistic Computing, Mots, TAL*.

2 Per i rapporti fra lingua e sue concrete possibilità d'analisi cfr. Guiraud (1954) ed Herdan (1964). La dimensione illimitata e sfuggente della lingua rende difficile associare alle parole una qualche "frequenza" in senso statistico-probabilistico. Quest'ultima è semmai misurabile su una raccolta di testi, intesi come spezzoni di lessici, ovvero come "campioni" particolari di un idioma. È così che ci si limita a considerare le occorrenze delle parole in un testo come un'approssimazione delle frequenze in un lessico, a patto che il corpus sia sufficientemente ampio (almeno 50.000 occorrenze).

3 Cfr. ad esempio, Muller (1977) e Brunet (1981, 1988).

4 In questo approccio l'attenzione sulla testualità del contenuto privilegia l'analisi statistica in forme grafiche (Lebart, Salem 1988 e Lebart *et al.*, 1998).

5 Recentemente si è visto che l'analisi dei dati testuali migliora con l'apporto di meta-informazioni di carattere linguistico (dizionari elettronici, lessici di frequenza, grammatiche locali) e con interventi sul testo (normalizzazione, lemmatizzazione e lessicalizzazione), cioè attraverso un'analisi integrata statistico-linguistica di tipo lessico-testuale.

Ripercorrendo rapidamente in modo schematico questa evoluzione, riconosciamo in G.K. Zipf[6] (1935), G.U. Yule (1944), alcuni fra i principali precursori della moderna *analisi quantitativa in ambito linguistico*, delle sue proprietà e applicazioni statistiche. Lo stesso J.P. Benzecri (1963) fonda sullo studio di dati linguistici (1981) le sue prime sperimentazioni di quella che sarà l'*analyse des données* (1973, 1982), contrapponendosi alle tesi di N. Chomsky[7] e inseguendo Z.S. Harris[8], che rappresenta, quanto a formalizzazione di strutture linguistiche della scomposizione sintagmatica della frase, un riferimento assai vicino ad un approccio statistico sul trattamento del linguaggio naturale.

Successivamente, Ch. Muller (1973) e P. Lafon (1984), sviluppano indici e misurazioni divenute classiche nella *statistica linguistica*, in cui fin dagli anni '30-'50 si studiano le proprietà della lingua, concentrando l'attenzione su lessemi, morfemi, n-grammi; o nella *statistica lessicale*, in cui l'analisi del linguaggio si fonda sullo studio dei lemmi (anni '60-'70).

In parallelo, in Italia linguisti come A. Zampolli e T. De Mauro, attraverso il loro interesse per le misure di frequenza d'uso delle parole a livello di *lemmi*, mettono le basi per una *linguistica quantitativa*, sviluppando le prime risorse statistico-linguistiche (*lessici di frequenza*: il Lif di Bortolini, Zampolli (1971); i VdB, Veli e Lip di De Mauro *et al.* (1980, 1989, 1993).

---

6 Cfr. il sito http://linkage.rockefeller.edu/wli/zipf/

7 Chomsky sostiene che la linguistica non può essere induttiva, nel senso che la grammatica non può essere dedotta da regole trovate di fatto su un insieme di testi (corpus), ma solo deduttiva, quindi solo partendo da assiomi essa genera dei modelli delle lingue concrete (Benzécri 1982, 102). Come noto, Chomsky sviluppa una teoria grammaticale completa ed organica, la cosiddetta grammatica generativa con relative teorie trasformazionali (*Syntactic structures*, 1957).

8 In *Elementary transformations* (1964), Harris chiama distribuzione di una parola l'insieme dei suoi possibili contesti locali. In *Le strutture matematiche del linguaggio* (1968), egli sostiene che il discorso si presta ad una analisi distributiva indipendentemente dal senso; egli propone di determinare le regole combinatorie della lingua allo scopo di rivelare le relazioni elementari fra differenti classi di concetti presenti in un corpus. A tal fine, "occorre integrare al trattamento quantitativo del corpus un'analisi morfo-sintattica dei dati testuali, ossia introdurre algoritmi di descrizione delle frasi che consentono di segmentare gli enunciati del testo nei loro costituenti sintagmatici, poi di identificarli e infine di esplicitare i loro rapporti interni" (Martinez, 2003, p. 275).

Via via dall'interesse per i *testi veri e propri* (come i classici della letteratura, sfruttati negli studi *stilometrici* sull'opera di un Autore: si vedano R. Busa 1974-1980; E. Brunet 1981, 1988; D. Labbé 1990, 2003[9]) si passa allo studio di *testi "artificiali" (non testi)* e all'interesse verso i *dati espressi in linguaggio naturale* provenienti dalle fonti più diverse: indagini sul campo (domande aperte o interviste); analisi di frammenti o testi corti (abstract, bibliografie, manifesti, messaggi), raccolti in una collezione di documenti costituente un *corpus di dati testuali*. Il corpus può essere studiato secondo la sua frammentazione in documenti o "records". Un vantaggio dell'analisi automatica su base statistica consiste nell'essere indipendente dall'ampiezza o dimensione dei testi che hanno originato la raccolta e nel consentire ogni possibile confronto fra loro successivi raggruppamenti in partizioni, secondo variabili categoriali associate a ciascun "frammento".

Alla fine degli anni '80, L. Lebart e A. Salem (1988) definiscono i confini della *statistica testuale* basata sull'analisi per *forme grafiche* (e non più per *lemmi*) ed in parallelo sviluppano *software* per l'analisi dei dati testuali. In particolare, *Spad_T* che fa impiego di metodi multidimensionali, come le analisi fattoriali su matrici sparse con calcolo degli autovalori in lettura diretta (Lebart, 1982); *Lexico* che consente l'individuazione nel corpus dei *segmenti ripetuti* e l'analisi delle *specificità* per l'estrazione di parole caratteristiche delle sub-parti grazie ad un test basato sulla legge ipergeometrica.

2. Le diverse unità di analisi del testo

Il problema essenziale per un'analisi automatica di un testo è operare il riconoscimento del senso ivi presente. Con il termine *parola* si indica convenzionalmente l'unità di analisi del testo. A seconda degli obiettivi,

---

9 «… Nous avons la preuve que Corneille a probablement écrit beaucoup des pièces de Molière…» (da Le Monde, 11/6/03) è ciò che afferma Labbé in un articolo del *Journal of Quantitative Linguistics* del dicembre 2001 a partire da una prossimità eccezionale del vocabolario tra una commedia di Corneille, Le Menteur, scritta nel 1644, e sedici pièces di Molière (Labbé, 2003).

tale unità può essere una forma grafica, un lemma, un poliforme o un'*unità mista* (lessia), in grado di catturare al meglio il contenuto presente nel testo.

Nella statistica testuale, le analisi basate sulle forme grafiche hanno il vantaggio di essere *indipendenti dalla lingua*. Si tratta di un *approccio puramente formale* che privilegia i *segni* (significanti) per arrivare al *senso* (in quanto insieme di significati) come rappresentazione del contenuto o del discorso.

Il *segno* linguistico, come noto, è composto di un *significante* distinto dal punto di vista "fonico" (parlato) e/o "grafico" (scritto) e di un *significato* a sua volta distinto dal punto di vista della "forma" (come classe "sintattica": grammatica, morfologia e sintassi) e della "sostanza" (come classe "semantica"). L'analisi statistica, secondo i cosiddetti *formalisti*, è condotta "a prescindere dal significato delle unità di testo".

Il *senso* (significato/accezione) di una parola è determinato dalle parole che la circondano (asse *sintagmatico*), ma anche dalla selezione delle altre parole che possono rimpiazzarla nella stessa frase (asse *paradigmatico*); ossia dall'insieme delle parole che possono essere sostituite fra loro nel sintagma, senza modificare la struttura dell'enunciato, poiché "funzionano" in maniera equivalente (Martinez, 2003). Il *senso* sottostante un testo/discorso, di cui s'intende dare una rappresentazione con metodi statistici, è costituito dal sistema dei significati che "si tiene" (come una sorta di ecosistema) sulla base dell'insieme delle co-occorrenze nell'intero corpus di dati testuali.

J.P. Benzécri (*Addad,* 1981), A. Salem (*Lexicloud,* 1987) e M. Reinert (*Alceste*, 1986-2003), con i loro software, mostrano che partendo da un'analisi puramente formale si arriva a cogliere la struttura del *senso* presente nel corpus di testi. Da un'analisi di tipo *paradigmatico*, in cui le parole sono listate in un qualche ordine (alfabetico, inverso, lessicometrico), si può ottenere una rappresentazione della struttura *sintagmatica* presente nel testo. L'ambiguità insita nel linguaggio viene risolta attraverso l'analisi complessa di *grandi matrici di dati testuali* grazie ai metodi e alle tecniche di *analisi multidimensionale* (analisi delle corrispondenze, cluster analysis, analisi discriminante, multidimensional scaling). Tali analisi, misurando la

similarità di profili lessicali, producono rappresentazioni contestuali dell'informazione testuale che si traducono in visualizzazioni nelle quali vale il principio gestaltico "vicinanza *vs* somiglianza" delle unità lessicali che consente di coglierne *l'accezione interna* al corpus investigato.

Attraverso un'analisi fattoriale, ad esempio è possibile in alcuni casi ricostruire dei sintagmi latenti o "frasi modali" (Bolasco, 1999), utilizzabili come veri e propri *modelli di senso* del contenuto del testo. Un altro esempio di utilizzo di assi semantici latenti è quello utilizzato nell'approccio detto *semiometrico* (L. Lebart *et al.* 2003): a partire da un set di 200 "parole-stimolo" ad alto contenuto simbolico, è possibile posizionare un campione di intervistati secondo alcune dimensioni di senso ricostruibili stabilmente nelle culture occidentali, molto utili nelle analisi di marketing.

Accanto a questa tradizione statistica di tipo "formalista", negli stessi anni, alcuni linguisti di tradizione harrisiana sistematizzano la formalizzazione linguistica di particolari classi di parole (ad esempio tavole dei verbi (Gross, 1968; Elia, 1984), di forme composte (avverbi, preposizioni e gruppi nominali) e sviluppano strumenti concreti di *lessicografia e linguistica computazionali*[10], quali dizionari elettronici e automi/trasduttori a stati finiti per la descrizione di grammatiche locali (si veda *Intex*[11]: Silberztein 1993; Fairon 1999; Vietri, Elia 2001). I linguisti *quantitativi*, cimentandosi nei primi tentativi di *lemmatizzazione automatica*, mettono a punto nuovi *lessici di frequenza*. In Italia, grazie ad un lemmatizzatore dell'IBM, T. De Mauro costruisce un prototipo di vocabolario elettronico della lingua italiana (Veli) e il lessico dell'italiano parlato (Lip).

