Università degli Studi di Napoli
Federico II

LE BASI DELLA STATISTICA TESTUALE

Michelangelo Misuraca

Dipartimento
di Matematica e Statistica
Università degli Studi di Napoli "Federico II"
via Cintia, Monte Sant'Angelo – 80126 Napoli

Capitolo 1

La statistica e lo studio del linguaggio

Il *linguaggio naturale* è un fenomeno complesso e in continua evoluzione, difficile da analizzare con procedure di tipo automatico. Lo sviluppo di metodologie che consentono il trattamento di dati qualitativi secondo una logica di confronto e non di misura, insieme alle enormi possibilità offerte dall'informatica, ha determinato il potenziamento delle tecniche di analisi dei testi, diffondendone l'uso in contesti disciplinari molto differenziati rispetto a quelli originari.

L'analisi delle informazioni testuali è infatti per sua natura estremamente interdisciplinare, e può progredire solo grazie all'interazione dei differenti ambiti coinvolti. Per tale ragione, è sempre più necessario il contributo congiunto di diversi domini scientifici (principalmente la linguistica, l'informatica e la statistica) ed il confronto dei risultati raggiunti, poiché spesso tali risultati non sono sufficientemente considerati al di fuori dell'area di ricerca in cui sono stati conseguiti.

I progressi registrati nel campo delle tecniche di analisi e dei software "dedicati" suggeriscono di curare sempre più la qualità e la rilevanza teorica delle applicazioni nel campo dell'analisi testuale, utiliz-

zando i risultati in un ambito meno circoscritto di quello sperimentato finora e sviluppando strategie mirate di ricerca.

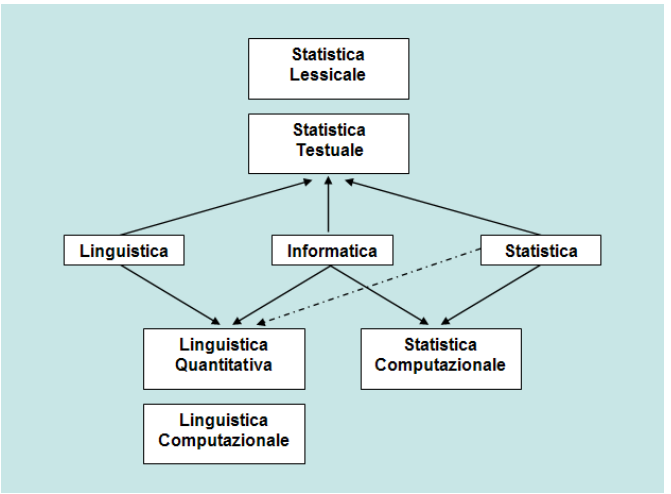

Figura 1.1: Interconnessioni disciplinari nello studio della lingua

L'*Analisi dei Dati Testuali* è un insieme di metodi, in forte progressione, che mira a scoprire l'informazione essenziale contenuta in una raccolta di testi, e può essere applicata in qualsiasi campo per il quale è possibile l'elaborazione di *dataset* di tipo documentale.

Un processo tipico di analisi statistica dei testi si compone di tre fasi: il *pre-trattamento* dei testi, l'*organizzazione* della base di dati, l'*analisi statistica* in senso stretto.

Il pre-trattamento ha lo scopo di trasformare l'informazione testuale in "dato", avvalendosi anche di strumenti linguistici, come dizionari elettronici e lessici di frequenza, per individuare le categorie sintattiche e grammaticali delle parole. Nella fase successiva i testi vengono codificati e quindi organizzati in matrici per permetterne il successivo trattamento statistico.

1.1 Il fenomeno linguistico come oggetto d'indagine

Il linguaggio naturale è la facoltà, esclusiva del genere umano, di esprimere sensazioni, sentimenti, riflessioni, giudizi; di narrare fatti, situazioni; di descrivere particolari aspetti della realtà mediante un *medium* che sia espressione di un determinato livello comunicativo. La lingua è lo strumento di comunicazione maggiormente utilizzato, ed è costituita da un complesso sistema di segni organizzati in struttura.

La scienza che studia tale sistema di segni e di strutture è la *linguistica*, come afferma C. Hagège "*la seule science actuelle dont l'objet coïncide avec le discours qu'elle tient sur lui*" [34].

Le origini di questa disciplina sono quelle di una scienza "prescrittiva", le prime grammatiche difatti si occupavano di descrivere le regolarità ricorrenti di una lingua tralasciando le eccezioni, per poi divenire in tempi recenti una scienza "descrittiva" in quanto si occupa di capire come gli elementi di una lingua si organizzino in sistema.

L'affermarsi della linguistica moderna, con F. de Saussure, sostituisce quindi agli interessi precedenti (la ricostruzione delle famiglie e delle parentele linguistiche, lo studio grammaticale al quale il lessico è subordinato), la visione della lingua come sistema in cui ogni elemento va studiato. Di una lingua possiamo dare una descrizione *diacronica*, ossia della sua evoluzione nel tempo, e *sincronica* ossia del suo stato attuale. Va da sé che per effettuare una corretta indagine diacronica occorre una completa conoscenza sincronica delle due situazioni da confrontare.

Esistono vari livelli di descrizione delle strutture linguistiche, e ognuno di questi livelli è approfondito da una disciplina specifica della linguistica. Le varie teorie concordano sull'esistenza di quattro domini principali: la *fonologia*, che descrive come i suoni di una lingua si organizzano in sistema; la *lessicologia*, che descrive la composizione

del lessico di una lingua; la *morfologia*, che descrive la struttura delle parole; la *sintassi*, che descrive le categorie di parole e le loro relazioni.

Le ultime due discipline, la morfologia e la sintassi, vengono fatte rientrare nel più ampio contenitore della "grammatica" di una lingua, e insieme alla lessicologia sono senza dubbio, in un'ottica di trattamento statistico, le più interessanti.

1.1.1 Alcune annotazioni sul concetto di parola

Ogni parola può essere considerata in relazione al suo significato o al ruolo che riveste nell'articolazione della lingua. Nel primo caso può essere scomposta in *lessema* e *morfema*, nel secondo essere considerata in qualità di *sintagma*.

Il lessema è l'unità di base del lessico e può essere una radice (ad esempio *cant-* in canto, cantare, cantante), una parola autonoma (madre, piede, aratro) o una sequenza di parole fissatasi nell'uso in modo tale che i suoi singoli elementi non possano più essere scambiati né sostituiti con sinonimi (per lo più, dopo cena, mulino a vento).

Il morfema è l'unità grammaticale di base, inteso come il più piccolo elemento di un enunciato che ha un significato. A seconda delle possibilità combinatorie i morfemi si dividono in *liberi*, se possono presentarsi isolati ed avere una propria autonomia di senso, e *legati*, se viceversa non hanno autonomia e di conseguenza non possono restare isolati. Vi sono inoltre i cosiddetti *marcatori sintattici*, morfemi che pur non avendo un significato specifico o definito forniscono informazioni sulla struttura delle proposizioni. Il morfema sintattico ha la sola funzione di indicare qual è il ruolo di un gruppo di parole (o di frasi, o di clausole) in un determinato contesto. Il tipo più importante di marcatore sintattico è la "parola-funzione", che ha molti tratti in comune con il morfema legato. Entrambi indicano tempo o relazione, ma le parole-funzione sono spesso elementi lessicali separati (ad esempio le preposizioni, gli articoli, i pronomi).

Il sintagma (dal greco $\sigma\acute{\upsilon}\nu\tau\alpha\gamma\mu\alpha$, disposizione) è l'unità sintattica autonoma. La sintassi si occupa dei modi in cui le parole si combinano mostrando connessioni di significato all'interno della frase. Per la caratterizzazione sintattica della frase, o di qualunque unità più piccola al suo interno, si parla di "costruzione". Una data entità sintattica può essere considerata da due punti di vista, nella sua interezza, per la funzione che ha isolatamente, o come parte di una unità più ampia.

Nell'ambito della linguistica testuale (che considera come oggetto di analisi un testo e non la frase) l'aspetto sintattico si occupa non della sintassi della frase in senso stretto ma piuttosto delle relazioni che intercorrono tra le unità testuali (frasi, gruppi di frasi, e così via).

1.1.2 Parti del discorso e processi morfologici

I linguisti raggruppano le parole proprie di una lingua in classi che mostrano un comportamento sintattico simile, e sovente una struttura semantica tipica.

Tali classi sono comunemente indicate con il nome di categorie grammaticali o categorie sintattiche, ma con maggior precisione si deve far riferimento ad esse come alle "parti del discorso" (d'ora in avanti POS, dall'Inglese *Part Of Speech*).

In genere è possibile riconoscere due differenti categorie di parole:

- le *categorie lessicali* (o aperte), che rappresentano la classe più numerosa e sono in costante aggiornamento, poiché in esse vi è un continuo processo di acquisizione e coniazione di parole "nuove" (neologismi, barbarismi, ecc.);
- le *categorie funzionali* (o chiuse), con un numero di elementi più limitato ma caratterizzate dal fatto di avere, all'interno di una grammatica, un ruolo ed un utilizzo definito.

In linea generale è possibile distinguere 8 differenti categorie di POS (come riportato nella tabella sinottica in figura 1.2), ma l'evolu-

zione della lingua e le diverse esigenze d'analisi richiedono, sempre più frequentemente, classificazioni ben più "raffinate".

POS Lessicali	**POS Funzionali**
- sostantivi - aggettivi - verbi	- articoli - preposizioni - congiunzioni - pronomi - avverbi

Figura 1.2: Classificazione delle Part Of Speech

Le categorie sono sistematicamente relazionate ai processi morfologici, quali la formazione del plurale di una parola dal singolare, o del femminile dal maschile (e viceversa). La morfologia è importante per il linguaggio naturale perché la lingua è "produttiva": in ogni testo analizzato è possibile infatti incontrare parole o forme flesse di parole non comprese nei dizionari cui si fa riferimento. Molte di queste parole nuove sono comunque morfologicamente connesse a parole note, da cui è possibile inferire le proprietà sintattiche e semantiche.

I principali processi morfologici da considerare sono la *flessione*, la *derivazione* e la *composizione*.

Le *flessioni* sono modifiche sistematiche della radice di una data parola, detta *lessema*, per mezzo di prefissi o suffissi. Tali flessioni non agiscono in modo determinante sulla categoria o sul significato della parola, ma su caratteristiche quali il genere (maschile/femminile), il numero (singolare/plurale) o il tempo (presente/passato/futuro/...).

Il processo di *derivazione* segue un criterio per certi versi meno preciso del precedente, anche se ogni lingua ha meccanismi caratteristici differenti. Il risultato è un cambiamento più radicale della categoria grammaticale e spesso anche del significato e dell'uso della parola. Esempi di derivazione sono la trasformazione dei verbi in sostantivi e aggettivi, e dei sostantivi e aggettivi in avverbi.

La *composizione* è la fusione di due parole distinte in una parola composta avente talvolta significato completamente diverso da quello delle singole parole costituenti[1]. Nella lingua italiana tale fenomeno è meno diffuso che in altre lingue, come ad esempio l'Inglese, e necessita comunque dell'utilizzo di preposizioni e congiunzioni. In generale si definisce *gruppo nominale polirematico*, o più semplicemente *polirematica*, un'espressione linguistica composta non modificabile che ha in un lessico l'autonomia di una parola singola.

