LED ON LINE
STUDI E RICERCHE

Luca Giuliano
Gevisa La Rocca

L'ANALISI AUTOMATICA E SEMI-AUTOMATICA DEI DATI TESTUALI

SOFTWARE E ISTRUZIONI PER L'USO

LED
Edizioni Universitarie di Lettere Economia Diritto

ISBN 978-88-7916-382-8

LED Edizioni Universitarie di Lettere Economia Diritto
Via Cervignano 4 - 20137 Milano
Catalogo: www.lededizioni.com - E-mail: led@lededizioni.com

In copertina:
R. van Marle, *Iconographie de l'art profane au Moyen-age et à la Renaissance II. Allégories et symboles*. Nijhoff, Den Haag, 1932

Stampa: Digital Print Service

INDICE

Il progetto del volume è equamente condiviso dai due autori. I capitoli 1, 7, 8, 9, 10, 11 sono stati scritti da Luca Giuliano; i capitoli 2, 3, 4, 5 e 6 sono stati scritti da Gevisa La Rocca.

1.
INTRODUZIONE ALL'ANALISI DEI DATI TESTUALI

Una della qualità fondamentali per un ricercatore o un professionista della comunicazione è quella di saper gestire le informazioni contenute nei testi per estrarne il contenuto e interpretarle. Sintetizzare la rassegna stampa su un argomento, leggere in modo sistematico una documentazione scientifica, esaminare dei rapporti di ricerca, riassumere i risultati di una documentazione, analizzare la trascrizione di un'intervista o di una discussione di gruppo, classificare le risposte a una domanda aperta in un questionario o le e-mail dei clienti di un'agenzia di servizi sono attività che, per portare a risultati convincenti, devono essere eseguite con procedure rigorose, pubbliche, controllabili e, entro certi limiti, replicabili.

Oggi una buona parte di questi testi sono disponibili in formato digitale. Questo ne facilita la memorizzazione, la visualizzazione e la stessa estrazione di informazioni e dati per mezzo di software appositamente sviluppati a questo scopo. L'informatica ha trasformato radicalmente non solo il nostro modo di scrivere e leggere i testi, ma anche di interpretarli rendendo molto più sfumato il confine tra "parole che contano" e "conteggio delle parole".

Privilegiare le parole rispetto ai numeri, significa assumere il punto di vista classico dell'ermeneutica, una pratica che nasce in Grecia (e che qualcuno vorrebbe collegare a Ermes, il messaggero degli Dèi dell'Olimpo) e poi si sviluppa in riferimento agli scritti aventi autorità, per esempio le Sacre Scritture. È stato Wilhelm Dilthey (1833-1911), il fondatore dello storicismo tedesco, soprattutto con *La costruzione del mondo storico nelle scienze dello spirito* (1910), a esporre la necessità di una scienza dell'essere umano nella sua interezza, il cui vissuto si esprime in un mondo di significati e valori che possono essere compresi solo con una filosofia ermeneutica che sia in grado di stabilire un rappor-

to intenso e vitale tra il ricercatore e il suo oggetto. "Comprendere", diversamente da "spiegare", vuol dire per Dilthey risalire dalla espressione dello spirito alla sua interiorità. Questo è l'obiettivo specifico delle scienze che si occupano della realtà dell'uomo. Pertanto l'attività umana, soprattutto in senso storico, secondo questa prospettiva è un "testo", una narrazione che deve essere interpretata.

Sottoporre le parole al dominio dei numeri è invece l'approccio di ricerca della tradizione americana iniziata da Harold Lasswell (1902-1978), che ha dato vita alla *content analysis* e il cui obiettivo è la classificazione logica dei contenuti e dei rispettivi valori semantici, attraverso l'individuazione delle unità di analisi (grammaticali, tematiche o formali) che ne rappresentano la base di interrogazione e quantificazione (Losito, 2002).

Questi due prospettive metodologiche hanno portato a una separazione tra comprendere e spiegare che ha condizionato per lungo tempo le scienze sociali e ha finito per rallentare lo sviluppo di procedure rigorose, controllabili e condivise che sappiano trarre profitto dalla complessità dell'interazione sociale, con il suo "mondo di significati", senza trascurare le relazioni strutturali e analitiche che ne rappresentano il contesto. Oggi nella "pratica di ricerca" l'integrazione è possibile attraverso due approcci che nella letteratura internazionale sono noti come:

- **Analisi dei dati qualitativi assistita dal computer** (*Computer Assisted Qualitative Data Analysis Softwares - CAQDAS*), un approccio semi-automatico che si è sviluppato soprattutto con la diffusione dei personal computer e ha trovato un terreno fertile nei paesi di lingua inglese, in forte connessione con i metodi qualitativi di ricerca. Il suo scopo è facilitare la lettura e interrogazione dei documenti (testi, immagini, rendiconti etnografici, bibliografie ecc.) per trarne sistematicamente delle risposte sulla base di domande "a priori", oppure essere di aiuto per la costruzione di ipotesi e teorie che emergono dalla esplorazione diretta delle fonti stesse.
- **Analisi statistica dei dati testuali** (*Analyse statistique des données textuelles*), un approccio di tipo lessicometrico che ha avuto origine nei paesi di lingua francese ed è, ancora oggi, fortemente radicato nell'Europa continentale. È basato principalmente sul confronto dei profili lessicali e quindi sulla distribuzione delle occorrenze delle parole senza passare attraverso la lettura diretta del testo (per questo è definito anche "automatico").

Questi due punti di vista metodologici hanno dato vita a potenti strumenti al servizio della ricerca, tanto più efficaci quanto più sono utilizzati per arricchire reciprocamente i risultati che si possono trarre sia dagli uni che dagli altri. Le scienze sociali non possono che trarre vantaggio da una integrazione tra quali-

tà e quantità, permettendo al ricercatore, secondo il campo micro-sociale o macro-sociale in cui si produce il testo, di afferrare la totalità dei significati per scendere nel particolare e controllare le ipotesi attraverso classificazioni e confronti, oppure di iniziare dalla frammentazione ultima delle parole per salire verso l'astrazione e la generalizzazione dei numeri e poi tornare al contesto in cui le parole ritrovano l'intreccio dei significati.

1. 1. INFORMAZIONI E DATI

La parola "informazione" si presta a numerosi fraintendimenti. Il termine latino *informatio* (da *informare*) per i romani significava "immagine", "nozione", "idea", sia come nozione preconcetta che come idea derivata dalla conoscenza. Per noi, oggi, "informazione" ha molti significati: oltre a quelli già indicati, è una notizia, un ragguaglio, una direttiva, un simbolo, un gesto, una quantità misurabile in bit (*binary digit*) fisici (l'informazione "metrica" della teoria dell'informazione classica) e, da ultimo, il contenuto di una sequenza di nucleotidi in una molecola di DNA (informazione genetica).

Gregory Bateson diede una definizione esemplare di **unità elementare di informazione** parlando di "differenza che produce una differenza" (1976, p. 493). Secondo Bateson l'informazione è differenza perché si produce solo là dove si presentano delle variazioni, delle differenze. Ma non tutte le variazioni producono informazione perché vi sono differenze irrilevanti. Un pezzo di gesso contiene una infinità di "fatti potenziali", offre una serie enorme di differenze che emergono sia dalle relazioni tra il pezzo di gesso e gli altri oggetti che dalle componenti stesse del pezzo di gesso. Tra l'infinità di differenze noi ne scegliamo alcune, pochissime rispetto al loro potenziale, e queste poche, limitate differenze diventano informazioni. C'è *una differenza che produce una differenza*. Ciò che deve essere sottolineato è che la differenza si manifesta in un contesto di relazioni, è sempre una differenza relativa, che nasce dal confronto e dalla distinzione. Nella teoria dell'informazione si esprime questo concetto definendo l'informazione come "varietà codificata" e il rumore come "varietà non codificata". L'informazione pura e il rumore puro non esistono. Un rumore interferisce con un segnale e un segnale è il codice che permette di escludere il rumore dalla comunicazione. Rumore e informazione dipendono dal contesto, dipendono da una scelta che produce una differenza. Un ricercatore sta ascoltando una registrazione telefonica e si ode il suono delle campane. Per lui è un rumore e si concentra sulle voci dei parlanti. Per un investigatore che sta cercando di identificare il luogo in cui si trova uno dei due interlo-

cutori, il suono delle campane diventa un'informazione mentre la conversazione è un rumore.

L'informazione, così definita, non deve essere confusa con i significati. I significati sono una sorta di "valori d'uso" dell'informazione, hanno a che fare con il senso che diamo alle informazioni quando esse sono inserite in un contesto di relazioni interpersonali e di relazioni tra noi e l'ambiente in cui viviamo. Alcune informazioni (e le cose che ne sono il veicolo) hanno una loro utilità e quindi diciamo che sono *dotate di significato*. Il profumo che viene da una cucina indica che c'è del cibo, ma esprime anche il nostro bisogno di mangiare e la gradevolezza che esso anticipa per il nostro palato. Lo scrosciare della pioggia che ascoltiamo mentre siamo in casa è una benedizione per la terra arida, ma è una scocciatura se stiamo per uscire. Per la prima volta nella storia l'umanità sta sperimentando uno stato di "opulenza informativa" (Maldonado, 1997, p. 88 sgg.). Il rischio è che tutto diventi rumore, perché non sappiamo *perché* dovremmo acquisire una sovrabbondanza di informazioni come quella che ci viene riversata oggi dai mezzi di comunicazione. L'*accesso* all'informazione è diventato un *eccesso* di informazione in cui è difficile discernere i valori d'uso e quindi i significati. Ad esempio lo *spam* che riempie la nostra casella di posta elettronica impedisce o disturba la ricezione dei messaggi che sono invece legati ai nostri interessi, a scelte e obiettivi. Nella visione molto critica di Tomàs Maldonado, il cittadino totalmente informato non solo non può esistere perché le sue capacità di elaborazione personale sono comunque limitate, ma non è nemmeno desiderabile perché questo cittadino "democratico" ideale avrebbe paradossalmente una coscienza individuale completamente annichilita nella sfera pubblica, sarebbe una persona priva di senso.

Il passaggio tra informazione e significato ci introduce nel cuore del problema. Le differenze di cui ci parla Bateson per essere efficaci in termini di conoscenza devono essere a loro volta differenziate e classificate, ossia devono costituire "classi di differenze". E questo non solo in riferimento alla distinzione tra rumore e informazione (che non è affatto oppositiva perché il rumore è situato logicamente all'interno della varietà e l'informazione emerge all'interno del rumore) ma anche in riferimento alle gerarchie di classificazione dell'informazione stessa. La **conoscenza del senso comune**, quella che ci serve nella vita di ogni giorno e che ci permette di utilizzare le informazioni nel modo che ci è più utile, è gerarchicamente posta a un livello superiore rispetto alla **conoscenza scientifica** che introduce nell'osservazione specifici vincoli derivati da interessi, ipotesi di lavoro, circostanze e condizioni in cui si esercita un'attività cognitiva orientata soprattutto a sapere come accrescere il nostro sapere.

Le informazioni raccolte dal ricercatore sono il risultato di un'os-

servazione volontaria, sistematica, finalizzata e controllabile che porta alla raccolta di dati.

Le conseguenze di questa definizione sono molteplici. Ci dice, ad esempio, che l'osservazione il più delle volte non è volontaria e tuttavia è produttrice di informazione e di senso. Una persona che chiede il nostro aiuto desta in noi l'attenzione per soccorrerla, sebbene l'osservazione fosse inaspettata. La conoscenza del senso comune nel migliore dei casi non porta alla raccolta di dati perché l'osservazione da cui scaturisce può essere anche volontaria e finalizzata ma certamente non è sistematica. La definizione ci dice anche che i dati non sono il risultato di una elaborazione dell'esperienza vissuta da un singolo ricercatore. Gli oggetti che fanno parte del nostro sapere hanno un grado abbastanza ampio di intersoggettività (sono controllabili). Quest'ultimo aspetto merita qualche considerazione aggiuntiva sulla natura dei dati.

Il nostro interesse è prima di tutto rivolto alla complessità delle esperienze umane, al vissuto che le persone esprimono di queste esperienze e alla forma che esse assumono come prodotti di individui e di collettività. La conoscenza non si sviluppa a partire dai "dati" per il semplice motivo che essi non esistono. Non ci sono "dati" che non siano costruiti o interpretati (Popper, 1984, p. 124). Nella ricerca scientifica questo processo di astrazione dal piano dell'esperienza e poi di ritorno al piano empirico dell'osservazione si chiama di solito **operativizzazione,** trasformazione di concetti in variabili. L'attenzione si sposta pertanto sulle procedure che il ricercatore mette in atto nella costruzione della **base empirica** della ricerca e sui criteri di selezione adottati. Le informazioni raccolte possono essere il risultato di tecniche diverse tra loro (inchieste, osservazioni sul campo, interviste a singoli o di gruppo, questionari, esperimenti di laboratorio) ed essere costituite da rapporti, trascrizioni, fotografie, videoregistrazioni, materiali di archivio come lettere, documenti ufficiali, autobiografie, articoli di quotidiani e riviste, fonti letterarie e storiche, senza dimenticare le più recenti versioni digitali dei documenti, come le pagine web, le e-mail, le discussioni nei forum, i log delle chat. L'elenco potrebbe essere infinito perché sono infinite e mutevoli le modalità attraverso le quali l'uomo lascia una traccia delle proprie relazioni sociali.

Il ricercatore, per raccogliere le informazioni, deve necessariamente stabilire rapporti e confronti tra esperienze (inclusa la *propria* esperienza di ricercatore) e quindi produrre astrazioni, **concetti** che rappresentano una determinata ricostruzione delle "esperienze elementari" da cui traggono origine (Carnap, 1966, p. 234).

I concetti, come aveva ben intuito Bruno De Finetti (2006, p. 101 sgg.), sono dedotti dagli eventi, sono *proposizioni* che rappresentano momenti di se-

parazione mentale all'interno di un flusso di esperienze. La proposizione "quel fiore è rosso" è il risultato di un'operazione di distinzione in cui un oggetto è assunto come analogo ad altri oggetti naturali diversi tra loro già classificati con un certo grado di approssimazione nella categoria dei "fiori" e a esso viene applicato un predicato "rosso" che, altrettanto approssimativamente, è il risultato di una analogia tratta da altri eventi-distinzione: "questa matita è rossa", "questo attaccapanni è laccato di rosso". L'analogia ovviamente non sta nel contenere la parola "rosso", ma nelle sensazioni (esperienze) che la parola suggerisce. Il concetto quindi ha solo un valore d'uso. È utile per osservare, comunicare e analizzare le esperienze che sono sempre esperienze complesse.

Per non dimenticare mai la relazione che c'è tra il risultato di un'osservazione, il soggetto osservato e l'osservatore è bene porre la massima attenzione al fatto che una variabile non è nulla di più che un concetto tradotto in una entità osservabile: sia essa il potere, il razzismo, l'intelligenza, la disoccupazione, il rischio, la fiducia o qualsiasi altro costrutto di interesse per una certa disciplina. In questo processo operativo il campo cognitivo dell'osservatore e del soggetto osservato interagiscono tra loro in una relazione asimmetrica nella quale il ricercatore detiene più responsabilità e più modalità di scelta, pertanto suo è il potere di definizione dei costrutti e suo è l'onere della prova che essi possano adattarsi al mondo dell'esperienza.

La procedure operative che permettono di tradurre le astrazioni concettuali in entità osservabili passano inevitabilmente attraverso una distinzione cruciale tra **qualità** e **quantità**. Si tratta di due termini oppositivi che hanno avuto un peso decisivo nella storia della filosofia e rappresentano certamente una tensione costante nello sviluppo delle scienze empiriche tra l'attenzione che si deve prestare a eventi/oggetti singolari e irripetibili rispetto a "classi di eventi/oggetti" ai quali è utile applicare un certo grado di generalizzazione. Tuttavia si è parlato troppo in passato (e il dibattito purtroppo non è ancora esaurito) di approcci, metodi, variabili, analisi ai quali di volta in volta è stato applicato l'aggettivo "qualitativo" o "quantitativo". Gli stessi dati sono stati definiti talvolta come quantitativi, qualitativi o, come si usa dire da qualche tempo, "quali-quantitativi".

I due approcci sono solidamente ancorati a due differenti paradigmi epistemologici che si presentano tuttora come dominanti nelle scienze sociali: l'**interpretativismo** sul versante delle tecniche qualitative e il **positivismo** sul versante delle tecniche quantitative. Senza addentrarci in una discussione lunga e complessa di questi problemi metodologici, sinteticamente si può dire che ciò che viene messo in evidenza di solito è il modo in cui il ricercatore si pone rispetto alla realtà oggetto del suo processo conoscitivo. Se l'accento viene po-

sto su un mondo sociale conoscibile in modo imperfetto ma sostanzialmente indipendente dall'agire degli individui, allora il ricercatore si pone all'interno di una scelta di campo positivista; se invece l'accento viene posto sul significato che gli individui attribuiscono alla realtà sociale e sulla interpretazione che essi ne danno, allora il ricercatore si pone in un ambito interpretativista.

In qualche caso si tende a presentare i due approcci come se facessero riferimento a metodi adatti per studiare determinati problemi piuttosto che altri. Il metodo qualitativo che fa largo uso di etnografie ed è orientato a studiare i casi singoli o i piccoli gruppi; il metodo quantitativo che utilizza essenzialmente le interviste formali e i questionari, ed è basato sullo studio di un gran numero di casi, nel rispetto delle regole di campionamento che permettono di generalizzare i risultati. Secondo Andrew Abbott, che ha dedicato un bel volume a questi temi suggerendo soluzioni che puntano ad arricchire i diversi punti di vista piuttosto che a evidenziarne arbitrariamente le differenze, non vi sono metodi "particolarmente buoni per particolari questioni" (Abbott, 2007, p. 27). Qualità e quantità non sono proprietà degli eventi ma proprietà definite dall'osservatore per ricondurre gli eventi all'interno di un mondo osservabile sulla base di scelte teoriche e tecniche. L'informazione, e quindi anche il "dato" che ne consegue, è sempre qualitativa prima di essere quantitativa. Dopotutto la quantità è una determinazione della qualità, mentre non vale il contrario.

Da ultimo, nelle scienze umane, si sta facendo strada la distinzione tra **approccio standard**, e cioè l'adesione a una visione della scienza che formula i suoi asserti sulla base di relazioni controllabili fra proprietà che sono indipendenti dalle valutazioni personali del ricercatore, e **approccio non-standard**, che invece postula una irriducibilità dei fenomeni sociali alle rappresentazioni troppo semplicistiche e riduttive della matrice "casi per variabili" per valorizzare invece una specificità delle scienze sociali che non possono prescindere dal contesto in cui si producono i materiali della ricerca empirica (Marradi, 2007, p. 79 sgg.). Sul piano dei principi le due posizioni appaiono inconciliabili e anche fortemente condizionate da prese di posizione aprioristiche e ideologiche fino al punto da contrapporre i ricercatori tra "qualifobici" e "quantifobici" (Boyatsis, 1998, p. viii). C'è qualcosa di irragionevole in tutto questo, ma nello stesso tempo di perfettamente comprensibile se si pensa a quello che è stato lo sviluppo della scienza moderna, alla sua separazione dal pensiero umanistico e al tentativo che essa ha compiuto di subordinare il qualitativo al quantitativo, fino a tracciare un confine tra interpretazione (soggettiva) e spiegazione (oggettiva), perpetuando quella demarcazione tra *scienze della cultura* e *scienze della natura* che ha indotto Lord Ernest Rutherford, premio No-

bel per la chimica nel 1908 e tra i padri fondatori della fisica nucleare, ad affermare con ironia: "*qualitative is nothing but poor quantitative*" (il qualitativo non è niente più che il quantitativo espresso in modo grossolano).

Il tentativo, compiuto soprattutto negli ultimi due secoli, di matematizzare il mondo, di ridurre la qualità in quantità, nonostante gli innegabili successi raggiunti dalle grandi leggi della meccanica e della fisica a partire da Galilei e Newton in poi, è stato messo in crisi proprio nella fisica matematica stessa da una rivalutazione e rinascita del pensiero qualitativo già con Poincaré alla fine del 1800 fino agli sviluppi più recenti della "teoria delle catastrofi" di Thom, degli studi sul "continuo geometrico" e della topologia, cioè quella geometria delle "figure elastiche" che studia le proprietà di figure che rimangono invariate anche quando vengono sottoposte a deformazioni così profonde da perdere le loro proprietà metriche e proiettive (Thom, 1980).

Si è assistito così al paradosso delle scienze "molli" che si sforzano di imitare la fisica e ambiscono a conseguire una precisione che sia almeno approssimabile a quella delle scienze esatte; mentre le scienze "dure" sono costrette a misurarsi sempre più spesso con problemi di tipo qualitativo. Di fatto sembrerebbe destinato all'insuccesso ogni schematismo che volesse sostituire completamente la descrizione linguistica del reale con la sua descrizione matematica. È sicuramente più prudente adottare un approccio empirico meno ambizioso che include il mondo dell'indeterminato in cui la conoscenza scientifica convive con il dubbio e l'incertezza (Feynman, 1998).

L'obiettivo di ogni ricercatore - e non da ora - è sempre stato quello di trovare una **spiegazione** rigorosa e convincente del fenomeno oggetto di studio. Nelle scienze fisiche e naturali la spiegazione trova il suo modello ideale nella ricerca delle cause, di cui il metodo sperimentale è la traduzione operativa più completa, sebbene non esente da limiti e problemi (Marradi, 2007, p. 83 sgg.). Nelle scienze sociali occorre allargare la sguardo ad altri "programmi esplicativi" e i diversi metodi di cui esse fanno uso provvedono a fornire diversi tipi di spiegazione che non si escludono reciprocamente ma conducono a forme di integrazione che possono contribuire a mettere in luce ciò che non sappiamo (Abbott, 2007, p. 29 sgg.).

Sul piano delle "pratiche di ricerca" in cui si colloca questo manuale, è necessario uscire definitivamente dalle ingannevoli dicotomie di cui si è dato brevemente conto. Occorre compiere delle scelte che permettano di circoscrivere i problemi che siamo in grado di trattare, di utilizzare gli strumenti più adeguati per tentare di risolverli, seppure provvisoriamente, e di accrescere quindi le nostre conoscenze.

Il richiamo alla inscindibile unità e complessità dei dati risponde a questo

obiettivo. I dati costituiscono delle basi empiriche che non sono né qualitative né quantitative; i dati che raccogliamo corrispondono a informazioni su eventi/oggetti materiali, psichici e culturali che tentiamo di tradurre in parole e/o in numeri. Ci dobbiamo chiedere perché lo facciamo.

In quanto ricercatori il fine che guida la raccolta delle informazioni è sicuramente l'**analisi**. D'altra parte il termine "analisi" è già nel titolo di questo libro e ci sarà pure una ragione per questo. Il pensiero scientifico, anche su questo argomento, ha manifestato dubbi e riserve, ma vorremmo evitare ancora una volta di porre questioni che si basano sulla critica dei fondamenti. Non ci sono informazioni che possano essere colte nella loro totalità. Per esempio, non riusciremo mai a definire il gioco, o qualsiasi tipo di gioco, cercando di trovare qualcosa che sia in comune con tutti gli eventi che chiamiamo "giochi". È stato Ludwig Wittgenstein a portare alle estreme conseguenze questa riflessione (1999, §66-67). Quello che possiamo fare è cercare qualcosa di comune tra i giochi di scacchiera, i giochi di carte, le gare sportive e così via. Nessuno di questi "oggetti-gioco" è identico all'altro, ma è possibile descriverli con una rete di somiglianze e di differenze, cioè di gruppi di somiglianze che differiscono tra di loro per qualche carattere. Wittgenstein parla di "somiglianze di famiglia". I giochi sono legati tra loro da parentele, formano una famiglia. Questo modo di procedere è analitico e non pretende di conseguire una descrizione completa degli eventi "giochi" più di quanto possa pretendere di giungere a definire un linguaggio ultimo e completo per descrivere il mondo. Non possiamo conoscere a priori confini che non siano già tracciati; però possiamo tracciare dei confini secondo scopi particolari che riteniamo importanti per il conseguimento degli obbiettivi che ci siamo posti.