Nella tradizione anglosassone, J. Sinclair (1991) e D. Biber (1998), autorevoli esponenti della *Corpus Linguistics*, propongono un approccio *corpus-based*, orientato all'analisi di vasti databases di esempi reali di linguaggio memorizzati su computer, dal quale trarre gli usi del

---

[10] Per una panoramica sugli sviluppi più recenti di queste aree di ricerca e relativi strumenti, si veda l'interessante contributo di Isabella Chiari (2004).

[11] Oggi trasformato in "Nooj": www.nooj4nlp.net

linguaggio scritto o parlato. Per la messa a punto di corpora di riferimento annotati si rimanda agli esempi riportati in nota[12].

In parallelo a questi contributi, nell'ambito della statistica testuale, cresce l'attenzione a considerare un'unità di analisi di tipo misto che ho chiamato *forma testuale* (forma/lemma/poliforme) (Bolasco, 1990) e che potremmo dire una *lessia* nel senso di B. Pottier (1992), come particella minimale di senso, ossia un'unità non più indipendente dalla lingua.

Nasce così un *approccio lessico-testuale*, nel quale è riconosciuta migliore una unità d'analisi di tipo "flessibile", come può essere appunto una *lessia* (semplice: <carta>; composta: <carta geografica>; complessa: <carta di credito>), che comprenda sia forme grafiche sia espressioni, ogni qualvolta queste ultime rappresentino delle unità minimali – atomi di senso – in grado di catturare il giusto significato. In questo caso, il parsing del testo è svolto ora per parole ora per *polirematiche*[13], come certi gruppi nominali di tipo Nome_Aggettivo (lavoro nero, carta bianca, economia sommersa), Aggettivo_Nome (terzo mondo, estratto conto, ampio respiro) o Nome_Preposizione_Nome (ordine del giorno, capo dello stato, anni di piombo, chiavi in mano) il cui significato è non composizionale, ossia diverso dalla somma dei significati elementari delle parole componenti. Le polirematiche e le locuzioni grammaticali (avverbiali, preposizionali, aggettivali) – una volta isolate – permettono di abbassare drasticamente il livello di *ambiguità* delle singole parole, prima della loro lemmatizzazione. Al fine di selezionare le espressioni più ricorrenti, viene messo a punto un lessico di frequenza anche di poliformi a partire da un corpus di testi di italiano standard (Bolasco, Morrone 1998).

---

12 Per un riferimento generale cf. http://helmer.hit.uib.no/corpora/sites.html; vedi anche WebCorp: http://www.webcorp.org.uk. Per un esempio di italiano televisivo cf. http://www.sspina.it/cit/annotazione.htm che rispetta gli standard della Text Encoding Initiative (TEI), nata nel 1987 in seno a tre associazioni accademiche che si occupano del rapporto tra studi umanistici e informatica (Association for Computers and the Humanities, Association for Computational Linguistics, e Association for Literary and Linguistic Computing). Nel 1994 la TEI ha pubblicato la prima versione delle sue Guidelines (P3) e nel 2000 la nuova versione (P4), compatibile con il linguaggio XML. Per l'italiano parlato, infine, si veda anche http://languageserver.uni-graz.at/badip/badip/home.php

13 cf. Bolasco (1999, p. 196).

3. Una filiera per l'analisi automatica dei testi

Per dare una adeguata rappresentazione del corpus, dopo il *parsing* del testo secondo un'opportuna unità di analisi, occorrono diversi *step* integrati fra loro in una *filiera*. Pensare a filiere in tale ambito non vuol dire "cristallizzare" le procedure possibili, in un contesto in cui se ne possono concepire infinite varianti, bensì fissare soltanto alcuni passi fondamentali per un'analisi automatica del testo.

Le principali fasi che individuano una filiera "ideale" sono quattro:

A) preparazione del testo, B) analisi lessicale, C) estrazione d'informazione, D) analisi testuale.

A) La fase di preparazione è essenziale per una corretta scansione del testo secondo l'unità di analisi prescelta. Questa fase andrebbe sempre più consolidata, per creare degli standard nel trattamento dei dati testuali, ancora lontani dall'essere comunemente condivisi. Essa consiste in primo luogo nella *pulizia* (definizione del set di caratteri alfabeto/separatori, spoliazione dei formati di gestione del testo (XML o altro) e nella *normalizzazione* del testo consistente nell'uniformare spazi, apostrofi e accenti, riconoscere a priori entità particolari (date, numeri, valute, titoli, sigle, abbreviazioni), nonché nomi, toponimi, società, personaggi o espressioni e locuzioni d'interesse. Per queste ultime, un problema consiste nella loro *fixedness*: la "stabilità", intesa come univocità di significato, non sempre può essere garantita (ad esempio: "una volta" o "a volte" hanno un senso variabile; diverso è il caso di polirematiche come "punto di vista", "carta di credito" che hanno un solo significato).

Ma fanno parte ancora di questa fase "preliminare" i differenti step di *annotazione del testo* che consistono nell'associare meta-informazioni alle parole (Bolasco 1998, 2002). Fra queste: la categoria grammaticale, il lemma di appartenenza, una eventuale etichettatura semantica, possibili tagging di tipo relazionale (quali sinonimie, iper/iponimie o altri link previsti nelle ontologie), il numero di occorrenze nel corpus, alcune caratteristiche morfologiche o altro, tutte annotazioni sfruttabili nelle tre fasi successive. Esistono *software* in grado di gestire questo

livello di meta-informazioni sul testo, in maniera trasparente rispetto alla lettura automatica del testo[14].

B) La fase di *analisi lessicale* fornisce una rappresentazione paradigmatica del corpus: lo studio del suo vocabolario, ossia del linguaggio. È un'analisi di tipo "verticale" in cui la rappresentazione del testo è fatta senza tener conto dello sviluppo del discorso ma solo estraendo le parole come da un'urna, che in questo contesto viene chiamata "*bag of words*". Ricostruire il lessico di un "corpus" vuol dire produrre statistiche sui verbi, avverbi, sostantivi, aggettivi, ossia le principali classi di parole cosiddette "piene" (di contenuto) evidenziandone le più frequenti, ma anche quelle appartenenti a determinati gruppi morfologici (enclitiche verbali unite ai pronomi personali; derivati; esotismi), utili per evidenziare alcune costanti di quel lessico, particolarmente significative. Alcuni esempi sono proposti nel paragrafo 4.

Un ulteriore livello di analisi "verticale" riguarda lo studio delle parole "vuote" (connettivi, preposizioni, congiunzioni, determinanti, interiezioni), degli incipit di frase, della punteggiatura, della lunghezza e struttura della frase o altre analisi d'interesse più strettamente linguistico.

In particolare, con gli strumenti della Statistica, l'analisi lessicale consente una descrizione di alcune costanti del linguaggio, in termini d'incidenza percentuale di alcune classi di parole (*imprinting*) in grado di differenziare i testi originari, di individuarne il livello e il tipo (l'incidenza del vocabolario di base (VdB), la presenza di discorso astratto/concreto, il tono positivo/negativo[15]).

C) La fase di *estrazione di informazione* (Bolasco *et al.*, 2004) costituisce un momento importante dell'analisi di un testo, in quanto porta a concentrare l'attenzione su quella parte del linguaggio che risulta particolarmente significativa. Tale fase è utile per selezionare il

---

[14] SATO: http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_033.pdf;
LEXICAL STUDIO: http://www.synthema.it/english/documenti/Prodotti_LexicalStudio_i.pdf;
TALTAC: www.taltac.it

[15] Cf. Bolasco, della Ratta-Rinaldi (2004).

cosiddetto *linguaggio peculiare*, ossia quel 12-15% di vocabolario in genere più rilevante per condurre l'analisi testuale. Potremmo distinguere due situazioni, rispettivamente generate *con* o *senza* una qualche *query*. L'estrazione di linguaggio peculiare senza l'input di una specifica query può condursi ricorrendo a risorse esogene (mediante calcolo di uno scarto standardizzato d'uso della parola, rispetto alla frequenza d'uso di riferimento in un lessico assunto come modello, ove queste ultime frequenze sono da assumersi come valori attesi) oppure ricorrendo a risorse endogene (mediante calcolo delle specificità) per selezionare il linguaggio specifico di ciascuna partizione (quello dei maschi rispetto alle femmine, o dei giovani/adulti/anziani ecc.). Quando invece si utilizza una query, il calcolo di un indice come il *TFIDF* (Salton 1989; Sebastiani, 2002) permette di selezionare[16] i termini più vicini alla richiesta, al fine di ordinare secondo un principio di rilevanza i documenti ripescati.

A livello di analisi di sequenze, l'estrazione di espressioni tipiche del corpus avviene, a partire dall'inventario dei segmenti ripetuti (Salem 1987), grazie al calcolo di un indice *IS* (Morrone 1993) che filtra i segmenti rilevanti secondo la loro capacità di assorbimento delle occorrenze delle parole componenti[17].

D) La fase di *analisi testuale* riguarda tutte le operazioni rivolte direttamente sul corpus, quindi in grado di fornire una rappresentazione sintagmatica del testo, sia puntualmente attraverso analisi di concordanze più o meno sofisticate a seconda del tipo di query, sia globalmente attraverso analisi di co-occorrenze. Queste ultime possono ricostruirsi sia direttamente dall'analisi statistica delle sequenze

---

16 L'indice *TFIDF* indicato con $w = tf * \log N/n$, ove *tf* è la frequenza del termine in ciascun documento, *n* il numero di documenti contenenti quel termine e *N* il numero totale dei documenti del corpus. Questo indice pondera le parole in funzione della loro rilevanza, ossia tanto più esse sono frequenti esclusivamente in pochi documenti.

17 L'indice *IS*

$$IS = \left[\sum_{i=1}^{L} \frac{f_{segm}}{f_{fg_i}} \right] \cdot P$$

somma i rapporti di composizione delle occorrenze delle *L* parole appartenenti al segmento (f_{segm} frequenza della parola nel segmento e f_{fg} frequenza della parola nel corpus), ponderando tale somma con il numero *P* di parole piene.

(predecessori/successori in un LAG predefinito) rispetto a parole pivot, sia indirettamente mediante ricostruzione di dimensioni semantiche latenti prodotte con tecniche di riduzione dimensionale di tipo: analisi fattoriale delle corrispondenze, *singular value decomposition*, *multidimensional scaling*.

Ma l'analisi testuale, quando non si fa uso di tecniche statistiche multidimensionali, consente di: i) rispondere a interrogazioni complesse sul corpus (analisi di concetti) estraendo i documenti più rilevanti che le verificano; ii) visualizzare le entità di interesse ricercate; iii) categorizzare i frammenti di testo creando nuove variabili "testuali", che poi alimentano campi di un *database* strutturato.