1.2 L'utilizzo della statistica nell'analisi del testo

Con l'affermarsi e il diffondersi di strumenti informatici adeguati, sia dal lato dell'hardware che dei software, è stato possibile sviluppare delle tecniche d'analisi della lingua sempre più sofisticate. Gli studi sul linguaggio naturale originariamente intrapresi da linguisti, sociologi e psicologi, sono stati affiancati dal lavoro che informatici e statistici, partendo da prospettive e problematiche diverse, hanno effettuato sui dati testuali.

I due orientamenti della ricerca, qualitativo e quantitativo, si muovono e si evolvono seguendo linee separate, ma non di rado le due metodologie si "confondono" nel tentativo di analizzare il fenomeno

[1]Benzécri riporta ad esempio, in relazione allo studio di testi in Cinese, che la forma *professore* è costituita dalle forme *nato* e *prima*

linguistico in tutti i suoi aspetti.

Gli approcci di Analisi Testuale che si basano su metodologie statistiche fanno riferimento a strumenti di tipo quantitativo per trattare le unità linguistiche contenute in una raccolta di testi. È in particolare alla scuola francese di *Analyse des Donneés* che va il merito di aver determinato un notevole salto di qualità nell'analisi dei dati testuali e di aver prodotto le prime proposte metodologiche.

1.2.1 La logica lessicale e quella testuale

Le prime applicazioni della statistica in ambito linguistico rientravano soprattutto in analisi di tipo letterario, come lo studio delle opere dei grandi autori o l'attribuzione di testi non firmati, che richiedevano complesse operazioni di preparazione del testo.

La *Statistica Lessicale* nasce alla fine degli anni '50. Il Centro Studi del Vocabolario della Lingua Francese di Beçanson aveva portato a termine una classificazione delle opere di Corneille e la loro trasposizione su supporto informatico, presentando tale lavoro nel 1957, a Strasburgo, nel corso di una conferenza con cui si voleva dare avvio al progetto *Trésor de la Langue Française*. La disponibilità di questa incredibile risorsa incoraggiò C. Muller a sfruttarla per effettuare le prime analisi lessicometriche con l'ausilio di strumenti statistici. La logica implicita era che il testo analizzato è un esemplare rappresentativo della lingua: dallo studio di una base di dati testuali è quindi possibile inferire alla lingua stessa alcuni risultati d'indagine.

Parallelamente, negli anni '60, J.P. Benzécri si interessò ai metodi di Analisi dei Dati non come strumento di ricerca in campo psicologico (ambito in cui tali strumenti erano nati e che inizialmente ha dato luogo agli sviluppi più numerosi), ma per l'applicazione degli stessi allo studio della lingua, ponendo le basi alla *Analisi dei Dati Linguistici*. L'idea portante era quella di aprire le porte ad una nuova linguistica, in un'epoca dominata dalla *linguistica generativa*, superando le tesi di

N. Chomsky [20] secondo cui non potevano esistere procedure sistematiche per determinare le strutture linguistiche a partire da un insieme di dati come una raccolta di testi.

Nel tentativo di confutare questo assunto, Benzécri propose un metodo "*avec á l'horizon l'ambitieux étagement des recherches successives ne laissant rien dans l'ombre des formes, du sens et du style*" [11].

Con le prime proposte metodologiche di L. Lebart e di A. Salem negli anni '80, si delinea nei suoi tratti fondamentali l'impianto teorico della *Statistica Testuale*, che a differenza della Statistica Lessicale pone una maggiore attenzione alla *testualità* della base di dati analizzata. La tendenza attuale è quella di una *Statistica Lessico/Testuale* che utilizzi un approccio "integrato", intervenendo *a priori* sul testo oggetto d'analisi e considerando a supporto delle meta-informazioni di carattere linguistico.

1.2.2 I fondamenti della statistica testuale

Volendo trasporre nella terminologia tipica della statistica le entità caratteristiche dell'Analisi dei Dati Testuali, è opportuno individuare innanzi tutto la base di dati oggetto d'analisi. Il *collettivo* analizzato è rappresentato da una raccolta coerente di materiale testuale, detta *corpus*, omogenea sotto un qualche punto di vista oggetto d'interesse.

Questa definizione di *corpus* è applicabile alle fonti testuali più disparate; nel corso degli anni i campi applicativi sono stati numerosi, dalle prime analisi sulle domande aperte contenute nei questionari ai discorsi politici, le annate di stampa periodica, i messaggi pubblicitari, per arrivare al linguaggio utilizzato nei siti Internet. Potenzialmente è possibile applicare le metodologie proprie della Statistica Testuale a qualsiasi ambito o disciplina che preveda l'utilizzo di un linguaggio più o meno specifico.

Il *dato* statistico rilevato è il numero di volte in cui una unità lessicale (detta *occorrenza*) si presenta nella raccolta in esame.

Non è scontato attibuire alle occorrenze di una data parola il significato statistico di *frequenza*, in particolar modo se il *corpus* considerato non è sufficientemente ampio. Per poter effettuare dei confronti tra *corpora* di ampiezza diversa risulta conveniente ricorrere alle *occorrenze normalizzate*, ossia delle frequenze relative ottenute dividendo le occorrenze di ogni parola per una quantità data, variabile in relazione alla dimensione del *corpus* (in genere 10000, 100000 o 1000000).

La *distribuzione statistica* delle parole all'interno del *corpus*, ossia il *vocabolario*, è ottenuta misurando per ciascuna parola il numero di volte in cui si presenta nella raccolta analizzata. L'ampiezza del vocabolario V è definita dal numero di parole presenti nel testo:

$$V = V_1 + \ldots + V_k + \ldots + V_{f_{max}} \tag{1.1}$$

dove V_1 rappresenta il numero di parole che si presentano una volta sola all'interno del testo (*hapax*), V_k il numero di parole che si presentano k volte, e $V_{f_{max}}$ la frequenza della parola con il maggior numero di occorrenze nel vocabolario.

1.3 La scelta dell'unità di analisi

Il modo in cui si rappresenta l'informazione testuale dipende dalla scelta dell'unità di analisi e dalle regole del linguaggio naturale ritenute significative per il riconoscimento e la combinazione delle stesse.

L'unità elementare del linguaggio, la parola, non si presta di per sé ad una definizione univoca. La lingua, infatti, difficilmente potrebbe essere vista come un universo in senso statistico, poiché sfugge a qualsiasi definizione operativa accettabile.

La variabilità del fenomeno "lingua" non è facilmente misurabile e, comunque, l'ampiezza del vocabolario è sensibilmente differente da idioma a idioma: per avere una pur limitata idea basti pensare, ad

esempio, che il verbo *parlare* ha 35 forme flesse, mentre il suo equivalente in lingua inglese, *to speak*, ne ha solo 5. Risulta allora non banale individuare, innanzi tutto, l'unità elementare d'analisi.

1.3.1 Forme grafiche e forme testuali

Una parola è, convenzionalmente, una *forma grafica*, ossia una sequenza di caratteri appartenenti ad un alfabeto predefinito, delimitata da due separatori (segni di interpunzione, spazi, o altri caratteri definiti *ad hoc*). Tale definizione, proprio perché frutto di convenzioni, risulta essere però arbitraria. L'operazione di riconoscimento all'interno del *corpus* delle forme grafiche che lo compongono, conduce ad una perdita di informazione sul significato, i contesti, lo stile, e più in generale di tutti quei fenomeni generati dalla combinazione di segnali linguistici. Le forme grafiche non consentono infatti di individuare, ad esempio, la presenza di sinonimi o antonimi, in particolare questi ultimi qualora espressi per anteposizione alla forma di una particella con valore di negazione.

Il problema della scelta dell'unità di testo si estrinseca quindi nel decidere quale tipo di riconoscimento adottare.

Una scelta opportuna è quella di far riferimento alle cosiddette *unità minimali di senso* nell'accezione data da M. Reinert [60], in modo da limitare le ambiguità delle forme omografe e dei polisenso (cfr. § 1.4) e in generale, nella fase d'analisi, a minimizzare il "disturbo" dovuto a quelle forme che presentano un contenuto informativo ridotto o nullo. Le unità di senso possono tanto essere delle forme grafiche, quanto dei segmenti di testo che esprimono un contenuto autonomo. I *segmenti ripetuti* sono disposizioni di $2, 3, ..., p$ forme che si ripetono più volte all'interno del *corpus*; tali sequenze possono essere *vuote* o *incomplete*, formate cioè solo da parole grammaticali o da parti di sintagmi, oppure *caratteristiche*, se costituiscono unità di senso indipendenti. In tale caso specifico si parla, nell'analisi della lingua, di *poliformi*.

Un particolare tipo di poliforme il cui significato è frutto di un "calcolo non composizionale" è, come visto, la *poliremática*, il cui senso globale non è risultante dalla somma dei significati delle singole forme grafiche componenti.

In generale è possibile considerare l'incidenza differente di due classi di poliformi, ottenute operando una distinzione tra gruppi nominali o verbi idiomatici, e poliformi a carattere grammaticale come ad esempio le locuzioni avverbiali, aggettivali, prepositive e congiuntive. Gli elementi della prima classe più specificamente evidenziano le tematica legate al *corpus* oggetto di studio, e dunque rigorosamente dipendenti del contesto. Gli elementi della seconda classe sembrano relativamente meno legate al contesto generale e sono presenti in modo più diffuso nei differenti tipi di discorso [18].

In un'ottica di Statistica Testuale è opportuno considerare come unità elementare d'analisi la *forma testuale* [17], una componente significativa minima del discorso non ulteriormente decomponibile, sia essa semplice, composta o complessa.

1.3.2 Forme strumentali e forme principali

I testi analizzati hanno una elevata variabilità e contengono numerose forme utili alla comprensione del testo ma prive o quasi di contenuto realmente informativo.

Esiste un'ampia classe di forme che non hanno significato autonomo una volta estrapolate dai contesti ed è pertanto inutile considerare nell'ottica del trattamento statistico. Tali forme, dette *strumentali* (articoli, preposizioni, congiunzioni, pronomi), sono generalmente indicate come "parole vuote" o *stop word*, e rappresentano i cardini di alcuni costrutti lessico-grammaticali. In particolare sono utili a discernere il senso generale del fenomeno analizzato, ma devono essere filtrate per semplificare l'analisi da un punto di vista computazionale, diminuendo al contempo la presenza di rumore nella base di dati.

Le *forme principali*, altresì note come "parole piene", sono invece portatrici di parti "sostanziali" del contenuto di un *corpus*, delle sue modalità di enunciazione o di azione.

La costruzione di un elenco di forme strumentali (*stop list*) è un problema delicato. È impossibile, infatti, compilare un elenco che vada bene per tutti gli scopi: non ci sono particolari problemi con le POS funzionali (si veda la figura 1.2), ma è necessario individuare di volta in volta (a seconda del contesto) quelle forme che risultano "banali", e quindi povere di contenuto informativo. Se si analizza ad esempio un *corpus* costituito da rapporti bancari, la forma *banca* non sarà particolarmente significativa per gli scopi d'analisi; se questa stessa forma è presente in una raccolta contenente articoli di giornale, può consentire la selezione di quelli che trattano argomenti finanziari o economici.