L'analisi è una procedura che ci permette, in vista di uno scopo definito, di passare dal complesso al semplice, dal costituito ai costituenti. Se lo scopo è quello di raccogliere le informazioni, una procedura di analisi adeguata è quella che ci permette di passare dall'evento che si presenta come un processo continuo alla nostra esperienza alle parti costituenti di esso che sono i dati. Forse è il caso di ricordare che abbiamo definito "dati" i prodotti informativi di un'osservazione volontaria, sistematica, finalizzata e controllabile. Un uomo che sta arando il campo non è un dato. La descrizione di un uomo che sta arando un campo è un dato; lo è la fotografia che ritrae questo evento; oppure la descrizione dell'aratro e i modi con cui lo si usa. La misura dell'area del campo è un dato.

I dati sono la ricostruzione in forma discreta nel dominio dell'osservazione di quello che accade nel continuo dell'esperienza.

I concetti di "continuo" e "discreto" rimandano, in prima battuta, a due

modelli matematici. Il discreto è classicamente rappresentato dalla successione dei numeri naturali, mentre il continuo è la retta reale, cioè il sistema dei numeri reali come totalità dei decimali infiniti. La definizione pertanto necessita di un approfondimento. Un dipinto di Vincent Van Gogh che ritrae il contadino che lavora con l'aratro è forse un dato? No, se c'è una fruizione delle qualità estetiche dell'opera. Sì, se vi è l'intenzione di ricostruire storicamente la figura del pittore, lo sviluppo dell'impressionismo oppure si sta osservando il dipinto come testimonianza della vita dei contadini nel XIX secolo. L'opera del pittore, analogamente a quella del musicista e del danzatore, si colloca nell'ambito del continuo: la comunicazione artistica è dotata di una semantica "mobile" e "sfumata" che contrasta con la pretesa semantica rigida della scienza (Manin, 1978, p. 965). La sinfonia in sol minore di Mozart nel momento in cui la ascoltiamo e prende forma nella nostra esperienza ci dice qualcosa di più dello spartito della sinfonia in sol minore. L'ascolto della musica non avviene solo con il cervello, coinvolge tutto il nostro organismo: il corpo e le emozioni. È impossibile distinguere la conoscenza dell'ascolto dall'esperienza dell'ascolto. La sinfonia in sol minore ci trasforma e prende vita dalla interazione tra l'orchestra e la nostra presenza attiva come ascoltatori. È un processo. Per *spiegare* questo processo, per dare una motivazione del perché ci piace quest'opera, dobbiamo fare ricorso a simboli, metafore e, inevitabilmente, anche a termini tecnici del linguaggio musicale. Eppure avremo sempre la consapevolezza di non essere riusciti a comunicare agli altri la nostra reazione emotiva nella sua completezza. Il processo di conoscenza è incomunicabile. Solo la conoscenza riflessiva, la consapevolezza del conoscere che è il frutto della condivisione di tratti dell'esperienza, permette di operare una riduzione di complessità che passa attraverso la selezione delle proprietà rilevanti.

Il continuo, cioè la possibilità di variazioni arbitrariamente piccole di una certa caratteristica, presuppone da una parte l'infinito matematico (che non è interpretabile direttamente in forma sperimentale) e, dall'altra, quella percezione di continuità che ciascuno di noi prova quando pensa al movimento nello spazio, alla durata nel tempo, al flusso delle emozioni. L'informazione e il linguaggio si presentano in sistemi di unità discrete. Il linguaggio "digitalizza" il processo continuo dell'esperienza perché provvede a isolare gli oggetti e gli eventi in vista di una classificazione. Nella percezione del cieco che usa il bastone per muoversi a tentoni dove inizia e dove termina il suo *io*? Il suo sistema mentale finisce con l'epidermide della mano che impugna il bastone? Oppure comincia con la punta del bastone? Secondo Bateson queste domande perdono di senso appena si tiene presente che ciò che si vuol spiegare è un determinato comportamento: la marcia del cieco con il bastone (Bateson, 1977,

p. 500). Per giungere a questa spiegazione occorre selezionare e introdurre nel sistema da analizzare l'uomo, il bastone, la strada, gli oggetti che incontra, i rumori che ascolta ecc. Ciò che importa nel definire il sistema mentale è che esso deve essere "delimitato", "ritagliato", in modo da non lasciare fuori ciò che potrebbe rendere non convincente il modello esplicativo adottato. In breve: i dati devono essere raccolti in modo adeguato alla spiegazione plausibile che si intende offrire e sottoporre alla prova dei fatti.

I dati si collocano nell'universo del discorso, sono l'espressione di un sapere selezionato, organizzato, controllato e controllabile. Il tema è affascinante e potrebbe portare molto lontano dagli argomenti di questo libro. Ciò che importa, in questo momento, è avere ben chiaro che la ricerca nelle scienze sociali ha a che fare con dati che provengono da fonti eterogenee: da interazioni personali trascritte, fotografate o video-registrate (**etnografie**); da modelli di rilevazione, somministrazioni di questionari o schemi di intervista, più o meno strutturati (**sondaggi**); da registri o documenti istituzionali di fonte amministrativa o burocratica come atti notarili, schede di registrazione, certificati anagrafici, fatture (**documenti**); da testimonianze come lettere, e-mail, diari, blog, memorie, oppure articoli di giornali, rapporti, verbali di riunioni, (**narrazioni**); da istruzioni su come si devono eseguire determinati compiti, regolamenti, leggi, statuti e costituzioni (**prescrizioni**); da interpretazioni di idee, commenti, forum, spiegazioni più o meno soggettive e documentate (**argomentazioni**). Questa classificazione non pretende di essere completa, dimostra anzi che molte di queste fonti di informazione, soprattutto quelle che non sono state create appositamente all'interno di una ricerca (**documenti artificiali**) ma sono costituite da materiali prodotti (**documenti naturali**) durante i processi di comunicazione (articoli, libri, epistolari, trasmissioni radio-televisive, pagine e documenti in web), permettono di rilevare dati molto eterogenei tra loro e che si presentano in forme diverse all'interno di una stessa nomenclatura.

Di tutta questa produzione di dati, questo manuale seleziona solo una parte: quelli che possono essere espressi in forma di testi e parole. Lo fa scegliendo due modalità di analisi ben distinte: l'analisi semi-automatica, che è in gran parte basata sull'unità concettuale del testo e sul complesso dei significati che il testo intende comunicare; l'analisi automatica, che è invece basata sulle parole e quindi sulla frammentazione del testo nelle sue unità minime costitutive. Vedremo, durante il percorso, che questi due momenti - separati nella tecnica – si possono e si devono integrare nel metodo perché, in un certo senso, **tutti i dati sono qualitativi**; tutti i dati che raccogliamo corrispondono a eventi/oggetti materiali, psichici e culturali che tentiamo di tradurre in parole ("Carlo sta piangendo"; Antonio è ubriaco") e/o in numeri ("il termometro

segna 24 gradi centigradi"; "su 100 persone, 51 hanno votato sì, 39 hanno votato no e 10 si sono astenute), al fine di costituire basi empiriche che ci siano utili per spiegare i fenomeni sociali e controllare le nostre teorie. Ma, con la stessa modalità provocatoria, potremmo affermare che **tutti i dati sono quantitativi** perché è sempre possibile convertire il linguaggio delle parole nel linguaggio dei numeri attraverso un processo di codifica e poi, a sua volta, riportare i numeri (o meglio le misure) e le relazioni individuate tra i numeri in interpretazioni e spiegazioni che non possono essere altro che sequenze ordinate di parole dotate di senso. Si tratta di raccogliere la sfida che aveva lanciato un maestro dell'ermeneutica del Novecento, Paul Ricoeur, secondo il quale era necessario rivedere radicalmente le due nozioni di spiegazione e interpretazione, la prima troppo debitrice alla logica causale di origine naturalistica e la seconda profondamente influenzata dalla psicologia della comprensione di Dilthey. Il testo è il luogo in cui, nel "circolo dell'interpretazione", si ricompone la dialettica tra spiegare e comprendere:

> (...) comprensione e spiegazione non sono due metodi tra loro opposti. A rigore, solo la spiegazione è metodica. La comprensione è il momento non metodico che precede, accompagna e ingloba la spiegazione. Per contro, la spiegazione sviluppa analiticamente la comprensione (P. Ricoeur, 1987, p. 90).

Con l'analisi dei dati testuali quello che cerchiamo di ottenere è uno schema interpretativo che soggiace al testo; una forma di gestione della conoscenza particolarmente adeguata in una situazione di sovraccarico informativo come quella che si è generata attraverso la digitalizzazione dei testi in Internet. In questo ambito necessariamente la ricerca tende a sviluppare meccanismi automatici di estrazione del contenuto che non necessitano della lettura diretta dei testi. Tuttavia non dobbiamo dimenticare che gli automatismi razionalistici non possono supplire da soli alla conoscenza tacita che si esprime nel contesto e nell'extra-testo. Sarebbe assurdo pensare di individuare uno schema interpretativo nelle opere di Shakespeare attraverso un'analisi automatica senza conoscere la mitologia classica, la storia dell'Inghilterra del XIV-XVI secolo e la poetica del teatro elisabettiano. Dobbiamo dare per scontato che nessun ricercatore si avventurerà ingenuamente nell'analisi automatica dei dati testuali senza una ricognizione della complessità cognitiva che i testi esprimono sia che si tratti di testi finzionali che di testi empirici. D'altra parte l'approccio automatico e semi-automatico all'analisi del contenuto non è in grado di aggirare il problema del rapporto tra teoria e osservazione imposto da Popper (1983, p. 452). Nemmeno i software del CAQDAS sono in grado far nascere la teoria interpretativa dall'analisi dei dati come Atena dalla testa di Zeus. La teoria pre-

cede l'osservazione così come l'apprendimento dei segni linguistici di base precede la comprensione della lingua e apre la strada a nuovi percorsi di apprendimento (Boyatzis, 1998). Anche quando i dati testuali assumono una codifica numerica e le parole vengono sottoposte a conteggio ciò che le relazioni tentano di misurare è il significato. Le parole sono soltanto gli elementi microscopici che compongono le unità di senso, sono come coriandoli colorati che si dispongono in modo ordinato fino a costituire delle forme riconoscibili. Tuttavia è il ricercatore con le sue **ipotesi di lavoro** e con le sue **scelte** che imprime una certa direzione all'osservazione della nuvola di coriandoli, facendosi guidare dai modelli statistici, dalle sue intuizioni e dal rigore delle argomentazioni.

L'analisi dei dati testuali non è più un metodo pionieristico ma è ancora – e forse lo sarà sempre – un metodo di frontiera. Non offre una soluzione unica e valida per tutti i problemi di ricerca. Il ricercatore è costretto a muoversi con agilità e perizia tra diversi software e discipline cercando un percorso adeguato agli scopi che si prefigge. Spesso il suo scopo principale è l'esplorazione preliminare, la navigazione nel corpus in cerca di un approdo alle proprie idee o di un punto di appoggio alle proprie convinzioni. È una frontiera affascinante proprio perché ogni volta si pone come una sfida, la sfida di chi tenta di raccogliere gli indizi per sciogliere l'intreccio, l'enigma del significato, l'interpretazione del testo.

1. 2. TESTO, SIGNIFICATO E INTERPRETAZIONE

Textus deriva dal latino, come participio passato di *texĕre*, una parola antichissima dalla radice TEKY che indica il lavoro del taglialegna e del carpentiere. In questo senso è stata rilevata nelle aree indoiranica, greca (*téktōn*, "carpentiere"), slava, germanica e celtica (Devoto, 1979).

'Tèssere', pertanto, va inteso come ordire una trama di fili, una tela, così come il carpentiere disponeva i blocchi di legno. Il verbo viene usato anche in senso figurato, "ordire una macchinazione o un inganno". Da qui 'tessuto', 'intreccio' e quindi "complesso linguistico del discorso" (Segre, 1981) così come appare nella *Institutio oratoria* (IX,4,13) di Quintiliano. È proprio dall'affermasi del *textus* nel latino dell'era cristiana che Cesare Segre vede il trionfo della scrittura, delle religioni del Libro a fronte della diffidenza che i Greci avevano per la parola scritta intesa come semplice trascrizione del discorso orale.

D'altra parte la società antica è caratterizzata dalla comunicazione orale,

una comunicazione in cui i messaggi vengono recepiti nella **situazione** in cui sono emessi. "Io parlo, tu ascolti." La comunicazione è prodotta nel mondo della vita e il discorso prende senso solo dai rapporti **particolaristici** tra i parlanti. È una **comunicazione sincronica**, delimitata nel tempo e nello spazio. La conoscenza nella società della comunicazione orale è esperienza diretta e trasmissione di esperienze tra le generazioni. L'esperienza ricondotta alla forma orale si trasmette in modo ciclico: è sempre uguale a se stessa e ritorna all'inizio di ogni ciclo generazionale per trasmettersi identica nel ciclo successivo. La società si racconta attraverso il mito e il racconto è dotato di una forza di conservazione eccezionale. Nessun racconto è eterno quanto il mito. Le società prive di scrittura conservano paradossalmente la propria cultura in modo più rigido e stabile di quanto non accada nelle società della scrittura. Il messaggio nella trasmissione orale è attuale nel momento stesso in cui viene emesso. Il testo del discorso non esiste. Il discorso è "in atto".

Nelle società della scrittura il discorso è trascritto per essere conservato. Il testo della trascrizione orale permette la comunicazione asincrona, la **comunicazione diacronica**. Il testo si separa dal mondo della vita, dalla situazione in cui viene creato il discorso. "Io scrivo, tu leggi." Un messaggio scritto su una tavoletta o su un papiro varca i limiti del tempo e dello spazio per essere letto a grande distanza e anche molti secoli dopo. Per il significato, l'interpretazione e la ricezione del messaggio, la situazione "in presenza" non è più una necessità. Diventa importante invece il **contesto** della trascrizione e la scelta che ne è all'origine: concepire messaggi che sono intenzionalmente **universali** (Stele di Rosetta). La scrittura nasce nelle società antiche perché i discorsi non sono più limitati alla immediatezza della situazione in cui vengono enunciati. I discorsi si trascrivono perché sono destinati a sopravvivere alla dimenticanza, a trasmettere l'identità sociale e la permanenza di una coscienza collettiva.

La scrittura al suo esordio annuncia la sacralità. Si formano identità collettive forti che organizzano la trasmissione delle esperienze in modo lineare e gerarchizzato. Sono società centralizzate, spesso dotate di una casta sacerdotale che è depositaria dell'interpretazione dei testi. Sono società che si raccontano attraverso il libro. È così che nascono le "religioni del Libro".

L'introduzione della scrittura ha determinato un distacco della parola dal "corpo vivo" in cui essa viene prodotta (Levy, 1997, p. 28). La scrittura non è soltanto la registrazione della parola detta, ma si presenta come una "virtualizzazione della memoria" e ne permette la trasmissione. Attraverso la scrittura la memoria si attualizza nel momento in cui il testo è esposto alla lettura. Tuttavia, il senso del testo non è indipendente dal lettore e dalla sue scelte. Il testo assume un significato solo in presenza di un lettore. Il lettore collabora al-

l'attualizzazione del testo mettendo in collegamento il testo, anche inconsapevolmente, con un mondo di significati che gli appartengono in quanto lettore. Vi è un extra-testo, qualcosa "fuori dal testo", che contribuisce al testo ma, essendo condiviso con il lettore, non ha bisogno di essergli spiegato (o meglio, l'autore del testo "ritiene" che non debba essergli spiegato). In un romanzo contemporaneo, ad esempio, la parola "automobile" normalmente non ha bisogno di altre spiegazioni perché chiunque sa di che cosa si tratti; non varrebbe la stessa cosa per il "landò", un tipo di carrozza elegante a quattro ruote in uso nell'Ottocento. Lo stesso testo letto da persone diverse produce diversi "significati" del testo. La lettura non è una attualizzazione del testo virtualmente rappresentato da segni o ideogrammi. La lettura è una **attualizzazione dei significati** del testo.

Ecco allora che nella scrittura sacra si pone il problema della tradizione critica e delle pratiche interpretative del testo (l'ermeneutica) che ne devono assicurare l'universalità. Da qui la grande importanza che viene affidata alla lettura del testo, alla sua memorizzazione e trasmissione da una generazione all'altra, per sottrarla alla soggettività e alle distorsioni che minerebbero la sua natura immutabile ed eterna. La religione musulmana, ad esempio, pone il Libro (il Corano) nel suo centro. L'assoluta superiorità del testo rappresentato dal Corano è un vero e proprio dogma di fede nell'Islam. Il Corano è la Parola di Dio. Il Corano è Dio che si fa testo. Il Corano viene memorizzato nelle scuole coraniche attraverso la recitazione salmodiante. Così come viene fatto per la recitazione della Torah nella sinagoga. La sacralità del Sefer Torah (il Pentateuco) si esprime anche nella sua trascrizione che avviene secondo regole molto rigide sulla preparazione dell'inchiostro, su come effettuare le eventuali correzioni ecc. Il Sefer Torah è sacro non solo nella enunciazione ma anche nella sua materialità: quando è danneggiato e inutilizzabile viene chiuso in un contenitore di terracotta e sepolto nel cimitero ebraico.

Il testo subisce una ulteriore trasformazione nella società della comunicazione digitale. La comunicazione digitale non conosce tempo e non conosce confini. È contemporaneamente sincronica e diacronica. Il testo, nella comunicazione digitale, appare là dove qualcuno lo richiede. La sua **universalità** dipende dalla compresenza di altri testi, dalla interconnessione dei messaggi tra loro. Il contesto non è più il mondo della vita cristallizzato in un sistema di conoscenze che ne permettono la ricostruzione. Il contesto è rappresentato da tutti gli altri testi che sono collegati al testo in una **rete**. Nelle società della comunicazione digitale l'**esperienza** è **reticolare**. Priva di centro, molteplice, diversificata.

La società digitale si racconta attraverso l'ipertesto. L'ipertesto è una vir-

tualizzazione del testo, un suo potenziamento. Il testo reticolare è l'insieme delle memorie virtualizzate, l'intelligenza collettiva. Il testo in Rete si attualizza in una forma particolare di lettura che è la navigazione ipertestuale, intertestuale e intratestuale.

Il "tessuto linguistico del discorso" (Segre, 1981, p. 269) è realizzato attraverso successioni di lettere e accenti, interrotti da spazi o da segni d'interpunzione, che costituiscono le parole disposte in righe parallele. La lingua si presenta all'osservazione attraverso i testi. Un testo scritto è composto di segni che sono ordinati in una certa sequenza per formare una catena. L'ordinamento è determinato: da sinistra a destra per l'alfabeto latino, da destra a sinistra per l'alfabeto ebraico, dall'alto in basso per l'alfabeto mongolo ecc.): "Chiamiamo testo la totalità di una catena linguistica così sottoposta ad analisi" (Hjelmslev, 1970, p. 111).

La parola "testo" viene utilizzata sia per indicare il contenuto di un discorso (il **significato**) quanto per indicare i segni da leggere, il veicolo materiale della trascrizione (il **significante**). Quando vi trovate di fronte a un testo scritto in una lingua straniera di cui non sapete nulla, nemmeno i segni dell'alfabeto, allora vi trovate di fronte a un significante allo stato puro. Naturalmente c'è un rapporto tra significante e significato, perché senza significante mancano le condizioni in cui si manifesta il significato. Senza significante non c'è possibilità di esprimersi e dunque non c'è più significato. Il significante è un'immagine acustica che, associata a un significato, forma un segno (Saussure, 1970, p. 83 e sgg.).

Il rapporto tra significante e significato è rappresentato dalla **significazione**. La significazione è indipendente dalla natura del significante sulla base del quale si manifesta. Alcuni autori hanno criticato questa impostazione (Abrham Moles, Marshall McLuhan), sostenendo che il modo di presentare qualche cosa (per esempio un prodotto pubblicitario) ha un impatto maggiore di ciò che viene presentato realmente. E chiamano questo (per esempio l'immagine a colori) come "contenente" che affascina il pubblico e agisce su di esso al di là del suo contenuto letterale. In realtà essi sembrano confondere due piani diversi. L'immagine non è significante, ma è già significato. Il vero significante è rappresentato dalla carta e dal contrasto cromatico dei colori.

Il testo ha un'oggettività? In un certo senso sì, la sua oggettività è rappresentata dalla materialità del significante. Un discorso orale non può che essere soggettivo, singolare, irripetibile. Padre Roberto Busa, il gesuita che con il suo studio lessicografico sull'opera di S. Tommaso d'Aquino (*Index Tomisticus*, 1974-1980) è stato tra i pionieri dell'analisi testuale, ha descritto il suo immenso lavoro come quello di un ricercatore che percorre il greto di un fiume or-

mai inaridito e raccoglie i sassi sui quali l'acqua ha lasciato i segni del suo passaggio. Lo scorrere dell'acqua rappresenta la voce melodiosa della discorsività di Tommaso; i sassi rappresentano quanto resta nei testi delle sue parole (testimonianza raccolta durante la presentazione a Roma, all'Università Gregoriana, di una ricerca sull'analisi testuale delle encicliche papali; 27 marzo 2001).

Solo in tempi recentissimi, in un tratto brevissimo della storia della parola, si è resa possibile la registrazione e la riproduzione della voce, ma non per questo l'oralità acquista oggettività. L'oggettività del testo attraverso la scrittura è potenzialmente possibile nel momento in cui la sua attualizzazione è differita, ma l'intervento del lettore, e quindi della significazione, porta con sé l'irrompere della soggettività (Segre, 1981, p. 272). Infatti è all'interno di questa intersoggettività, attraverso la competenza dei codici e dei riferimenti al contesto e all'extra-testo, che si organizza l'**interpretazione**: il confronto, il dibattito, la critica, la problematicità continua che è alla base del lavoro di una comunità scientifica.

L'interpretazione del testo è una approssimazione alla verità nella consapevolezza che la verità è una meta che potrebbe non essere mai raggiunta. La verità non si può trovare da qualche parte, non si può "scoprire" ma si "produce" attraverso i meccanismi dell'interpretazione (Rorty, 1979). Ogni testo ospita contemporaneamente più testi. Ogni testo è soggetto di più interpretazioni. Ma non solo per la intersoggettività del testo. Un testo autografo di solito contiene aggiunte, correzioni, varianti sulle quali il critico esercita le sue scelte per stabilire il "testo autentico". Il problema della autenticità del testo non è eliminato nella comunicazione digitale. Un errore di ortografia, una punteggiatura messa nel posto sbagliato, una decodifica errata compiuta nel passaggio da un sistema operativo all'altro, possono far convivere più testi in un unico testo digitale.

1. 3. CLASSIFICAZIONE DEI TESTI E FORMAZIONE DEL CORPUS

Nel parlare di "testi" che cosa intendiamo? Ci sono sicuramente i testi poetici e letterari, gli articoli dei giornali, ma anche gli atti notarili, come il contratto d'acquisto di una casa, gli atti pubblici come i bandi di concorso, oppure i documenti legislativi. Ci sono le lettere che si scambiano i privati, le lettere in formato elettronico (e-mail), oppure i verbali delle assemblee societarie. Ci sono testi che sono trascrizioni di interrogazioni verbali (gli interrogatori giudiziari oppure le interviste non direttive) e testi che sono semplicemente schede di rilevazione di dati, come il modello di rilevazione del censimento.

Per classificare i testi prendiamo come riferimento una ricerca internazionale promossa dell'OCSE (Organizzazione per lo Sviluppo e la Programmazione Economica) nel 2000 che aveva come scopo la rilevazione delle competenze linguistiche, matematiche e scientifiche dei ragazzi di 15 anni (PISA – *Programme for International Students Assessments*). In questa ricerca sono stati assunti tre criteri principali di classificazione:

- *genere*: letteratura di fantasia e letteratura empirica;
- *categoria*: testi descrittivi, narrativi, informativi, argomentativi, conativi, documenti, ipertesti;
- *struttura*: testi continui e testi discontinui.

La prima è una distinzione di *genere* e riguarda i testi che si riferiscono alla *fiction* e alla *non fiction*. La distinzione è tutt'altro che semplice. I **testi finzionali** sono prodotti di attività estetiche sottoposti a convenzioni letterarie. I testi non finzionali, definiti anche come **testi empirici**, fanno riferimento a dati di esperienza. *Le ultime lettere di Jacopo Ortis* (1806) di Ugo Foscolo è un testo finzionale, mentre le lettere degli immigrati polacchi analizzate da W. Thomas e F. Znaniecki in *Il contadino Polacco in Europa e in America* (1918) sono testi empirici.