4. Analisi lessicale e analisi testuale

Operazioni dello stesso tipo possono applicarsi sia in analisi lessicale alle unità di testo (parole o lessie) costituenti il vocabolario (V), sia in analisi testuale alle unità di contesto (documenti o frammenti del "discorso") costituenti il corpus come insieme totale delle occorrenze (N). La Tabella 1 illustra in parallelo queste "analogie" nei due tipi di analisi: dalle operazioni di base o di Text Mining alla ricerca di concordanze, dall'utilizzo di meta-informazioni alla estrazione d'informazione con risorse sia interne sia esterne, dagli output primari frutto delle suddette investigazioni agli output secondari utili per successive analisi statistiche multidimensionali.

Mostriamo in questo paragrafo alcuni esempi dei due tipi di analisi, tratti da nostre precedenti ricerche. Una di queste riguarda uno studio sul lessico eno-gastronomico svolto a partire dall'analisi delle Guide dei Vini (GVi) e dei Ristoranti (GRi) del GamberoRosso[18]. Altri esempi provengono dallo studio di dieci annate del quotidiano "*La Repubblica*",

[18] Il corpus nel complesso è formato da oltre 700.000 occorrenze (*tokens*, N); in particolare, la GRi comprende 320.000 tokens e la GVi 380.000. Il vocabolario complessivo è di 35.000 parole diverse (*types*, V), di cui oltre 10.000 sono nomi di luoghi, aziende, persone e prodotti (Bolasco & Bolasco 2004). Uno studio sui messaggi pubblicitari di vini italiani pubblicati sulla rivista GamberoRosso (1992-1994) è apparso sul primo numero di questa rivista (Balbi, 1998).

raccolte in un corpus denominato "Rep90"[19] che è servito di base per la costruzione delle risorse statistico-linguistiche presenti nel software Taltac (www.taltac.it; Bolasco 2002).

Tabella 1 – Sinottico sulle caratteristiche proprie dell'analisi lessicale e dell'analisi testuale (*)

Tipo di analisi ==>	**Analisi lessicale**	**Analisi testuale**
Livello di analisi	paradigmatico ("verticale")	sintagmatico ("orizzontale")
Ricerche su	vocabolario	corpus
Unità di analisi	unità di testo: "Parole" -- > Lessie (ULT)	unità di contesto: "Frammenti" / Records, Documenti
Operazioni di base	categorizzazione grammaticale, lemmatizzazione fusioni per classi di unità di testo, imprinting ponderazione: dispersione, uso, TFIDF	etichette / annotazioni sulle singole occorrenze individuazione di sequenze, di strutture disambiguazioni ponderazione: TFIDF
Text Mining	query semplici query predefinite (complesse), piani di lavoro (insiemi di query predefinite) query per tipi/classi di unità di testo	Information Retrieval (recupero dei frammenti che verificano la query) Information Extraction (visualizzazione delle unità di testo oggetto della query nei frammenti selezionati)
Ricerca di concordanze	semplici - IR sul vocabolario (per disambiguare le parole) per tipi, classi o gruppi di unità di testo	ricerche *full text* di parole o entità d'interesse (date, numeri, valute, misure, ...) ricerche *full text* di *entità note* (nomi, toponimi, società, …)
Utilizzo di meta-informazioni	categorizzazione delle ULT da tagging grammaticale / semantico	categorizzazione dei frammenti da dizionari tematici e da regole
Estrazione di informazione con : risorse interne	parole rilevanti nel vocabolario da TFIDF parole caratteristiche in una partizione da analisi di specificità	frammenti rilevanti da TFIDF (IR) rispetto all'intero vocabolario o a specifiche query (forme selezionate)
risorse esterne	linguaggio peculiare da lessici di frequenza incidenza d'uso del vocabolario di base "terminologia" da dizionari tematici (positivo/negativo, cibo ecc.)	categorizzazione dei frammenti da dizionari o da regole con popolamento di campi di un DB tradizionale
Output primari	indici / liste con ordinamento alfabetico, lessicometrico, inverso …	ricostruzione del corpus "annotato" con etichettatura grammaticale / semantica
Output secondari su matrici per analisi multidimensionali	matrice "forme x testi" (da partizione del corpus in sottoinsiemi di frammenti secondo variabili categoriali)	matrice "frammenti x forme" con filtri su singole sezioni del corpus (sub-corpus) o sulle forme (selezionate secondo un criterio predefinito) matrice "forme x forme" di co-occorrenze semplici o "pesate"

(*) la maggior parte di queste funzionalità sono presenti nel software TALTAC_2 (www.taltac.it)

19 Cf. Balbi, Bolasco, Verde (2002); Bolasco, Canzonetti (2005); Bolasco (2005).

4.1 Le parole più frequenti

Per quanto riguarda l'analisi lessicale, un primo screening viene svolto di solito sui termini più frequenti (a livello di lemmi), distintamente per singole "parti del discorso". Nello studio sul linguaggio eno-gastronomico, l'analisi degli aggettivi evidenzia un eccesso di qualificazione. Come è naturale aspettarsi in una Guida, vi è una marcata tendenza alla positività (*buono, ottimo, grande, bello* sono gli aggettivi più frequenti) e, nella GVi, una multi-aggettivazione che arriva fino a raccogliere 5-6 qualificazioni intorno ad un solo sostantivo[20]. La forte concomitanza di superlativi (forme in "–issimo" o associate al "molto") e l'abbondanza di avverbi (*davvero, leggermente, decisamente, rigorosamente*) conferma questi eccessi.

La lista dei sostantivi più frequenti rivela non poche sorprese, soprattutto quando si osserva la graduatoria delle citazioni dei cibi o dei piatti. Si scopre che le guide specialistiche dell'eccellenza raccontano di carne e pesce secondo la logica del mangiare "la domenica", che la carne più frequente è l'*agnello*, che il *risotto* è citato più delle *tagliatelle*, che *gamberi, tonno e scampi* sono più frequenti di *spigola* e *orata*, che il *tortino al cioccolato* è diventato fra i *dolci* il più comune. In buona sostanza che al ristorante si va seguendo categorie ben precise: cose "speciali", inseguendo le mode, preferendo il *pecorino* al *parmigiano* ecc., e nel contempo aspettandosi la cura e le attenzioni che si trovano "in casa".

Per i verbi si rimanda per brevità alla Tabella 9, dove i lemmi più frequenti sono raffrontati a quelli peculiari.

4.2 Le costanti del discorso

L'*imprinting* delle principali classi grammaticali nelle guide enogastronomiche è netto e chiaro: rispetto al linguaggio comune, (Tabella 2) esistono quasi il doppio di aggettivi (in occorrenze), quasi il 50% in più di sostantivi (a questi vanno aggiunte le citazioni di nomi di

---

[20] Un esempio: "*Il Gewürztraminer convince per le sue note varietali, è minerale e intenso, fine e complesso, molto elegante e lungo; un grande vino*".

persone, luoghi, vini e prodotti che raggiungono un terzo [11.280] dell'intero vocabolario utilizzato) e un conseguente sottoutilizzo di verbi (dovuto al linguaggio proprio di una "scheda").

Tabella 2 – Imprinting delle principali classi grammaticali: un confronto del lessico eno-gastronomico con l'italiano comune (lessico della stampa) per varietà di forme (types) e per occorrenze (tokens).

	Less stampa		2 Guide GRi e GVi				
	types	tokens	types tot		tok tot	tok GVi	tok GRi
Cat gramm	%	%	v.a.	%	%	%	%
A	16,5	5,1	2.759	16,8	**9,2**	**10,6**	7,4
AVV	2,6	3,3	457	2,8	3,2	**3,6**	2,7
N	35,8	31,6	7.708	46,9	**43,7**	39,5	**48,5**
V	44,5	25,7	5.021	30,6	**13,1**	14,9	11,0
PREP, CONG ...	0,7	34,3	479	2,9	30,8	31,4	30,2
totale	100,0	100,0	16.424	100,0	100,0	100,0	100,0

Fra le costanti più tipiche della Guida dei Ristoranti, come illustrato in Tabella 3, emerge la tendenza a denominare i piatti con alterati, (*lasagnetta, ravioloni*) soprattutto diminutivi/vezzeggiativi (*lumachine, aragostelle, ricottina*).

Tabella 3 – Esempi di alterati dei cibi o "piatti" con occorrenze nella Guida dei ristoranti del GamberoRosso

Primi piatti		Carni		Pesci		Verdura e frutta		Diversi	
gnocchetti	105	maialino	74	calamaretti	78	pomodorini	145	sformatino	48
cavatelli	40	polpettine	47	seppioline	31	insalatina	71	sfogliatina	38
lasagnette	39	straccetti	22	polpetti	16	finocchietto	50	fagottini	31
raviolini	20	nervetti	8	scampetti	15	verdurine	29	frittatina	23
spaghettini	20	mocetta	7	totanetti	9	cannellini	27	crostatina	20
pennette	18	porcetto	6	alicette	4	fragoline	18	croccantino	19
passatina	16	coscette	6	sardelle	3	pomodorino	12	fagottino	19
lasagnetta	15	costicine	6	trigliette	3	caponatina	11	torroncino	16
ravioloni	15	arrosticini	5	aragostine	3	carciofini	9	canestrelli	11
raviolone	8	tordelli	5	aragostelle	3	puntarelle	7	ricottina	11
tagliatelline	8	lombetto	5	granchietti	3	scorzette	6	salsina	7
spaghettoni	8	guancette	5	gobbetti	3	cicorietta	3	frittelline	7
tubettini	6	rognoncino	4	rombetto	2	cicorielle	3	cassatina	7
chitarrina	5	tacchinella	4	tomette	2	fragolina	3	filettini	6
ravioletti	4	lumachine	4	polipetti	2	finocchietti	3	trancetti	5
risottino	3	quaglietta	4	merluzzetti	2	spinacetti	2	rotolini	5
bavettine	3	stufatino	2			bietoline	2	schiacciatina	5
spaghettino	2	guancialino	2			zucchinette	2	fritturina	4

L'incidenza degli alterati è il 30% superiore a quella presente nel linguaggio comune (rispettivamente il 4,6% contro il 3,5%).

L'abbondanza di derivati è legato a due aspetti: le variabilità linguistiche territoriali (nomi propri a tutti gli effetti: *tortellino*) o di specie (*calamaretti*); l'uso vezzoso delle specialità, tipico del linguaggio gastronomico di questi anni (*verdurine*, *sfogliatine, passatina*). Questo abuso in una guida gastronomica porta persino a superare le frequenze di *tortino* (310) a quelle di *torta* (290), quando in genere il derivato copre il 20% della forma base (*spaghettini*/*spaghetti*).