1.3.3 La codifica dell'informazione testuale

Lo schema maggiormente utilizzato per codificare *corpora* testuali in linguaggio naturale è il cosiddetto *Bag-of-Words* (BOW). Tale codifica consente di trasformare ogni documento (o frammento di testo) contenuto nel *corpus* così da strutturare i dati e poterli sottoporre a trattamento statistico.

Ogni documento D_j è visto come un vettore nello spazio delle forme del vocabolario:

$$D_j = (w_{1,j}, w_{2,j}, w_{3,j}, \ldots, w_{p,j}) \tag{1.2}$$

dove ogni termine $w_{i,j}$ è il peso della *i-ma* forma nel *j-mo* documento.

È possibile considerare, a seconda del tipo di analisi effettuata, differenti schemi di ponderazione, anche se nell'analisi dei dati su dataset testuali si preferisce in genere la frequenza della forma, ossia il numero di volte in cui questa è presente nel dato documento (si veda in proposito la sezione § 2.2).

Seguendo lo schema BOW, i documenti sono organizzati in una matrice $\mathbf{T}$, detta *tabella lessicale*, con p righe e q colonne. Sulle righe si trovano le p forme di maggior interesse selezionate dal vocabolario del *corpus* dopo aver effettuato le operazioni di pre-trattamento, mentre sulle colonne vi sono i q documenti considerati. In questa matrice $\{forme \times documenti\}$ (come nell'esempio riportato nella figura 1.3) ogni cella f_{t_i,d_j} contiene l'occorrenza della forma *i-ma* nel documento *j-mo*. Se le occorrenze sono ponderate in accordo a qualche sistema di pesi allora la cella f_{t_i,d_j} non contiene più la semplice frequenza ma una funzione della stessa.

	doc 1	doc 2	doc 3	$\cdots$	doc q
forma 1	1	0	0	$\cdots$	1
forma 2	0	2	1	$\cdots$	0
forma 3	1	0	1	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
forma p	1	0	0	$\cdots$	0

Figura 1.3: La tabella lessicale

Il principale vantaggio della codifica Bag-of-Words è la relativamente bassa complessità computazionale, derivante dal fatto di poter trattare i testi analizzati come vettori e quindi calcolare facilmente delle misure statistiche d'interesse. Per contro, le matrici coinvolte nell'analisi sono matrici *sparse*, ossia con molte celle nulle, poiché non è pensabile ritrovare in ogni documento tutte le forme del vocabolario considerate rilevanti ai fini della ricerca svolta.

Qualora sia possibile registrare, per ogni documento, delle informazioni esterne di tipo strutturale, diverse a seconda del fenomeno del quale si sta analizzando il linguaggio peculiare, è possibile ottenere delle tabelle lessicali *aggregate*. In tali matrici non vengono più consi-

derate le occorrenze delle forme contenute da ogni singolo documento, ma da classi di documenti raggruppati secondo una data variabile di classificazione, limitando in tal modo il numero delle celle nulle.

Di seguito, nel capitolo 2, i problemi connessi alla codifica e alla riduzione della dimensionalità delle tabelle lessicali saranno trattati più approfonditamente, con particolare attenzione alle criticità connesse all'analisi multidimensionale dei dati.

1.4 La problema della disambiguazione

Gli strumenti informatici sviluppati per l'analisi dei dati testuali consentono di effettuare molte delle operazioni di pre-trattamento automaticamente. È necessaria, comunque, una attenta supervisione per evitare che informazioni importanti per l'analisi del fenomeno oggetto d'interese vengano distorte (errata lessicalizzazione di poliformi e poliremathiche, errata lemmatizzazione) o perdute (errata definizione delle stop word).

Uno dei problemi più complessi da gestire nella preparazione della base di dati da analizzare è dato dalla presenza, all'interno dei *corpora* considerati, di forme che presentano una certa *ambiguità*. Questa evenienza ricorre, principalmente, per la necessità di associare ad ogni data forma un significato definito, che sia chiaramente distinguibile dagli altri potenzialmente attribuibili alla stessa.

Le ragioni dell'ambiguità possono essere di natura *lessicale* (o morfosintattica) e *semantica*:

- ⋆ le forme possono essere identiche a livello di rappresentazione lessicale e distinguersi unicamente nei loro contesti sintagmatici, qualora siano forme flesse di lemmi differenti;
- ⋆ le forme possono essere nettamente distinte a livello lessicale ma mostrare relazioni di natura extralinguistica, di tipo semantico o meno, perché possono riferirsi a più concetti differenti.

L'ambiguità può quindi avere diversa natura: può essere dovuta al fatto che talune forme sono identificate da una medesima stringa di caratteri dell'alfabeto o dalla medesima pronuncia (*omonimia*), o che una stessa forma, con un unico significante, può assumere significati differenti a seconda dei contesti in cui è usata (*polisemia*).

1.4.1 Omonimia e Polisemia

L'omonimia è un fenomeno particolare causato sia dai fattori diacronici che dal contatto linguistico (bilinguismo o *diglossia*).

Raramente due codici in contatto presentano un'assoluta parità, è più frequente la dominanza dell'uno sull'altro o sugli altri. Nella lingua italiana l'omonimia si manifesta solo sotto forma di *omografia*, e la maggiore difficoltà che si incontra nell'individuare le diverse forme grafiche sta nel fatto che, a differenza di altri idiomi, gli accenti tonici non vengono indicati convenzionalmente nella grafia delle forme. Esemplare è il caso delle forme *pèsca* e *pésca*: la prima indica il frutto, la seconda l'attività del pescare, ma omettendo l'accentazione non sono immediatamente distinguibili.

La polisemia, uno dei fenomeni più studiati dell'ambiguità lessicale, è frutto dello sviluppo nel tempo di una cultura e della lingua che la esprime: quando una comunità necessita di nuovi segni linguistici per creare nuovi concetti, di rado si rifà a segni totalmente nuovi anche sul piano del significante, ma di frequente aggiunge nuovi significati a significanti preesistenti per mezzo di procedimenti metonimici e metaforici. Nasce dal linguaggio figurato, il quale conferisce a una parola nuovi significati, oppure dall'influsso straniero che può assegnare ad un termine che già esiste un nuovo significato.

È quindi un meccanismo basilare necessario per il buon funzionamento della lingua: se ogni forma avesse solo un significato l'uomo dovrebbe rifornire la propria mente di molti altri termini, tanti quanti sono i significati di cui ha bisogno. Grazie all'esistenza della polisemia

si è in grado di rappresentare i vari significati tramite un'unica forma, realizzando un'economia indispensabile per l'efficienza della stessa lingua e aumentando il potere simbolico del linguaggio.

È possibile rappresentare polisemia e sinonimia (nella quale si ha medesimo significato ma diverso significante) attraverso una matrice $\{significati \times forme\}$, indicata talvolta come *matrice lessicale*.

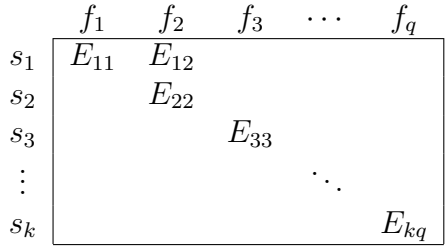

Figura 1.4: Rappresentazione matriciale di polisemia e sinonimia

Dalla matrice lessicale riportata nella figura 1.4, ad esempio, si deduce che la forma f_2 è polisemica, mentre invece le forme f_1 e f_2 sono sinonimi rispetto al significato s_1.

1.4.2 Le tecniche di disambiguazione automatica

Negli ultimi dieci anni, i tentativi di mettere a punto tecniche di disambiguazione automatica si sono moltiplicati, grazie soprattutto alla disponibilità di grandi raccolte di testi in formato elettronico e al corrispondente sviluppo di metodi statistici per identificare regolarità e strutture peculiari nella distribuzione di tali dati.

La disambiguazione automatica è stata oggetto d'interesse da parte degli studiosi fin dagli albori del trattamento computerizzato della lingua negli anni '50. Tale problema è stato definito come "AI-complete", perché può essere risolto solamente chiarendo *a priori* tutti i vincoli dell'*intelligenza artificiale* (AI), quali la rappresentazione del senso

comune e la conoscenza enciclopedica.

In termini generali, la disambiguazione lessicale e semantica (d'ora in poi WSD, dall'Inglese *Word Sense Disambiguation*) comporta, come visto, l'associazione di un determinato senso ad una forma, distinguibile da tutti gli altri significati potenzialmente attribuibili a quella data forma.

Il procedimento tipico comporta perciò necessariamente due passi: la determinazione per ogni forma di tutti i diversi sensi attinenti (per lo meno) al testo considerato, e un mezzo per assegnare ad ogni forma il senso adatto rispetto al contesto in cui si trova.

La definizione precisa di un *senso* è una questione considerevolmente dibattuta nella comunità scientifica. La varietà di approcci ha sollevato di recente la preoccupazione circa la reale comparabilità di molte tecniche di WSD e determinato notevoli difficoltà nella definizione dei sensi. Ad ogni modo, c'è sempre stato accordo sul fatto che i casi di disambiguazione morfo-sintattica e di disambiguazione semantica possano essere trattati separatamente, essendo diversa la natura del problema.

Riguardo agli strumenti con cui assegnare ad ogni forma un senso univoco, ci si affida principalmente a due fonti di informazione, discriminanti rispetto alle famiglie di tecniche da utilizzare:

- ⋆ il contesto della forma che deve essere disambiguata, includendo sia informazioni sul contenuto del *corpus* analizzato sia informazioni extralinguistiche, relative ad esempio alla situazione nella quale i testi contenuti nel *corpus* sono stati prodotti;
- ⋆ fonti informative esterne, quali ad esempio risorse lessicali o enciclopediche, dizionari elettronici, e l'apporto di conoscenza esperta.

In base a questa distinzione è possibile quindi distinguere tra metodi *data-driven* (o *corpus-based*) e metodi *knowledge-driven*, a seconda che le informazioni utilizzate per disambiguare siano interne o esterne al *corpus* analizzato [37].

Uno dei sistemi più utilizzati nei metodi basati "sulla conoscenza" è il *WordNet*, sviluppato dalla Princeton University a partire dagli anni '90 [53]. Si tratta di una risorsa linguistica in formato elettronico che organizza, definisce e descrive i concetti rilevanti della lingua; per ognuno di questi le differenze di senso sono definite mediante relazioni tassonomiche e associative distinte, così che è possibile confrontare i risultati prodotti da analisi diverse. Con il progetto *EuroWordNet*, finanziato dall'Unione Europea dal 1996 al 1999, sono stati sviluppati lessici di tipo WordNet per le diverse lingue europee, collegate in un database multilingue. Rispetto al paradigma inglese, di cui adottano la struttura base, i lessici europei considerano una nozione più allargata di equivalenza di significato, estesa anche a differenti categorie sintattiche, ed una più ampia classe di relazione di senso, atte a trattare in modo approfondito la polisemia. Il limite maggiore è comunque quello di non fornire risultati apprezzabili se il linguaggio analizzato è eccessivamente tecnico.