Il secondo criterio è riferito alle *categorie di organizzazione del contenuto* e cioè lo scopo per cui i testi sono stati scritti.

I **testi descrittivi** rispondono alla domanda "che cosa?". Possono contenere descrizioni "soggettive" che esprimono il punto di vista di chi scrive (o di chi parla, se il testo è una trascrizione) oppure possono essere descrizioni tecniche e scientifiche che, almeno nelle intenzioni, puntano ad essere "oggettive".

I **testi narrativi** rispondo alle domande "quando?" e "in che ordine?". Possono essere "racconti" (in cui il punto di vista è quello del narratore), "rapporti" (le informazioni contenute nel testo possono essere verificate/falsificate da altri che non sono il narratore), "testi di attualità" (la narrazione viene effettuata da un giornalista e permette al lettore di farsi una sua opinione della realtà in cui vive).

I **testi informativi** sono prevalentemente orientati a rispondere alla domanda "come?". Possono prendere la forma del "saggio esplicativo" (presentazione di concetti più o meno complessi), "riassunti" (sintesi di informazioni contenute in un testo originario), "verbali" (trascrizioni di quanto è stato detto durante una riunione), "interpretazione di testi" (commento a un testo al fine di dare una spiegazione di quanto vi è contenuto).

I **testi argomentativi** rispondono soprattutto alla domanda "perché?". Si tratta di "argomentazioni scientifiche" (interpretazione di idee o sistemi di pensiero e spiegazione di eventi su cui è possibile effettuare controlli di validi-

tà) oppure "commenti" (interpretazioni e spiegazioni che hanno un carattere del tutto personale e soggettivo).

I **testi conativi** forniscono indicazioni su come si devono svolgere determinati compiti. Appartengono a questa categoria le "istruzioni" (come i libretti che accompagnano le strumentazioni tecnologiche) e i "regolamenti" (dagli statuti associativi fino alle leggi istituzionali).

I **documenti** sono testi che servono a conservare le informazioni in una forma predefinita (certificati anagrafici, contratti di compra-vendita, fatture ecc.).

Gli **ipertesti** costituiscono la forma più recente di testi e sono costituiti da parti di testo collegati tra loro in modo che il lettore possa costruire un suo percorso personale durante la lettura.

Il terzo criterio è riferito alla *struttura fisica* del testo.

I **testi continui** sono quelli più consueti, come quello che state leggendo in questo momento. Vi sono delle frasi organizzate in capoversi e in paragrafi. Vi sono (a volte) dei titoli dei paragrafi e i paragrafi sono raggruppati in capitoli o in sezioni dotate a loro volta di titoli che ne riassumono il contenuto o suggeriscono qualche idea. Alcune parole sono evidenziate in qualche modo (tra virgolette, in corsivo, i grassetto, ecc.) e alcune parti del testo sono organizzate in modo da facilitare la lettura (elenchi) o da renderne riconoscibile la funzione speciale (note, indicazioni bibliografiche, commenti tra parentesi ecc.).

I **testi discontinui** sono quelli che normalmente non vengono considerati testi ma che nel programma PISA sono stati comunque classificati (elenchi, moduli, questionari, tagliandi, tabelle ecc.).

Per noi è ancora degno di nota un altro criterio di classificazione che non è stato considerato rilevante nel programma PISA ma che oggi è di assoluta preminenza: **testo digitale** o **testo a stampa.** Probabilmente la distinzione non è del tutto ortodossa perché oggi gran parte dei testi assumono o tendono ad assumere la forma digitale. Tuttavia vi è una differenza sostanziale tra i testi che sono già in origine digitali (e-mail, pagine web, blog, messaggi in forum e newsgroup, conversazioni in chat, ecc.) e testi che sono destinati alla stampa, provengono dalla digitalizzazione del testo a stampa oppure sono solo occasionalmente redatti in forma digitale ma sono orientati alla lettura stampata (come questi appunti che sto scrivendo sul monitor utilizzando la tastiera del computer). In alcuni casi vi sono testi, come le pagine dinamiche generate in web, che non sono pubblicabili in una forma diversa da quella elettronica.

I testi, di per sé, non sono immediatamente oggetti di analisi scientifica. Noi ci avviciniamo a essi con l'intento di analizzarli mediante strumenti quantitativi, di dare una interpretazione e classificazione logica del loro contenuto a partire dalla spiegazione dei valori semantici delle parole. Questo è un approc-

cio **logico-semantico** che va tenuto distinto da un approccio più propriamente **linguistico** rivolto all'analisi delle tipologie discorsive, dello stile, del "come" viene prodotto un discorso anziché del "che cosa" contiene.

Da un punto di vista operativo, per la costituzione di un oggetto di ricerca l'osservazione non è indipendente dalla teoria. Non possiamo delimitare il campo di interesse delle nostre osservazioni senza formulare le ipotesi che ci guideranno nella raccolta e l'analisi dei dati. I testi, pertanto, hanno un interesse e sono analizzabili solo se costituiscono un **corpus** di testi: "Per corpus s'intende un qualsiasi insieme di testi, fra loro confrontabili sotto un qualche punto di interesse"(Bolasco, 1999, p. 182).

Il corpus è l'insieme dei testi sui quali si deve effettuare l'analisi. In molti casi il corpus si costituisce facilmente: per esempio, la totalità delle risposte ad una domanda aperta in un questionario sottoposto a un numero definito di persone; ciascuna delle risposte costituisce un testo. I testi possono essere raggruppati secondo le caratteristiche dei soggetti intervistati (maschi e femmine, classe di età ecc.). In un'analisi comparata degli articoli dei giornali quotidiani su un certo avvenimento, il corpus sarà rappresentato dalla totalità degli articoli pubblicato nel corso di un certo periodo di tempo. Anche in questo caso gli articoli (o semplicemente i titoli) sono testi che possono essere classificati, oltre che secondo la testata, secondo la posizione e secondo il rilievo che hanno sulla pagina.

Ci sono casi in cui il corpus è più difficile da determinare. Il corpus è sempre una conseguenza di decisioni operative in un certo contesto di ricerca: discorsi politici, testi di interviste o storie di vita, messaggi pubblicitari, trascrizione di focus group, raccolte di e-mail su un certo argomento ecc. Il corpus è un insieme ragionato di testi che corrispondono ad un obiettivo e quindi a precise ipotesi di ricerca. È impossibile dire a priori se un corpus è costituito adeguatamente senza assumere come riferimento lo scopo per cui verrà analizzato (Habert, Fabre, Issac, 1998, p. 35). Una classificazione utile e operativa dei corpora è quella che distingue i corpora chiusi dai corpora campionati (Mellet, 2002, p. 6 sgg.).

Un corpus può essere **chiuso**, delimitato e quindi esaustivo, solo nell'ambito di una monografia (per esempio, il corpus dei *Sonetti* di Shakespeare). In questo caso il corpus è analizzato in quanto tale nella sua interezza (oppure nella sua incompletezza dichiarata a priori dal ricercatore) senza alcuna pretesa di generalizzazione e sempre all'interno dei criteri, arbitrari ma argomentati, che hanno guidato la scelta dei testi. Il corpus chiuso, di solito, è molto omogeneo.

Il corpus **campionato** deve rispondere a criteri di rappresentatività. Il

problema non è più quello di essere esaustivi, ma di costituire un campione rappresentativo di una popolazione in senso statistico. Questo è il caso delle ricerche socio-linguistiche che mettono a confronto, per esempio, il parlato contemporaneo tra diverse aree linguistiche. Un corpus linguistico rappresentativo è sempre un'impresa ardua. Come definire la popolazione di riferimento? Quante sono le modalità espressive del linguaggio parlato? Quanto deve essere esteso il corpus? Come facciamo a essere sicuri che le diverse trascrizioni sono complessivamente rappresentative della varietà del parlato? Spesso, il corpus campionato è il frutto di campionamenti per quote piuttosto che di campioni casuali rigorosi, e ancora una volta devono essere le ipotesi della ricerca a dettare i criteri di formazione delle quote e il grado di generalizzazione di confronti e modelli di analisi.

APPROFONDIMENTI TEMATICI

Questo capitolo introduttivo ha lo scopo di definire l'oggetto specifico dell'analisi semi-automatica e automatica dei testi. Il primo punto sul quale fermare l'attenzione è il concetto di "analisi". Per una ricostruzione completa di questo argomento che coinvolge i processi di conoscenza dalla formulazione del problema fino alla spiegazione, il riferimento d'obbligo è Jean-Michel Berthelot, *Les vertus de l'incertitude. Le travail de l'analyse dans le sciences sociales*, Paris, Presses Universitaires de France, 1996. Il tema è stato ripreso, in modo del tutto indipendente e con una radicamento maggiore nello sviluppo di una teoria sociologica, da Peter Hedström, *Anatomia del sociale. Sui principi della sociologia analitica*, Milano, Bruno Mondadori, 2006.

Poiché, nel nostro caso, si tratta di "analisi dei testi", evidentemente era necessario precisare che rapporto vi fosse tra "dati" e "testi". La visione del "dato" come un oggetto che si presenta come tale al ricercatore che lo riceve quasi passivamente è stata a lungo discussa nel corso del dibattito sulla separazione/unione tra scienze umane e scienze naturali (Luciano Gallino, *L'incerta alleanza. Modelli di relazioni tra scienze umane e scienze naturali*, Torino, Einaudi, 1992).

Ormai è da tempo un fatto acquisito che la conoscenza non si costruisce sulle percezioni, sulle osservazioni o sulla raccolta dei dati, ma sui problemi e sulle aspettative che ciascun ricercatore ha rispetto alle soluzioni proposte (Gaetano Calabrò, "Dato", in *Enciclopedia*, IV, Torino, Einaudi, 1978, pp. 376-388). Pertanto a fondamento dei processi di osservazione empirica vi sono le procedure di formazione delle "asserzioni-base" che le diverse discipline scientifiche ritengono adeguate per assolvere sia alle funzioni euristiche che di controllo delle teorie. Si tratta delle procedure, più o meno formalizzate, di "operazionalizzazione" (Alessandro Bruschi, *Metodologia delle scienze sociali*, Milano, Bruno Mondandori, 1999, pp. 55-79) o di "operativizzazione" (Piergiorgio Corbetta, *Metodologia e tecnica della ricerca sociale*, Bologna, il Mulino, 1999,

pp. 81-129) che consentono di tradurre il linguaggio naturale in linguaggio scientifico. Il passaggio dal piano delle esperienze al piano dei concetti e di qui al "linguaggio delle variabili" è presente in tutti i testi di metodologia ma, ai fini di un eventuale approfondimento, va certamente segnalato il saggio di Alberto Marradi, *Concetti e metodi per la ricerca sociale*, Firenze, la Giuntina, 1984 (ripubblicato in Mario Cardano e Renato Miceli [a cura di], *Il linguaggio delle variabili. Strumenti per la ricerca sociale*, Torino, Rosenberg & Sellier, 1991, pp. 120).

Le procedure di definizione dei concetti e la loro traduzione sul piano empirico si incontrano (o si scontrano) con un problema di difficile (e forse impossibile) soluzione come quello della "qualità/quantità".

I termini del dibattito (o della contesa) sono ampiamente rappresentati nel volume a cura di Costantino Cipolla e Antonio De Lillo, *Il sociologo e le sirene. La sfida dei metodi qualitativi*, Milano, FrancoAngeli, 1996; recentemente il tema è stato ripreso da Alberto Trobia, *La ricerca sociale quali-quantitativa*, Milano, FrancoAngeli, 2005, con un ampio apparato di riferimenti bibliografici che permette di ricostruire lo stato dell'arte su un tema ampiamente condiviso nella letteratura internazionale (John W. Creswell, *Research design: Qualitative and Quantitative approaches*, Thousand Oaks - CA, Sage, 1994; William M.K. Trochim, *The Research Methods Knowledge Base*, Cincinnati - OH, Cornell University, 2001; Mattew B. Miles, A. Michael Huberman, *Analyse des données qualitatives*, - rev. de J.J. Bonniol, Paris, De Boeck, 2003).

I termini oppositivi "positivismo/interpretativismo" e "spiegazione/ comprensione" possono sembrare delle semplificazioni, ma sono spesso utili sul piano didattico (Piergiorgio Corbetta, *op. cit.*, in particolare il Cap. I e Cap. X); il tema è discusso in modo più approfondito e problematico in Mauro Palumbo, Elisabetta Garbarino, *Strumenti e strategie della ricerca sociale. Dall'interrogazione alla relazione*, Milano, FrancoAngeli, 2004, e in Massimo Borlandi, Loredana Sciolla (a cura di), *La spiegazione sociologica. Metodi, tendenze, problemi*, Bologna, il Mulino, 2005 (in particolare, nel saggio di Antonio M. Chiesi e nella riflessione empirica di Alessandro Cavalli). Una visione critica ma costruttiva del rapporto tra interpretazione e spiegazione, etnografia sociale e sociologia quantitativa, è alla base della riflessione di John H. Goldthorpe, *Sulla sociologia*, Bologna, il Mulino, 2006. La convergenza di spiegazione e comprensione all'interno del "realismo critico" nelle scienze sociali è al centro dell'originale punto di vista di Peter T. Manicas, *A Realist Philosophy of Social Science: Explanation and Understanding*, Cambridge, Cambridge University Press, 2006.

Il passaggio dalle informazioni ai dati è tipico della "teoria dell'informazione" (Antony Wilden, "Informazione", in *Enciclopedia*, VII, Torino, Einaudi, 1979, pp. 562-628). Il tema del rapporto tra dato e informazione, in specifico per la ricerca sociale, è discusso in Daniele Nigris, *Informazione e intervento sociale. Prospettive metodologiche e operative*, FrancoAngeli, 2000 (in particolare pp. 70-75). In questo manuale si è scelto di raccogliere in un unico contenitore (i dati) ciò che altrove (A. Bruschi, *op. cit.*, pp. 215-240) è tenuto distinto (testi e le collezioni di dati); è la diretta conseguenza di un ragionamento che prende avvio da riflessioni recenti intorno alla costruzione della base empirica di una ricerca (Renato Miceli [a cura di], *Numeri, dati, trappole. Elementi di psi-*

cometria, Roma, Carocci, 2004) e alla sua "ispezionabilità" (Daniele Nigris, *Standard e non standard nella ricerca sociale. Riflessioni metodologiche*, Milano, FrancoAngeli, 2003): è la "pratica di ricerca", ovvero la ricerca concreta, che produce basi empiriche con diverse finalità rispetto alle quali possiamo produrre asserti e relazioni tra asserti. In questo manuale si parte dall'assunto che le basi empiriche della ricerca siano "ispezionabili" perché l'intenzione del ricercatore è di produrre delle argomentazioni controllabili. Ciò non toglie che i testi, come altre modalità di raccolta delle informazioni, possono essere utilizzati con finalità argomentative diverse che arricchiscono l'interpretazione e la spiegazione senza perdere di vista il rigore tipico della ricerca scientifica.

RIFERIMENTI BIBLIOGRAFICI

ABBOTT A. (2007) *I metodi della scoperta. Come trovare buone idee nelle scienze sociali*, Milano, Bruno Mondadori (ed. or. 2004).

BATESON G. (1977) *Verso un'ecologia della mente*, Milano, Adelphi (ed. or. 1972).

BOLASCO S. (1999) *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'intepretazione*, Roma, Carocci (II ed. 2004).

BOYATSIS R. E. (1998) *Transforming qualitative information*, Thousand Oaks - CA, Sage.

CARNAP R. (1966) *La costruzione logica del mondo*, Milano, Fratelli Fabbri Editori (ed. or. 1961).

DE FINETTI B. (2006) *L'invenzione della verità*, Milano, Raffaello Cortina Editore (manoscritto presentato alla Reale Accademia D'Italia nel 1934).

DEVOTO G. (1979) *Avviamento alla etimologia italiana*, Milano, Mondandori.

FEYNMAN R. P. (1999) *Il senso delle cose*, Milano, Adelphi (ed. or. 1998).

HABERT B., FABRE C., ISSAC F. (1998) *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*, Paris, Interedition-Masson.

HJELMSLEV L. (1970) *Il linguaggio*, Torino, Einaudi (ed. or. 1963).

LEVY P. (1998) *Il virtuale*, Milano, Raffaello Cortina Editore (ed. or. 1995).

LOSITO G. (2002) *L'analisi del contenuto nella ricerca sociale*, Milano, FrancoAngeli (IV ed.).

MALDONADO T. (1997) *Critica della ragione informatica*, Milano, Feltrinelli.

MANIN J. I. (1978) "Continuo/discreto", in *Enciclopedia*, III, Torino, Einaudi, pp. 935-986).

MARRADI A. (2007) *Metodologia delle scienze sociali*, Bologna, il Mulino.

MELLET S. (2002) "Corpus e recherches Linguistique. Introduction", *Corpus*, 1, pp. 5-12.

POPPER K. R. (1984) *Poscritto alla logica della scoperta scientifica. I. Il realismo e lo scopo della scienza*, Milano, il Saggiatore (ed. or. 1956-1983).

POPPER K.R. (1984) *Conoscenza oggettiva. Un punto di vista evoluzionistico*, Roma, Armando Editore ("Il recipiente e il faro: due teorie della conoscenza", conferenza per il Forum Europeo del Collegio Austriaco dell'agosto 1948).

RICOEUR P.(1987) "Logica ermeneutica?, in *Aut Aut*, 217-218, 1987, pp. 64-100 (ed. or. 1981).

RORTY R. (1979) *Philosophy and the Mirror of Nature*, Princeton (N.J.), Princeton University Press.

SAUSSURE F. DE (1970) *Corso di linguistica generale*, Bari, Laterza (ed. or. 1922).

SEGRE C. (1981) "Testo", in *Enciclopedia*, vol. XIV, Torino, Einaudi, pp. 269-291.
THOM R. (1980) "Qualità/quantità", in *Enciclopedia*, XI, Torino, Einaudi, pp. 460-476.
WITTGENSTEIN, L. (1999) *Ricerche filosofiche*, Torino, Einaudi (ed. or. 1953).

2.
TESTI ON LINE:
LUOGHI E PROCEDURE

I nuovi media siano essi elettrici o elettronici devono essere intesi come piattaforme da cui i gruppi sociali possono fronteggiarsi creando una serie di tribune per la discussione di problemi cruciali per lo svolgimento della vita sociale stessa (Marvin, 1994, p. 5). L'introduzione di un nuovo media viene così a rappresentare un'opportunità per i "gruppi" di riesaminare, mettere in discussione gli schemi radicati nei vecchi media e posti alla base degli scambi sociali. Nell'organizzazione dei vari pubblici attorno a questi strumenti ha inizio la storia del medium stesso; ne discende che le nuove consuetudini, i nuovi usi messi in atto dal pubblico rappresentano la modificazione di vecchie abitudini non più funzionali nei nuovi contesti.

In questi luoghi la comunicazione è digitata: scritta. Le informazioni scambiate attraverso le e-mail, le opinioni pubblicate su un blog o su un forum, i testi con cui si sceglie di popolare le pagine di un sito web possono diventare una base dati da poter analizzare.

Ognuno di questi strumenti ha delle caratteristiche proprie derivate dalla commistione fra contesti sociali e tecnologici. L'inizio di queste nuove modalità comunicative è senza dubbio rappresentato dalla comparsa del *World Wide Web*, il cui padre è Tim Berners-Lee, che giunse alla sua progettazione iniziando a lavorare a uno dei primi programmi: *Enquire*, il quale gli venne suggerito dal titolo di un vecchio librone trovato a casa dei suoi genitori *Enquire Within upon Everything*, che altro non è se non l'antesignano delle nostre Pagine Gialle. La vision da cui mosse fu quindi il titolo "entrate pure per avere informazioni su ogni argomento": un portale su un universo di informazioni.

2. 1. I LUOGHI DELLA RETE

L'idea del Web sembra essere scaturita da un crogiolo di esperienze. Berners-Lee (2001) racconta che un giorno tornando a casa dal liceo trovò il padre impegnato nella lettura di un libro sul cervello, alla ricerca di indizi su come creare un computer intuitivo in grado di realizzare collegamenti come il cervello biologico. Da allora questo pensiero e le poche chiacchiere scambiate al riguardo con il padre non lo lasciarono più.

La domanda che Berners-Lee si pose non è di molto differente da quella che guidò Vannevar Bush nell'articolo "As We May Think" (Bush, 1945) e Ted Nelson nel progetto Xanadu (Nelson, 1981). Elementi futuribili di cui Tim venne a conoscenza e che studiò e, come lui stesso ammise, ebbero la fortuna di arrivare in un momento propizio, quando gli ipertesti e Internet erano già grandi e lui dovette semplicemente unirli. Il suo lavoro si svolse principalmente al CERN (Comité Européen pour la Recherche Nucléaire) di Ginevra, affiancando agli incarichi ufficiali che gli venivano assegnati lo sviluppo del progetto Enquire.

Il programma era composto da schede che contenevano informazioni, ogni pagina era rappresentata come un **nodo** nel programma e l'unica possibilità di implementare le informazioni e quindi le schede era di inserirle aprendo un collegamento da un nodo già esistente. Tutti i **link** "da e per un" nodo erano infine visualizzabili ai piedi della pagina, come le note a un testo, e l'unico modo per trovare un'informazione era di iniziare a sfogliare le schede dalla prima pagina.

Enquire era dotato di due tipi di link:

- uno interno: che muove da una pagina o da un nodo all'altro ed è visualizzabile su entrambi i nodi ai quali è collegato;
- e uno esterno: che permette di saltare tra i vari file e procede in una sola direzione.

L'idea della connessione fra frammenti di informazione portò Berners-Lee a puntare di più sulla struttura dei **collegamenti fra le informazioni**. Da qui nacque il programma Tangle (intrico, nodo). Le informazioni sono immagazzinate dai computer come connessioni tra caratteri. Tangle era in grado, una volta che ricorreva una certa sequenza di caratteri, di creare un nodo in grado di rappresentarla. Così quando tale connessione di caratteri riappariva, il programma semplicemente creava un rimando al nodo principale. Così facendo, man mano che altre frasi venivano assimilate come nodi e altri puntatori le indicavano nelle ricerche, esse si trasformavano in una serie di collegamenti. In seguito Tangle si dimostrò molto complicato e venne accantonato; restava pe-

rò la necessità di condividere le informazioni che i gruppi di ricercatori del CERN utilizzavano su sistemi di supporto differenti. Tale problema sembrò trovare una soluzione nella scrittura di un programma RPC (*Remote Procedure Call*) grazie al quale un programma pensato per un modello di computer poteva essere reso compatibile con altri.

> Immaginai di combinare i link esterni di Enquire con l'ipertesto e con gli schemi di interconnessione che avevo sviluppato per RPC. Un programma Enquire capace di link esterni significava la differenza che passa tra la galera e la libertà, tra la notte e il giorno. In questo modo avrei potuto creare nuove reti per collegare computer distinti, e tutti i nuovi sistemi sarebbero stati in grado di andare verso gli altri. Per giunta, chiunque li stesse scorrendo avrebbe potuto aggiungere un nuovo nodo collegato tramite un nuovo link (Berners-Lee, 2001, p. 28).

La segnalazione dell'inserimento di un link ipertestuale all'interno di un documento sarebbe stata evidenziata attraverso la sottolineatura delle parole relative al link. In modo che non appena qualcuno avesse cliccato su una parola sottolineata il sistema lo avrebbe portato verso quel link. Ciò che mancava ora era un acronimo che potesse indicare in maniera inconfondibile questo processo, Berners-Lee scelse di iniziare ogni programma relativo a questo sistema con HT, *hypertext*. Bisognava però trovare un sistema per indicare un ipertesto globale. L'aiuto venne dalla matematica, dove per indicare un complesso di nodi e maglie in cui ogni nodo può essere collegato a un altro si utilizza *World Wide Web*. Decise che la sigla WWW rispecchiava in pieno la natura distributiva delle persone e dei computer che il sistema poteva mettere in collegamento (Berners-Lee, 2001, p. 34).

Quasi visionariamente si incominciava a delineare lo spazio dell'informazione così come noi oggi lo conosciamo. Tuttavia passare al *World Wide Web* non era così semplice, bisognava in qualche modo convincere gli utenti o gli altri ricercatore del CERN a utilizzare HT. La soluzione stava nell'URL (*Uniform Resource Locator*), ovvero nell'indirizzo che ogni documento possiede per essere ritrovato. Oggi gli indirizzi su Internet non sono molto diversi da come Berners-Lee li aveva progettati.