Altre analisi per individuare "costanti" statistiche nel linguaggio sono quelle tese a rilevare ad esempio il tono positivo/negativo di un testo. Nelle guide del GamberoRosso risulta molto rara la qualificazione negativa (Tabella 4): un'incidenza del 7% del rapporto NEG/POS degli aggettivi è davvero bassa (Bolasco, dellaRatta-Rinaldi 2004). Si fatica a trovare termini negativi: il primo aggettivo, con frequenza peraltro molto bassa, è "*difficile*" seguito da altri come "*stucchevole, banale, mediocre*".

Tabella 4 – Incidenza della qualificazione negativa su quella positiva

	media		GVi	GRi
	V	N	N(v)	N(r)
% NEG/POS	28,3	7,4	10,8	2,9
negativo	145	703	583	121
positivo	512	9508	5396	4112

4.3 La produttività delle parole

Quando una "parola" è molto frequente in un corpus è altamente probabile che la sua produttività morfologica in quel testo sia elevata. Per *produttività* s'intende la capacità di generare una varietà di forme a partire dal suo lessema o radice (Figura 1).

Nel grafo si illustra il caso del lessema <politic_> che, nel corpus "Rep90", produce una varietà di 198 forme grafiche diverse per un totale di 344.930 occorrenze. Di queste il 99,4% riguarda le quattro

forme base (politica/o/i/he) e lo 0,6% le altre formazioni che, espresse in lemmi, si articolano in 128 prefissi (2.530 occ.) e/o 28 suffissi (1.869 occ).

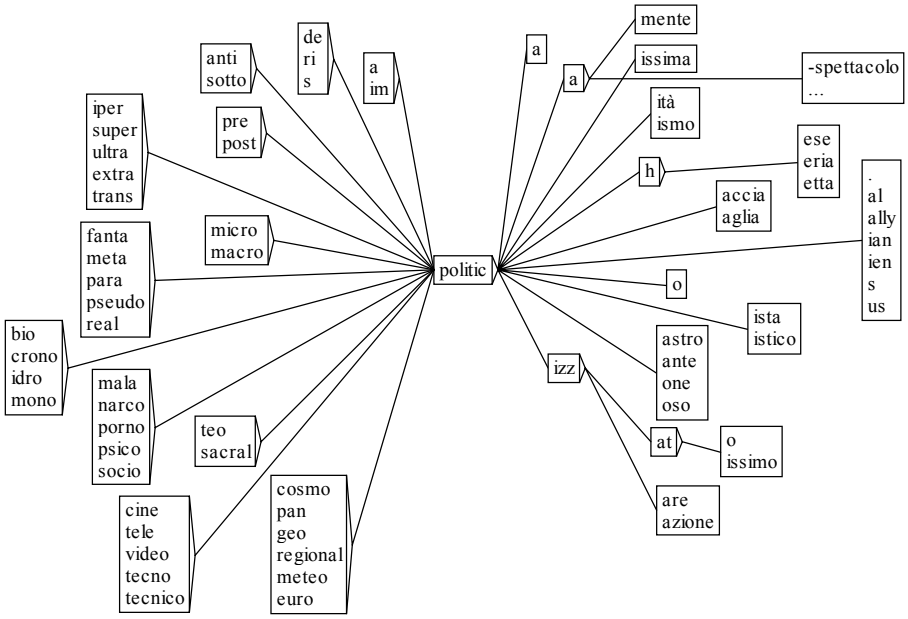

Figura 1. Grafo dei prefissi e suffissi della base <-politic-> in Rep90.

Talvolta lo studio della produttività linguistica dei nomi propri, ad esempio di personalità politiche, può essere particolarmente significativo. In un precedente lavoro (Bolasco, 2005), abbiamo rilevato come la "discesa in campo" di Silvio Berlusconi è stata puntualmente registrata con un picco di occorrenze del 1994 che, a confronto di personaggi nello stesso ruolo, non ha avuto eguali: in "Rep90" 24.000 occorrenze contro le 10.000 in media degli omologhi Dini, Prodi e D'Alema (Tabella 5).

È naturale attendersi quindi un fiorire di derivati e di neoformazioni lessicali incentrate sulla base <-berlusc->. Dalle più comuni derivazioni (-iano, -ista, -ismo, -izza) con/senza prefisso fino a svariate creazioni *ad hoc* (berlusconite, berluscomunista, berlusconcino), come ricostruito in dettaglio nella Tabella 6.

Tabella 5 – Citazioni in numero di occorrenze dei nomi dei Presidenti del consiglio nel quotidiano "La Repubblica" (corpus "Rep90").

NOME	OCC TOT	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Presidenti del Consiglio dei Ministri											
Andreotti	28.521	**6.523**	**6.905**	3.611	3.301	1.145	2.082	1.062	969	601	1.492
Amato	24.868	1.079	909	**5.012**	**4.080**	1.342	1.126	1.138	742	884	2.045
Ciampi	22.673	770	643	853	**4.267**	**2.552**	550	2.387	2.686	2.794	3.244
Berlusconi	77.796	2.464	1.626	2.348	3.773	**23.795**	**14.602**	9.017	7.279	7.358	5.534
Dini	24.501	116	122	218	250	1.786	**10.602**	**5.748**	2.090	1.468	1.329
Prodi	41.308	193	127	274	995	1.029	4.450	**9.175**	**8.416**	**8.453**	6.205
D'Alema	43.304	467	525	501	559	2.367	4.207	5.129	6.627	**7.517**	**10.774**

Questo esempio di linguistica del corpus testimonia quanto Berlusconi costituisca un caso "originale" non soltanto in politica ma anche dal punto di vista linguistico: non è facile infatti trovare un assortimento così ricco di varianti riferite ad una sola stessa base.

Tabella 6 – Produttività morfologica della base <berlusc> *in "Rep90"*

		filo-	anti-					
berlusconiano	1173	17	25	**iper**berlusconiano	2	**berluscones**	52
berlusconiana	903	5	26	**super**berlusconiano	1	berlusca	4
berlusconiani	485	10	16	**ultra**berlusconiana	1	berluscon	4
berlusconiane	269	2	3			berluscone	2
berlusconian	11			**pre**berlusconiana	1	berlusconiser	1
berlusconianamente	2			**neo**berlusconiano	1		
berlusconianus	2			**post**berlusconiani	1	**berluschino**	9
berlusconianissimi	1					berluschini	4
berlusconissimo	1					berlusconini	3
						berlusconina	1
berlusconismo	259	1	16	**social**berlusconismo	1	berlusconcino	1
berlusconismi	2						
berlusconesimo	1					**berlusconidi**	7
berlusconista	7					berlusconide	5
berlusconisti	4			**berlusconità**	2	berluscoide	1
				berlusconite	1		
berlusconizzazione	18			berlusconume	1	**berlusconesca**	1
berlusconizzata	8			berlusconia	1		
berlusconizzarsi	6			berlusconeria	1		
berlusconizzato	5			berlusconeide	1	**altre:**	
berlusconizzante	1					berluscomunista	1
berlusconizzarlo	1			**berlusconare**	1	berlusconcratici	1
berlusconizzati	1			berlusconata	3	similberlusconi	1
deberlusconizzata	1			berlusconese	1	fuoriberlusconi	1
deberlusconizzato	1			berlusconeggiano	3	berluschidolatriche	1

4.4 Analisi delle concordanze

Si tratta dell'esempio più classico ed elementare di analisi testuale. L'analisi delle concordanze è praticata nell'ambito delle analisi linguistiche ben prima degli studi svolti nel dopoguerra da Busa su S. Tommaso d'Aquino. Come noto, essa fornisce l'insieme dei *co-testi* destro e sinistro di una predefinita parola "*pivot*" ed è ancor oggi assai utile per discernere il significato reale di ogni occorrenza di un vocabolo; è quindi quasi indispensabile per la disambiguazione delle forme, sia dal punto di vista grammaticale che semantico. Oggi è possibile fornire *concordanze complesse* operanti su gruppi di parole[21]. La Tabella 7 riporta esempi di *query* applicabili sia a unità di testo sia a unità di contesto. In quest'ultimo caso, il risultato produce l'estrazione dei documenti che le verificano, con la evidenziazione delle singole occorrenze in modalità *fulltext*.

Tabella 7 – Esempi di queries per concordanze su unità di testo (vocabolario) e unità di contesto (documenti del corpus).

A - Query sul vocab.		**B** - Query complessa sui frammenti del corpus (unità di contesto)		
Ricerca del lessema ***auto****		Ricerca dei frammenti contenenti un elemento del concetto: ***parentela***	flessioni	forme attualizzate
auto	1	padre/i OR madre/i OR mamma/e OR babbo/i	8	5
automobile	2	papà OR papa'	2	1
autobus	3	figlio/a/e/i OR figliola/o/e/i	8	6
autovettura	4	marito/i OR moglie OR mogli	4	3
autostrada	5	fratello/i OR sorella/e OR frat(sor)ellino/a/e/i	12	7
autocarro	6	suocero/a/i/e	4	3
autogrill	7	genitore/i	2	2
autodromo	8	nonno/a/i/e OR bisnonno/a/i/e	8	5
autosalone	9	nipote/i OR nipotino/a/i/e	6	6
autotreno	10	zio/a/i/e	4	3
autofficina	11	cognato/a/i/e OR cugino/a/i/e OR cuginetta/o/i/e	12	12
autolavaggio	12	genero/i OR nuora/e	4	2
automezzo	13	parente/i OR familiare/i OR famigliare/i	6	6
		totale flessioni e forme attualizzate	80	61

[21] Ad esempio, nel vocabolario del corpus delle Guide enogastronomiche sono state etichettate 2400 forme relative ad un cibo o piatto su un totale di 38.000 parole. Con una sola query, richiamante l'etichetta "*cibo*", si producono le concordanze delle 2400 forme in questione, consistenti in oltre 98.000 co-testi.

La concordanza complessa su unità di testo relativa al lessema <auto*> (Tabella 7A), nel vocabolario tratto dalle risposte di un'indagine Istat, produce un insieme di 13 parole che permettono di visualizzare complessivamente migliaia di co-testi, relativi a mezzi di trasporto o luoghi.

Una concordanza complessa generata da un set di query può consentire di *analizzare un concetto*. Ad esempio, per cercare in un testo tutte le forme indicanti una *parentela* (*padre, figlia, fratello, …* fino al più generico *parente*) occorre l'insieme di query elementari, descritto in Tabella 7B. Con un'unica espressione regolare[22] che cumula queste query (corrispondenti alle varie figure parentali) da 80 flessioni teoriche, in grado di catturare in un corpus tutte le occorrenze relative al concetto (incluse derivazioni come "*nipotino*, *cuginetto*, *figliolo*"), si ottengono in una sperimentazione su dati di un'indagine Istat 61 forme attualizzate, per un totale di decine di migliaia di occorrenze. Applicando questa espressione alle unità di contesto, vengono estratti i frammenti contenenti almeno una citazione di parentela, che sono così categorizzati rispetto al concetto in questione. È interessante notare come la quota di frammenti nei quali al contrario non viene citato nessun tipo di parentela, ovvero nella fattispecie alcun comportamento che abbia a che fare con qualche familiare, costituisce a sua volta un sub-corpus di frammenti non categorizzati, sul quale poter indagare per interessanti analisi di contenuto specifiche.