1.4.3 La disambiguazione nella statistica testuale

Le ambiguità di natura semantica possono essere parzialmente eliminate ricorrendo a meta-dati di ordine stilistico e contestuale.

Se si analizzano *corpora* di grandi dimensioni è impensabile non ricorrere a metodologie di disambiguazione automatica, ma in taluni casi il costo in termini di perdita di informazione è considerevole. Le distorsioni dovute al raggruppamento degli omografi ambigui sono talvolta notevoli, e il ricorso a valori extralinguistici per ricostruire la motivazione polisemica pone seri problemi nell'analisi concreta di un *corpus* testuale.

Il primo passo da compiere, una volta individuate le possibili fonti di ambiguità, è quello di ricorrere all'*Analisi delle Concordanze*, con cui si compie uno studio sistematico dei *contesti locali*.

Per ogni forma ambigua, indicata come *pivot*, si considera un dato

insieme di forme adiacenti, in modo da migliorare la monosemia della stessa.

1.	istica economia e commercio ed	economia	aziendale per l inserimento ne
2.	discipline tecniche e i laureati in	economia	che desiderano acquisire prof
3.	ata verso le problematiche della	economia	di impresa prevede una rotazi
4.	ormazione fisica matematica ed	economia	e commercio ad indirizzo ban
5.	canica ingegneria gestionale ed	economia	e commercio con il massimo
6.	tica o scienze dell informazione	economia	e commercio denso manufactu
7.	nica gestionale informatica ed in	economia	e commercio e a diplomati ad
8.	le diplomati geometra laureati in	economia	e commercio ecc in questo ca
9.	i marketing produzione logistica	economia	e commercio ed economia azi
10.	ste principalmente ingegneria ed	economia	e commercio ma senza alcuna
11.	progetti erp si richiede laurea in	economia	e commercio o ingegneria ges
12.	ale scienze dell informazione ed	economia	e commercio requisito fondam
13.	lmente laureati in ingegneria ed	economia	e commercio senza alcuna pr
14.	ng per internet start up laurea in	economia	e commercio statistica knowle
15.	vi e formativi aziendali laurea in	economia	e commercio statistica scienz
16.	oni o dell informatica o attinenti l	economia	e la gestione d impresa viene
17.	marketing e vendite la laurea in	economia	è la più indicata per diventare
18.	enze dell informazione statistica	economia	è necessaria una conoscenza
19.	zzazione informatica ingegneria	economia	giurisprodenza informatica ric
20.	arnig center lauree più richieste	economia	ingegneria gestionale meccan
21.	temente ai laureati in ingegneria	economia	matematica fisica ed informati
22.	ere laureati in giurisprudenza in	economia	o in scienze politiche le tecno
23.	tunità per brillanti neolaureati in	economia	o ingegneria senza alcuna pre
24.	ovani laureati in giurisprudenza	economia	scienze dell informazione o di

Figura 1.5: Analisi delle Concordanze

Nell'esempio in figura 1.5 sono analizzati i contesti locali della forma *economia* in un *corpus* costituito da annunci di lavoro: l'ambiguità è data, in questo caso, dal fatto che la forma suddetta può tanto fare riferimento al tipo di laurea richiesta ai candidati quanto all'attività dell'azienda che propone la posizione lavorativa.

Dopo aver selezionato i frammenti che contengono la forma ambigua, si procede all'individuazione e all'analisi delle forme che con essa "co-occorrono" più frequentemente. Riprendendo l'esempio della forma *economia*, è possibile vedere in figura 1.6 una rappresentazione sintetica delle co-occorrenze e della loro posizione rispetto al pivot.

Per procedere alla disambiguazione in senso stretto è possibile mettere in atto, manualmente, una procedura di *categorizzazione seman-*

	Collocate	4.	3.	2.	1.	Node	1.	2.	3.	4.	Total
1.	commercio			1		economia		12			13
2.	ingegneria		1	2	2	economia	1	1		2	9
3.	laureati	2	2	2		economia					6
4.	scienze	2	1			economia	2		1		6
5.	informatica	1	1	1		economia		1		1	5
6.	informazione	1		2	1	economia			1		5
7.	laurea			4		economia					4
8.	gestionale	1		1		economia		1			3
9.	statistica				1	economia			2		3
10.	fisica		1			economia		1			2
11.	giurisprudenza			1	1	economia					2
12.	matematica			1		economia	1				2
13.	meccanica	1				economia			1		2
14.	aziendale					economia	1				1
15.	aziendali		1			economia					1
16.	laureato			1		economia					1
17.	vendite	1				economia					1

Figura 1.6: Analisi delle Co-occorrenze

tica delle forme ambigue, al fine di non perdere alcuna informazione interessante, marcando le stesse con una etichetta *ad hoc*: ad esempio, si marcano le forme che fanno riferimento al tipo di laurea postponendo alle stesse il codice `_lau`.

Altra strategia praticabile è quella della *lessicalizzazione*. La selezione sistematica delle unità minimali di senso (o lessie) all'interno del *corpus* risulta essere, infatti, una soluzione relativamente semplice e di facile applicazione, e consente di ridurre la possibile ambiguità data dalla singola forma grafica.

Con l'individuazione dei poliformi e delle poliremathiche d'interesse, sotto forma di segmenti ripetuti, si corre per contro il rischio di sottrarre informazioni necessarie alla comprensione del fenomeno. È possibile tenere sotto controllo l'importanza del segmento rispetto alle forme semplici che lo compongono, attraverso l'uso di un indice, detto di *assorbimento* (IS):

$$IS = \left[\sum_{i=1}^{L} \frac{f_{SEGM}}{f_{FG_i}} \right] \cdot P \tag{1.3}$$

Mettendo in relazione l'incidenza delle occorrenze del segmento rispetto a quelle delle L forme che lo compongono, e moltiplicando la somma dei quozienti per il numero P di parole piene presenti nel segmento, si ottiene una misura della significatività del segmento considerato. Dopo aver verificato quali sono i segmenti a maggior contenuto informativo si procede alla loro lessicalizzazione, sostituendo agli spazi tra le forme componenti il segno convenzionale "_" (*underscore*).

1.5 Le fasi di preparazione dei *corpora*

Per poter analizzare un *corpus* testuale, da un punto di vista statistico, è opportuno effettuare previamente una serie di operazioni, riportate nello schema della figura 1.7.

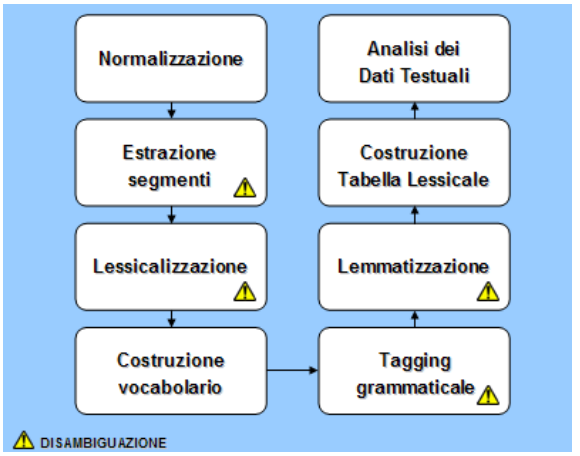

Figura 1.7: Pre-processing e Analisi dei Dati

Tali operazioni, riunite sotto il nome di *pre-trattamento del corpus*, sono indispensabili perché i testi scritti in linguaggio naturale non pos-

sono essere trattati direttamente per mezzo di algoritmi, essendo non strutturati e inoltre con un basso livello di standardizzazione.

È necessario predisporre poi, una volta completata la fase di preparazione, una procedura di codifica che consenta di rappresentare l'informazione contenuta nel *corpus* in modo compatto e in un formato tale da poter essere trattata statisticamente.

1.5.1 Acquisizione, normalizzazione, lessicalizzazione

Il testo è considerato come una successione di simboli appartenenti ad un codice, in cui è necessario identificare un insieme di caratteri *separatori* e un insieme complementare di caratteri non separatori che definiscono l'*alfabeto*.

Il primo passo è quindi quello di individuare le successioni di caratteri dell'alfabeto comprese tra i separatori attraverso una procedura detta di *parsing*, nella quale il testo è scansionato e ricondotto ad un elenco di forme grafiche. Attraverso la *normalizzazione* si agisce sull'insieme dei caratteri non separatori per eliminare possibili fonti di "sdoppiamento" del dato, eliminando ad esempio le differenze tra caratteri maiuscoli e minuscoli, o uniformando la grafia delle forme che presentano forte variabilità, come nomi propri, sigle e date[2].

Una volta completata la fase di normalizzazione si procede alla costruzione del vocabolario delle forme grafiche, sul quale è possibile calcolare degli indici di misura di tipo lessicometrico. Un indice che consente un riscontro immediato della variabilità del fenomeno ogget-

[2]Uno dei problemi sempre più comuni e di non facile trattazione è quello della *composizione*, ossia la costruzione di forme derivate o composte a partire dalle forme semplici, utilizzando il segno "-" (in Inglese *hyphen*). Mentre per alcune lingue esistono regole precise, in Italiano coesistono spesso grafie diverse di una stessa forma, evenienza sempre più aggravata dal linguaggio giornalistico e dai linguaggi specialistici

to di studio è il cosiddetto *type/token ratio*, ottenuto dal rapporto tra l'ampiezza del vocabolario V e quella del *corpus* N, e che consente di avere una misura della "ricchezza lessicale". Per limitare la dipendenza dalle dimensioni del *corpus* risulta comunque più interessante il *coefficiente di Giraud*, che considera al denominatore $\sqrt{N}$.

Nella lessicalizzazione l'obiettivo è quello di riconoscere e isolare all'interno del testo le unità minime di senso. È quindi necessario innanzi tutto selezionare una lista di segmenti ripetuti, fissando *a priori* una soglia di frequenza per le forme da considerare e quindi una soglia di frequenza per i segmenti. A questo punto è possibile scegliere i segmenti a maggiore contenuto informativo, in particolare poliformi e polirematiche, e marcarli all'interno del *corpus*, per poi procedere alla costruzione di un nuovo vocabolario di forme testuali.

1.5.2 La lemmatizzazione

Il passaggio obbligato, a questo punto, è quello del riconoscimento delle POS e del loro marcaggio attraverso il *tagging grammaticale*. La fase di lemmatizzazione consiste, una volta riconosciuta la POS di appartenenza, nel ricondurre ogni forma flessa al *lemma* di appartenenza.

Per lemma si intende la "forma canonica" con cui una data voce è presente in un dizionario: si considera quindi l'infinito per i verbi, il singolare per i sostantivi, il singolare maschile per gli aggettivi. Il principio alla base di tale strategia è che le varianti morfologiche più comuni di una forma hanno significato simile e sono usati in contesti simili, perciò è possibile individuare delle *invarianti semantiche* [16], termini monosemici il cui senso non varia con la flessione.