L'ultimo tassello per sviluppare e diffondere il Web in Rete fu rappresentato dall'introduzione dell'HTML (*HyperText Markup Language*), che rappresentò una sorta di linguaggio comune che permetteva una navigazione all'interno degli ipertesti e che di lì a poco sarebbe diventata la trama del Web.

Oggi la presenza in Rete può essere raggiunta attraverso:

- il *sito di presenza* - detto anche sito vetrina o sito di primo livello, si propone di presentare un'impresa o un'istituzione delineando la sua storia, la sua at-

tività, le sue caratterizzazioni e la gamma di servizi e prodotti offerti;
- il *sito di informazione e comunicazione* – si indicano come tali tutti quei siti per lo più istituzionali, associativi, ma anche aziendali che assolvono a un unico compito, cioè quello di mettere a disposizione informazioni per gli utenti;
- il *sito di promozione* - si propone di integrare la promozione off line dell'azienda, per questo indirizza i visitatori del sito verso i punti vendita tradizionali ubicati nel mercato fisico; il sito non si propone finalità di vendita diretta on line;
- il *sito di vendita* o *negozio virtuale* - attraverso questo sito è possibile vendere a dei clienti on line, ricevere i loro ordini, perfezionare transazioni elettroniche;
- il *portale* e i *motori di ricerca* - sono i siti utilizzati dagli utenti allorché iniziano una navigazione; dispongono di potenti motori di ricerca sono sostanzialmente una via d'ingresso e guida nella navigazione in Rete;
- il *mall* o *centro commerciale* - un sito che permette a più aziende di presentare e vendere prodotti o servizi ai clienti su tutta la Rete.

I siti appena elencati rappresentano le possibili modalità di essere in Rete, a seconda degli obiettivi prefissatisi nel fare e-business. Con il termine e-business ci si riferisce a quell'insieme di interventi on line, quali la vendita, il *trading*, l'e-banking, la fornitura di servizi o più semplicemente l'offerta di informazioni e il supporto nella ricerca (Foglio, 2002).

Ai siti web - primi elementi cui si pensa quando si parla di Internet – si affianca per notorietà la **posta elettronica** (e-mail), che rappresenta la forma più diffusa di comunicazione mediata da computer (CMC) asincrona.

Lo stile testuale dell'e-mail si può far risalire ai memorandum utilizzati nelle aziende alla fine del '900 per documentare tutto ciò che avveniva al loro interno. Il memorandum era il corrispettivo della *business letter* con la differenza che era destinato alla comunicazione interna anziché alla comunicazione esterna come la lettera. Il memorandum riprendeva lo stile della *business letter* ma poiché era rivolto verso l'interno risultava meno ricercato e più colloquiale; il suo uso aumentò con l'avvento della macchina da scrivere, grazie alla quale fu più semplice evidenziare alcune parti di testo o codificare una struttura da utilizzare universalmente. Per esempio si idearono le nuove forme di intestazione "a", "da" che oggi ritroviamo nell'*header* del testo di una e-mail con le forme *from*, *to*, CC (*Carbon Copy*), BCC (*Blind Carbon Copy*) e *Subject*.

Normalmente l'uso della posta elettronica implica la presenza di un *client* specifico come *Outlook* o *Eudora*, anche se la necessità di poter gestire le proprie e-mail da un altro computer ha sviluppato la realizzazione di servizi di webmail che consentono la lettura dei messaggi attraverso una semplice pagina web.

Una variante della posta elettronica è la **mailing list**. In questo caso il messaggio è spedito a una lista prestabilita potenzialmente interessata a riceverlo avendo espresso questo assenso mediante una precedente sottoscrizione. Una seconda variante è il **newsgroup** che permette la visualizzazione dei messaggi dei diversi utenti in una bacheca virtuale. Questo strumento permette la condivisione di documenti tra tutti gli individui che accedono al server.

Nei **forum**, invece, si ha la possibilità di postare un messaggio o più messaggi partecipando al dibattito che si attiva in una sezione dello stesso oppure prendendo parte a più discussioni; qualora il forum fosse organizzato in sotto argomenti. Naturalmente è possibile interagire con gli altri utenti solo previa registrazione al forum effettuando un login. I forum, in realtà, sono dei veri e propri "momenti" di discussione se non di dibattito su argomenti prestabiliti, che avvengono tra due o più persone che condividono un interesse e decidono di scambiarsi idee, pareri ed esperienze in proposito, al fine di ampliare la propria conoscenza e condividerla con altre persone. Importante in queste discussioni in Rete è la figura e il ruolo svolto dal "moderatore". Il suo compito è quello di monitorare costantemente i dibattiti e intervenire quando gli animi si infervorano oppure può incitare e animare la discussione nel caso in cui questa fosse statica. La sua presenza non è determinante per l'esistenza di un forum, tuttavia la sua figura garantisce che venga rispettata una linea di pensiero consona con l'istituzione del forum stesso. Effettuando l'accesso in un qualunque forum di qualsiasi sito, si può notare che i messaggi "postati" hanno tutti una struttura simile, ciascuno con:

- autore e indirizzo e-mail,
- argomento o *subject*,
- data,
- corpo o *body* del messaggio,
- spesso concluso con la firma e impreziosito con qualche *emoticon*.

Mediante la funzione "inserisci nuovo messaggio" o *new post* appare un *form*[1] da compilare. Una volta riempito e accettato dal moderatore, che dovrebbe passare in rassegna ogni messaggio – ma ciò non sempre avviene -, il messaggio viene inserito in testa a quelli scritti fino a quel momento; in caso contrario viene cestinato. In questo senso si può dire che si è in presenza di una discussione controllata.

Ognuno è libero di esprimere il proprio parere e scambiare idee con un interlocutore virtuale ma, se ritenuto opportuno, il moderatore può mettere a

[1] Si definisce *form* il modulo utilizzato per preparare un nuovo intervento da sottoporre al moderatore.

tacere un'idea, un pensiero, un'opinione di qualunque genere senza dare importanza alla sua provenienza. I principi con i quali prendere tali decisioni sono a discrezione del moderatore o sono stati stabiliti e resi noti precedentemente dal moderatore stesso o dal soggetto che promuove il sito o il forum.

Come variante del forum si può segnalare il *guestbook* dove gli individui postano i messaggi motivati da comuni linee di pensiero o preferenze sui contenuti proposti dal sito. Il *guestbook* è un libro degli ospiti attraverso il quale i visitatori di un sito possono annotare pareri e impressioni, inserendovi i propri dati. Le dinamiche di funzionamento sono le stesse, ma rispetto al forum cambia l'approccio degli interlocutori che sono liberi di scrivere qualsiasi cosa; il moderatore qui si limita a rispondere qualora gli vengano poste eventuali domande o richieste di precisazioni (Metitieri, 2003).

2. 2. I BLOG

Se inizialmente il *weblog* è nato come strumento per tenere traccia dei percorsi di ricerca effettuati nella Rete, oggi il blog consente a tutti di pubblicare senza ricorrere necessariamente a un editore. È nella contrazione del suo nome da *weblog* a blog che è racchiusa la storia e i cambiamenti d'uso di questo strumento.

Letteralmente *weblog* vuol dire "traccia su rete" - termine ottenuto dalla fusione di due vocaboli *web* e *log* – questa espressione fu utilizzata inizialmente dalle comunità informatiche per registrare, fra gli altri, gli accessi dei server Web ospitanti dei siti Internet, in pratica le macchine sulle quali vengono archiviate le pagine che consultiamo navigando in Rete. Quindi, *log* significa "traccia" o "registrazione" e *web* "rete"; conseguentemente l'utilizzo congiunto dei due termini indica una registrazione che è avvenuta nell'ambito di Internet.

La prima funzione delle registrazioni *weblog* è stata, quindi, quella di tener traccia degli accessi a un server in ordine cronologico in modo che chi amministra il sistema possa monitorare costantemente e verificare i tentavi di intrusione nei siti ospitati dal server.

È nel Dicembre del 1997 con Jorn Barger che il termine cambia il suo contesto d'uso. Barger annunciò su numerosi newsgroup di Usenet che avrebbe tenuto un log pubblico delle sue navigazioni nella Rete, scrivendo quotidianamente qualcosa sulle pagine web in cui si sarebbe, man mano, imbattuto. Nel messaggio postato, Barger sostenne che il suo modo di organizzare le informazioni avrebbe preso rapidamente piede e si sarebbe diffuso in tutta la

Rete rapidamente come un *meme*[2]. Barger accluse anche il link mediante il quale poter accedere al suo esperimento; <weblog.html> era la parte finale dell'indirizzo (Jerz, 2003). La comunità di tecnici informatici smise di utilizzare il termine *weblog* per indicare l'accesso ai siti monitorati passando a *server log* e così chi dopo Barger si affacciò a questo nuovo strumento lo conobbe nell'accezione da lui datagli. Stando alle note raccolte da Jerz, le conseguenze che il *weblog* avrebbe apportato a Internet e alla scrittura on line erano chiare al suo ideatore.

Nell'Agosto del 2001 Barger – un amante di James Joyce e dell'*interactive fiction* – postò un messaggio dal titolo "Theory: Write a web-book in a day" con il quale proponeva di raccogliere dei link, di assemblarli e di produrre delle descrizioni in grado di legarli insieme in una forma antologica. L'idea non fu un *meme*; è però evidente che i concetti di reticolarità e di connessione dei network informativi e semantici che possono produrre un *weblog* erano già allora piuttosto chiari.

Nel 1999 Peter Merholz propose di pronunciare il termine in una nuova maniera *we-blog*; da lì ci vollero solo poche settimane a dismettere il *we* e passare alla forma abbreviata *blog*. Bargers si preoccupò anche di dare una definizione di cosa fosse un *weblog*. Questa si trova attualmente on line[3] nella pagina delle FAQ (*Frequently Asked Questions)* del suo *Robot Wisdom Weblog*, alla voce "What is a weblog?". Per l'autore un *weblog* (chiamato talvolta blog, o pagina delle news o filtro) è una pagina web dove un *weblogger* (chiamato talvolta *blogger* o *pre-surfer*) "annota" (nel testo *logs*) tutte le altre pagine che trova interessanti. Solitamente si aggiungono le nuove annotazioni in cima alla pagina, in maniera che il visitatore frequente possa immediatamente percepire quali siano le novità scorrendo la pagina dall'alto sino a incontrare le annotazioni che aveva già letto nella sua ultima visita[4].

Le caratteristiche sopra attribuite ai *weblog*s sono in fondo le medesime delle prime pagine web. Tale è per esempio la pagina "What's New[5]" costruita da Tim Berners-Lee sui server del CERN nella quale presentava le novità prodotte dalla Rete e per la Rete. D'altra parte la tecnologia prende sempre forma e valore dal nostro modo di utilizzarla; così è stato anche per i blog. Inoltre la semplificazione delle procedure di registrazione e strutturazione dei blog ne ha

---

[2] Il termine indica un'idea che ha un forte potere virale in Rete ed è stato ripreso da Dowkins, 1995.

[3] <http://robotwisdom.com/weblogs/>.

[4] La definizione risale al Settembre del 1999.

[5]<http://www.w3.org/History/19921103-hypertext/hypertext/WWW/News/9201.html>.

permesso una capillare diffusione. All'inizio del 2002 in Italia se ne contavano soltanto 300, ma dopo sei mesi erano già diventati mille. Nessuno oggi è in grado di dire quanti sono i blog italiani ma, secondo alcune stime, hanno superato quota trecentomila, quanto basta per considerarli un fenomeno di massa a metà tra diario personale, newsgroup e sito di informazione.

La definizione di diario giunge a questo media dal termine inglese *log book* che indica il diario di bordo, per cui per estensione e appropriandosi della metafora della navigazione, *weblog* è un diario virtuale. Ma il termine diario può essere attribuito al blog solo se lo si intende nella forma di "diario intellettuale", poiché pur essendo organizzato per tematiche è difficile che si rimanga fedeli a uno specifico argomento; i blogger riportano, infatti, la loro opinione anche su altri argomenti. Una suddivisione tematica così rigida può essere fatta solo per i singoli post o argomenti (Granieri, 2005, pp. 27-28).

L'insieme di questi blog genera la blogosfera, un insieme di cluster connessi fra di loro, in cui a volte un blog può occupare una posizione centrale, di leader o star secondo la terminologia della Social Network; oppure può trovarsi in una posizione periferica. Tutto dipende da come i blogger linkano le informazioni.

A seconda delle modalità di ricerca delle informazioni, Giorgio Nava nel suo blog[6] individua e presenta tre tipi di blogger: i cacciatori, i tessitori e gli sciamani.

I **cacciatori** sono coloro che esplorano la Rete alla ricerca della "preda informazione"; nonostante tutti i blogger operino in questo modo, costoro sono quelli maggiormente specializzati in questa attività. Attraverso il loro lavoro strutturano e codificano anche i meccanismi di citazione dell'informazione, che costituisce una delle architetture cooperative della Rete. La loro è un'operazione di giornalismo di secondo livello, cioè costruiscono percorsi personali nel magma della Rete; possono scrivere più cose al giorno, anche se brevi. La lettura di questi blog tuttavia non rappresenta un sostituto al giornale tradizionale, ma un'integrazione, un approfondimento.

I **tessitori** invece aggregano informazioni diverse, costruiscono trame, ordiscono un tessuto: chi tra le informazioni, chi tra i blog. La loro scrittura è meno frequente dei cacciatori e i loro brani più lunghi. Attraverso la loro opera di aggregatori e ripetitori di informazione sono tra i principali artefici della stesura dei nessi di interazione secondaria tra i blog.

Gli **sciamani** dell'informazione sono coloro che sono in grado di costruire trame e tessuti sulla base di limitatissimi dati di partenza. Potremmo

[6] < http://falsoidillio.splinder.com/1043518643#32422> (25/1/2003).

quasi chiamarli degli osservatori partecipanti capaci di estrapolare il significato emotivo dell'azione traducendola nello scritto. Così facendo delineano la nostra collocazione nel mondo, rispondono agli interrogativi esistenziali, giungendo a trovare risposta all'annoso quesito del "chi siamo". Gli sciamani realizzano questo processo partendo da una base informativa esigua, costituita spesso soltanto dai propri moti interiori e dalle proprie osservazioni temprate dalla propria cultura. Secondo Nava costoro sono capaci di usare la scrittura non come uno strumento in vista di un "oggetto da dire", ma come "oggetto in sé", come oggetto della scrittura stessa, reificando attraverso la parola il mondo di cui dovrebbero parlare. Mentre cacciatori e tessitori elaborano una massa voluminosa di informazioni, gli sciamani cooperano con la blogosfera a un livello differente, fornendo un punto di vista, un'angolazione, uno sguardo, talmente singolare, che riesce ad alludere alla molteplicità, che dagli altri è invece accumulata in modo sequenziale e linkata su infiniti spazi.

Determinante è quindi la citazione della fonte dell'informazione che viene realizzata attraverso i "permalink". Il termine è la contrazione di *permanent link*, ovvero link permanente. È il link che un utente può usare per mettere il segnalibro a un blog. La sua importanza è determinata dalla possibilità che dà di poter ritrovare il blog in futuro, anche qualora questo venisse archiviato. Linkando le informazioni il blog realizza un perfetto sistema di scrittura ipertestuale in cui è possibile approfondire un'informazione seguendo i vari rimandi oppure ci si può fermare a una prima conoscenza.

Disposto in una barra laterale del blog c'è il *blogrolling*, una lista di blog che l'autore legge quotidianamente o quasi, rendendo così espliciti i legami sociali esistenti fra i blogger. I blogger hanno a disposizione la lettura dei *referrer*, ovvero dei siti che loro hanno portato nuovi lettori per poter conoscere la loro audience. Si crea un network, una comunità in cui tutti possono esprimere una loro opinione, con attenzione a non violare le regole in uso. L'inserimento in una discussione già avviata avviene attraverso il *trackback*, un meccanismo attraverso il quale un blog può riportare automaticamente le segnalazioni a un determinato post avvenute su un'altro blog. In pratica si crea una vera rete di collegamenti attorno a un argomento che aiuta a tenere traccia dell'evolversi della discussione.

Se è così facile aprire un blog e poi esprimere la propria opinione, sembra inevitabile chiedersi chi garantisce per ciò che viene scritto sui blog. È estendibile ai blog il concetto di "reputazione" elaborato per la compravendita on line a partire, in special modo, dall'esperienza di eBay (Dellarocas, 2003). Su eBay acquirenti e venditori lasciano un feedback sulla qualità delle relazioni che hanno intessuto fra di loro; questo diario funge da deterrente nell'adozio-

ne di comportamenti scorretti, poiché un guadagno immediato basato sull'inganno si tradurrebbe in una perdita futura. In pratica i feedback lasciati sul sito istituzionalizzano o digitalizzano il *Word of Mouth*: il passaparola. I blog sono a libero accesso così come il diario storico di eBay e legano il comportamento del blogger all'identità che costui ha deciso di adottare in Rete. La qualità del comportamento e quindi la reputazione sui blog è misurata in link che vi puntano, quindi allo stesso modo che su eBay si potrebbe decidere di non accettare le regole della comunità ma questo vorrebbe dire essere tagliati fuori. Guadagno a breve termine contro perdita nel lungo periodo.

Attraverso il blog l'identità assume una valenza "storica", il che vuol dire essere in un qualche modo presenti in Rete esponendo la propria persona. Prima dei blog, quindi con i forum o le mailing list, si poteva ricominciare facilmente cambiando *nicknmame*. Il blog rende più complesso tutto ciò, poiché anche nei casi in cui l'identità sia celata dietro uno pseudonimo ricominciare da capo significherebbe riavviare il motore della visibilità e del linkaggio per il proprio blog. La perdita di visibilità e l'importanza della reputazione, in qualche modo garantiscono sull'attendibilità/veridicità dell'informazione.

2. 3. IL DOWNLOAD E LA SUA ETICA

Una tematica centrale nella ricerca socio-etnografica consiste nel definire gli spazi sia del mondo on line che di quello off line e di cercare di individuare le relazioni intercorrenti fra i due e il ruolo giocato, nell'uno e nell'altro spazio, dall'analisi qualitativa. Negli spazi virtuali, siano essi dediti a modalità di comunicazione sincrona o asincrona, è possibile svolgere ricerca qualitativa servendosi di diverse tecniche: dall'osservazione partecipante al focus group.

All'interno degli strumenti asincroni che costituiscono – per noi – quelli di maggior interesse si può effettuare un ulteriore tipo di suddivisione considerando il tipo di informazione che essi producono per l'analisi testuale (tab. 2.1), ovvero individuando se si tratta di una comunicazione:

- *statica* quindi preparata da qualcuno e pubblicata, resa disponibile a beneficio di molti con scarsa possibilità di interazione;
- *dinamica* frutto di uno scambio di comunicazioni, messaggi fra più utenti.

Le tecniche proposte per l'analisi di questo tipo di informazione sono quelle proprie della ricerca qualitativa off line, tuttavia le tecnologie della CMC sollevano un nuovo e importante quesito correlato alla raccolta dei testi. L'esempio più semplice e chiarificatore è dato dalle interviste.

Tab. 2.1 – Tecnologie della CMC e loro uso nella ricerca qualitativa

Strumenti asincroni
Comunicazione "statica"
Siti web: utilizzati per promuovere un prodotto o servizio. Adatti per le ricerche documentali o per l'analisi dell'evoluzione nel tempo di linguaggi o tematiche.
Blog: utilizzato come diario personale o pagina d'opinione. Adatto per seguire l'evoluzione nel tempo di linguaggi, tematiche o immagini.
Comunicazione "dinamica"
Forum: utilizzata come strumento di veicolo di comunicazione tematica. Adatta per i focus group on line, per l'osservazione partecipante dei membri e per l'analisi dell'evoluzione nel tempo di linguaggi o tematiche.
e-mail: utilizzata per messaggi di testo con possibilità di allegarvi file di diversa estensione. Adatta per le interviste one-to-one on line.
Mailing list (server): sono utilizzate per diffondere il messaggio a più utenti iscritti alla lista. Le liste e quindi anche i messaggi possono essere postati da un moderatore. Adatta per i focus group on line, per l'osservazione partecipante dei membri e per l'analisi dell'evoluzione nel tempo di linguaggi o tematiche.
UseNet/Newsgroup: sono gruppi di discussione tematici. Adatti per i focus group on line, per l'osservazione partecipante dei membri e per l'analisi dell'evoluzione nel tempo di linguaggi o tematiche.

In una intervista *face to face* l'intervistatore al termine della somministrazione dei suoi quesiti si ritrova con un nastro da sbobinare e da interpretare. La CMC ci permette di saltare la fase della trascrizione e ci regala un testo già scritto; a volte ci permette addirittura di scaricare materiale, testi, messaggi, home page senza dover intervistare nessuno. Si raccoglie l'opinione di alcuni individui senza doverli informare. Questo apre numerosi dibattiti sulle implicazioni di analisi su di un testo già scritto e raccolto con tale facilità. La CMC offre una "scorciatoia" alla raccolta dei dati testuali, nella quale l'interazione con il soggetto/oggetto dell'analisi diventa opzionale e la trascrizione del testo è già data. Tuttavia, la qualità del dato rinvenuto non è sempre alta, poiché in ambiti come le chat la comunicazione può essere superficiale, poco strutturata, a scarso contenuto informativo; inoltre la CMC resta sempre un ibrido di oralità e scrittura all'interno della quale il ricercatore deve sapersi muovere.

I luoghi di comunicazione asincrona, come ad esempio le e-mail o le pagine web, nel momento in cui vengono inviate o pubblicate rappresentano una fonte documentale cumulabile e confrontabile nel tempo, in grado di restituire informazioni su usi, culture, tecniche di promozione. Il ruolo dell'analista diventa sempre più centrale, la direzione di ricerca è imprimibile al dato, i risultati – a posteriori – forniranno risposte agli interrogativi di ricerca. Inoltre, l'analista che si pone al lavoro su dati o informazioni raccolti o scaricati dalla

Rete deve individuare percorsi di aggregazione da seguire nel predisporre la propria collezione di testi.

Le pagine web sia che contengano informazioni "statiche" o "dinamiche" sono accompagnate da diversi link. È, infatti, importante aver presente che la struttura di un testo scritto mediante supporto multimediale ha una sequenzialità e una forza d'impatto che va distinta dalla linearità di un testo a stampa tradizionale.

"Leggere è seguire un cammino tra quelli suggeriti dalla disposizione del testo" (Bolter, 2002, p. 135). Secondo Jay David Bolter, un testo scritto è dotato di una struttura spaziale che implica anche una struttura temporale: un testo scritto è, infatti, paragonabile a una partitura musicale dotata da un lato di una rappresentazione visiva delle note sul pentagramma e dall'altro di una sua verticalità, che è propria del musicista che riesce a leggere le note negli spazi senza bisogno di suonarle. Il testo multimediale o ipertesto acuisce la libertà di movimento del lettore e dell'autore permettendo finalmente di andare zigzagando tra parole e concetti. Attraverso i link si può quindi passare da un nodo (pagina o paragrafo) all'altro.

La segnalazione della presenza di un collegamento avviene attraverso le "parole attive". La parola - attiva o meno che sia - sembra recuperare la sua forza e la sua connotazione intrinseca; porla al centro di uno studio che la vede protagonista nel momento in cui essa, in forma "muta", esprime un concetto adagiandosi su un foglio, reale o virtuale che sia, sembra una via per confrontarsi con le modalità e le tecniche di analisi dei contenuti di una comunicazione mediata dal computer.