*4.5 Ricerca di "*entità di interesse*"*

Cercare ogni entità "impresa" presente in un documento non è banale. Può costituire un buon esempio di ricerca di una *named entity* mediante un criterio *ibrido* (*dizionario* + *regola*). Si supponga di voler ripescare tutte le citazioni di una qualsivoglia impresa in una base

---

[22] L'espressione regolare è la seguente: (mpb)a(dmb)(mrb)(aieo) OR pap(aà)? OR figli? OR figli OR figliol* OR marit(oi) OR moglie OR mogli OR (fs)(ro)(ar)*ell? OR (fs)(ro)(ar)*ellin? OR suocer? OR genitor? OR *nonn? OR nipot? OR nipotin? OR zi? OR c(ou)g*(nt)(aoie) OR gener(oi) OR nuor(ae) OR parent(ei) OR fami*(gl)iar(ei). Una verifica *ex-post* di alcune forme flesse porta ad escludere alcune occorrenze: ad esempio <generi> in realtà figura sempre come "generi alimentari"; al contrario, <genero> non è mai una voce verbale.

documentale: ad esempio quella dei provvedimenti emessi dall'Antitrust sulle concentrazioni. Il corpus in questione (Baiocchi *et al.* 2005), di oltre 3500 provvedimenti, viene sottoposto preliminarmente al riconoscimento di tutte le imprese la cui ragione sociale è citata in maniera completa. Questo avviene grazie ad un "dizionario" di imprese contenente la "ragione sociale" (forma giuridica inclusa) di ogni società, ad esempio: FINDOMESTIC BANCA SPA. Successivamente a questo primo step, si sottopone il corpus all'applicazione di una regola. Dal momento che i testi contengono svariate varianti "incomplete" della ragione sociale di un'azienda, si definisca *A*=incipit, *B*=NOME e *C*="forma giuridica". La regola prescelta sarà pertanto la seguente:

$$(A+B+C)\cup(A+B)\cup(B+C)$$

Infatti nei testi possiamo trovare: *La FINDOMESTIC SPA* o *società FINDOMESTIC ...* o *... FINDOMESTIC SPA*. Ogni nuova impresa riconosciuta dalla regola può alimentare il dizionario, creando l'autoapprendimento del sistema.

Affinché la regola sia efficiente ed esaustiva, occorre inventariare un insieme di possibili *incipit* e di possibili forme giuridiche (previamente normalizzate: senza punti e in maiuscolo) e prevedere che il nome (sconosciuto) sia scritto in una sequenza di caratteri in maiuscolo. Il grafo di Figura 2 illustra la *grammatica locale* su cui è basata la regola.

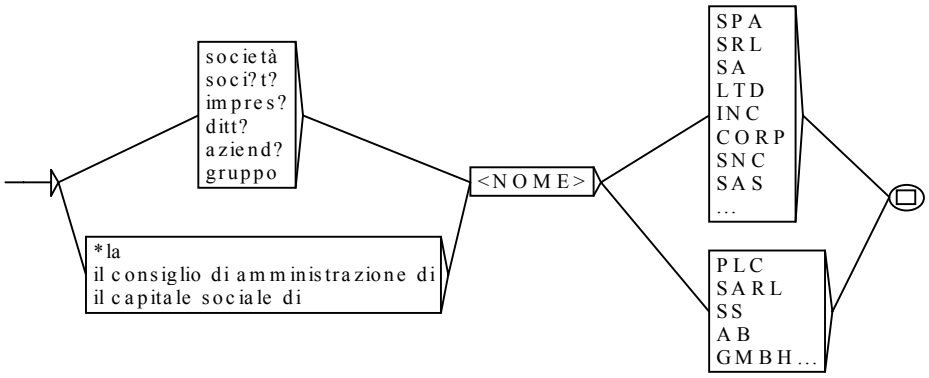

Figura 2 – Grafo per la ricerca in full text dell'entità <impresa>.

4.6 I sintagmi latenti come modelli di senso

A partire dall'analisi di una matrice <frammenti x forme>[23], mediante applicazione di un'analisi delle corrispondenze, è possibile dare una rappresentazione complessiva del contenuto del corpus in una mappa fattoriale, secondo dimensioni semantiche latenti. Gli assi fattoriali, analizzando la similarità fra profili lessicali (vedi nota a piè di pagina), consentono di fare induzioni sui fenomeni investigati. Se vi sono le condizioni di un'alta replicabilità nel linguaggio (Reinert, 2003) e di un numero elevato di micro-frammenti, attraverso la disposizione dei punti sugli assi è possibile ricostruire il discorso "modale", ossia quei sintagmi ricorrenti in molte parti del corpus. È il caso delle risposte libere a domande aperte nelle survey (Bolasco 1999, p. 235). In questo contesto, è interessante riconoscere che il sintagma (come disposizione delle parole in una frase) spesso rivela un aspetto *latente* del discorso: quindi qualcosa che va al di là del contenuto del testo e quindi dà forma a quel processo induttivo che permette di definire l'asse fattoriale in quanto struttura latente (è infatti una combinazione lineare) come il "modello" del senso sottostante al contenuto.

5. Estrazione d'informazione con risorse endogene ed esogene

5.1 Il linguaggio peculiare

Nel lessico della stampa, gli aggettivi più frequenti sono: *elettorale, attuale, economico, televisivo. recente, finanziario, culturale, famoso, straordinario*. In quello delle Guide dei ristoranti e dei vini sono:

---

23 Nel caso di una *survey*, questa matrice potrebbe essere costruita, dall'insieme delle risposte degli intervistati ad una domanda aperta (frammenti o righe della matrice) incrociate con l'insieme delle parole da questi utilizzate (lessie o colonne della matrice). Le parole sono codificate 1/0 rispettivamente se esistenti o meno nel singolo frammento (risposta dell'individuo). La tecnica in questione visualizza la similarità dei profili lessicali fra i vettori riga/colonna, ossia fra le risposte degli intervistati o fra le parole espresse nell'intero campione e tradotte in altrettanti punti sul piano. Tanto più due parole sono vicine nel piano, tanto più "probabilmente" esse vengono associate (co-occorrono) nei discorsi degli intervistati.

buono, molto, ottimo, grande, bello, *piacevole, ricco, elegante, interessante, semplice, gradevole, valido, delicato* e così via.

Entrambi i casi mostrano con evidenza alcuni contenuti prevalenti dei testi di provenienza. Tuttavia ci si può chiedere come demarcare meglio la differenza fra un testo di giornale (A) e un testo di enogastronomia (B), ossia qual è la "peculiarità" del secondo rispetto al primo? Quest'ultima è ricavabile come somma di 2 componenti: gli elementi di linguaggio "in comune", che risultano sovrautilizzati in B rispetto ad A, sommati agli elementi "originali" di B, ossia non presenti in A ma altamente pertinenti in termini dei contenuti di B[24]. La misura del sovrautilizzo è espressa da uno scarto standardizzato fra le occorrenze di B e di A (Muller 1977, p. 49; Bolasco 1999), dove queste ultime sono considerate le frequenze "attese" nel senso classico della statistica. Il linguaggio peculiare rappresenta una sorta di *lessico dei "termini"* del vocabolario di B. In Tabella 8 sono riportate le prime 20 forme in comune con l'italiano standard sovrautilizzate nelle Guide in ordine decrescente di scarto d'uso dal linguaggio di riferimento. Parole come *vini, cantina, etichette* sono frequenti in entrambe le Guide, mentre le altre sono tipiche di una sola delle due.

Alle parole in comune fra A e B vanno aggiunte le forme "originali" di B. Troviamo in ordine di occorrenze decrescenti: *sentori, tannini, antipasti, degustazione, cabernet, sauvignon, merlot, salumi, sangiovese, gamberi, barrique, tortino, ragù* ecc., ossia tutti "*termini*" caratteristici dell'enogastronomia che, pur riconoscibili come parole nel nostro idioma, non si ritrovano tuttavia nel corpus che è stato alla base della costruzione del lessico di frequenza dell'italiano standard[25].

Si noti che parole come "*sentori*" o "*tannini*" hanno centinaia di occorrenze nelle Guide, mentre non è bastato un campione di 4 milioni di occorrenze per considerarle parole diffuse nell'italiano.

---

24 Quest'ultima specificazione è dovuta al fatto che fra gli "originali" possono trovarsi "refusi" ossia parole con errori ortografici o numeri, nomi ecc.

25 Tale corpus è generato da varie fonti (linguaggio scritto/parlato, formale/informale per un totale di 4 milioni di occorrenze) atte a definire l'uso prevalente di parole dell'italiano standard. Il lessico di frequenza che ne è derivato contiene 50.000 forme flesse con almeno 2 occorrenze e tali da avere dispersione non nulla nelle fonti considerate, quindi parole presenti in almeno due testi diversi, ad esempio di linguaggio scritto e parlato (Bolasco, Morrone 1998).

Tabella 8 – Prime 20 forme sovrautilizzate nelle Guide del GamberoRosso rispetto all'italiano standard

Scarto	Parola	Occ_totali	occ_vini	occ_rist
2118,0	vini	3513	1771	1742
1143,3	uve	821	820	1
878,3	dessert	515	12	503
748,7	bonus	822	0	822
712,8	ricotta	418	0	418
695,7	vigneti	408	398	10
690,1	cantina	1405	942	463
652,2	cioccolato	663	99	564
620,6	ravioli	364	0	364
598,4	acidità	351	350	1
596,7	chef	350	1	349
554,1	beva	325	325	0
526,2	verdure	578	1	577
486,9	crema	572	13	559
448,5	formaggi	559	3	556
446,5	tartufo	262	1	261
442,2	ettari	688	687	1
438,0	vigneto	257	255	2
398,5	etichette	642	172	470
392,8	pesce	1087	3	1084

È interessante confrontare le parole più frequenti di un corpus con quelle peculiari. Proponiamo un esempio relativo ai verbi sempre tratto dal linguaggio enogastronomico. La lista per occorrenze decrescenti (Tabella 9) mostra necessariamente, fra i primi 20, lemmi di verbi ausiliari o verbi supporto e soltanto pochi verbi di contenuto specifico. Al contrario, l'estrazione secondo lo scarto decrescente d'uso rispetto all'italiano standard fa emergere dai verbi *peculiari* molti temi ora delle Guide dei vini ora di quella dei ristoranti ora di entrambe (come riportato in neretto in Tabella 9). Ma soprattutto la graduatoria anche di quelli più frequenti, fra i peculiari, è stravolta: *offrire, proporre, assaggiare* cambia in *assaggiare, segnalare, offrire*. È interessante che emergano verbi diversi, altrimenti trascurati come: *gustare, spaziare, donare, impreziosire, rivisitare, apprezzare.*

Per quanto riguarda gli aggettivi, i primi 15 peculiari delle guide individuano caratteri del vino/cibo (*rubino-rosso*, *ripieno, fresco, intenso, cotto, crudo*), della situazione/esperienza (*gustoso, gradevole, piacevole, accattivante, premuroso, godibile*) ed elementi tematici (*enologico, casalingo, gastronomico*). Informazione ben diversa da quella espressa dai più frequenti, visti in precedenza.