La lemmatizzazione automatica è effettuata per mezzo di particolari algoritmi che scompongono la forma grafica, utilizzando ad esempio il sistema degli *n-grammi*. Tali modelli sono fondati sull'*assunzione markoviana* per la quale la probabilità di avere un dato carattere dipende solo dai k caratteri precedenti:

$$\begin{aligned} &\mathrm{P}\left(X_n = x_n | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\right) \quad = \\ &\mathrm{P}\left(X_n = x_n | X_k = x_k, \ldots, X_{n-1} = x_{n-1}\right) \end{aligned} \tag{1.4}$$

Scomponendo le forme da lemmatizzare con i *trigrammi* (sequenze di tre caratteri) è possibile isolare il lessema e quindi riconoscere il lemma cui fare riferimento. Tale soluzione è però praticabile per lingue facilmente modellizzabili come l'Inglese, nel quale le flessioni sono regolate da un numero limitato di regole; per le lingue romanze è necessario ricorrere ai dizionari e ai *lemmari* elettronici e procedere per confronto.

L'ambiguità di talune forme può essere ulteriormente limitata attraverso il processo di lemmatizzazione, benché rimanga il problema dell'accorpamento delle forme omografe. È necessario effettuare allora, in taluni casi, una lemmatizzazione *interna* [47], intervenendo non sul *corpus* ma sul vocabolario. In tal modo è possibile operare da un lato una maggiore riduzione della dimensionalità dello spazio delle forme, e dall'altro l'attuazione di un processo di *fusione* anche per i sinonimi, ossia quelle forme che hanno significante diverso ma medesimo significato. Tale operazione è comunque rischiosa perché comporta una profonda modifica del vocabolario e rischia di alterare irrimediabilmente i risultati dell'analisi.

1.5.3 Selezione del linguaggio peculiare

Terminato il pre-trattamento del *corpus* è necessario selezionare le forme con il maggior contenuto informativo in relazione allo scopo della ricerca. Tale operazione è senza dubbio la più delicata, poiché da essa dipende in gran parte la significatività della successiva analisi.

L'apporto di procedure automatiche è in questo caso limitato, e quindi grande importanza rivestono l'esperienza del ricercatore e il supporto di conoscenza esperta rispetto al fenomeno indagato.

Un aiuto oggettivo al lavoro dell'analista è dato dal confronto del vocabolario relativo alla raccolta di testi considerata con *lessici di frequenza* relativi al tipo di linguaggio oggetto di studio.

Un lessico di frequenza è un particolare tipo di vocabolario ottenuto da una raccolta di testi, relativi ad uno specifico fenomeno, di dimensione notevole. In tal caso, facendo riferimento a *corpora* con un'ampiezza superiore alle 500000 occorrenze, si può considerare il numero di occorrenze normalizzate di forme molto frequenti come una buona approssimazione della probabilità [56]: più precisamente, il numero di occorrenze di una forma all'interno di un lessico di frequenza può essere considerato come l'espressione della sua frequenza nello specifico linguaggio considerato.

Per individuare le unità "peculiari" all'interno del *corpus* analizzato è possibile confrontare, quindi, la lista di forme alla fine del processo di pre-trattamento con il modello di riferimento offerto dal lessico di frequenza relativo. La specificità di ciascuna forma è valutabile attraverso lo scarto standardizzato z_i

$$z_i = \frac{f_i - f_i^*}{\sqrt{f_i^*}} \tag{1.5}$$

nel quale f_i rappresenta il numero di occorrenze normalizzate dell'*i-ma* forma nel *corpus* considerato e f_i^* il corrispondente valore nel lessico assunto come paradigma. Tale indice è di fatto la radice di una statistica Chi-quadro, nel quale si considera la frequenza della forma nel lessico di riferimento come frequenza teorica, e consente di valutare la *sovra/sotto-utilizzazione* delle forme rispetto alla loro frequenza nei normali contesti d'uso o negli ambiti settoriali.

Come specificato in precedenza è determinante, in ogni caso, la scelta soggettiva del ricercatore rispetto al tipo di forme da considerare, scelta influenzata fortemente dal grado di conoscenza del fenomeno che si sta analizzando.

Operativamente è possibile utilizzare come metodo per la riduzione del vocabolario e la selezione delle forme più significative il calcolo dei *quartili*. Dopo aver ordinato le forme in base al numero di occorrenze nel *corpus* oggetto d'analisi, ed aver calcolato le frequenze cumulate, si individuano le forme corrispondenti al primo e terzo quartile, che escludono a sinistra e a destra ripettivamente il primo e l'ultimo 25% delle forme del vocabolario.

In tal modo si considerano solo le forme diffusamente presenti nel *corpus*, escludendo tanto gli *hapax* e le forme poco frequenti, caratteristiche di pochi documenti, quanto le forme "ovvie", caratteristiche della maggior parte dei documenti, e quindi indicative dell'ambito generale di analisi, ma con un basso potere discriminante. Tale operazione risulta comunque utile solo nel trattamento di *corpora* di notevoli dimensioni, per i quali è interessante analizzare, in un'ottica esplorativa, il tipo di linguaggio utilizzato.

Capitolo 2

L'analisi multidimensionale dei dati testuali

La linguistica e la statistica sono due discipline profondamente differenti tra loro per storia e principi avendo ciascuna subito numerose mutazioni profonde, fino all'ultima importante dovuta all'utilizzo dell'informatica. Poiché ogni produzione letteraria è generalmente strutturata in modo logico, è possibile trovare tra le due scienze dei nodi di interconnesione e quindi porre in essere uno studio dei *corpora* testuali attraverso un approccio propriamente tecnico quale quello statistico.

La famiglia di metodologie comunemente alla base dell'analisi statistica dei testi rientra nell'ambito delle tecniche di Analisi Multidimensionale dei Dati.

Gli psicologi sono stati, all'inizio del secolo, i pionieri dell'esplorazione dei dati multivariati; C. Spearman cercò, analizzando i legami tra i risultati scolastici e le attitudini sensoriali di alcuni studenti, di dimostrare l'esistenza di un fattore generale di attitudine o di intelligenza. In seguito, a partire da dati sempre più numerosi, furono ricercati non solo uno ma parecchi "fattori", definendo così il presupposto logico che si ritrova in modo differente nell'*approccio fattoriale*.

All'inizio, la mancanza di strumenti di calcolo adeguati ha costituito un ostacolo insormontabile alla piena affermazione delle proposte metodologiche, per lo più sviluppati in un contesto quasi esclusivamente teorico [3].

A partire dagli anni '60 la diffusione dei primi elaboratori elettronici facilitò la rapida diffusione di tali tecniche in ambito applicativo. In particolare, si deve alla scuola francese di *Analyse des Données* di Benzécri lo sviluppo di una nuova impostazione basata su ipotesi di tipo strutturale piuttosto che di tipo distribuzionale.

Per lungo tempo la scuola anglofona di analisi multivariata e quella francese di analisi multidimensionale hanno percorso strade parallele ma separate, ma il dibattito scientifico degli ultimi anni ha evidenziato di volta in volta meriti e limiti delle due impostazioni.

Lo sviluppo delle tecniche di Analisi dei Dati in Francia è strettamente connesso allo studio del fenomeno linguistico. Lo stesso Benzécri ricorda come "*c'est principalement en vue de l'étude des langues que nous nous sommes engagés dans l'analyse factorielle* [...]" [12].

Le specificità che fanno preferire ad altri un approccio "multidimensionale" risiedono nel fatto che il fenomeno oggetto di studio (il linguaggio naturale) è di tipo osservazionale e che non sono possibili riferimenti statistico-probabilistici sul tipo di distribuzione delle variabili considerate.

Rispetto alla tradizionale analisi del contenuto (cfr. § 2.1), le strategie di analisi della scuola francese si differenziano per un approccio marcatamente induttivo e descrittivo, basato sulla disponibilità di un ampio sistema di informazioni elementari capace di cogliere il fenomeno nella sua complessità. L'analisi del contenuto si propone, invece, senza attardarsi sul materiale testuale propriamente detto, di accedere direttamente ai significati dei differenti segmenti che compongono il testo. La sua riuscita presuppone quindi che il sistema di categorie definite *a priori* sia allo stesso tempo coerente e pertinente, cosa che è difficile assicurare nella pratica, specialmente quando si hanno poche

informazioni a priori sulla base documentale oggetto d'analisi.

Lo scopo principale dell'Analisi dei Dati è quello di evidenziare la struttura latente sottostante al testo in esame tramite una riduzione della dimensionalità dello spazio di rappresentazione delle variabili linguistiche o di quello dei frammenti. Particolare rilievo viene dato alla visualizzazione su sottospazi di migliore approssimazione delle relazioni tra forme o tra le unità documentali (*metodi fattoriali*), o all'individuazione di classi di equivalenza delle unità con l'ausilio di grafi o strutture ad albero (*metodi di classificazione*).

2.1 Analisi qualitativa e quantitativa del linguaggio

Un testo scritto in linguaggio naturale può essere analizzato utilizzando metodologie proprie della ricerca qualitativa o quantitativa.

Nel primo caso l'attenzione è rivolta all'analisi delle parole e delle espressioni presenti all'interno del testo, con lo scopo di individuare i diversi contesti peculiari e classificarli in base ad un dato criterio legato agli obiettivi della ricerca effettuata. Nel secondo caso l'obiettivo è quello di scomporre il testo in unità elementari ed effettuare quindi delle misurazioni relative alla ricorrenza di alcuni termini e delle loro combinazioni.

L'analisi qualitativa dei testi è uno strumento di ricerca frequentemente utilizzato in campo sociologico e psicologico per lo studio del comportamento e della psiche umana. Risulta spesso utile, ad esempio, al di là dell'utilizzo di questionari strutturati con domande chiuse o aperte, l'analisi dei testi prodotti dai soggetti intervistati in relazione ad un fenomeno d'interesse. Tale analisi può esplicitarsi nell'utilizzo di un metodo *positivista*, qual è l'*Analisi del Contenuto*, o invece lasciare spazio ad un approccio al testo di tipo *interpretativista*.

L'Analisi del Contenuto, nella definizione data da K. Krippendorf, può essere considerata come un metodo per inferire i significati contenuti in un testo rispetto al contesto in cui lo stesso è stato prodotto [39]. È di fatto una tecnica sistematica e replicabile con la quale classificare i contenuti di un testo mediante l'utilizzo di categorie definite.

Nella codifica dei dati testuali è possibile seguire due differenti approcci, la cosiddetta codifica *emergente* e la codifica *a priori*.

Nella codifica *emergente* è necessario un esame preliminare dei testi per individuare le categorie di contenuto; tale esame è generalmente effettuato da due o più ricercatori, in modo da individuare una lista di possibili categorie che sia sintesi dei diversi approcci individuali al testo. Una volta testata l'affidabilità della lista (in termini di stabilità e riproducibilità dei codici), si estende la classificazione a *corpora* di dimensioni maggiori, verificando periodicamente la qualità del sistema di categorie approntato. Nella codifica *a priori* la definizione delle categorie è invece fondata su un impianto teorico prestabilito, cercando di rispettare i criteri dell'esaustività e dell'esclusività.