Tuttavia è da chiedersi secondo quale criterio empirico sia necessario ricomporre il testo, quindi riconoscere il tipo di informazione con il quale si lavora, quando si analizza il prodotto della comunicazione veicolata da questi strumenti. Si può cominciare a distinguere fra:

- opinioni "a mezzo stampa": rappresentano un pensiero strutturato intorno a un determinato argomento, poiché sono degli articoli che prima di essere pubblicati sono soggetti a revisione: ci si accerta che quanto si vuole veramente esprimere attraverso l'uso di quelle parole filtri verso il ricevente della comunicazione. Si tratta quindi di "documenti eletti e chiusi", deputati a svolgere una funzione comunicativa e di diffusione, tali sono i testi di un blog o di un sito web di un'azienda o di una Pubblica Amministrazione;
- opinioni "a mo' di commento": rappresentano invece il commento, la risposta allo stimolo offerto dagli articoli pubblicati. La loro è una forma espressiva di "oralità scritta"; chi lascia un commento lo fa in maniera istintiva senza porsi troppi filtri comunicativi, riservandosi il diritto di ritornare su

quanto postato qualora qualcuno decida di rispondere al proprio messaggio e si apra un dibattito; sono quindi scritti in continuo divenire. Si tratta di "documenti estorti e aperti", attraverso i quali si sviluppa una comunicazione fra più soggetti intorno a uno specifico argomento. Tali per esempio sono i commenti postati su di un blog;

- opinioni "appassionate": sono quelle espresse su di un forum tematico che rappresentano le idee che alcuni "esperti" - nel senso di "dedicati" all'argomento - lasciano su di un forum, con l'obiettivo di far sentire la propria voce. Anche questa è una forma ibrida di oralità e scrittura: chi posta un messaggio lo fa per far sentire la propria voce, per aprire un dibattito su un particolare aspetto riguardante quello specifico tema. Siamo in presenza di un'opinione che, nell'idea di chi la rilascia, è quella di un cultore. Diversamente dai commenti lasciati in un blog qui si tratta di "documenti visibili, aperti e accorati", attraverso i quali si sviluppa una comunicazione fra più appassionati intorno a uno specifico argomento.

Nel costruire la "collezione di testi" si deve prestare attenzione a come organizzare le opinioni raccolte; esse possono essere aggregate secondo un "criterio di pertinenza" ovvero apponendo, per esempio, i commenti vicino agli articoli cui si riferiscono o secondo un "criterio di appartenenza". Secondo quest'ultimo criterio nel ricostruire, per esempio, una vicenda riportata su di un quotidiano on line si può – dopo aver letto l'introduzione – proseguire cliccando sul link indicato ed evitando di selezionare il testo non desiderato.

Accanto alla questione della raccolta del dato testuale la CMC apre un dibattito sull'etica dell'uso di tali messaggi e sulla correttezza delle procedure di ricerca. Da un lato, infatti, essa offre ampio margine di manovra al ricercatore dall'altro, permette l'inevitabile ovvero lo sviluppo di una ricerca irresponsabile (Lindlof e Taylor, 2002). Inoltre, l'oggetto di studio - sia esso una mailing list, una comunità, un forum - dovrebbe avere la possibilità di sapere di essere studiato, osservato; per esempio attraverso la comunicazione, mediante invio di un messaggio al forum, dell'avvenuto download dei messaggi postati o comunicando al moderatore o *owner* della lista che si intende realizzare uno studio sullo specifico ambiente. In questo senso la CMC erode il tradizionale significato dell'anonimato e della confidenzialità dell'intervista o della dichiarazione e apre d'altro canto la ricerca di una definizione della validità per tali dati.

2. 4. DOCUMENTO-TESTO, SELEZIONE E PRE-TRATTAMENTO

Qualunque sia lo strumento utilizzato per comunicare in Rete è inevitabile notare come vi sia un denominatore comune: il testo scritto, a volte rivisitato e adattato. La scrittura in Internet richiede una grande rapidità. Si deve abbreviare al massimo la lunghezza delle frasi e delle parole. Il testo elettronico è più spezzato, i periodi corti, con le coordinate in netta prevalenza sulle subordinate. L'impossibilità di far trasparire immediatamente componenti quali il tono di voce, le espressioni del viso, gli umori e gli stati d'animo, ha spinto il popolo della Rete a sviluppare nuove forme di scrittura capaci di tradurre il linguaggio cosiddetto non verbale, creando una notevole quantità di innovazioni comunicative. La prova sono le *emoticons* o "faccette", utilizzate per esprimere le proprie emozioni, inventate il 18 settembre 1982 da Scott Fahlman, ricercatore di informatica presso l'Università di Pittsburg. Questi nuovi strumenti linguistici permettono ai navigatori del ciberspazio di evitare fraintendimenti senza dover rinunciare alla velocità comunicativa.

Tutti questi "accessori" comunicativi sono familiari e facilmente riconoscibili nel momento in cui si legge un messaggio singolo; più difficile è valutarne l'impatto nel momento in cui si trattano i messaggi come un insieme di informazioni. In quel caso, per l'analisi testuale essi costituiscono "rumore" (Giuliano, 2004).

Una buona norma nelle collezioni di documenti on line è di tenere traccia delle modalità di raccolta e del percorso di ricerca effettuato, al fine di rendere possibile una valutazione da parte della comunità scientifica e al tempo stesso iniziare a sedimentare modalità d'uso e di analisi.

Bruschi (1999, pp. 216-218) classifica i documenti[7] in base a specifiche variabili:

- le ragioni della produzione documentale,
- le modalità secondo cui l'informazione è stata prodotta,
- il supporto informativo,
- il contenuto informativo.

In base alla prima variabile possiamo avere: documenti naturali o artificiali. I **documenti naturali** sono prodotti per un qualche fine della vita sociale e secondariamente vengono utilizzati dai ricercatori ai fini scientifici. I **documenti artificiali** sono documenti prodotti specificatamente per la ricerca, quali per

[7] Per "documenti", in questo caso, non si intendono solo i testi scritti, ma insiemi di informazioni registrate su una qualche forma di supporto fisico, dalle fotografie ai murali delle città, dalle storie di vita ai certificati anagrafici.

esempio i dati di un sondaggio o le risposte date a un'intervista.

Distinguendo, invece, mediante la modalità utilizzata per produrre l'informazione abbiamo: i **testi**, che corrispondono all'informazione contenuta in un testo scritto o in un'immagine e nei suoni; le **collezioni di dati**, quando si tratta di un'informazione organizzata secondo matrici di casi descritti da variabili.

Per supporto informativo si intende la fonte dalla quale il messaggio proviene, pertanto possiamo distinguere tra **fonti segniche** (testi *scritti*, testi *orali*, testi *iconici* e testi *audiovisivi*) e **fonti non segniche** (*manufatti* della cultura materiale e *tracce* lasciate dalle attività dell'uomo).

Il contenuto informativo distingue l'informazione in tre categorie: descrittiva, espressiva-valoriale, d'uso. Un documento che contiene resoconti su eventi, siano essi scritti o audiovisivi, avrà un carattere prevalentemente **descrittivo**; un documento che palesa sentimenti e stati affettivi sarà caratterizzato da un contenuto **espressivo-valoriale**. E, infine, se il documento contiene un'informazione d'uso, quindi che si riferisce a oggetti della vita sociale, siamo in presenza di un contenuto informativo **d'uso**.

È opportuno ricordare che quando si distingue fra le tre categorie si stabilisce, semplicemente, quale delle tre sia maggiormente prevalente rispetto alle altre; poiché nella realtà dei fatti esse convivono.

Applicare la classificazione di Bruschi ai documenti o informazioni provenienti dalla Rete non è sufficiente; accanto alle variabili proposte dall'autore proviamo ad affiancarvi nuove modalità (tab. 2.2). Per comodità proponiamo di distinguere due dimensioni: una **concettuale** e l'altra **applicativa**. Nella prima si attribuisce la tipologia di Bruschi alle diverse forme di comunicazione che si possono rintracciare in Rete (statica e dinamica); nella seconda si dà conto delle modalità di selezione e creazione di una collezione di testi.

La dimensione applicativa si articola in cinque variabili:

- *declinazione del contenuto informativo*: dà conto del tipo di contenuto lì espresso, approfondendo la variabile contenuto informativo proposta da Bruschi;
- *percorso di ricostruzione del testo*: esplicita il percorso, la scelta dei link che si è deciso di seguire nella navigazione ipertestuale;
- *criteri di ricostruzione del testo*: individua i criteri attraverso i quali un insiemi di testi è divenuto una collezioni di testi unici;
- *etica di download del testo*: ottenere dei testi che provengono da Internet sembra essere molto più semplice del farsi rilasciare un'intervista; tuttavia anche le parole che fluttuano nel web sono di qualcuno ed è necessario chiedere o quanto meno informare dell'analisi che si sta per compiere chi le ha emesse. Le modalità di richiesta variano, poi, a seconda dell'impersonalità o della personalizzazione del supporto web che viene utilizzato per la comunicazione;

- *pre-trattamento*: questa variabile merita attenzione, poiché il pre-trattamento di un testo è dipendente dal tipo di software che si utilizza per l'analisi e dal tipo di testo con cui si lavora. Pertanto le procedure di pre-trattamento indicate rispettivamente per ogni fonte sono da intendersi come determinate dalla tecnica di analisi - semi-automatica o automatica - del contenuto scelta e dal grado di rumore presente nel testo.

Tab. 2.2 – Classificazione teorica/pratica dei testi contenuti nei diversi strumenti

	Comunicazione statica	**Comunicazione dinamica**
	Dimensione concettuale	
Motivi della produzione documentale	Documenti naturali.	Documenti naturali.
Modalità di formulazione dell'informazione	Testi.	Testi.
Supporto informativo	Testo segnico multimediale.	Testo segnico multimediale.
Contenuto informativo	Descrittivo. Espressivo-valoriale. D'uso.	Descrittivo. Espressivo-valoriale.
	Dimensione applicativa	
Declinazione del contenuto informativo	Opinioni a mezzo stampa. Opinioni a mo' di commento. Opinioni appassionate e personali.	Opinioni appassionate e personali.
Percorso di ricostruzione del testo	Continua a leggere l'articolo o il soggetto. Leggi i commenti all'articolo.	Per singola comunicazione. Come insieme di comunicazioni.
Criteri di ricostruzione del testo	Criterio di pertinenza. Criterio di appartenenza.	Criterio di appartenenza.
Etica di *download* del testo	Per strumenti come i blog che equivalgono a diari personali - anche se pubblici - è buona norma porsi in contatto con il blogger. Per strumenti come le pagine web di aziende o società che equivalgono a documenti pubblici non si richiede alcuna autorizzazione	Il forum è uno strumento messo a disposizione per l'incontro di molti utenti. Occorre quindi chiedere un'autorizzazione: nel caso del forum basta postare un messaggio, nel caso di una mailinglist si può chiedere al moderatore
Modalità di pre-trattamento	Creazione di una collezione di testi. Trattamento dell'*header* presente nel testo.	Trattamento dell'*header*, del *quoting*, delle firme digitali.

È, inoltre, importante preparare il testo in modo che possa essere "assimilato" da un determinato software. Per esempio, le firme digitali e il *quoting* (la citazione di messaggio del testo di un messaggio precedente) in un'analisi automatica con TaLTtaC2 o con Lexico3 possono costituire rumore e, in certi casi, esercitare un'azione di "disturbo" rispetto alle analisi da svolgere; è bene quin-

di preparare – pre-trattare - il file tenendo conto degli obiettivi e della compatibilità fra le caratteristiche del testo e le specifiche del software. Con i software sviluppati all'interno della *Grounded Theory* il lavoro è semi-automatico e l'analista ha maggiore possibilità di azione, può, quindi, non prestare molta cura a questi elementi. Mano a mano che elabora i suoi dati può decidere come operare sul *quoting* o sulle firme digitali.

È obiettivo di questi software garantire un'interpretazione dei testi, siano essi interviste o documenti, e ricondurli a specifici significati; costruire categorie mediante l'estrapolazione dei significati in essi contenuti e stabilire attraverso il loro studio le associazioni e le relazioni tra i significati ivi rinvenuti, in modo da pervenire alla costruzione di teorie generali e particolari. L'analisi quantitativa del lessico consente, invece, di valutare l'aspetto morfologico e sintattico del testo, nonché di produrre un'analisi semantica.

Tab. 2.3 – Classificazione dei software

Analisi	Software	Tipo	Dimensioni del testo	Trattamento
Semi-automatica	Atlas.ti5	A supporto	Determinante	Semi-manuale
	NVivo7			
Automatica	Lexico3	A sostegno	Irrilevante	Automatico
	TaLTaC²			

I pacchetti di lavoro appartenenti a queste due matrici teoriche si possono distinguere per:

- *tipo*: si intende il contributo che il software dà all'analisi e può essere **a supporto** qualora non sia determinante e **a sostegno** quando dall'output dipende l'intera interpretazione;
- *dimensioni del testo*: quando l'uso del software è limitato dalla misura del corpus di analisi, si distingue in **irrilevante** e **determinante**;
- *trattamento*: questa variabile distingue il software utilizzato in base alla componente **manuale**, **semi-manuale** o **automatica** in esso utilizzata. Per esempio, lavorando con software come Atlas.ti5 ci si accorge di farne un uso semi-manuale, in quanto il software serve quasi da *block notes*, ma allo stesso tempo ci aiuta a stabilire relazioni fra categorie. È importante notare che il margine d'azione della "mano" del ricercatore è prevalente rispetto alla "forza" del software, cosa che non accade con Lexico3 e TaLTaC², che sono, prevalentemente, software di analisi automatica.

2. 5. IL CORPUS UTILIZZATO NEGLI ESEMPI: "BULLISMO"

Il modo più semplice e diretto per intraprendere un percorso didattico sull'analisi testuale è quello di esplorare e analizzare un corpus con il sussidio di alcuni software tra i più noti in questo campo (Atlas.ti5, Nvivo7, Lexico3 e TaLTaC[2]).

Il corpus, che chiameremo convenzionalmente "Bullismo", è costituito da 277 messaggi (310 kb) inviati al forum del quotidiano *la Repubblica* dedicato a "Il bullismo nelle scuole" <http://forum.repubblica.it/> aperto in occasione di un episodio di cronaca che riguardava il video di uno studente disabile picchiato dai suoi compagni a scuola e inviato su YouTube dagli stessi studenti. Il messaggio di lancio del forum fu il seguente: "Il video dello studente disabile picchiato a Torino ha riproposto il problema del bullismo nelle scuole. Il ministro è intervenuto e si promettono punizioni esemplari. Ma molti mettono sotto accusa la distrazione di genitori e prof. Cosa ne pensate?".

I messaggi di commento, inviati dal 16 novembre 2006 al 29 aprile 2007 (di cui 262 inviati nei primi 4 giorni di "discussione") sono anonimi ma, sulla base degli indizi che si possono trarre dai *nicknames* dalla loro lettura, è stato possibile classificarli secondo due variabili: "genere" (modalità: maschio, femmina, indefinito) e "operatore" (modalità: sì, no, incerto), dove per "operatore" si intende un insegnante, un docente o un dirigente scolastico (tab. 3.4).

Tab. 3.4 – Descrizione dei messaggi del corpus (Totale = 277)

Variabili	**Modalità**	**Valori assoluti**
Genere	Maschio	163
	Femmina	39
	Indefinito	75
Operatore	Sì	45
	No	123
	Incerto	109

Questa classificazione permette di analizzare il corpus effettuando alcuni confronti esemplificativi sotto l'ipotesi, semplice ma – come vedremo - efficace, che il genere e la professione nell'ambito scolastico condizionano la rappresentazione del fenomeno "bullismo".

I messaggi del forum sono stati salvati in formato ASCII (*American Standard Code for Information Interchange*) detto anche "testo puro". Il codice ASCII, introdotto nel 1963 negli Stati Uniti, nel 1988 è stato adottato come norma in-

ternazionale sotto il nome di ISO 646 (*International Standards Organisation*). Successivamente alcuni produttori di computer e di software hanno sviluppato delle estensioni del codice, portando i caratteri codificabili a 256. Così il codice ASCII, che permetteva di codificare solo i testi in inglese (senza accenti e senza lettere alfabetiche speciali), ha assunto diverse varianti. Tra queste varianti la più diffusa è la tabella ISO -8859-1 detta anche ISO-LATIN o LATIN-1 che permette di codificare le principali lingue dell'Europa occidentale. I codici Windows, MS-DOS e IBM sono estensioni diverse del codice ASCII.

Senza addentrarci troppo in questo argomento, occorre tenere presente che la conversione tra un codice e l'altro e da un sistema operativo all'altro (da Macintosh a Windows, per esempio) comporta dei cambiamenti di carattere, specialmente nella conversione delle lettere accentate. Per esempio, il codice 233 in Macintosh restituisce *é* mentre lo stesso codice in ANSI Windows restituisce *Ú* (i caratteri sono richiamabili direttamente dalla tastiera numerica attivando BlocNum e digitando Alt + il numero decimale corrispondente). Pertanto è necessario porre la massima attenzione nel salvataggio dei file quando si passa da un sistema operativo all'altro.

In generale, è sempre bene salvare il corpus in formato "testo puro", cosa che in Windows è facilmente eseguibile con una delle opzioni di salvataggio di un documento Word. Per la gestione dei testi in formato ASCII, specialmente quando si tratta di testi di grandi dimensioni, è preferibile utilizzare un "editor testuale", cioè un programma che non si presenta come un *word processing* ma è piuttosto una versione avanzata di Notepad.

Un eccellente editor di testi è TextPad, scaricabile per il consueto periodo di prova dal sito *Helios Software Solutions* <http://www.textpad.com/>. TextPad opera solo in formato ASCII e permette una gestione semplice ed efficace delle principali funzioni di modifica del testo. Con file di grandi dimensioni è più veloce di Word e permette di compiere agevolmente anche le operazioni più sofisticate.

APPROFONDIMENTI TEMATICI

L'attenzione della linguistica per i cambiamenti che Internet e la comunicazione digitale hanno provocato nel linguaggio è tuttora limitata. Il lavoro più sistematico e completo in questo senso è quello di David Crystal, *Language and the Internet*, Cambridge, Cambridge University Press, 2006 (agg. di una edizione del 2001), che prende in esame le e-mail, le comunità di chat e le comunità virtuali. Molta più attenzione hanno avuto gli studi sui cambiamenti culturali indotti dalle nuove tecnologie, come in Jay David Bolter e Richard Grusin, *Remediation. Competizione e integrazione tra media vecchi e nuovi*, Milano, Guerini Studio, 2005 (ed. or. 1999) e Lev Manovich, *Il linguaggio dei nuovi media*, Milano, Olivares, 2002 (ed. or. 2001). In generale sulla ricerca sociale che utilizza le fonti Internet: Piero

Rondanini e Norma Rech, *Internet per la ricerca sociale: dalle indagini via email all'analisi dei newsgroup*, Milano, CUSL, 2006.

Sulla comunicazione in chat è da segnalare uno studio pionieristico di Christopher C, Werry, "Linguistic and Interactional Features of Internet Relay Chat", in Susan C. Herring (ed.) *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspective*, Amsterdam, John Benjamins, 1996, pp 47-63, sebbene limitato a due sessioni di 10 minuti, in cui sono evidenziati gli sforzi degli interlocutori per simulare il linguaggio parlato con codici, abbreviazioni e l'uso creativo delle maiuscole. Il mondo della chat italiane, in una prospettiva di analisi più sociologica e con osservazione etnografica, è esaminato da Antonio Roversi, *Chat line*, Bologna, il Mulino, 2001. Con lo stesso metodo, Guy Merchant, "Teenagers in Cyberspace: An Investigation of Language Use and Language Change in Internet Chatrooms", *Journal of Research in Reading*, 2001, 24 (3), pp. 293-306, analizza l'interazione in chat delle ragazze utilizzando il concetto di "capitale linguistico" elaborato da Pierre Burdieu. Con un approccio di analisi della conversazione, è da segnalare anche lo studio di Marino Bonaiuto, Cristina Buffone, Elio Castellana, "La struttura conversazionale della comunicazione scritta via chat-line", in Marino Bonaiuto (a cura di) *Conversazioni virtuali*, Milano, Guerini e Associati, 2002, pp. 89-125. Più recente e con maggiore attenzione al cambiamento linguistico, è il lavoro di Elena Pistolesi, *Il parlar spedito. L'italiano di chat, e-mail e SMS*, Padova, Esedra, 2004, svolto sulla base di corpora piccoli ma significativi.

Il linguaggio ibrido della posta elettronica, a metà strada tra il parlato (confidenziale) e lo scritto (formale), è al centro dell'analisi dell2 e-mail di Naomi S. Baron, "Who Sets E-Mail Style? Prescriptivism, Coping Strategies, and Democratizing Communication Access", *The Information Society*, 2002, 18 (5), pp. 403-413 e di Sara Peticca (2002) *Il linguaggio dell'e-mail*, Soveria Mannelli, Rubbettino.

I newsgroup, nonostante il loro rilievo per l'evoluzione del linguaggio e la disponibilità di un immenso database in tutti i gruppi linguistici, hanno ricevuto meno attenzione, forse perché Usenet è stato sempre percepito come una sede di discussione per gli iniziati e i fanatici di Internet. Da segnalare, con approccio conversazionale, lo studio di Michel Marcoccia, "Parler politique dans un forum de discussion", *Langage et Société*, 2003, 104, June, pp. 9-55, e con approccio lessicometrico: Luca Giuliano (2006) "Analysis of the Content of Newsgroup Messages: Methodological and Technical Issues", in P.-L. Law, L. Fortunati, and S. Yang (eds), *New Technologies in Global Societies*, World Scientific, Singapore, pp. 107-124. Un interessante tentativo di porre metodologicamente il problema della costituzione di un corpus dei messaggi da utilizzare come base empirica di ricerca è il lavoro di Sebastian Hoffmann, "Processing Internet-derived Text-Creating a Corpus of Usenet Messages", *Literary and Linguistic Computing*, 2007, 22 (2), 151-165. Sui newsgroup italiani c'è anche Vera Gheno, "Prime osservazioni sulla grammatica dei gruppi di discussione telematici di lingua italiana", *Studi di Grammatica Italiana*, 2004, 22, pp. 267-308.

I blog dal punto di vista linguistico non sono ancora stati esaminati ma vi sono alcuni interessanti interventi sulle loro specifiche modalità di comunicazione: Carlo Baldi e Roberto Zarriello, *Penne digitali. Dalle agenzie ai blog: fare informazione nell'era di in-*

ternet, Roma, Centro di documentazione giornalistica, 2005; Guido Di Fraia (a cura di) *Blog-grafie: identità narrative in rete*, Milano, Guerini Studio, 2007; Robert Scoble e Shel Israel, *Business blog: come i blog stanno cambiando il modo di comunicare dell'azienda con il cliente*, Milano, Il sole-24 ore, 2007.

Un gruppo di ricerca dell'università Paris V - René Descastes ha realizzato un lessico di frequenza del francese contemporaneo su fonti Internet basato su un corpus di 31 milioni di parole: Boris New, Christophe Pallier, Ludovic Ferrand, Rafael Motos, "Une Base données lexicales du francais contemporain sur Internet: LEXIQUETM", *L'Annee psychologique*, 2001, 101 (3), pp. 447-462, disponibile gratuitamente a questo indirizzo: <http://www.lexique.org> (10/3/2008).

RIFERIMENTI BIBLIOGRAFICI

BERNERS-LEE T. (2001) *L'architettura del nuovo Web. Dall'inventore della rete il progetto di una comunicazione democratica*, *interattiva e intercreativa*, Milano, Feltrinelli (ed. or. 1999).

BOLTER J.D. (2002) *Lo spazio dello scrivere. Computer*, *ipertesto e la ri-mediazione della stampa*, Milano, Vita e Pensiero (ed. or. 2001).

BRUSCHI A. (1999) *Metodologia delle scienze sociali*, Milano, Bruno Mondadori.

BUSH, V. (1945) "As We May Think", *Atlantic Montly*, 176 (July), pp. 101-108 (tr. it. "Come possiamo pensare", in T.H. Nelson, *op. cit.*, pp. 1/49).

DAWKINS R. (1995) *Il gene egoista*, Milano, Mondadori-De Agostini (ed. or. 1976)

DELLAROCAS C. (2003) "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms", *Management Science*, 49 (10), pp. 1407-1424; <http://ebusiness.mit.edu/research/papers/173Dellarocas_Word_of_Mouth.pdf> (7/3/2008).

FOGLIO A. (2002) *E-commerce e web marketing : strategie di web marketing e tecniche di vendita in internet* , Milano, FrancoAngeli, 2002

GIULIANO L. (2004) "L'analisi automatica dei testi ad alta componente di rumore", in in E. Aureli Cutillo e S. Bolasco (a cura di), *Applicazioni di analisi statistica dei dati testuali*, Roma, Casa Editrice La Sapienza, pp. 41-54.

GRANIERI G. (2005) *Blog Generation*, Bari, Laterza.