Tabella 9 – Confronto fra i lemmi dei verbi più frequenti e i verbi peculiari nel linguaggio eno-gastronomico.

Verbi più frequenti					Verbi peculiari					
Lemma	Forme Flesse	Occ	Guida vini	Guida ristor	Scarto da linguaggio Standard	Lemma	Forme Flesse	Occ	Guida vini	Guida ristor
essere	25	8922	5759	3163	230,91	***assaggiare***	16	**385**	134	**251**
avere	34	1388	904	484	105,87	gustare	13	178	7	**171**
potere	25	1033	401	632	84,34	spaziare	4	133	53	80
trovare	25	528	284	244	69,45	donare	15	148	**144**	4
offrire	12	497	246	251	57,86	***segnalare***	10	**319**	102	**217**
proporre	16	462	164	298	56,20	impreziosire	7	58	**45**	13
venire	17	437	192	245	37,20	rivisitare	8	60	3	**57**
assaggiare	16	385	134	251	36,45	apprezzare	6	168	**109**	59
esprimere	11	352	**340**	12	35,80	prenotare	5	80	0	**80**
fare	35	348	186	162	33,07	sprigionare	4	45	**44**	1
accompagnare	13	347	93	254	32,92	***offrire***	12	**497**	246	251
andare	18	341	176	165	32,74	abbinare	4	49	22	27
seguire	10	319	97	222	30,71	incentrare	4	57	20	37
segnalare	10	319	102	217	30,43	***accompagnare***	13	**347**	93	**254**
ottenere	11	317	**307**	10	28,80	meritare	11	192	69	**123**
chiudere	7	311	132	179	28,35	***proporre***	16	**462**	164	**298**
ricordare	11	300	136	164	28,17	profumare	2	24	22	2
sembrare	11	267	190	77	28,02	***chiudere***	7	**311**	132	179
cominciare	10	260	90	170	27,66	sfiorare	7	112	**110**	2
arrivare	11	235	129	106	26,44	spiccare	3	59	21	38

5.2 Analisi delle specificità

Un altro esempio di estrazione di informazione, è la cosiddetta *analisi delle specificità* (Lafon 1980), che permette di estrarre diciamo pure il linguaggio "peculiare" relativamente alle singole parti di una partizione. Nel nostro caso, se *assaggiare, offrire* sono verbi peculiari per entrambe le Guide (Tabella 9), al contrario altri verbi risultano caratteristici (*specifici*) dell'una o dell'altra. Nelle GRi, infatti, *gustare, provare, scegliere, consigliare, rivisitare, accompagnare, accogliere,* descrivono le azioni dell'ospite, del critico gastronomico o del servizio da parte del ristoratore. Mentre nelle GVi *degustare, donare, esprimere, rivelare, colpire, aprire*, … sono caratteristici poiché descrivono da un lato le proprietà organolettiche percepite dal degustatore e dall'altro rimandano alle fasi di lavorazione del vino (*ottenere, produrre* e così via).

Una lettura d'insieme, basata sulla specificità, di tutti i termini caratteristici delle due Guide ci dice quali sono le discriminanti tra un testo qualsiasi e un testo sull'"*assaggiare*". Dall'analisi emerge che

sono parole caratteristiche (e *originali*) d'un testo sul "*degustare*" un vino: *sentori, tannini, vaniglia, aromi, uve, acidità, beva, rubino, spezie, annata, fruttato, intenso, sensazioni, elegante, piacevoli, colore …* . Mentre quelle di un testo sul "*gustare*" un piatto sono: *antipasti, degustazione, tortino, dessert, ricotta, cioccolato, ravioli, chef, crema, tartufo, etichette, pesce, tonno, coniglio, manzo, balsamico, scampi*.

5.3 Estrazione di neologismi dalla cronologia di un testo

Nell'analisi di un corpus talvolta interessa studiare il ciclo di vita delle parole: si pensi ai discorsi lungo l'intero arco di un processo giudiziario o alla nascita/scomparsa di certe parole seguendo la "cronaca" registrata giornalmente dalla stampa. Salem propose in *Lexico* l'analisi delle specificità cronologiche per estrarre le parole che appaiono o scompaiono lungo una partizione che scandisca periodi temporali (mesi, anni). In un precedente lavoro (Bolasco, Canzonetti, 2005) ci siamo proposti, tramite un indice statistico di studiare il *ciclo di vita* di una singola unità lessicale.

Per stabilire se una parola ha un trend crescente/decrescente si considerano gli scarti tra le occorrenze normalizzate in ciascun anno (Occ_j) e le occorrenze che si avrebbero nel caso di equidistribuzione (ovvero in media, Occ_M), scarti indicati qui sotto come $Scarto\_norm_j$, nonché i prodotti tra gli scarti normalizzati adiacenti[26] costruendo il seguente indice, che denominiamo IT_0:

$$IT_0 = \frac{\prod_{j=2}^{n}\left[\left(Scarto\_norm_j\right)*\left(Scarto\_norm_{j-1}\right)\right]-1}{2}$$

dove n è il numero degli anni dell'intero periodo considerato. I valori possibili di IT_0 sono –1 e 0, che discriminano rispettivamente i trend crescenti/decrescenti dai trend "misti". IT_0 viene poi moltiplicato per $(Occ_1 - Occ_n)/(Occ_1 + Occ_n)$, al fine di ottenere l'indice IT vero e proprio

26 Ciò permette di individuare gli "attraversamenti della media". Questa produttoria dà luogo ad un valore 1, se il numero di attraversamenti della media è pari, e ad un valore –1, se il numero di attraversamenti della media è dispari.

con un campo di valori variabile senza soluzione di continuità fra –1 e 1. Nei suoi valori limite, l'indice IT evidenzia rispettivamente i *neologismi* (IT = 1) e gli *obsolescenti* (IT = –1), mentre nei suoi valori intermedi evidenzia varie tipologie di trend[27].

Attraverso i valori dell'indice IT, è possibile creare una graduatoria delle parole – ponderandole per intensità all'interno della tipologia di appartenenza – in funzione del gap di frequenza fra inizio e fine periodo. Limitandoci qui in Tabella 10 ad evidenziare solo neologismi e obsolescenti, per i quali è possibile anche ricavare l'anno di nascita/morte[28], riportiamo alcuni esempi dall'analisi del corpus "Rep90".

Tabella 10 – Esempi di neologismi/obsolescenti (in forme grafiche) secondo l'anno di inizio/fine del ciclo di vita in "La Repubblica" nel periodo 1990-2000.

	Neologismi	Obsolescenti
1991	ceceno, G8, politically, cossighiani, picconatore, scafisti, tlc	
1992	www, cd-rom, e-mail, on-line, clintoniano, euroscetticismo, tangentisti, transgenici, ciberspazio	
1993	airbag, coordinator, pentium, snowboard, outsourcing, inciuci, satanisti, cartolarizzazione	antirachena, effetto-golfo
1994	forzista, dalemiano, mediatici, creatina, governance, piercing, html, http, multiplex	mediocrediti, superfortezze, motocorazzata, antisandinisti, dopoborsa, bushismo
1995	diessino, buonismo, buonista, prodiano, taliban, ematocrito, provider, browser	interaraba, anticraxiani
1996	ulivista, premierato, diniani, dipietristi, gabbianella, contendibilità	demoproletari, forzanovista, aspromontano
1997	riccometro, sanitometro, antiproporzionale	supercannone, poll-tax, eurolira, padrinaggio, nicaraguensi, gavianei, coupons, sandinismo
1998	kosovaro, e-commerce	dopolistino, stairs, reaganomics, polimeri, kolkhoz, cheque, narcotraficantes
1999		forlaniano, sandinista, assegnatario, governo-ombra, eurolire, vicedirezione

Questa analisi in realtà evidenzia non sempre dei neologismi veri e propri (1995: *buonismo*), ma solo "neologismi" citati la prima volta nel

[27] In particolare, l'intervallo –1 < IT < 0 i trend in declino; mentre l'intervallo 0 < IT < 1 i trend in crescita. Dunque i trend sono crescenti quando l'indice è positivo, e decrescenti quando è negativo (ove entrambi i tipi di trend possono avere andamenti non necessariamente monotòni).

[28] Cfr. Bolasco 2005, p. 346.

quotidiano "La Repubblica" (1991: *ceceno*) nel periodo considerato (anni Novanta).

6. Applicazioni e sviluppi del Text Mining

In questi ultimi anni, suscita molto interesse, nell'area della Statistica che riguarda l'analisi dei dati testuali, un nuovo indirizzo noto con il termine di *Text Mining* (TM).

L'incessante crescita delle risorse informatiche dimostra che ogni 2-3 anni le *dimensioni dei testi* analizzabili con un personal computer si decuplica[29]. Con queste prospettive "esponenziali" di crescita, solo lo studio in profondità del significato del testo può dare robustezza all'analisi automatica del testo.

Un tale obiettivo è una realtà praticabile per pochi[30], ma lo sarà di più in futuro a patto che aumentino le risorse linguistiche condivisibili. Oltre ai dizionari elettronici, occorre costruire *basi di conoscenza* (wordnet: http://www.cogsci.princeton.edu/~wn/), *dizionari multilingue* per la traduzione automatica (eurowordnet: http://www.illc.uva.nl/EuroWordNet/), *thesauri*, ed in qualche caso anche *ontologie*, indispensabili a rappresentare domini particolari. Per la costruzione di queste ultime, contributi significativi derivano dai lavori sull'*Information Extraction* (fra gli altri: T. Poibeau 2003; M.T. Pazienza 2003).

In questo contesto, dalla metà degli anni '90, si sviluppano "soluzioni" di *Text Mining* che servono a far fronte all'eccesso di

---

29 Per esperienza diretta, nel 1995 analizzavo un corpus di 400mila occorrenze (discorso programmatico di governo), nel 1998 di 4 milioni (corpus di un campione di italiano standard), nel 2000 di 25 milioni (l'annata di un quotidiano) e nel 2003 di oltre 250 milioni di occorrenze ("Rep90"; Bolasco, Canzonetti 2005). Quest'ultimo corpus produce un vocabolario di oltre 1 milione di forme grafiche diverse (non tutte necessariamente parole) e un inventario con 4,5 milioni di segmenti ripetuti (non tutti poliformi), a soglia di 20 occorrenze: un'immensa miniera di dati su cui sviluppare la linguistica del corpus. È evidente che una ricerca in Internet può fondarsi su corpora ancor più vasti.