Spesso per ovviare ai poblemi dovuti ad un sistema di codici non idoneo, si utilizza un piccolo campione di documenti per testare la validità del tipo di codifica adottato. È utile comunque, in entrambi i casi considerati, il ricorso a procedure quali il *Key Word In Context* (KWIC) e il *Key Word Out of Context* (KWOC) per definire la consistenza all'interno del testo analizzato delle forme potenzialmente più interessanti. Nel primo caso si fa riferimento alla stessa idea di base dell'Analisi delle Concordanze, definendo una lista di possibili contesti d'utilizzo; nel secondo caso invece, ogni forma interessante è collegata ad un indice in base al quale si ritrovano poi i diversi contesti.

Una volta effettuata la classificazione delle forme è possibile ad esempio dedurre informazioni sullo stile di chi ha prodotto i testi, sul tipo di linguaggio utilizzato, ma non è possibile estrarre ulteriori informazioni sugli aspetti latenti dei *corpora* analizzati, come invece avviene ricorrendo a tecniche quali quelle dell'Analisi dei Dati.

2.2 Il *peso* delle parole

Nell'implementazione di un qualsiasi strumento per l'analisi di *corpora* testuali è rilevante (come accennato nella sezione § 1.3.3) il tipo di codifica utilizzato per la *numerizzazione* dell'informazione in essi contenuta. Tale codifica deve tener conto dell'importanza, in termini di contenuto informativo, di ogni forma presente nei documenti.

Il problema della ponderazione delle forme, e sovente anche dei documenti analizzati, è centrale nel trattamento statistico dei dati testuali. In un'ottica esplorativa, essendo notevole la mole di informazioni trattate nel corso dell'analisi, è fondamentale considerare in modo corretto l'importanza di ogni unità coinvolta, a costo di aumentare la complessità computazionale, per non incorrere in distorsioni fuorvianti nell'interpretazione del fenomeno linguistico.

Solitamente, per analizzare in modo automatico un set di documenti si ricorre alla loro trasformazione in "vettori". In generale, un vettore/documento può essere rappresentato come:

$$D_j = (w_{1,j}, w_{2,j}, w_{3,j}, \ldots, w_{p,j}) \tag{2.1}$$

dove il valore assunto da $w_{i,j}$ rappresenta l'importanza della *i-ma* forma nel *j-mo* documento.

2.2.1 Gli schemi di ponderazione

Lo schema più semplice da adottare è quello *booleano*, nel quale si valuta esclusivamente la presenza o l'assenza di una certa forma all'interno di ogni singolo documento; in particolare, il peso $w_{i,j}$ assume valore 1 se la forma i è espressa nel documento j, e valore 0 se invece non lo è. Il limite evidente di tale sistema di ponderazione è che l'importanza di ogni forma è mal misurata, perché espressa in egual modo tanto nei documenti fortemente caratterizzati da essa quanto nei documenti in

cui la stessa forma non ha un contenuto informativo caratterizzante.

L'alternativa più diffusa allo schema di presenza/assenza è quella che tiene conto del numero di occorrenze di ogni forma, come già visto nel caso dello schema Bag-of-Words. Tale quantità, assimilabile alla frequenza statistica, è computata considerando il numero di volte in cui la forma i occorre all'interno di ogni documento.

I vettori/documento ottenuti tanto con il primo quanto con il secondo schema possono essere giustapposti per costrire delle tabelle lessicali del tipo $\{forme \times documenti\}$. L'informazione presente sulle distribuzioni marginali di riga assume nei due casi significato diverso: nello schema booleano infatti rappresenta il numero di documenti nel quale è presente ogni data forma, mentre nell'altro caso è il numero di volte in cui ogni forma è presente all'interno del *corpus* analizzato.

Chiaramente, il tipo di codifica prescelto dipende dal tipo di analisi che si vuole effettuare sui dati. In talune strategie di trattamento del linguaggio naturale, ed in particolar modo nelle tecniche connesse all'*Information Retrieval*, si preferisce utilizzare dei sistemi di pesi "complessi", che tengano conto allo stesso tempo dell'importanza di ogni forma rispetto ad uno specifico documento e rispetto a tutti i documenti contenuti nel *corpus*.

2.2.2 L'indice TF-IDF

Il *term frequency/inverse document frequency* (TF-IDF), proposto da G. Salton e C. Buckley [61], è alla base di una famiglia di strategie di ponderazione delle forme che ha trovato largo impiego in letteratura per la loro robustezza e la relativa semplicità di calcolo.

L'idea di fondo nell'utilizzo di questo schema scaturisce da alcune considerazioni sul trattamento dell'informazione testuale.

Le forme più frequenti all'interno di un documento, escluse le forme strumentali, sono generalmente indicative del suo contenuto. Per tener conto dell'importanza relativa di ogni forma è opportuno utiliz-

zare, come fattore di normalizzazione, il numero di occorrenze della forma che appare più volte all'interno del documento. Tale rapporto rappresenta il cosiddetto *term frequency*, espresso sovente come:

$$\mathrm{tf}_{ij} = 0.5 + 0.5\frac{f_{ij}}{\max f_j} \tag{2.2}$$

In corrispondenza di livelli più alti dell'indice, considerando un campo di variazione compreso tra 0.5 e 1, si individuano le forme con un contributo informativo maggiore per la descrizione del testo. In molte applicazioni il TF è "smorzato" utilizzando una funzione del tipo $f(\mathrm{TF})$, ad esempio $\sqrt{\mathrm{TF}}$ o $1 + \log(\mathrm{TF})$, così da assegnare una importanza "relativa" alle forme più frequenti.

Le forme caratterizzanti per alcuni documenti possono essere presenti, con minore importanza, anche negli altri documenti del *corpus*. Per poter valutare il livello di "discriminazione" delle forme all'interno della collezione analizzata è allora opportuno introdurre un altro indice, l'*inverse document frequency* [63]. Indicato con df_i il numero di documenti in cui la forma *i-ma* appare, l'espressione più diffusamente utilizzata per computare l'IDF è:

$$\mathrm{idf}_i = \log\left(\frac{n}{\mathrm{df}_i}\right) \tag{2.3}$$

dove n è il numero totale dei documenti nella collezione; l'uso del logaritmico è giustificato dalla necessità di "compensare" l'effetto del TF.

Combinando in vario modo il TF e l'IDF si ottiene l'indice completo: una delle peculiarità del TF-IDF è, infatti, che non esiste uno schema ideale e molto è lasciato all'esperienza del ricercatore e alla validità degli esperimenti empirici condotti negli anni[1].

Una delle possibili formulazioni (il cosiddetto *best fully weighted*

[1]N. Oren ha proposto il ricorso alla *programmazione genetica* per ottenere automaticamente nuovi schemi basati sul TF-IDF [58]

system), che consente tra l'altro di comparare i risultati ottenuti da *corpora* differenti, è data da:

$$\frac{f_{ij} \cdot \log\left(n/\mathrm{df}_i\right)}{\sqrt{\sum_i \left[f_{ij} \cdot \log\left(n/\mathrm{df}_i\right)\right]^2}} \tag{2.4}$$

dove la quantità al denominatore normalizza l'indice considerando la lunghezza del documento considerato.

La 2.4 è senz'altro più indicata in uno schema di ponderazione dei documenti; il TF-IDF è infatti adoperato prevalentemente nei processi di Information Retrieval, dove è necessario distinguere il sistema di pesi utilizzato per i documenti dal sistema di pesi utilizzato per le *query*. In letteratura la suddetta distinzione non è sempre sufficientemente sottolineata, preferendo per semplicità l'utilizzo congiunto della 2.2 e della 2.3. Nell'introdurre il TF-IDF in una strategia di analisi dei testi basata su metodi di tipo fattoriale (come evidenziato in particolare nella sezione § 3.3) è possibile seguire tale consuetudine, non essendo necessaria l'adozione di due sistemi di pesi differenti.

2.3 Misure di similarità e distanza

Nell'analisi statistica dei testi è importante stabilire dei criteri per poter valutare il livello di *similarità* tra i documenti e tra le forme. Tale concetto è sovente di intuitiva e semplice modelizzazione per i dati numerici, ma lo stesso non accade quando si analizzano dei testi.

In un approccio di tipo qualitativo è possibile leggere testi differenti e quindi, una volta individuati i concetti principali e i temi chiave, formulare un giudizio sulla loro similarità o dissimilarità. Nel trattamento automatico di *corpora* è invece necessario approssimare i concetti espressi con delle semplificazioni, e conseguentemente adattare gli schemi usualmente utilizzati in statistica per valutare la similarità.

La maggior parte degli indici di misura della cosiddetta *similarità semantica* è espressa in termini di similarità tra vettori, poiché come detto è necessario numerizzare l'informazione testuale.

2.3.1 Similarità tra vettori/documento

I documenti, come visto, possono essere codificati ricorrendo ad uno schema di tipo booleano, nel quale si assegna valore 0 alle forme del vocabolario del *corpus* non presenti nello specifico documento e valore 1 alle forme presenti (quale che sia la loro frequenza).

Dato un *corpus* con p forme, la misura di similarità tra due vettori/documento $\mathbf{D}_j$ e $\mathbf{D}_h$ più semplice da utilizzare, in termini computazionali, è il cosiddetto *coefficiente di matching*:

$$\mathrm{S}_M(\mathbf{D}_j, \mathbf{D}_h) = |\mathbf{D}_j \cap \mathbf{D}_h| \equiv \sum_{i=1}^{p} w_{ij} \cdot w_{ih} \tag{2.5}$$

Tale misura considera di fatto il numero di dimensioni non nulle di entrambi i vettori, ma ha il limite di non tener conto della diversa lunghezza dei documenti e delle forme presenti in ciascuno di essi.

Il *coefficiente di Dice*, a differenza del precedente, è normalizzato per la lunghezza dei documenti ed è ottenuto dall'espressione:

$$\mathrm{S}_D(\mathbf{D}_j, \mathbf{D}_h) = \frac{2\,|\mathbf{D}_j \cap \mathbf{D}_h|}{|\mathbf{D}_j| + |\mathbf{D}_h|} \equiv \frac{2 \sum\limits_i w_{ij} \cdot w_{ih}}{\sum\limits_i w_{ij} + \sum\limits_i w_{ih}} \tag{2.6}$$

con un numeratore pari al doppio della 2.5, in modo da mantenere un campo di variazione compreso tra 0 e 1 (vettori/documento identici).

Il *coefficiente di Jaccard*, frequentemente utilizzato come misura di similarità anche in altre discipline, "penalizza" i documenti con un numero di forme in comune relativamente piccolo in proporzione a tutte le forme presenti. Per contro ha il vantaggio di fornire valori più bassi

rispetto alle altre misure proposte per i casi di bassa sovrapposizione tra i documenti (*overlap*). Dati due vettori/documento è pari a:

$$\mathrm{S}_J(\mathbf{D}_j, \mathbf{D}_h) = \frac{|\mathbf{D}_j \cap \mathbf{D}_h|}{|\mathbf{D}_j \cup \mathbf{D}_h|} \equiv \frac{\sum_i w_{ij} \cdot w_{ih}}{\sum_i w_{ij} + \sum_i w_{ih} - \sum_i w_{ij} \cdot w_{ih}} \tag{2.7}$$

Il *coseno* rappresenta una misura di similarità estremamente interessante, soprattutto per il significato che riveste all'interno dello spazio vettoriale nel quale i documenti sono rappresentati; per calcolare il coseno dell'angolo formato da due vettori/documento si ricorre all'espressione:

$$\mathrm{S}_C(\mathbf{D}_j, \mathbf{D}_h) = \frac{|\mathbf{D}_j \cap \mathbf{D}_h|}{\sqrt{|\mathbf{D}_j| \times |\mathbf{D}_h|}} \equiv \frac{\sum_i w_{ij} \cdot w_{ih}}{\sqrt{\sum_i w_{ij}^2 \cdot \sum_i w_{ih}^2}} \tag{2.8}$$

Se i vettori/documento hanno lo stesso numero di forme espresse, allora il coseno coincide con il coefficiente di Dice. Al contempo, fornisce una misura meno penalizzante qualora il numero di forme espresse nei due documenti sia molto diverso. Tale proprietà è di indubbia importanza nel trattamento statistico del linguaggio naturale, poiché nel comparare documenti con una diversa lunghezza non sempre è utile valutare la loro similarità (o dissimilarità) soltanto in base a questa informazione.