JERZ, D.G (2003) "On the trail of the Memex. Vannevar Bush, Weblogs and the Google Galaxy" <http://www.dichtung-digital.org/2003/1-jerz.htm> (14/3/2008).

LINDLOF T.R., TAYLOR B.C. (2002) *Qualitative Communication Research Methods*, London, Sage Pub.

MARVIN C. (1994) *Quando le vecchie tecnologie erano nuove*, Torino, UTET (ed. or. 1988).

Metitieri F. (2003) *Comunicazione personale e collaborazione in rete. Vivere e lavorare tra email*, *chat*, *comunità e groupware*, Milano, FrancoAngeli.

NELSON T.H. (1981) *Literary Machine*, Swarthmore, pubb. in proprio, 1981 (tr. it. dell'ed. del 1990: *Literary Machine 90.1. Il progetto Xanadu*, Padova, Muzzio, 1992).

ZOPPETTI A. (2003) *PerQuenau?: la scrittura cambia con Internet*, Roma, L. Sossella Ed.

3.
LA *GROUNDED THEORY*

La *Grounded Theory* è una teoria sociologica che nasce dai dati sistematicamente ottenuti da una ricerca (Glaser e Strauss, 1967, p. 21). Questa iniziale definizione apre il testo *The Discovery of Grounded Theory: Strategies for Qualitative Research*, nel quale si legge che per produrre questo "tipo di teoria" non è necessario ricorrere né all'elaborazione statistica di dati o di informazioni raccolte nel corso dell'indagine, né a un'analisi delle interviste o delle osservazioni usufruendo di un qualsiasi supporto che sia di tipo statistico-matematico (Strati, 1997).

La *Grounded Theory* è infatti tale perché è una teoria che emerge dal basso, dal "suolo" ed è intenzione dichiarata di Glaser e Strauss sottolineare in questo modo – ovvero con la scelta del participio passato del verbo *to ground* – la sostanziale differenza e lontananza della loro teoria dalla *grand theory*, con la quale i due studiosi intendono il "grandioso" approccio sviluppato in seno al metodo ipotetico-deduttivo.

3. 1. LE ORIGINI DELLA *GROUNDED THEORY*

La costruzione di un paradigma all'interno del quale possano riconoscersi gli studiosi che affrontano ricerche di matrice qualitativa è uno degli obiettivi dei due autori che negli anni Sessanta danno vita alla *Grounded Theory*.

Barney Glaser e Anselm Strauss hanno alle spalle dei dipartimenti con solide tradizioni sociologiche: il primo proviene dal Dipartimento di Sociologia della Columbia University dove si era sviluppata la metodologia sociologica quantitativa di Paul Lazarsfeld e le teorie di medio raggio di Robert Merton; Strauss proviene, invece, dall'Università di Chicago, ovvero dalla scuola del-

l'interazionismo simbolico e delle interviste qualitative in profondità.

L'eco di queste tradizioni traspare dagli scopi del loro primo testo. Le pagine iniziali di tale volume sono, infatti, dedicate a segnalare i punti di vicinanza:

- dalle teorie di medio raggio e
- dal concetto di *serendipity* di Merton;
- e a sviluppare la necessaria distanza:
- dal modello ipotetico-deduttivo e
- dal linguaggio delle variabili di Lazarsfeld.

Comprendere quali siano gli elementi costituenti delle teorie prodotte in seno alla *Grounded Theory* e i tipi a essa riconducibili significa valutare, innanzitutto, i punti di vicinanza e di distacco dai paradigmi allora esistenti.

Il punto di vista adottato da Glaser e Strauss in merito al rapporto fra teoria e mondo empirico appare chiaramente dalla difesa da loro praticata dell'opera di Thomas e Znaniecki *The Polish Peasant in Poland and America* (1920) dagli attacchi di Blumer. Sostanzialmente la critica mossa da Blumer agli autori de *Il contadino polacco* è sulla validità che storie di vita, quindi prove e dati raccolti sulla base di "documenti umani" (Madge, 1966, p. 125 e seg.), siano in grado di supportare le generalizzazioni da loro formalizzate.

La questione viene affrontata all'interno della *Grounded Theory* come la ricerca di una risposta a una sorta di dubbio amletico "*verification or generation?*". Glaser e Strauss sostengono che Blumer ha prodotto una critica dell'opera in base al criterio di verificazione di una teoria non in base al criterio di generazione di una teoria che invece è l'elemento su cui occorre – per loro – puntare, poiché altrimenti si corre il rischio di vedere un proliferare di teorie a sfondo imitativo; conseguenza del lavoro di alcuni ricercatori che si limitano a testare e applicare in piccole aree della sociologia quanto è stato scoperto da sociologi quali Weber, Durkheim e altri grandi (Glaser, Strauss, 1967, p. 13 e seg.).

In questo periodo il dibattito fra ricerca qualitativa *versus* ricerca quantitativa vede quest'ultima avere la meglio sulla prima: è il tempo di Lazarsfeld, Guttman e altri. Costoro discutono della conferma di una teoria attraverso i risultati empirici favorevoli o della falsificazione di una teoria come conseguenza del riscontro empirico di risultati contrari a essa.

A questi modi di affrontare la questione corrispondono due diverse tradizioni epistemologiche: quella americana che rimanda ai lavori di Hempel e Carnap e quella anglosassone che rimanda ai lavori di Popper e dei suoi allievi (Boniolo, Vidali, 1999). Sia Hempel che Popper si muovono all'interno della *Subsumption Theory of Explanation*. La vicinanza è tale che fra i due sussiste una controversia sull'attribuzione di paternità del modello di spiegazione scientifica. Secondo Popper "dare una spiegazione causale" significa dedurre un'asser-

zione che lo descrive, usando come premesse della deduzione una o più leggi universali, insieme con alcune asserzioni singolari dette condizioni iniziali (Popper, 1934). Stando a quanto sostenuto da Hempel si può considerare la spiegazione come un'argomentazione secondo cui ci si doveva aspettare il fenomeno da spiegare (definito come *explanandum*), in virtù di certi fatti esplicativi (*explanans*). Questi ultimi si distinguono in due gruppi: i fatti particolari e le uniformità esprimibili mediante leggi generali (Hempel, 1942). Si identifica la teoria della spiegazione scientifica con la teoria hempeliana della spiegazione per via della sua sistematicità e compiutezza.

Appare già chiara la distanza fra l'approccio nomotetico-deduttivo e la "teoria fondata" proposta dalla *Grounded Theory*; questo distacco diventa ancora più tangibile se a questa breve disamina si affiancano dei cenni al modello statistico-induttivo di Lazarsfeld.

Il paradigma di Lazarsfeld chiarisce in quattro brevi passaggi come si arrivi dai concetti alle variabili e quindi alla misurazione dei fenomeni. Lazarsfeld individua quattro fasi distinte nel passaggio dai concetti alle variabili: 1) rappresentazione figurata del concetto, 2) individuazione delle dimensioni; 3) scelta degli indicatori, 4) costruzione dell'indice.

Siamo quindi in presenza di una struttura ipotetico-deduttiva che sostanzia la spiegazione scientifica di un fenomeno secondo procedure codificate attraverso le quali è possibile operazionalizzare i concetti e poi verificare o falsificare quanto ipotizzato. In questo sistema le teorie scientifiche sono secondo Hempel (1952) un insieme di un insieme non interpretato, sviluppato deduttivamente, e di un'interpretazione conferente significato empirico ai termini e alle proposizioni di tale sistema. E ancora:

> Una teoria scientifica è pertanto paragonabile a una complessa rete sospesa nello spazio. I suoi termini sono rappresentati dai nodi, mentre i fili colleganti questi corrispondono, in parte, alle definizioni e, in parte, alle ipotesi fondamentali e derivate dalla teoria. L'intero sistema fluttua, per così dire, sul piano dell'osservazione, cui è ancorato mediante le regole interpretative. Queste possono venire concepite come fili non appartenenti alla rete, ma tali che ne connettono alcuni punti con determinate zone del piano d'osservazione. Grazie a siffatte connessioni interpretative, la rete è utilizzabile come teoria scientifica: da certi dati è possibile risalire, mediante un filo interpretativo, a qualche punto della rete teorica, e di qui procedere attraverso definizioni e ipotesi, ad altri punti, dai quali, per mezzo di un altro filo interpretativo, si può ridiscendere al piano dell'osservazione (Hempel, 1952, pp. 46-47).

Il processo secondo il quale si arriva alla formulazione di teorie dal basso è per Glaser e Strauss meno strutturato e – come loro stessi sostengono in una nota

contenuta nelle prime pagine del loro testo – quasi per *serendipity*. La parola sta a indicare l'atto di trovare qualcosa di prezioso mentre si cerca qualcos'altro oppure trovare qualcosa che si andava cercando ma in un luogo o in un posto del tutto inaspettato, quindi fortunosamente.

La storia di questa parola comincia con il carteggio intrattenuto dal 1740 al 1786 fra i due cugini Horace Walpole e Horace Man; nella realtà dei fatti la parola *serendipity* è apparsa soltanto una volta e in una sola delle lettere che questi si scambiarono ma l'importanza travalicò lo scritto (Merton e Barber, 2002). Nella lettera Walpole per spiegare una scoperta fatta in maniera fortunata, scrisse al cugino che la spiegazione sarebbe stata più esaustiva se gli avesse raccontato la favola dei tre principi di Serendip (*The Three Princes of Serendip*).

> (…) nel corso dei loro viaggi, le loro Altezze scoprivano continuamente, per caso e per sagacia, cose che non andavano cercando: ad esempio uno di loro scoprì che un mulo cieco dall'occhio destro era passato dalla loro strada di recente, perché l'erba era mangiata sul lato sinistro, dov'era più brutta che sul destro – ora capisci cos'è la *Serendipity?* (*ivi*, p. 30).

L'elemento di questa "sagacia accidentale" riconducibile all'intento di generare una teoria fondata opposta alla *grand theory* è quello di considerare come importante quanto emerge per caso, dove "per caso" indica un processo di emersione della teoria dal dato riscontrato e non una sua formulazione aprioristica. Il concetto di *serendipity* nella *Grounded Theory* può essere visto come confinante con l'atto dell'induzione proprio della teoria fondata dal basso. L'induzione è infatti la via per conoscere a partire dalla realtà studiata: la teoria deriva dai dati.

L'induzione di cui si sostanzia la *Grounded Theory* è diversa da quanto per induzione si intende sia nel neo-positivismo che nel post-positivismo (Popper, 1934; Reichenbach, 1951). Non si tratta del passaggio dal particolare all'universale attraverso leggi generalizzate; qui per induzione si intende la comprensione degli atti interattivi effettuata sia dai soggetti che operano nel contesto in esame che dai ricercatori che pongono in essere quella situazione (Strati 1997, p. 129).

Dalla distanza tra la definizione fornita da Hempel per le teorie scientifiche e quella data da Merton per la scoperta effettuata in maniera serendipitosa e dalla sua articolazione delle teorie di medio raggio, possiamo pian piano arrivare a sostanziare cos'è una teoria empiricamente fondata. Merton dichiara di occuparsi

> (…) di quelle teorie che ho chiamato di *medio raggio*: teorie intermedie fra le ipotesi di lavoro che si formulano abbondantemente durante la routine quotidiana della ricerca

> e le speculazioni onnicomprensive basate su uno schema concettuale unificato che mirano a spiegare tutte le uniformità, empiricamente osservabili, del comportamento sociale, dell'organizzazione sociale e del mutamento sociale (Merton, 1949, p. 76).

Le teorie di medio raggio si configurano con una serie limitata di presupposti da cui, però, è possibile derivare e verificare delle ipotesi specifiche. L'intento di Merton è quello di riempire una sorta di spazio vuoto che sussiste fra l'empirismo bruto e le teorie generali e onnicomprensive. È compito quindi del ricercatore sviluppare teorie speciali che possano essere verificate empiricamente e che progressivamente diano forma a uno schema concettuale più generale capace di articolare gruppi di teorie speciali (Merton, 1949, p. 86).

Quella che Merton propone è una costruzione paradigmatica che proceda per sovrapposizioni di teorie di medio raggio verificate; secondo Merton questo è anche il procedere delle teorie nelle scienze fisiche e naturali. In fisica esistono – egli sostiene – un insieme di teorie specifiche di ampiezza variabile e i fisici sono animati dalla speranza – dimostratasi valida nel corso del tempo – che queste teorie si uniscano in famiglie teoriche.

La creazione di una teoria generale fondata sulla stratificazione di teorie speciali costituisce un elemento di vicinanza delle teorie di medio raggio con la *Grounded Theory*. Nella *Grounded Theory* si trovano due tipi di teorie: "teorie evidenti o reali" (*substantive*) e "teorie ufficiali o formali" (*formal*) ed entrambe possono essere definite come teorie di medio raggio.

Alcune caratteristiche di queste teorie vogliono che:

- emergano entrambe dai dati;
- si trovino a un livello distinguibile di generalizzazione;
- differiscano fra di loro in termini di gradi di generalizzazione;
- le teorie evidenti o di primo livello costituiscono il link che permette di generare dai dati le teorie ufficiali o di secondo livello (Glaser, Strauss, 1967).

Gli elementi di cui si costituiscono le teorie sono le categorie concettuali e le proprietà concettuali delle categorie stesse. Come la categoria è un elemento concettuale proprio di una teoria così le proprietà sono, a loro volta, aspetti concettuali delle categorie.

3. 2. LA COSTRUZIONE DELLE TEORIE

La generazione di teorie avviene, soprattutto, avvalendosi del metodo comparativo, il quale può essere applicato su unità di analisi – fenomeni sociali – di diverse dimensioni.

Gli elementi ritenuti fondamentali nella *Grounded Theory* per la generazio-

ne della teoria (*generating theory*) sono stati individuati in:

- *comparative analysis:* il confronto degli avvenimenti e degli elementi applicabili a ciascuna categoria che avviene a più stadi dello studio;
- *categories and properties:* l'individuazione delle categorie e delle loro proprietà;
- *hypothesis:* l'emersione della relazione fra categorie e loro proprietà dal fenomeno empirico;
- *integration:* l'integrazione delle categorie e delle loro proprietà e la conseguente delimitazione della teoria;

scrittura della teoria.

I due autori – Glaser e Strauss – hanno condotto principalmente studi su pazienti ospedalieri, come riportato anche in altri testi (Strati, 1997, p. 140). Seguire la generazione della teoria fondata attraverso gli esempi da loro citati è molto più semplice.

Uno di questi studi esamina la cura di pazienti terminali; l'obiettivo è pervenire attraverso i dati empirici alla costruzione della categoria "perdita sociale" (*social loss*). Questa categoria si riferisce alla perdita della famiglia o dell'occupazione; elementi, questi ultimi, che a loro volta influiscono su come il personale infermieristico percepisce i malati. Le infermiere si "formano un'idea" del paziente morente in base a delle caratteristiche evidenti di questo, quali il colore della pelle, e a delle caratteristiche personali del paziente che emergono mano a mano che le infermiere ne approfondiscono la conoscenza.

Tuttavia, attraverso le indagini i ricercatori potrebbero accorgersi che quanto indicato non accada, ovvero che ci siano degli ospedali in cui i pazienti ricevano cure simili a prescindere dalla posizione sociale o dalla classe sociale, ma anche dal colore della pelle. In questo caso le conseguenze per la teoria non sarebbero negative, poiché è dal contesto empirico che emergono concetti e categorie e queste possono essere modificate. Il centro dell'analisi è rappresentato dal concetto di "perdita sociale" e non dal modo in cui questa si manifesta.

Le categorie possono essere modificate a seconda di quanto emerge dalla comparazione con diversi gruppi di analisi. Le categorie emergono a un primo basso livello di raccolta dei dati e a un livello di specificità maggiore durante la fase di codifica e analisi dei dati. La categoria "perdita sociale" appare immediatamente nella fase di ricerca ed emerge dal confronto delle risposte date dalle infermiere alle domande relative all'eventuale scomparsa dei loro pazienti. Le risposte date a queste domande sottolineano una correlazione tra il concetto di perdita percepito dalle infermiere con la perdita accusata dalla famiglia per la morte del proprio caro. Attraverso i dati si stabiliscono, quindi, delle relazioni tra i concetti e le categorie.

In questo momento si è nella fase definita delle "ipotesi", si utilizza que-

sto termine con un certo distacco dal metodo ipotetico-deduttivo. Le ipotesi, infatti, emergono come una sorta di "relazione suggestiva" fra le categorie e le loro proprietà: non sono testate empiricamente bensì è il dato empirico che le fa emergere.

Dal confronto delle informazioni raccolte durante il loro studio, i due autori si accorsero che la perdita di alcuni malati era percepita – dalle infermiere – come una grande privazione, mentre per altri malati la sofferenza dimostrata era minore. Questo permise di mettere in luce la relazione positiva fra cura e perdita sociale; quindi fra un tipo di cure prestate a un malato piuttosto che a un altro. Emerse anche una correlazione positiva tra la conoscenza che le infermiere sviluppano del paziente con il passare del tempo e la percezione della perdita. Lo studio del comportamento delle infermiere portò alla luce un altro elemento: l'importanza per queste ultime di perdere la "compostezza" innanzi al pianto della famiglia per la morte del loro caro.

Riassumendo, le categorie emerse nello studio possono essere raggruppate in due gruppi principali:

- quelle di primo livello che emergono dai dati: in questo caso sono da attribuire al linguaggio specifico del contesto in cui si svolge la ricerca, per esempio la *compostezza* delle infermiere;
- e quelle poste a un livello più alto: costruite dal ricercatore, quali la *perdita sociale* o il *calcolo della perdita sociale*.

Le categorie costruite dal ricercatore forniscono delle spiegazioni ai fenomeni rilevati ma anche alle categorie costruite dalla raccolta dei dati, come la compostezza. Fra questi due tipi di categorie sembra sussistere la medesima relazione che intercorre fra i due tipi di teorie (*substantive* e *formal*). Il metodo della comparazione costante viene applicato durante tutto il percorso di ricerca e nella fase della "integrazione" esso viene utilizzato per comparare le categorie e le loro proprietà.

Nel caso delle infermiere e dei loro malati Glaser e Strauss notarono che mano a mano che le infermiere approfondivano la conoscenza del malato ricalcolavano la perdita sociale. Integrando le categorie e le loro proprietà si scoprì che il calcolare e il ricalcolare l'importanza della perdita sociale da parte delle infermiere avveniva mano a mano che queste ultime andavano ricostruendo una sorta di storia della perdita sociale del paziente. Questa ricostruzione, così come l'iniziale calcolo della perdita sociale era una strategia adottata dalle infermiere per mantenere la loro compostezza sul lavoro. Il calcolare e il ricalcolare può essere inteso come un modo per attribuire il giusto valore alla perdita sociale, per giustificare o determinare la propria compostezza da parte delle infermiere.

Prima di arrivare alla "scrittura della teoria" occorre "delimitare la teoria",

ciò vuol dire che vi è un momento in cui si smette di creare categorie e qualora ne emergano di nuove, queste non mettono in discussione la teoria fino a quel momento costruita, ma se ne segue l'evoluzione da quel punto in poi della ricerca. Costruire la teoria fondata è l'ultimo stadio di questo iter. Produrre una teoria fondata non vuol dire riportare come in una sorta di diario tutto ciò che accade, bensì partire dalle categorie evidenziate per poter arrivare a generare una teoria, tornando – qualche volta – ai dati per validare quanto si scrive (Strati 1997, p. 143). Nel caso esaminato, Glaser e Strauss ritornano agli elementi salienti dello studio: "il calcolare la perdita sociale"; "la storia della perdita sociale"; "l'impatto della perdita sociale sulla compostezza delle infermiere". Partendo da questi concetti si cercò di illustrane i contenuti mettendo in luce sia i punti deboli che quelli di forza, secondo il principio del metodo della comparazione che prende in considerazione tanto le similarità che le differenze. Alla fine dell'esposizione del loro caso gli autori forniscono un modello di interpretazione in grado di guidarci nella comprensione degli elementi di generazione della teoria e di esplicazione degli elementi che permettono di passare da una teoria di primo livello alla teoria formale (tab. 3.1).

Tab. 3.1 – Elementi costituenti i due differenti tipi di teorie

	Tipi di teoria	
	Sostantiva	*Formale*
Categorie	Perdita sociale del paziente morente.	Valore sociale delle persone.
Proprietà delle categorie	Calcolo della perdita sociale in base alle caratteristiche apprese e a quelle apparenti del paziente.	Calcolo del valore sociale delle persone in base alle caratteristiche apprese e a quelle apparenti.
Ipotesi	Maggiore è la perdita sociale per il paziente morente, 1) migliori sono le cure, 2) e le infermiere sviluppano una maggiore razionalizzazione della spiegazione della morte.	Maggiore è il valore sociale di una persona minore è l'attesa per ricevere cure dagli esperti.

Fonte: Glaser e Strauss, 1967, p. 42 (nostra trad.).

Gli elementi che nella cella delle *substantive theory* sono attribuite ai pazienti morenti passando a un livello di generalizzazione maggiore (*formal theory*) sono attribuite alle persone in generale. La categoria "perdita sociale del paziente" si eleva a un grado maggiore di generalizzazione concettualizzando il "valore so-

ciale delle persone". Allo stesso modo se il calcolo e il conseguente ricalcolo della perdita sociale da parte delle infermiere avviene in base alle caratteristiche apprese e a quelle apparenti del paziente le stesse proprietà vengono applicate nel determinare il valore sociale delle persone. Dallo studio empirico della relazione fra le categorie e le loro proprietà emergono, come "relazioni suggestive", le ipotesi che, apprese nello specifico campo di studio, sono poi estendibili alle persone nella teoria formale.

La tabella 3.1 chiarisce in modo significativo i punti di contatto con le teorie di medio raggio ma anche con quell'intuizione-suggestione – vicina alla *serendipity* – che emerge dal basso e che viene qui definita come ipotesi; allo stesso tempo siamo lontani dalla spiegazione scientifica del modello nomologico-deduttivo e statistico-inferenziale.

3. 3. IL PROCESSO DI CODIFICA E DI CONCETTUALIZZAZIONE

La codifica dei dati consiste nella sua prima fase nell'analisi *line-by-line* di segmenti, parole, paragrafi, porzioni di testo. Questo tipo di micro analisi è necessaria all'inizio dello studio per poter attivare il processo di concettualizzazione e generazione delle categorie e delle loro proprietà.

L'analisi "riga per riga" dei dati richiede un dispendio di energie non indifferente ma produce un dettaglio di studio maggiore rispetto a qualsiasi altro tipo di indagine condotta sui dati qualitativi. Ovviamente anche per questa modalità di ricerca è necessario avere un qualche interrogativo da cui muovere, un quesito cui voler dare risposte e poi procedere con il metodo comparativo (Strauss e Corbin, 1996, p. 73).

Secondariamente i dati qualitativi sono codificati secondo tre modalità distinte:

- la codifica aperta;
- la codifica assiale;
- la codifica selettiva.

La codifica aperta è il processo analitico attraverso il quale i concetti vengono identificati e le loro dimensioni emergono dai dati (Strauss, Corbin, 1996, p. 101). Il cuore della codifica aperta è rappresentato dai concetti; del resto – come sostengono Anselm Strauss e Juliet Corbin – non esiste scienza senza concetti.

Open Coding vuol dire quindi "aprire" un testo e far emergere da esso le idee, le forme comunicative che contiene. In questo senso il primo passo di questo approccio è la "concettualizzazione": un concetto è un fenomeno etichettato (*labeled phenomenon*) (Strauss e Corbin, 1996, p. 103).

Nel processo di concettualizzazione c'è molto dell'astrazione: i dati vengono spezzati in frazioni di avvenimenti, separati gli uni dagli altri e analizzati nella loro unicità. Nell'etichettare il fenomeno il ricercatore può attribuire un proprio nome, una propria etichetta a quanto l'intervistato dice o a quanto emerge da un testo oppure può utilizzare le parole stesse del soggetto, quest'ultimo processo di codifica è spesso definito come "*in vivo codes*".

Per illustrare il processo di concettualizzazione e la modalità di codifica aperta i due autori propongono l'esempio – nella seconda edizione del volume *Basic of Qualitative Research* (1996) – delle interviste condotte su giovani donne con meno di vent'anni in merito all'uso delle droghe da parte dei *teenagers*. A partire dalle risposte date dalle ragazze alla domanda posta dall'intervistatore "Mi parli degli adolescenti e dell'uso delle droghe" si dà il via al processo di concettualizzazione e di codifica dell'intervista, che richiede molta creatività.