30 Dal centro ricerche IBM di Pisa sono nate negli anni 1980-1990, a livello d'industrializzazione della lingua, società (Synthema, Expert System e Celi) in grado di sviluppare risorse assai costose (valutabili in alcune decine di anni-uomo) per l'elaborazione del linguaggio naturale (NLP).

informazione. Si tratta di tecnologie e procedure utili soprattutto alle aziende/istituzioni (Bolasco *et al*. 2005) che mettono in concatenazione operazioni di *Information Retrieval* e *Information Extraction.* Tali tecnologie, tendenti a catturare la sola informazione rilevante presente nei testi, integrano in maniera intrinsecamente interdisciplinare metodi statistici propri del Data Mining (DM) e tecniche di Intelligenza Artificiale, al fine di creare, a partire da fonti non strutturate, conoscenza utilizzabile in svariati settori dell'attività produttiva. Per informazione *rilevante* s'intende quella parte significativa del testo riutilizzabile al momento opportuno, in quanto pertinente rispetto a specifiche *queries* o ricerche d'interesse.

Una procedura di TM prevede, in genere, sia la individuazione ed l'estrazione automatica dai testi di argomenti inerenti concetti predefiniti, di nomi di persone, società, luoghi, città e altre "*named entities*", nonché di numeri, misure, sigle o altro; sia la categorizzazione dei documenti e l'archiviazione delle informazioni estratte in un database strutturato per successive fasi di utilizzo. Ciò presuppone, nel caso di una azienda, l'esistenza di un *document warehouse* (DW) come corpus sul quale investigare (Sullivan, 2001). L'interesse è *trasformare l'insieme dei testi non strutturati in un insieme di dati strutturati*, allocati successivamente in un database tradizionale.

Una *filiera* di text mining prevede i seguenti passi:

A) Fase di *pre-processing* dei testi (in cui prevale l'Informatica) consistente nel reperimento dal web o da Intranet delle fonti dei testi (es.: news o articoli di stampa, contenuto di siti web, messaggi, chat, forum o altre basi documentali), nella loro formattazione (es. trasformazione in XML) e nella costituzione del *document warehouse*.

B) Fase di *lexical processing* (in cui prevale la Linguistica) consistente nel riconoscere i vocaboli (con uso di dizionari e basi di conoscenza, reti semantiche, sensigrafi o altro), individuare parole chiave o concetti già noti (con uso di regole e di ontologie), effettuare lemmatizzazioni (riconoscimento delle principali parti del discorso, soprattutto sostantivi, aggettivi e verbi). Questa non è una fase necessaria a tutte le applicazioni, perché a volte non viene effettuato un trattamento linguistico del testo.

C) Fase di *Text Mining* vero e proprio (in cui la Statistica e le tecniche di Data Mining hanno un ruolo cruciale) consistente in uno o più dei seguenti passi:

1. *Categorizzazione automatica* di documenti per recupero successivo d'informazioni,

2. *Ricerca di entità* (*termini*) in testi anche *multilingue*, quindi anche indipendentemente dalla lingua di origine dei termini (ciò presuppone la disponibilità e l'allineamento di specifiche risorse linguistiche nelle diverse lingue investigate [31]).

3. *Interrogazioni in linguaggio naturale*, interpretato da processi di NLP basati anche su algoritmi di intelligenza artificiale.

In generale, le soluzioni di Text Mining per l'estrazione d'informazione rilevante fanno uso, dal punto di vista statistico, dei seguenti tipi di procedure:

1) Categorizzazione e classificazione automatica di documenti articolata attraverso la:

i) identificazione delle tematiche principali dei documenti;

ii) individuazione di relazioni fra entità di interesse rilevanti ai fini della gestione della conoscenza, e popolamento con le relative informazioni di campi di un database strutturato, dal quale procedere per ulteriori successive analisi;

iii) classificazione dei documenti in classi precedentemente definite (classificazione *supervised*).

2) Processi di clusterizzazione dei testi basati sulla similarità del vocabolario (classificazione *unsupervised*), per ricavare tipologie utili a individuare aree concettuali o per enucleare "comportamenti omogenei" (ad esempio, tipi di opinioni dell'utenza/clientela intorno a reclami e segnalazioni su prodotti o servizi).

I *campi applicativi* privilegiati nel TM sono:

• *Customer Relationship Management* (CRM): classificazione e indirizzamento automatico delle e-mail, mediante integrazione di tecnologie statistiche di classificazione (basate su parole chiave e/o su concetti)

---

[31] Cfr. F. Neri, R. Raffaelli (2005, p. 71-74).

e tecnologie linguistiche di estrazione della informazione, basate sulla comprensione del testo contenuto nel messaggio.

• *Customer Opinion Survey*: analisi automatica delle segnalazioni e/o reclami pervenuti per telefono o posta elettronica; monitoraggio costante delle opinioni espresse dai clienti in forum di discussione virtuale, come newsgroup e chat; analisi di domande aperte nelle survey quali/quantitative.

• *Gestione delle risorse umane*: controllo della motivazione aziendale a partire dall'analisi automatica delle opinioni espresse dai dipendenti in occasione di apposite rilevazioni; analisi dei curriculum vitae on-line per l'estrazione di specifici skills professionali.

• *Osservazioni sulla concorrenza e sull'utenza*: monitoraggio della situazione del mercato – sia in termini di potenziali clienti che di concorrenti – mediante il reperimento sul Web di liste di aziende, corredate dalle informazioni desiderate; analisi dell'immagine dell'azienda così come emerge dall'esame automatico di notizie e articoli.

• *Technology Watch* e analisi dei brevetti: ricerca e archiviazione sistematica di informazioni sulle tecnologie esistenti per l'identificazione dei settori in maggiore sviluppo; analisi automatica delle informazioni testuali contenute nei brevetti per identificare settori di ricerca emergenti.

• *Analisi di basi documentali settoriali* (economico-finanziarie, giuridiche, epidemiologiche, medico-farmaceutiche ecc.) con estrazione automatica di contenuti, riconoscimento di argomenti e relativa categorizzazione semantica.

• *Natural Language Processing*: costruzione di risorse linguistiche e di basi di conoscenza specifiche (dizionari, grammatiche, reti semantiche) e predisposizione di sistemi per la gestione di interrogazioni in linguaggio naturale, ad esempio nell'ambito di sistemi di *e-government*.

Anche nelle attività di *Intelligence* riguardanti problemi di sicurezza nazionale è sempre più diffuso l'utilizzo di tecnologie di TM. In particolare, ad esempio nelle analisi multilinguistiche di vasti giacimenti di informazioni sul web e nell'identificazione del "parlante" (o autore del testo).

I *settori* maggiormente interessati dal TM sono quelli dell'editoria e dei media (archivi multimediali automatizzati di grandi gruppi editoriali); delle telecomunicazioni, energia e altre aziende di servizi (call-center, portali web per servizi alle piccole e medie imprese); dell'Information Technology e Internet (NLP, risorse linguistiche *on-line*, traduttori automatici); delle banche, assicurazioni e mercati finanziari (CRM, analisi del rischio finanziario e della comunicazione finanziaria d'impresa); delle istituzioni politiche, della Pubblica Amministrazione e della documentazione giuridica (analisi documentale, informazione istituzionale *on-line*, interrogazioni in linguaggio naturale); e, infine, il settore farmaceutico e sanitario (estrazione automatica dei dati da abstracts a contenuto biomedico, gestione dei dati clinici).

Dalle applicazioni di TM finora sviluppate nelle aziende emerge che la messa a punto dei supporti al NLP è fortemente *time consuming*: le basi di conoscenza, le grammatiche locali, le ontologie sono dipendenti dal dominio applicativo e devono essere costruite *ad hoc*. Una volta popolato il database strutturato a partire dal *document warehouse* non strutturato, non sempre in azienda si utilizzano tecniche statistiche di sintesi e ulteriore estrazione dell'informazione, adeguate allo sforzo messo in atto per strutturare l'informazione.

7. Conclusioni

Come si è capito il *Text Mining* è un'applicazione specifica di *Text Analysis* ed in sostanza costituisce solo una delle possibili finalizzazioni di un'analisi testuale in forma automatica. Le procedure e le tecniche di text mining tendono in sostanza a trattare i materiali testuali in formato libero, quindi "dati non strutturati", estraendo da questi informazioni specifiche da riportare in databases tradizionali e quindi creando dati codificati in campi strutturati, dai quali trarre *informazione* che crei valore, nel senso della *business e competitive intelligence*.

La *statistica testuale* riveste una funzione cruciale nel TM per il successo dell'applicazione, ma dipende da quanto, a monte di essa, viene posto in essere per realizzare l'analisi automatica del testo, ossia il riconoscimento in profondità del significato delle parole.

L'ambito scientifico che comprende le applicazioni di *analisi statistiche dei dati testuali*, che chiamiamo appunto statistica testuale, è fortemente multidisciplinare, in quanto per analizzare dati espressi in linguaggio naturale non può prescindere da un adeguato trattamento delle unità di analisi di volta in volta considerate. Questa area statistica necessita quindi di risorse e strumenti offerti dalla linguistica computazionale e dall'informatica, ma al tempo stesso è fortemente intrecciata con la cosiddetta Intelligenza Artificiale, per la messa a punto di alcuni processi di estrazione di informazione.

Tuttavia una tradizione in questo settore è ormai consolidata e testimoniata dai contributi che in ambito europeo sono presentati da 15 anni nelle giornate internazionali JADT (quasi interamente disponibili on-line: http://www.cavi.univ-paris3.fr/lexicometrica/jadt/index.htm). In Italia il settore è particolarmente attivo e una recente testimonianza è raccolta in Aureli e Bolasco (2004). Il futuro è quasi interamente legato ai progressi della capacità d'elaborazione e alla possibilità di disporre in tempo reale di sofisticate informazioni sul senso delle parole, anche grazie allo sviluppo futuro di nuovi modelli statistici.

Riferimenti bibliografici

Anastex S. J. (ed.), (1993), *JADT93 - Actes des Secondes Journées Internationales d'Analyse Statistique de Données Textuelles,* ENST-Telecom, Paris.

Aureli E., Bolasco S. (eds.) (2004), *Applicazioni di analisi statistica di dati testuali*, Casa Editrice Università "La Sapienza", Roma.

Baiocchi F., Bolasco S., Canzonetti A., Capo F. M. (2005), *Estrazione di informazione da testi per la classificazione automatica di una base documentale: la soluzione di Text Mining per l'Authority della Concorrenza*, in S. Bolasco, A. Canzonetti, F. Capo (2005), 45-54.