Il coseno può essere utilizzato per valutare il livello di similarità tra documenti anche se questi sono numerizzati per mezzo della codifica Bag-of-Words, che consente da un punto di vista linguistico un maggiore apporto informativo rispetto allo schema booleano.

2.3.2 Il concetto di distanza

Per poter valutare la relazione tra due entità, in termini di *distanza*, è necessario introdurre alcuni concetti generali.

Sia d un'applicazione che assegna un numero reale non negativo a ogni coppia di osservazioni (e_i, e_h) appartenenti ad un dato spazio $\mathcal{R}$, e si considerino le seguenti condizioni:

$$\begin{array}{cc} \textit{separabilità} & \rightarrow \quad d(e_i,e_h) = 0 \Leftrightarrow e_i = e_h \\ \textit{simmetria} & \rightarrow \quad d(e_i,e_h) = d(e_h,e_i) \\ \textit{disuguaglianza triangolare} & \rightarrow \quad d(e_i,e_h) \leq d(e_i,e_k) + d(e_k,e_h) \end{array} \tag{2.9}$$

Se l'applicazione d soddisfa le condizioni di separabilità e simmetria può essere assunta come una misura di similarità[2] tra e_i e e_k; se viene soddisfatta anche la cosiddetta "disuguaglianza triangolare" per tutte le triplette di entità (e_i,e_k,e_h) allora d è una misura della distanza. La matrice delle distanze $\mathbf{D}$ è definita come *metrica* dello spazio $\mathcal{R}$.

Una espressione generale della distanza tra due vettori $\mathbf{x}$ e $\mathbf{y}$, in forma matriciale, è data da:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\mathbf{M}(\mathbf{x}-\mathbf{y})} \tag{2.10}$$

dove la metrica $\mathbf{M}$ è una matrice simmetrica definita positiva.

A seconda del tipo di metrica utilizzata nella 2.10 vengono definite differenti misure di distanza. Se si considera $\mathbf{M} \equiv \mathbf{I}$ con $\mathbf{I}$ matrice unitaria, si ottiene allora la *distanza Euclidea*:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})} \tag{2.11}$$

[2]Dato l'indice di similarità s_{e_i,e_h}, è possibile ricavare il corrispondente indice di *dissimilarità* d_{e_i,e_h} come complemento a 1

Se si utilizza come metrica $\mathbf{M} \equiv \mathbf{\Sigma}$, con $\mathbf{\Sigma}$ matrice di varianza-covarianza, si ha la cosiddetta *distanza di Mahalanobis*:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})} \tag{2.12}$$

Un particolare tipo di metrica, di fatto una metrica euclidea *ponderata*, è la metrica del *Chi-quadro*, basata sulla statistica χ^2. Nelle tabelle di contingenza, la generica cella in corrispondenza della *i-ma* modalità del carattere I e della *j-ma* modalità del carattere J contiene la frequenza f_{ij}, cui corrispondono le frequenze marginali $f_{i.}$ e $f_{.j}$, rispettivamente di riga e di colonna. La metrica del Chi-quadro definisce la distanza tra due righe o tra due colonne come:

$$d^2(i, i') = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \tag{2.13}$$

$$d^2(j, j') = \sum_i \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 \tag{2.14}$$

Rispetto ad altre metriche, la metrica del Chi-quadro gode dell'importante proprietà della *equivalenza distributiva*, risultando invariante rispetto ai criteri di codifica o al modo di aggregare le entità in gruppi, a condizione che le unità aggregate siano omogenee. Come evidenziato nella figura 2.1, è possibile considerare due punti molto prossimi, e quindi con un profilo simile, come un unico punto che abbia per massa la somma delle frequenze dei punti originari.

Il vantaggio/svantaggio dell'utilizzo di tale metrica è strettamente connesso al fatto che le modalità meno frequenti e quelle più frequenti, ponderate per il reciproco delle loro frequenze marginali, sono ugualmente ben rappresentate. Nel trattamento statistico delle basi documentali il dataset iniziale è filtrato a monte, perché l'attenzione del ricercatore è rivolta principalmente all'analisi e alla successiva rappresentazione delle forme a maggior contenuto informativo. È necessario

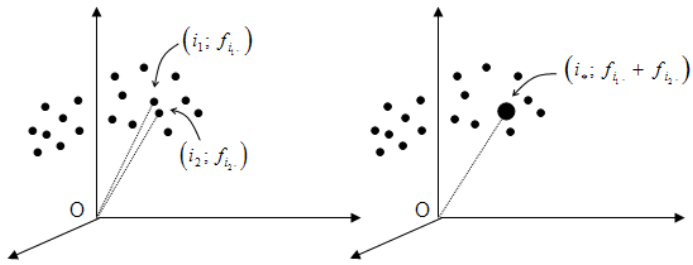

Figura 2.1: Equivalenza distributiva del χ^2

allora considerare di volta in volta, a seconda dell'obiettivo dell'analisi, la convenienza dell'utilizzo di metodi basati sul Chi-quadro.

2.4 L'Analisi delle Corrispondenze su dati testuali

Lo scopo principale delle tecniche di analisi multidimensionale è quello di interpretare e visualizzare la struttura di fenomeni complessi. Per misurare o descrivere un particolare aspetto del fenomeno d'interesse si ricorre a variabili di diversa natura; in generale, le variabili quantitative sono indicative dell'intensità di un dato aspetto, mentre le variabili qualitative evidenziano la presenza/assenza dello stesso.

L'*Analisi delle Corrispondenze* (AC) è una tecnica attraverso la quale è possibile descrivere da un punto di vista sia geometrico che algebrico le relazioni tra le distribuzioni, espresse in forma matriciale, delle modalità di due o più caratteri in un set di unità statistiche. Considerando infatti simultaneamente due diverse partizioni di una popolazione o di un campione da essa estratto, è possibile studiare le variazioni nella distribuzione dei dati per *categorie di risposta*.

Il principio è quello della ricerca di sottospazi sui quali rispettivamente le righe e le colonne della matrice siano meglio rappresentate, per poi visualizzare graficamente le suddette relazioni. L'AC permette quindi di analizzare in maniera organica il linguaggio utilizzato in una raccolta di documenti e di visualizzare le principali associazioni tra i concetti in essi espressi.

2.4.1 Lo schema classico dell'AC

Si consideri un campione di n unità statistiche sul quale è misurato un dato fenomeno per mezzo di due variabili qualitative x e y, rispettivamente con p e q modalità. Sia $\mathbf{N}$ una tabella di contingenza con dimensioni (p,q) in cui si incrociano le due variabili, dove n_{ij} esprime il numero di unità che presenta contemporaneamente la *i-ma* modalità di x e la *j-ma* modalità di y.

L'obiettivo è quello di calcolare a partire dalle variabili originarie una serie di fattori, ciascuno dei quali rappresenta un aspetto latente del tipo di associazione presente nei dati, ricorrendo a delle procedure di decomposizione della tabella dei dati. La successiva rappresentazione in forma grafica dei fattori consente un'interpretazione semplice della struttura dell'associazione e permette di evidenziare aspetti non direttamente rilevabili dalla lettura diretta dei dati.

Per procedere all'analisi si costruiscono, innanzi tutto, la matrice delle frequenze relative $\mathbf{F}$ (con $f_{ij} = n_{ij}/n_{..}$) e i vettori delle frequenze marginali di riga $\mathbf{r}$ e colonna $\mathbf{c}$ (con $f_{i.} = n_{i.}/n_{..}$ e $f_{.j} = n_{.j}/n_{..}$); da queste ultime si ricavano le matrici $\mathbf{D}_p \equiv diag(\mathbf{r})$ e $\mathbf{D}_q \equiv diag(\mathbf{c})$, ossia le matrici diagonali delle distribuzioni marginali di riga e colonna.

Effettuando il prodotto della matrice $\mathbf{F}$ con $\mathbf{D}_p^{-1}$ e $\mathbf{D}_q^{-1}$ si hanno:

$$\tilde{\mathbf{R}} \equiv \mathbf{D}_p^{-1}\mathbf{F} \equiv \begin{bmatrix} \tilde{\mathbf{r}}_1' \\ \vdots \\ \tilde{\mathbf{r}}_p' \end{bmatrix} \qquad \tilde{\mathbf{C}} \equiv \mathbf{F}\mathbf{D}_q^{-1} \equiv \begin{bmatrix} \tilde{\mathbf{c}}_1' \\ \vdots \\ \tilde{\mathbf{c}}_q' \end{bmatrix} \tag{2.15}$$

dove $\tilde{\mathbf{R}}$ rappresenta la matrice dei *profili riga* e $\tilde{\mathbf{C}}$ la matrice dei *profili colonna*, ossia delle distribuzioni condizionate. Si vuole rappresentare la nube dei profili riga $\tilde{\mathbf{r}}_1 \dots \tilde{\mathbf{r}}_p$ in uno spazio $\mathcal{R}^{q-1}$ con pesi $\mathbf{D}_p$ e metrica $\mathbf{D}_q^{-1}$; analogamente, si vuole rappresentare la nube dei profili colonna $\tilde{\mathbf{c}}_1 \dots \tilde{\mathbf{c}}_q$ in uno spazio $\mathcal{R}^{p-1}$ con pesi $\mathbf{D}_q$ e metrica $\mathbf{D}_p^{-1}$.