La prima risposta dell'intervistata a questa domanda è che l'uso di droghe da parte degli adolescenti è un atto di liberazione dai genitori. Il ricercatore pone accanto a questa sentenza/frase l'etichetta "atto di ribellione". L'intervistata aggiunge che, ovviamente, dice ciò guardando al fenomeno in generale; se dovesse parlare per sé lo definirebbe come un'esperienza. Il ricercatore codifica quindi il fare uso di droghe nell'adolescenza come "esperienza". In questo caso utilizza un termine che viene dal testo dell'intervista, quindi un "*in vivo codes*".

L'intervistata continua dicendo che si sente molto parlare di droga (discorsi sulla droga[8]), e che spesso in questi discorsi è definita come qualcosa di cattivo (connotazioni negative). Si comincia a farne uso perché è facilmente accessibile (acquisto facile), e perché tutti sono attratti dalle cose nuove (nuova esperienza). È un fenomeno di moda e anche se si è a conoscenza di tutto ciò che su di essa si dice di negativo (connotazioni negative), si considerano questi "no" che la circondano come dei tabù. Tutti gli adulti sono contro le droghe (posizione negativa degli adulti), ma quando si è adolescenti la prima cosa che si vuol fare è provarla (cambiamento di atteggiamento nei confronti delle droghe nell'età adulta).

L'intervistatore continua approfondendo la tematica e chiedendo alla ragazza se lei ha avuto molte esperienze di droga. In realtà lei ne ha fatto poco uso (esperienze limitate), ovviamente questo dipende anche da quanto è accessibile la droga (grado di accessibilità). La maggior parte degli adolescenti non sperimenta l'*hard-core*, l'uso pesante delle droghe.

Per l'analista questa ultima affermazione è molto importante, poiché gli

[8] Si inseriscono qui fra parentesi tonde le etichette date.

permette di utilizzare un concetto *in vivo codes* in opposizione a un'etichetta da lui attribuita: l'esperienza limitata (*hard-core use vs. limited experimentig*).

Ovviamente, l'uso di droghe leggere come l'*hashish* o la *marijuana* dipende anche dalla fase della vita in cui ci si trova (sviluppo delle tappe della vita). Si comincia con le droghe leggere, per poi proseguire con le droghe più pesanti, fra queste gli allucinogeni. L'intervistatore si informa anche sui luoghi di consumo e sull'opinione che la gente ha nei riguardi di chi fa uso di droghe.

Un passaggio interessante per analizzare il processo di codifica è rappresentato dalla risposta alla domanda sulle esperienze personali della ragazza in merito alle droghe. Fare uso di droghe è secondo lei più un'esperienza di condivisione nel momento in cui viene praticata piuttosto che un argomento di cui si può parlare: rappresenta quindi un tema di comunione fra gli adolescenti che la consumano. Il nostro analista pone qui in contrapposizione due concetti, due etichette date: "prendervi parte vs. parlarne".

La differenza tra parlarne e prendervi parte è data anche dal livello di uso che ne si fa (uso-pesante vs. esperienza limitata). Infatti, aggiunge la ragazza, il primo approccio con le droghe avviene durante le scuole secondarie superiori, lì provare la droga non è un modo per scoprire se stessi (no scoperta del sé) ma un modo per seguire la folla (imitazione dei propri coetanei). Anche questi due ultimi concetti scoperta del proprio io e imitazione dei propri coetanei sono in opposizione (Strauss e Corbin, 1996, pp. 106-109).

Ultimata l'analisi dell'intervista il ricercatore ha quindi effettuato una micro analisi del testo estrapolando concetti su una tematica e creando etichette. Si pone quindi il problema successivo di riflettere e sintetizzare quanto trovato attraverso la codifica aperta.

Il ricercatore può utilizzare per la codifica dei *tools* che lo aiutino e appuntarsi le informazioni che trova interessanti su dei *memos*. Il *memo* è definibile come una sorta di post-it dove il ricercatore annota i suoi pensieri, le interpretazioni date e utili spunti per le successive analisi.

Nell'ottica di costruire un set maneggevole di dati è necessario raggruppare i concetti e le categorie a essi attribuite. Si passa quindi alla codifica assiale.

L'*Axial Coding* è il processo che collega le categorie alle sub-categorie, collegando le categorie alle proprie proprietà e dimensioni (Strauss e Corbin, 1996, p. 123).Nella codifica aperta si lavora sui concetti che emergono dal testo, nella codifica assiale si lavora sulle relazioni fra categorie e loro dimensioni. L'obiettivo della codifica assiale è ricostruire i dati frammentati durante la prima operazione di codifica.

Nel suo svolgersi la codifica assiale segue delle specifiche procedure che consistono:

- nell'ipotizzare determinate relazioni tra le proprietà e le dimensioni afferenti a una specifica categoria, che erano apparse durante la codifica aperta;
- nell'identificare e verificare le relazioni, le condizioni, le interazioni/azioni che in forma di ipotesi sono state attribuite al fenomeno;
- nello specificare come sono state create le relazioni fra categorie, dimensioni e proprietà, indicandone frequenza, intensità, durata, direzione, potenzialità;
- nel cercare una categoria maggiore, riassuntiva, alla quale le altre possano essere collegate e specificare il legame fra queste.

Collegare le categorie alle proprie dimensioni è nella pratica molto più semplice di quanto possa sembrare. Strauss e Corbin sottolineano come questa attività sia già in *nuce* nella codifica aperta.

Per esempio, quando si dice che il sé è "liberato" attraverso l'uso delle droghe, che l'accesso è "facile", che si tratta di una esperienza "nuova", si sta dando una qualificazione a un concetto. Ora – nella codifica assiale – si tratta di sussumere l'insieme di dimensioni individuate in categorie multiple onnicomprensive: più grandi.

Continuando con l'analisi del testo sulle droghe emergono due categorie, la prima è «sperimentare attraverso le droghe» (*experimenting with the drugs*) cui si possono collegare concetti quali la liberazione dai genitori, le nuove esperienze, le nuove amicizie, l'accesso, la connotazione negativa di alcune chiacchiere sulle droghe, il cambiamento d'opinione quando si passa all'età adulta; la seconda può essere rappresentata dalla «esperienza limitata» che di essa si ha (*limited experimenting*), quando per esempio la ragazza racconta che c'è un uso limitato delle droghe pesanti.

Le due categorie rappresentate possono – in un qualche modo – riassumere una parte del testo dell'intervista. In forma grafica si hanno due opzioni di visualizzazione: o un *Mini-Framework* oppure un diagramma.

I *Mini Frameworks* sono dei diagrammi più piccoli che emergono dai dati come risultato della nostra opera di codifica praticata attorno a un concetto (Strauss e Corbin, 1996, p. 141). Le rappresentazioni grafiche ottenute mediante diagrammi e *Mini-Frameworks* non sono da considerarsi orpelli da aggiungere alle conclusioni cui si è pervenuti attraverso la codifica, bensì rappresentano un modo per giungere alle conclusioni dell'analisi.

L'ultimo processo di codifica è rappresentato dalla codifica selettiva che è il processo di integrazione e rifinitura della teoria.

La *Selective Coding* è il momento in cui si individua una categoria principale e si decide di far ruotare attorno a essa l'interpretazione che dei dati si vuole fornire. Anche in questo momento della *Grounded Theory* è necessario una volta indi-

viduata la categoria attenersi alla comparazione costante tra questa categoria centrale e le altre o ulteriori elementi che possano emergere dai dati qualitativi.

Centrale in questa fase è l'individuazione della categoria principale, del *focus* attorno al quale far ruotare la narrazione di quanto trovato. La categoria principale è quella che appare più di frequente nei dati; ha più connessioni con le altre categorie e la spiegazione/interpretazione che essa fornisce ai dati appare logicamente dagli stessi, non mediante una forzatura. Inoltre, la frase o le parole utilizzate dal ricercatore per indicare questa categoria, quindi il concetto attraverso il quale la si designa, deve porsi a un livello di astrazione tale da poter essere attribuito, senza subire cambiamenti alcuni, sia alla teoria evidente che alla teoria formale[9]. In questo modo si accresce il potere esplicativo della teoria fondata.

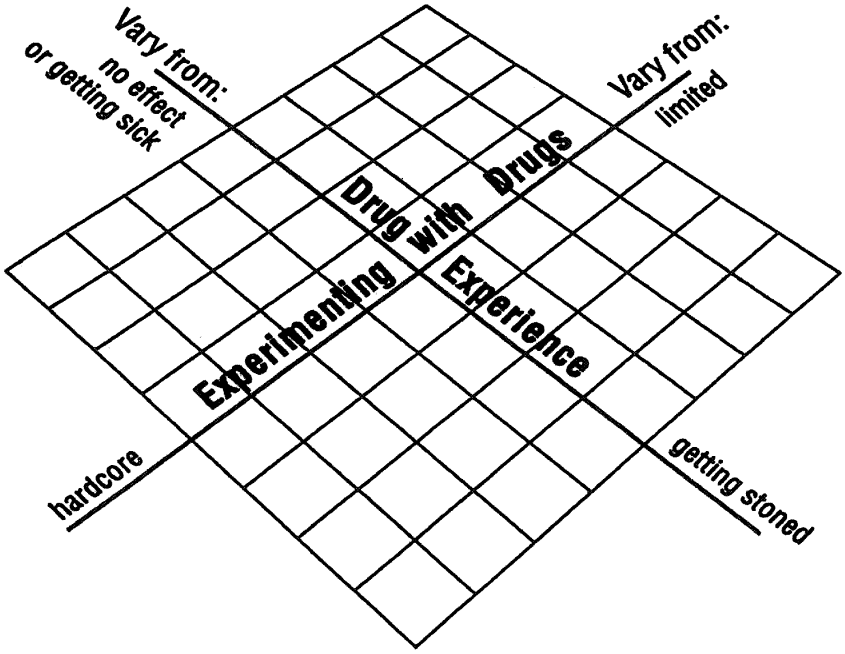


Fig. 3.1 – Esempio di *Mini-Framework* ottenuto incrociando due concetti principali

[9] Si pensi al concetto di "perdita sociale" illustrato nei paragrafi precedenti.

Nell'esempio fin qui seguito sull'uso delle droghe da parte dei *teenagers* emerge dai dati come centrale sia la categoria: sperimentare attraverso la droga; l'idea, quindi, che la droga nell'età adolescenziale rappresenti più che altro un esperimento, uno sperimentarsi (*experimenting with the drugs*).

Nell'adolescenza è più importante condividere il momento di consumo delle droghe leggere piuttosto che parlarne; si ha una concezione positiva del consumo, che poi si modifica con l'età adulta; l'uso è limitato alle droghe leggere, è uno strumento di integrazione in un gruppo di propri pari. Attraverso un processo di astrazione e utilizzando i *memos* che il ricercatore ha man mano prodotto e astraendo l'evento analizzato si individua come categoria principale il "rituale di passaggio".

Identificare il consumo di droghe come un momento di passaggio nell'adolescenza, che radicalmente ne cambia il corso e la forma, rappresenta un modo per narrare la storia ottenuta dai dati qualitativi analizzati. È evidente come l'identificazione in "rituale di passaggio" renda questa categoria valida sia per le teorie di livello *substantive* che *formal*.

APPROFONDIMENTI TEMATICI

Questo capitolo sulla *Grounded Theory* ha l'obiettivo principale di definire le origini dell'approccio allo studio dei testi dal basso, specificando le modalità proprie dell'operare utilizzando un approccio puramente induttivo. Il punto di partenza per lo studio della *Grounded Theory* è rappresentato indubbiamente dall'opera di Barney G. Glaser e Anselm L. Strauss *The Discovery of Grounded Theroy: Strategies for Qualitative Research*, Chicago, Aldine, 1967.

Successivamente le procedure per l'applicazione empirica della *Grounded Theory* sono state delineate da Anselm Strass e Juliet Corbin (*Basic of Qualitative Research. Techniques and Procedures for developing Grounded Theory*, London, Sage Pubblication, 1996) che in *Grounded Theory in Practice*, London, Sage Pubblication, 1997 forniscono anche esempi pratici di studi in cui è stata utilizzata tale metodologia.

In Italia si è dedicato allo studio di questo approccio soprattutto Antonio Strati che nel 1997 pubblica un saggio dal titolo "La Grounded Theory", in Luca Ricolfi (a cura di) *La ricerca qualitativa*, Roma, Carocci, pp. 125-163 e poi Elvira Cicognari con "L'approccio qualitativo della Grounded Theory in psicologia sociale: potenzialità, ambiti di applicazione e limiti", in Bruno Mazzara (a cura di) *Metodi qualitativi in psicologia sociale*, Roma, Carocci.

Più di recente allo studio della *Grounded Theory* si è dedicato Massimiliano Tarozzi con *Che cos'è la Grounded Theory*, Roma, Carocci, 2008. A lui si deve la prossima uscita del celebre testo di Glaser e Strauss (1967) in lingua italiana per l'editore Armando di Roma.

Oltre oceano, nuovo vigore ha dato all'uso della *Grounded Theory* il lavoro di Kathy Charmaz, che intende tale metodo come un insieme flessibile di indicazioni procedurali (*Constructing Grounded Theory*, London, Sage Pubblication, 2006) allontanandosi in questo modo dall'uso classico fattone da Glaser e Strauss. In ultimo, va citato il lavoro sempre della Charmaz con Antony Bryant in *The Sage Handbook of Grounded Theory*, London, Sage Pubblication, 2007.

RIFERIMENTI BIBLIOGRAFICI

BERNERS-LEE T. (2001) *L'architettura del nuovo Web. Dall'inventore della rete il progetto di una comunicazione democratica*, *interattiva e intercreativa*, Milano, Feltrinelli (ed. or. 1999).

BOLTER J. D. (2002) *Lo spazio dello scrivere. Computer*, *ipertesto e la ri-mediazione della stampa*, Milano, Vita e Pensiero (ed. or. 2001).

BONIOLO G., VIDALI P. (1999) *Filosofia della scienza*, Milano, Bruno Mondadori.

BRUSCHI A. (1999) *Metodologia delle scienze sociali*, Milano, Bruno Mondadori.

BUSH, V. (1945) "As We May Think", *Atlantic Montly*, 176 (July), pp. 101-108 (tr. it. "Come possiamo pensare", in T. H. Nelson, *Literary Machine 90.1. Il progetto Xanadu*, Padova, Muzzio, 1992, pp. 1/49).

HEMPEL C.G. (1961) *La formazione dei concetti e delle teorie della scienza empirica*, Milano, Feltrinelli (ed. or. 1952).

MADGE J. (1966) *Lo sviluppo dei metodi di ricerca empirica in sociologia*, Bologna, Il Mulino (ed. or. 1962).

MERTON R. K. (1959) *Teoria e struttura sociale,* Il Mulino, Bologna (ed. or. 1949).

MERTON R. K., BARBER E. G. (2002) *Viaggi e avventure della serendipity*, Il Mulino, Bologna (ed. amer. 2004).

POPPER K. R. (1970) Logica della scoperta scientifica, Torino, Einaudi, (ed. or. 1934).

REICHENBACH H. (1961) *La nascita della filosofia scientifica*, Il Mulino, Bologna (ed. or. 1951).

4.
LAVORARE CON ATLAS.TI5

Atlas.ti è un software per l'analisi dei dati qualitativi assistita dal computer. Si presenta come un potente banco da lavoro, poiché permette di trattare in modalità semi-automatica dati testuali, immagini audio e video.

La possibilità di trattare dati testuali ma anche dati visuali deriva ad Atlas.ti dalle diverse funzioni che il software incorpora in sé che permettono di gestire, estrarre, comparare, esplorare e riassemblare porzioni e/o insiemi di dati in maniera creativa, ma soprattutto sistematica. La versione qui utilizzata è la versione 5 ed è reperibile all'indirizzo <http://www.atlasti.com>.

4. 1. LA BARRA DEGLI STRUMENTI

Le icone presenti sulla barra degli strumenti permettono un accesso facilitato alle funzionalità del software. La prima barra degli strumenti cui si accede è quella dell'unità ermeneutica (HU). Successivamente ogni modulo è dotato di una propria barra che tuttavia presenta alcune icone simili a quelle dell'HU, richiamando le medesime funzionalità presenti in quella principale.

Editare i documenti: si attiva appena si apre un documento e permette di scrivere all'interno del documento aperto e di salvare o eliminare le modifiche effettuate. Attivata tale modalità (*Enter Edit Mode*) è possibile accedere alle altre funzionalità proprie dell'editing dei testi: grassetto (**B**), corsivo (*I*), sottolinea (U), ingrandisci il carattere (V), rimpicciolisci (Λ), evidenzia, allinea a sinistra, a destra, giustifica il testo, inserisci punto elenco, taglia, copia e incolla. Effettuate le

modifiche nel testo è possibile salvarle e disattivare le funzionalità di scrittura (*Save and Leave Edit Mode*), salvare soltanto le modifiche effettuate (*Save Only*), uscire dalla funzionalità non salvando le modifiche effettuate (*Discard Changes and Leave Edit Mode*) oppure non salvare le modifiche e lasciare la funzionalità attiva (*Discard Changes Only*). Per disattivarla, non avendola utilizzata, occorre selezionare *Leave Edit Mode* che appare non appena ci si pone sull'icona.

Crea o visualizza un network: permette di assegnare un nome al nuovo network (insieme di relazioni) che si sta per creare e/o di accedere all'editor per la realizzazione dello stesso.

Scrivi un commento: apre l'editor di testo per attaccare un commento all'unità ermeneutica.

Salva l'unità ermeneutica: permette di salvare il lavoro sotto il nome dell'unità ermeneutica di lavoro corrente.

Esplora gli oggetti: visualizza in ordine gerarchico i documenti, le unità, i codici creati fino al momento corrente.

Query Tool: modulo di analisi in cui i codici e/o le famiglie vengono messi in relazione. Il rapporto fra gli stessi può essere definito secondo criteri di tipo logico, semantico, strutturale.

Assegna un documento: si accede ai file testo, audio o video cui si intende lavorare.

Trova gli oggetti *(Object Crawler)*: permette, digitando una query o selezionando gli elementi, di cercare file o unità di lavoro.

Crea una lista di parole: modulo per la creazione di un vocabolario delle parole presenti nel documento esportabile anche in formato Excel.

Convertitore: converte i file con estensione XML in HTML.

Impostazioni: permette di accedere e modificare le proprietà del software: caratteristiche dell'editor dell'unità ermeneutica, margini, *fonts*, tempo d'attesa prima dello *stand by*, percorsi, risoluzione della stampa dei documenti, *memos*.

Torna indietro: quando si lavora con le citazioni e/o frammenti (*quotations*) permette di tornare alla precedente.

Da cliccare quando si è frustrati: una voce incoraggia a continuare con il lavoro.

4. 2. LA PREPARAZIONE DEI DOCUMENTI

Atlas.ti5 permette di lavorare con documenti di testo di diverse estensioni. Qui si considerano solo i formati di testo, quali *txt*, *rtf* e *doc*. Nella scrittura dei *memos* o di altri editor presenti nel software, il file di testo è in formato *rtf*; gli editor che si presentano sono vuoti, ovvero bianchi, da scrivere quando si accede a tali funzionalità.

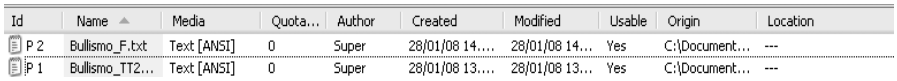
La lavorazione del e sul testo rappresenta una parte fondamentale dell'analisi semi-automatica supportata dai software CAQDAS; non di meno accade con Atlas.ti5. La possibilità di organizzare il lavoro sui testi è garantita dal modulo *Primary Documents (PDs)*, che permette di gestire più documenti di testo e operare su di essi in base a filtri.

Per accedere a questo modulo di gestione dal menu **Documents** si seleziona *Assign/Primary Document Loader*.

Atlas.ti5 non richiede che il testo sia ripartito secondo chiavi funzionali al riconoscimento da parte del sistema, come avviene per altri software; la suddivisione operabile all'interno del documento e quindi l'organizzazione dei file si piega alle esigenze del ricercatore. Il file sul bullismo che si adopera in questo manuale, per esempio, può essere ripartito secondo la variabile di genere: "maschio", "femmina". All'interno della modalità "maschio", per comodità, si fanno confluire i messaggi per i quali non è chiara, bensì "incerta" l'attribuzione di genere. Si preparano quindi in formato *txt* due file che saranno poi caricati nella stessa unità ermeneutica.

Una volta caricati i file, dal menu **Documents** si seleziona *Primary Doc Manager* e si visualizzano gli elementi di dettaglio (fig. 4.1): i documenti inseriti sono due, l'ID sarà quindi rappresentato da P1 e P2. Segue il nome del file; il

tipo di supporto multimediale (*Media type*), in questo caso *Text*; il numero di citazioni fino a quel momento estratte (*Quotation*); il nome dell'utilizzatore di Atlas.ti5 (*Author*); la data di creazione del file (*Created*); quella relativa all'ultima modifica operata (*Modified*); la possibilità di accedere (*Yes*) o non poter accedere al file (*No*), indicata come *Usable*; il percorso da dove è stato caricato il file (*Origin*) e quello in cui è stato salvato (*Location*).



Id	Name	Media	Quota...	Author	Created	Modified	Usable	Origin	Location
P 2	Bullismo_F.txt	Text [ANSI]	0	Super	28/01/08 14....	28/01/08 14...	Yes	C:\Document...	---
P 1	Bullismo_TT2...	Text [ANSI]	0	Super	28/01/08 13....	28/01/08 13...	Yes	C:\Document...	---

Fig. 4.1 – Dettaglio del *Primary Documents*

Un modo alternativo e veloce per selezionare le procedure del modulo *Primary Documents* è ovviamente l'utilizzo della *toolbar* (barra degli strumenti).

Richiama o assegna il documento principale.

Tasto per ritornare in cima.

Scrivere un commento.

Aprire un network.

Rimuovere il documento.

Visualizzare i file mediante la disposizione per icone.

Permette di accedere al modulo di gestione delle famiglie di codici.

4. 3. CREAZIONE DI UNA UNITÀ ERMENEUTICA

Non appena si clicca sull'icona di Atlas.ti5 per avviare il programma, il software ci dà il benvenuto e ci chiede di scegliere fra le varie opzioni di lavoro:

- aprire una unità ermeneutica da una lista selezionata,
- aprire l'ultima HU di lavoro,
- crearne una nuova,
- continuare con il lavoro già avviato.

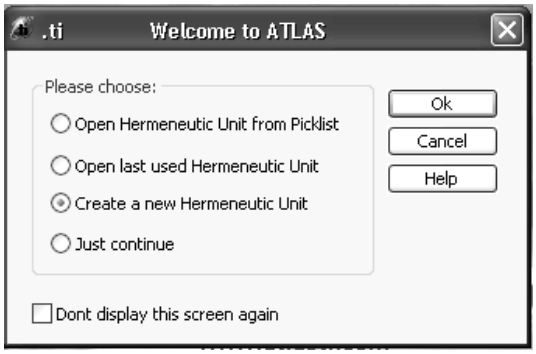


Fig. 4.2 – Finestra d'avvio dell'unità ermeneutica

Questa ultima opzione si presenta quando si lascia il computer in *stand by* per un po'; il tempo di non lavoro prima dello *stand by* può essere modificato dalle impostazioni.

L'unità ermeneutica rappresenta la nostra sessione di lavoro. In questo caso spunteremo "Crea una nuova unità ermeneutica"; il passo successivo è rappresentato dall'inserimento del documento.

Dal menu **Documents** si sceglie *Assign*, si apre la maschera *Primary Document Loader*, si va a recuperare la cartella all'interno della quale si è riposto il file di lavoro e si carica il file di testo. Avendo più file si caricano uno alla volta. L'ordine di inserimento determina il numero progressivo attribuito dall'identificativo. Quindi il documento caricato per primo sarà contraddistinto da P1, quello per secondo da P2 e così via.