Balbi S. (1998), Lo studio dei messaggi pubblicitari con l'analisi dei dati testuali, *Quaderni di Statistica*, 1, 155-171.

Balbi S., Bolasco S., Verde R. (2002), Text Mining on Elementary Forms in Complex Lexical Structures in A. Morin, P. Sébillot (eds.) *JADT 2002*, IRISA-INRIA, Rennes, vol. 1, 89-100.

Benzécri J. P. (1963), *Cours de linguistique mathématique*, Rennes: Université de Rennes, Rennes.

Benzécri J. P. (1973), *L'Analyse des données (2 tomes)*, Dunod, Paris.

Benzécri J. P. *et al*. (1981), *Pratique de l'analyse des données - Linguistique et lexicologie*, Dunod, Paris.

Benzécri J.P. (1982), *Histoire et préhistoire de l'analyse des données,* Dunod, Paris.

Biber D. *et al.* (1998), *Corpus linguistics*, Cambridge University Press, Cambridge.

Bolasco S. (1990), Sur différentes stratégies dans une analyse des formes textuelles: une expérimentation à partir de données d'enquête*,* in M. Bécue, L. Lebart, N. Rajadell (eds.) *JADT 1990 Jornades Internationals D'Analisi de Dades Textuals,* UPC, Barcellona*,* 69-88.

Bolasco S. (1998), Meta-data and strategies of textual data analysis: problems and instruments, in Hayashi *et al.* (*eds*.) *Data Science, Classification and related methods*, (proceedings *V IFCS* - Kobe, 1996) Springer-Verlag Tokio, 468-479.

Bolasco S. (1999), *Analisi multidimensionale dei dati,* Carocci Ed., Roma.

Bolasco S. (2002), Integrazione statistico-linguistica nell'analisi del contenuto, in B. Mazzara (ed.) *Metodi qualitativi in psicologia sociale*, Carocci Ed., Roma, 329-342.

Bolasco S. (2005), La reperibilità statistica di tendenze diacroniche nell'uso delle parole, in I. Chiari e T. DeMauro (eds.) *Parole e Numeri - Analisi quantitativa dei fatti di lingua*, Aracne, Roma, 335-354.

Bolasco S., Bisceglia B., Baiocchi F. (2004), Estrazione di informazione dai testi, *Mondo Digitale*, III, 1, 27-43.

Bolasco S., Bolasco M. (2004), Il gusto delle parole: il lessico della critica enogastronomica, relazione al Convegno "*Comunicare il Gusto*", Dipartimento di Sociologia e Comunicazione, Università di Roma "La Sapienza", 19 aprile 2004.

Bolasco S., Canzonetti A. (2005), Some insights into the evolution of 1990s' standard Italian using Text Mining techniques and automatic categorisation, in M. Vichi, P. Monari, S. Mignani e A. Montanari (eds.) *New developments in classification and data analysis*, Serie *Studies in*

Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, Berlin, 293-302.

Bolasco S., Canzonetti A., Capo F. (2005), *Text Mining - Uno strumento strategico per imprese e istituzioni*, Cisu Editore, Roma.

Bolasco S., della Ratta-Rinaldi F. (2004), Experiments on semantic categorisation of texts: analysis of positive and negative dimension, in Purnelle G., Fairon C., Dister A. (eds.), *Le poids des mots, Actes des 7es journées Internationales d'Analyse Statistique des Données Textuelles,* UCL, Presses Universitaires de Louvain, 202-210.

Bolasco S., Morrone A. (1998), La construction d'un lexique fondamental de polyformes selon leur usage, in S. Mellet (ed.), *JADT Proceedings*, Université de Nice, 155-66.

Bortolini U., Zampolli A. (1971), Lessico di frequenza della lingua italiana contemporanea: prospettive metodologiche, in *Atti del Convegno Internazionale di Studi "L'insegnamento dell'italiano in Italia e all'estero",* Vol. 2, Bulzoni, Roma, 639-648.

Brunet E. (1981), *Le vocabulaire français de 1789 à nos jours*, Slatkine & Champion, Genève Paris.

Brunet E. (1988), *Le vocabulaire de Victor Hugo*, Champion & Slatkine, Paris-Genève.

Busa R. (1974-1980), *Index Thomisticus: Sancti Thomae Aquinatis operum omnium Indices et Concordantiae*, Frommann-Holzboog, Stuttgart, 56 voll.

Chiari I. (2004), *Informatica e lingue naturali - Teorie e applicazioni computazionali per la ricerca sulle lingue*, Aracne, Roma.

Chomsky N. (1957), *Syntactic structures*, Mouton & Co., The Hague.

Cipriani R., Bolasco S. (eds.) (1995), *Ricerca qualitativa e computer*, Franco Angeli, Milano.

De Mauro T. (1980), *Guida all'uso delle parole*, Editori Riuniti, Roma.

De Mauro T. (1989), I Vocabolari ieri e oggi, in "*Il vocabolario del 2000*" a cura di IBM Italia, Roma.

De Mauro T., Mancini F., Vedovelli M., Voghera M. (1993), *Lessico di frequenza dell'italiano parlato,* EtasLibri, Milano.

Elia A. (1984), *Le verbe italien - Les completives dans les phrases à un complement*, Shena-Nizert, Fasano di Puglia - Parigi.

Fairon C. (ed.) (1999), Analyse lexicale et syntaxique: le système Intex, in *Linguisticae Investigationes*, Tome XXII / 1998-1999.

Gross M. (1968), *Grammaire transformationnelle du français: 1) Syntaxe du verbe*, Cantilène, Paris.

Guiraud P. (1954), *Les caractères statistiques du vocabulaire*, Puf, Paris.

Harris Z. S. (1964), *Elementary transformations*, TDAP 54, University of Pennsylvania, Philadelphia (ristampato nel 1970 in *Papers in Structural and Transformational Linguistics*, Reidel, Dordrecht, 482-532).

Harris Z. S. (1968), *Mathematical structure of language*, Wiley, New York.

Herdan G. (1964), *Quantitative linguistics*, London, Butterworth & Co. Publishers (traduzione italiana 1971, Il Mulino, Bologna).

Labbé C., Labbé D. (2001), Inter-textual distance and authorship attribution Corneille and Molière, *Journal of Quantitative Linguistics*, 8, 212-231.

Labbé D. (1990), *Le vocabulaire de François Mitterand*, Presses de la Fondation Nationale de Sciences Politiques, Paris.

Labbé D. (2003), *Corneille dans l'ombre de Molière,* Les Impressions Nuovelles, Paris.

Lafon P. (1980), Sur la variabilité de la fréquence des formes dans un corpus, *Mots*, 1, 127-165.

Lafon P. (1984), *Dépouillements et statistique en lexicométrie*, Ed. Slatkine & Champion, Genève-Paris.

Lebart L. (1982), Exploratory analysis of large sparse matrices, with application to textual data, *COMPSTAT*, Physica Verlag, Vienna, 67-76.

Lebart L., Piron M., Steiner F. (2003), *La sémiométrie - Essai de statistique structurale*, Dunod, Paris.

Lebart L., Salem A. (1988), *Analyse statistique des données textuelles*, Dunod, Paris.

Lebart L., Salem A. (1994), *Statistique textuelle*, Dunod, Paris.

Lebart L., Salem A., Berry L. (1998), *Exploring textual data*, Kluwer Academic Publishers, Dordrecht (The Netherlands).

Martinez W. (2003), *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels,* (Thèse de doctorat) Université de Paris 3.

Morrone A. (1993), Alcuni criteri di valutazione della significatività dei segmenti ripetuti, in S. J. Anastex (ed.) (1993), 445-53.

Muller, C. (1973), *Initiation aux méthodes de la statistique linguistique*, Hachette, Paris (ristampa 1992, Champion, Paris).

Muller, C. (1977), *Principes et méthodes de statistique lexicale*, Hachette, Paris (ristampa 1992, Champion, Paris).

Neri F., Raffaelli R. (2005), Una nuova procedura multilingue di Text Mining basata sulla rilevazione della terminologia principale, delle memorie di traduzione e sul clustering, in S. Bolasco, A. Canzonetti e F.M. Capo (2005), 71-74.

Pazienza M.T. (ed.) (2003), Information Extraction, in *The Web Era-Lecture Notes in Artificial Intelligence 2700*, Springer-Verlag, Berlin Heidelberg.

Poibeau T. (2003), *Extraction automatique d'information: du texte brut au web semantique*, Hermes-Lavoisier, Paris.

Pottier B. (1992), *Théorie et analyse en linguistique*, Hachette, Paris.

Reinert M. (1986), Un logiciel d'analyse lexicale: ALCESTE, *Les Cahiers de l'analyse des données*, XI, **4**, 471-484.

Reinert M. (1992), I mondi lessicali di un corpus di 304 racconti di incubi attraverso il metodo "Alceste", in R. Cipriani, S. Bolasco (eds.) (1995), 203-223.

Reinert M. (1993), Quelques problèmes méthodologiques posés par l'analyse de tableaux "Enoncés x Vocabulaire", in S. J. Anastex (ed.) (1993), 523-534.

Reinert M. (2003), Le rôle de la répétition dans la representation du sens et son approche statistique par la méthode "ALCESTE", *Semiotica* 147, **1/4**, 389-420.

Salem A. (1987), *Pratique des segments répétés - Essai de statistique textuelle,* Klincksieck, Paris.

Salton G. (1989), A*utomatic text processing: the transformation, analysis and retrieval of information by computer*, Addison-Wesley, Reading, MA.

Sebastiani F. (2002), *Machine learning in automated text categorization,* ACM Computing Surveys, 34, **1**, 1-47.

Silberztein M. (1993), *Dictionnaires électroniques et analyse automatique de textes - Le système INTEX,* Masson, Paris.

Sinclair J. (1991), *Corpus, concordance and collocation*, Oxford University Press, Oxford.

Sullivan D. (2001), *Document Warehousing and Text Mining - Techniques for improving business operations*, *Marketing and sales*, Wiley, New York.

Vietri S., Elia A. (2001), Analisi automatica dei testi e dizionari elettronici, in E. Burattini e R. Cordeschi (eds.), *Intelligenza artificiale*, Carocci, Roma.

Yule G. U. (1944), *A statistical study of vocabulary*, Cambridge University Press, Cambridge.

Zampolli A., Calzolari N. (1995*),* Problemi, metodi e prospettive nel trattamento del linguaggio naturale: l'evoluzione del concetto di risorse linguistiche, in R. Cipriani, S. Bolasco (eds.) (1995), 51-68.

Zanasi A. (ed.) (2005), *Text mining and its applications to intelligence, CRM and knowledge management*, WIT Press, Southampton.

Zipf G. K. (1935), *The psychobiology of language - An introduction to dynamic philology,* Houghton-Mifflin, Boston, (traduzione francese La psychobiologie du language, Paris, RETZ-CEPL, 1974).