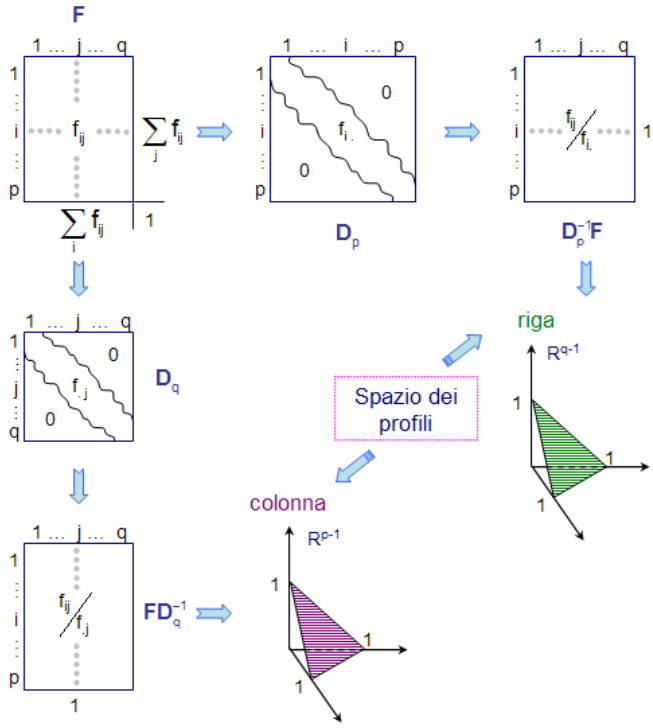

Figura 2.2: Le matrici dell'Analisi delle Corrispondenze

I sottospazi n-dimensionali che meglio approssimano (in termini di minimi quadrati) le proiezioni dei profili riga e colonna sono ottenuti dalla *decomposizione in valori singolari generalizzata* (gSVD) della

matrice $\mathbf{F}$ [31]; gli n vettori singolari destri e sinistri, rispettivamente con metrica $\mathbf{D}_q^{-1}$ e $\mathbf{D}_p^{-1}$, definiscono gli assi principali degli spazi in cui sono rappresentate le nubi dei profili riga e colonna:

$$\begin{aligned} \mathbf{F} &= \mathbf{U}\mathbf{D}_\mu\mathbf{V}' \\ \text{con } \mathbf{U}'\mathbf{D}_p^{-1}\mathbf{U} &= \mathbf{V}'\mathbf{D}_q^{-1}\mathbf{V} = \mathbf{I} \end{aligned} \tag{2.16}$$

dove $\mathbf{D}_\mu$ è la matrice dei valori singolari e le colonne delle matrici $\mathbf{D}_p^{-1/2}\mathbf{U}$ e $\mathbf{D}_q^{-1/2}\mathbf{V}$ rappresentano gli assi principali degli spazi dei profili colonna e dei profili riga[3].

Le coordinate dell'*i-mo* punto riga e del *j-mo* punto colonna sull'α-*mo* asse, dette *coordinate principali*, sono ottenute rispettivamente dalle relazioni:

$$\begin{aligned} \psi_{\alpha i} &= \sqrt{\lambda_\alpha}\, f_{i.}^{-1/2} u_{\alpha i} \\ \varphi_{\alpha j} &= \sqrt{\lambda_\alpha}\, f_{.j}^{-1/2} v_{\alpha j} \end{aligned} \tag{2.17}$$

La bontà dell'approssimazione è valutata in termini di variabilità del fenomeno spiegata dai sottospazi individuati con l'analisi fattoriale, ricorrendo al rapporto tra λ_α, il quadrato dell'α-*mo* valore singolare, e la somma dei quadrati di tutti i valori singolari ottenuti dalla gSVD; quest'ultima quantità è deducibile anche da:

[3]Altro procedimento spesso utilizzato per il calcolo dei fattori è quello della ricerca degli *autovalori* e dei corrispettivi *autovettori* della matrice $\mathbf{F}'\mathbf{D}_p^{-1}\mathbf{F}\mathbf{D}_q^{-1}$. Tra la matrice degli autovalori e quella dei valori singolari vale la relazione

$$\mathbf{D}_\mu \equiv \mathbf{D}_\lambda^{1/2}$$

$$\begin{aligned} \mathrm{tr}\left[\mathbf{F}'\mathbf{D}_p^{-1}\mathbf{F}\mathbf{D}_q^{-1}\right] &= \sum_{\alpha} \lambda_\alpha = \\ &= \sum_i {f_{i.}}^{-1} \sum_j {f_{.j}}^{-1} (f_{ij} - f_{i.}f_{.j})^2 \end{aligned} \tag{2.18}$$

La 2.18 rappresenta una statistica χ^2 e fornisce una misura dell'interdipendenza delle variabili x e y. Il *contributo assoluto* dei punti riga e colonna indica il ruolo giocato da ciascun elemento nella determinazione dell'α-*mo* asse principale, ossia la percentuale di variabilità dell'asse spiegata da ogni singolo punto, ed è dato rispettivamente da:

$$\begin{aligned} \mathrm{ca}_\alpha(i) &= f_{i.}\psi_{\alpha i}^2 \\ \mathrm{ca}_\alpha(j) &= f_{.j}\varphi_{\alpha j}^2 \end{aligned} \tag{2.19}$$

I risultati dell'analisi possono essere rappresentati su grafici piani ottenuti utilizzando coppie di assi principali; in particolare, l'Analisi delle Corrispondenze, mettendo in relazione le modalità di un carattere con le modalità dell'altro, recupera una simmetria complessiva e consente quindi con i dovuti accorgimenti una "ideale" sovrapposizione dei risultati visualizzati nei due sottospazi $\mathcal{R}^{p-1}$ e $\mathcal{R}^{q-1}$.

L'AC gode della cosiddetta *proprietà baricentrica*, per la quale è possibile calcolare la posizione dei profili riga come media ponderata dei "vertici" definiti dai profili colonna, dove i pesi sono pari proprio agli elementi del profilo riga stesso. Allo stesso modo, la posizione dei profili colonna può essere ottenuta come media ponderata dei vertici definiti dai profili riga. La rappresentazione ottenuta considerando simultaneamente i due spazi è detta β-baricentrica per sottolineare la diversità dalla situazione ideale (e impossibile) in cui ogni riga è baricentro delle colonne e viceversa; nello specifico si ricerca il valore $\beta > 1$, con β pari al reciproco del corrispondente α-*mo* valore singolare, tale che siano verificate le relazioni:

$$\begin{aligned} \psi_{\alpha i} &= \beta \mathbf{D}_p^{-1} \mathbf{F} \varphi_{\alpha j} \\ \varphi_{\alpha j} &= \beta \mathbf{D}_q^{-1} \mathbf{F}' \psi_{\alpha i} \end{aligned} \tag{2.20}$$

È importante sottolineare come la rappresentazione nei due spazi separati consenta di intepretare la natura delle similarità tra i profili riga e tra i profili colonna, in termini di distanza del Chi-quadro, e inoltre a visualizzare graficamente la dispersione nelle due nubi. La rappresentazione congiunta al contrario è indicativa della "corrispondenza" tra le due nubi (per le quali vale separatamente quanto detto sopra), non avendo esplicitato una metrica specifica per valutare le similarità tra i punti riga ed i punti colonna, e tale relazione è governata dalla natura "pseudo-baricentrica" delle formule di transizione riportate nella 2.20. Di seguito, nella sezione § 2.4.2, sono riportate le regole di lettura delle mappe fattoriali ottenute dall'AC, adattate al caso dei dati testuali ma valide per ogni caso generale.

2.4.2 L'Analisi delle Corrispondenze Lessicali

Si consideri un *corpus* composto da n documenti relativi ad uno specifico ambito rispetto al quale si ha un dato bisogno informativo. Dopo aver effettuato tutte le procedure di pre-trattamento e analizzato il vocabolario delle forme che compongono i testi, si selezionano le p forme testuali a maggior contenuto informativo. Si ottiene così una tabella lessicale $\mathbf{T}$ di dimensioni (p,n) in cui l'elemento generico t_{ij} rappresenta il numero di occorrenze della *i-ma* forma nel *j-mo* documento.

Come nello schema classico dell'Analisi delle Corrispondenze, è possibile rappresentare i profili colonna (qui indicati come *profili lessicali*) nel sottospazio definito dai profili riga e viceversa. In un'ottica di analisi testuale ciò significa analizzare le relazioni tra le forme nello spazio definito dai documenti, e allo stesso modo di analizzare le relazioni tra documenti nello spazio definito dalle forme del *corpus*.

Poiché i frammenti considerati hanno generalmente lunghezza variabile e sono costituiti da combinazioni di forme differenti, come già precedentemente evidenziato, si ha che la tabella lessicale **T** ha dimensioni elevate e molte celle nulle, cioè la tabella è una matrice sparsa: ne consegue che in taluni casi il contenuto informativo complessivo di **T** è limitato e risulta difficile individuare aspetti latenti del *corpus*.

È allora utile raggruppare i frammenti rispetto a caratteristiche comuni, in modo da ottenere delle tabelle lessicali aggregate. Data una matrice **Q** di dimensioni (n,q) in codifica disgiuntiva completa, che ha in colonna le q modalità di una variabile qualitativa espressione di una caratteristica riscontrabile nei documenti analizzati, è possibile ottenere la tabella $\mathbf{Z} = \mathbf{T} \cdot \mathbf{Q}$ che ha in riga le p forme ed in colonna le q classi di documenti e come elemento generico z_{ik}, il numero di occorrenze della *i-ma* forma nell'*k-ma* classe (Figura 2.3).

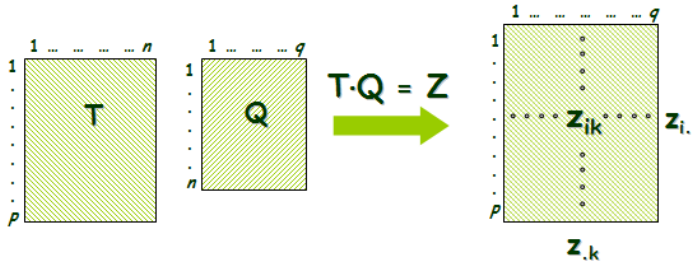

Figura 2.3: Costruzione della tabella lessicale aggregata

Nel sottospazio $\mathcal{R}^{q-1}$, con metrica $\mathbf{D}_q^{-1}$, vengono rappresentate le forme riportate in tabella. Allo stesso modo nel sottospazio $\mathcal{R}^{q-1}$, con metrica $\mathbf{D}_q^{-1}$, vengono rappresentati i documenti.

Nell'interpretare le mappe fattoriali ottenute dall'Analisi delle Corrispondenze Lessicali, è necessario seguire alcune regole di lettura:

- la dispersione dei punti intorno all'origine degli assi principali mostra la forza dell'associazione nella tabella lessicale;
- se due forme sono vicine sono utilizzate in maniera simile;
- se due modalità della variabile di partizione sono vicine, allora vuol dire che le corrispondenti categorie di documenti utilizzano un vocabolario simile;
- non si può leggere la prossimità di una forma ad una categoria (o viceversa), ma valutare la sua posizione rispetto all'intera nube delle categorie (o delle forme), secondo la logica della rappresentazione β-baricentrica;
- l'importanza di un punto rispetto alla spiegazione di un asse principale è letta in termini di contributi assoluti.

L'utilizzo di una metrica χ^2 ha come conseguenza quella di assegnare alle forme rare, ossia quelle con un minor numero di occorrenze nel *corpus*, una importanza eccessiva rispetto all'obiettivo dell'analisi: questo se da un lato mette in evidenza le forme con un uso peculiare o che caratterizzano in modo particolare un documento o un gruppo di documenti, rischia al contempo d'indurre una distorsione nell'interpretazione dei risultati ottenuti.

In taluni casi, a seconda dell'obiettivo dell'analisi e del bisogno informativo connesso, può risultare utile il passaggio da un approccio simmetrico, come nello schema classico dell'Analisi delle Corrispondenze, ad un approccio *non simmetrico*.