A questo punto è, però, bene ricordare che nessun file viene fisicamente inserito all'interno di Atlas.ti5. Questa operazione di inserimento del file serve per creare dei collegamenti fra l'HU e il file di testo sul quale si intende lavorare. Quindi i file sul bullismo rimarranno nella cartella in cui erano stati posizionati inizialmente dall'utente (sul proprio desktop, su una risorsa esterna ecc.); non subiranno nessun spostamento. Questa precisazione è importante, perché ci segnala che se vogliamo spostare una unità ermeneutica da un supporto digitale a un altro dobbiamo trasferire anche i file dati. L'attribuzione dei percorsi ai file è verificabile dalla barra degli strumenti mediante l'icona a forma di martello che indica le **Impostazioni**.

Il salvataggio del lavoro svolto può avvenire in due modi o dal menu **File**/*Save as* o dal menu **Tools**/*Copy Bundle*/*Create Bundle*. Con la prima forma di salvataggio si salvano solo le informazioni relative all'unità ermeneutica in lavorazione, che rimangono però sul medesimo supporto sul quale si è iniziato il lavoro. La seconda procedura consente di trasferire tutti i file lavorati in un nuovo supporto. Per accedere a tali file da un altro supporto occorrerà avviare una unità ermeneutica e quindi dal menu **Tools** selezionare *Copy Bundle*/*Install Bundle* e seguire poi la *directory* di localizzazione del file.

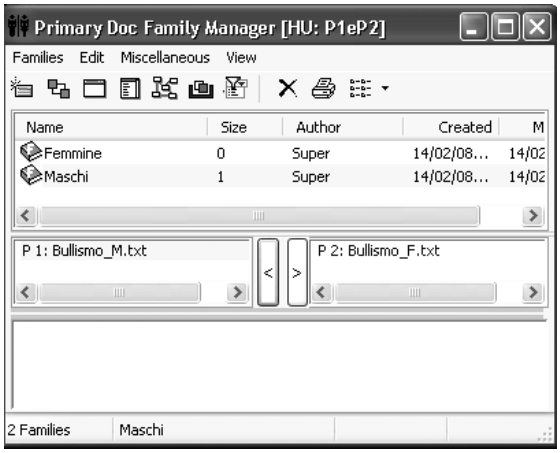


Fig. 4.3 – Modulo per l'assegnazione delle variabili nominali

La suddivisione per genere attribuita ai documenti va riportata all'interno del modulo *Primary Doc Family Manager* (fig. 4.3). Mediante tale modulo si assegnano delle variabili nominali al documento; nel nostro caso è la ripartizione per genere, ma potevano essere anche informazioni quali la zona di provenienza o le classi d'età, ecc. Dal menu **Documents**/*Edit Families*/*Open Family Manager* si accede al modulo di gestione. Da *Families*/*New Families* si assegna il nome alla variabile e si inseriscono i documenti corrispondenti. È così possibile attribuire delle variabili al testo. Eseguite queste procedure preliminari di organizzazione dei documenti è possibile tornare all'unità ermeneutica.

I file introdotti nel software sono visualizzabili dalla barra attraverso **P-Docs:** la finestra a tendina permette di visualizzare l'identificativo e i nomi dei file P1: *Bullismo_M.txt* e P2: *Bullismo_F.txt*.

Fig. 4.4 – Dettaglio della barra dell'unità ermeneutica: P-Docs

Cliccando sull'icona posta accanto a **P-Docs** si apre l'unità di gestione dei documenti (*Primary Doc Manager*); si clicca due volte sul documento e si visualizza il testo prescelto che è ora pronto per essere analizzato.

I file si aprono uno alla volta all'interno di una interfaccia di lavoro che permette di codificarli. Le codifiche assegnate alle porzioni di testo vengono visualizzate a destra della schermata, mentre a sinistra la colonna interna visualizza i paragrafi in cui il programma suddivide il testo e la colonna verso l'esterno riporta le icone di alcune funzionalità specifiche.

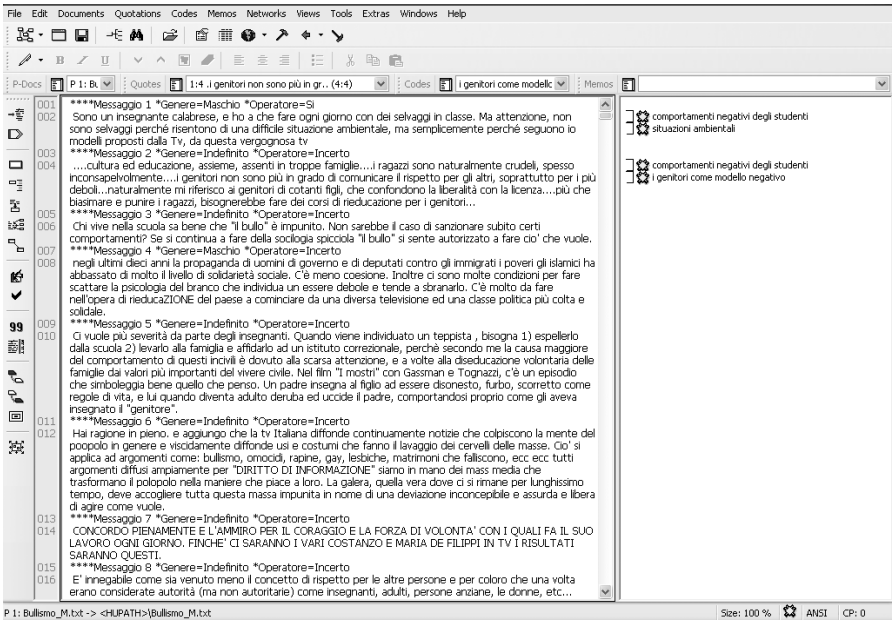


Fig. 4.5 – Interfaccia di lavoro HU

In alto accanto a **P-Docs** troviamo **Quotes** (citazioni): è l'area in cui vengono conservate le frasi o porzioni di testo che il ricercatore ha scelto come significative.

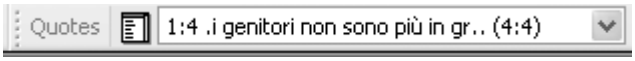


Fig. 4.6 – Dettaglio della barra dell'unità ermeneutica: Quotes

Codes è, invece, l'area in cui vengono conservati i codici creati.

Fig. 4.7 – Dettaglio della barra dell'unità ermeneutica: Codes

La terza finestra a tendina è rappresentata dai *memos: Il* modulo *Memo Manager* include tutti i commenti che il ricercatore produce e si annota al momento della codifica.

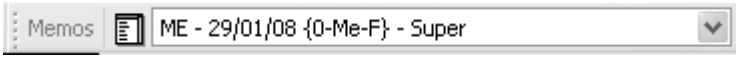


Fig. 4.8 – Dettaglio della barra dell'unità ermeneutica: Memos

Per creare un *memo* occorre selezionare con il cursore del mouse la porzione di testo cui si riferisce, poi cliccare sull'icona del *block notes* dei **Memos**: a questo punto si apre l'unità di gestione *Memo Manager*. Oppure si può procedere dal menu **Memos**/*Create free memo* e si accede a uno spazio di scrittura.

Fig. 4.9 – Modulo di gestione dei Memos

Lo spazio sottostante alla prima schermata del modulo *Memo Manager* serve per visualizzare il contenuto dei *memos* sopra elencati. Per visualizzarli basta cliccare due volte sulla stringa corrispondente. Una volta generato il *memo*, questo può essere collegato ad altri elementi, dal menu **Memos**/*Link memo to* e scegliere fra *Quotations*, *Codes*, *Memos*.

Il *memo* ha la funzione di un post-it; si digita la riflessione che la lettura del testo ha suscitato nel ricercatore e poi da **Memos**/*Save* si salva quanto scritto che appare in questo modo sia nell'unità di gestione che nella barra degli strumenti in alto (**Memos),** ma anche accanto alla porzione di testo corrispondente.

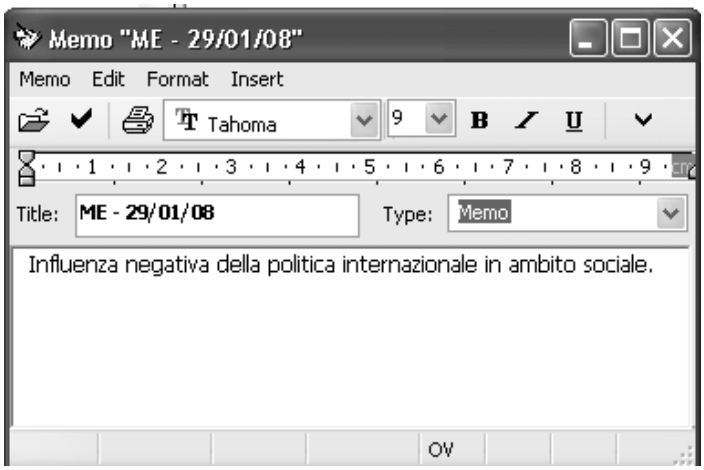


Fig. 4.10 – Editor dei Memos

Essendo una sorta di banco da lavoro, Atlas.ti5 è dotato di numerose funzionalità; alcune, come le icone poste a sinistra dell'interfaccia dell'unità ermeneutica, garantiscono un facile accesso alle risorse del software.

Cerca una riga di testo: digitando il paragrafo o il numero di una linea di testo ci permette di trovarla.

Cerca parole, espressioni o famiglie di parole: trova nel testo la parola digitata o un'espressione selezionata dal menu a tendina

Crea una citazione (Quotes): selezionando il testo permette di creare velocemente una citazione.

Crea codici: permette di creare uno o più codici.

Crea un Code In Vivo: genera un codice nominandolo con la stessa espressione selezionata nel testo.

Crea un codice attingendo da una lista: genera un codice permettendo di scegliere da una lista di codici già esistenti.

Applica un codice: applica alla porzione di testo selezionata l'ultimo codice utilizzato.

Crea un memo: apre l'editor di testo per la scrittura dei *memos*.

Modifica i margini di una citazione: selezionando una citazione dal menu a tendina ed evidenziando una nuova porzione di testo, applica alla nuova citazione la prima citazione selezionata.

Seleziona/Deseleziona linee e paragrafi: permette di paragrafare o di eliminare la paragrafazione dal testo.

Imposta margini: permette di ampliare o ridurre i margini dell'interfaccia di lavoro.

Crea una legame fra le citazioni: stabilisce un collegamento ipertestuale fra le citazioni visibile nello spazio a destra dell'interfaccia dell'unità ermeneutica. L'utilizzo di questa funzione è propedeutico all'icona successiva.

Crea una relazione fra le citazioni: la relazione creata è di tipo concettuale. Si chiede infatti di determinare il tipo di relazione intercorrente fra la citazione selezionata e una precedente (selezionata mediante l'utilizzo della precedente icona). La relazione può essere di diverso tipo ed è indicata da simboli, quali:

1. >>>> se la citazione **continua** la precedente;
2. X> se **contraddice** la precedente;
3. ->| se **critica** la precedente;
4. :> se **argomenta** la precedente;
5. ? se **amplia** la precedente;
6. ?> se **spiega** la precedente;

7. !> se **giustifica** la precedente;
8. *> se **supporta** la precedente.

Ciascuno di questi simboli definisce il tipo di relazione che intercorre fra le diverse citazioni. Tale relazione può essere di tipo asimmetrico o transitivo; è asimmetrica nel primo e nel terzo caso, di tipo transitivo in tutti gli altri. L'ultimo tasto della finestra del comando permette di aprire una finestra di editor per le relazioni (*Hiperlink-Relations Editor*).

Visualizza le citazioni: evidenzia le citazioni presenti nell'intero paragrafo selezionato.

Zoom: ingrandisce i caratteri del testo.

4. 4. CODIFICARE UN TESTO

La prima attività di analisi operabile sul testo è la sua codifica. Essa si realizza leggendo il testo riga per riga e creando per le parti ritenute importanti dei codici, delle etichette interpretative.

Posti davanti allo schermo e all'interfaccia dell'unità ermeneutica mediante il cursore del mouse si evidenzia la porzione di testo ritenuta importante. Mediante l'icona **Crea una citazione** o **Crea codici** si stabilisce cosa fare del testo evidenziato.

4. 4. 1 ESTRARRE CITAZIONI DAL TESTO

Il lavoro di analisi del testo si svolge mediante l'estrazione delle citazioni. Per realizzarlo è possibile operare in tre maniere distinte.

Una prima possibilità è rappresentata dal tasto **Crea una citazione** presente nell'editor dell'unità ermeneutica. In questo caso occorre selezionare il testo e cliccare su tale icona.

Una seconda modalità consiste nell'uso del **tasto destro** del mouse che apre la finestra sottostante (fig. 4.11). L'opzione *Create Free Quotation* rende il testo selezionato una citazione visualizzabile dalla finestra a discesa **Quotes**.

Create Link Source e *Create Link Target* hanno la stessa funzione di **Crea un legame** e/o **una relazione fra le citazioni.**

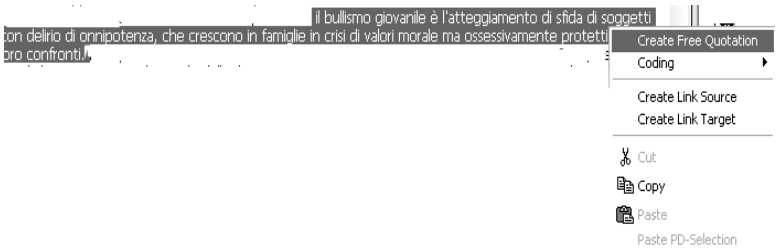


Fig. 4.11 – Esempio di estrazione delle citazioni

Una terza opzione è rappresentata dalla possibilità di selezionare la porzione di testo e di trascinarla all'interno della finestra a discesa **Quotes.**

Le citazioni e i legami stabiliti fra le stesse sono visualizzabili alla destra della schermata del testo.

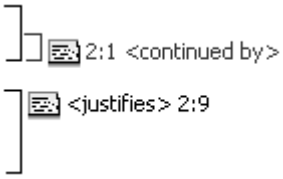


Fig. 4.12 – Esempio di visualizzazione delle citazioni e dei loro legami

Per visualizzare il contenuto di una citazione (quindi la parte di testo corrispondente) basta cliccare sull'icona della citazione ed essa appare.

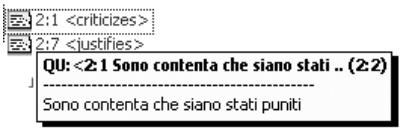


Fig. 4.13 – Esempio di visualizzazione del contenuto di una citazione

Il legame fra più citazioni e fra le citazioni e i codici può essere rappresentato anche mediante un output grafico o "network tematico", in cui si offre una sorta di identikit tematico di una citazione.

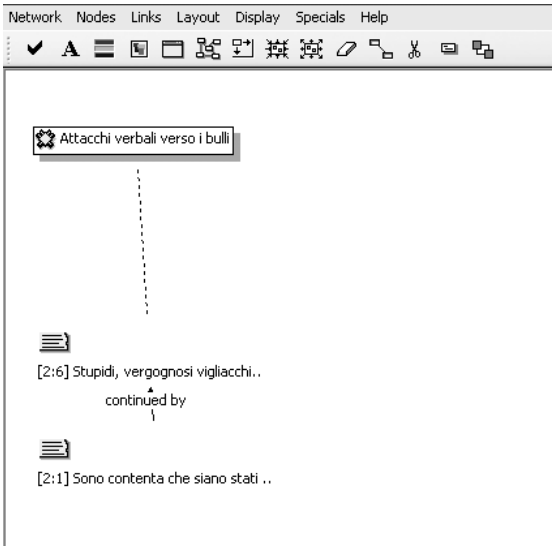


Fig. 4.14 – Network delle relazioni stabilite fra più citazioni e un codice

A esso si accede evidenziando la citazione posta ai margini dell'interfaccia con il tasto destro e selezionando *Open Network* (fig. 4.14). La citazione "sono contenta che siano stati puniti" prosegue nella "stupidi, vergognosi, vigliacchi" ed entrambe sono state codificate con il codice "attacchi verbali verso i bulli".

La gestione delle citazioni è garantita dal modulo *Quotation Manager* (fig. 4.15) cui si accede dal menu **Quotations.** L'*Id* identifica le coordinate della citazione; il primo numero si riferisce al documento mentre il secondo è il progressivo della citazione. La presenza di una relazione fra le citazioni è indicata con *Density*; *Start* indica il paragrafo corrispondente alla citazione stessa. In totale per i due documenti sul bullismo sono state estratte 275 citazioni.

La visione del modulo di gestione *Quotation Manager* ci dà la possibilità di controllare le citazioni realizzate e di verificare se alcune di esse sono sovrapponibili o a posteriori non rilevanti. In questo caso si può optare per estendere e/o ridurre i margini di un frammento di testo precedentemente selezionato. Per farlo si ha a disposizione o l'icona **Modifica i margini di una citazione** che si trova nell'editor dell'interfaccia dell'unità ermeneutica oppure il menu **Quotation**/*Miscellaneus*/*Merge Quotations*. In questo modo due citazioni vengono unite in una sola.

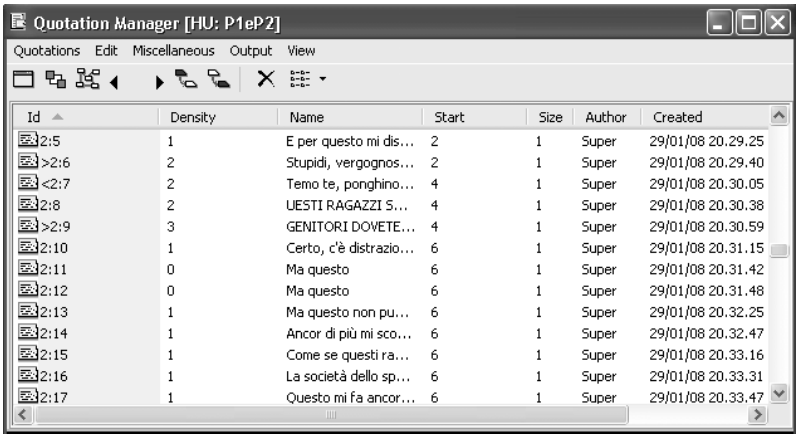


Fig. 4.15 – Modulo di gestione delle citazioni

Una reportistica per le citazioni è visualizzabile mediante **Quotation**/*Output* (fig. 4.16). È poi possibile scegliere se visualizzare l'informazione per una sola citazione (*Selected Quotation*), per tutte le citazioni in forma estesa (*All Quotation*) oppure sempre per tutte le citazioni ma solo come elenco (*All Quotation list*).

L'editor dell'output permette di visionare l'informazione sulle citazioni, riportando il documento cui si riferisce, il frammento di testo e il codice che si è dato a quel frammento. Alla citazione "i genitori non sono più in grado di comunicare il rispetto per gli altri, soprattutto per i più deboli" è stato attribuito il codice "assenza del ruolo educativo della famiglia"; il documento da cui è stata estratta è il P1, quindi è quello che raggruppa gli individui di genere maschile.

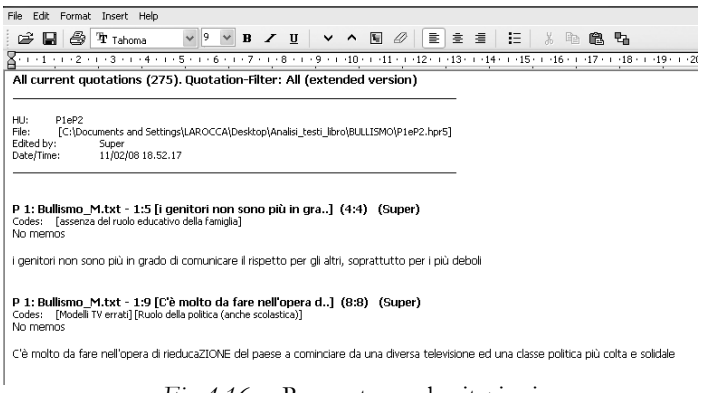


Fig. 4.16 – Rapporto per le citazioni

La seconda citazione è invece un esempio di attribuzione di due codici "modelli TV errati" e "ruolo della politica (anche scolastica)" allo stesso frammento di testo.

Le relazioni stabilite fra le varie citazioni possono essere visualizzate mediante un output grafico che dà vita a un "network testuale" (cfr. § 4.7).

4. 4. 2 CREARE E ATTRIBUIRE CODICI

Una citazione si lega, quasi sempre, a un codice. Mediante il menu **Codes**/*Coding* si realizza una codifica, che può essere un *Open Coding* quando si digita una etichetta di propria attribuzione; *Code In Vivo* se la stessa citazione viene utilizzata come codice; *Code by List* se si attinge a una lista di codici generata man mano nel processo di codifica; *Quick coding* applica l'ultimo codice utilizzato alla nuova citazione.

Realizzato il lavoro di codifica, che ovviamente è molto soggettivo, attraverso **Codes**/*Code Manager* si accede al modulo di gestione dei codici. I codici sono disposti in ordine alfabetico (*Name*); accanto è posta la frequenza (*Grounded*) e la densità che indica la presenza di una relazione fra il codice in oggetto e gli altri presenti nel testo (*Density*).

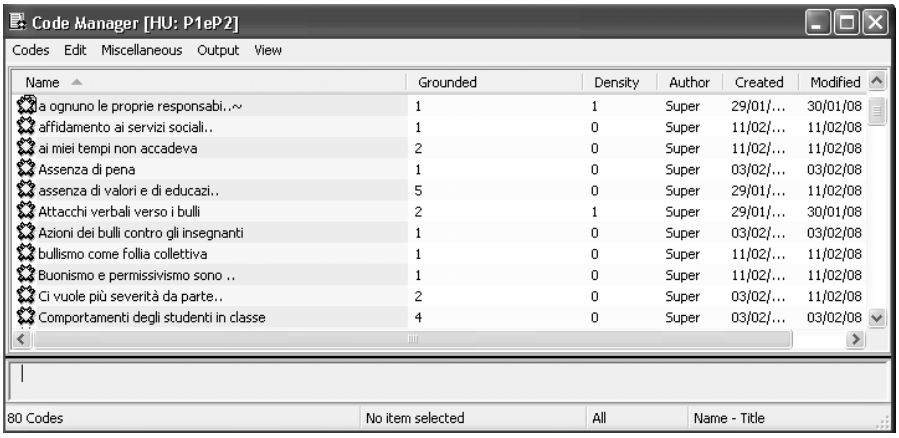


Fig. 4.17 – Modulo di gestione dei codici

Il modulo *Code Manager* permette di avere una panoramica sul lavoro svolto fino a quel momento. Ci si potrebbe accorgere che alcuni codici creati sono simili fra di loro, come nel nostro caso "ruolo della scuola", "ruolo dei genitori" e "contrapposizione scuola/genitori nella responsabilità del bullismo". In que-

sti casi è possibile far confluire i codici in uno solo; si potrebbero sommare, per esempio, i primi due nell'ultimo.

Evidenziamo dal modulo il codice nel quale vogliamo far confluire gli altri codici; quindi dal menu selezioniamo **Miscellaneus**/*Merge Codes* e dalla finestra uno o più codici - in questo caso due - e diamo l'*OK*. In questo modo si somma la frequenza del codice "ruolo della scuola" e "ruolo dei genitori" in "contrapposizione scuola/genitori nella responsabilità del bullismo" che avrà come frequenza la somma delle frequenze di tutti e tre i codici.

A questo punto chiuso il *Code Manager* è possibile avere delle statistiche sulla frequenza dei codici nel primo e nel secondo documento: **Codes/***Output/Codes-Primary-Documents-Table/Standard Report;* l'output può essere visualizzato anche in formato Excel.

In una nuova finestra *Send output to* ci viene chiesto se si vuole l'output come *Editor*, *Printer*, *File* o *File & Run*; va bene nel formato *Editor*.
L'output riportato ci permette di leggere, come in una distribuzione di valori assoluti, per ogni codice il totale per documento e il totale cumulato (fig. 4.18). Il lavoro di codifica sui due file ha prodotto 240 codici di cui 125 nel primo documento e 115 nel secondo.

```
Code-Filter: All
PD-Filter: All

-----------------------------------
                       PRIMARY DOCS
CODES                  1   2 Totals
-----------------------------------
"il bullo" è impunit   1   0   1
a ognuno le proprie    0   1   1
affidamento ai servi   1   0   1
ai miei tempi non ac   2   0   2
Assenza di pena        0   1   1
assenza di valori e    2   3   5
Attacchi verbali ver   0   2   2
Azioni dei bulli con   0   1   1
bullismo come follia   1   0   1
Buonismo e permissiv   1   0   1
Ci vuole più severit   2   0   2
Comportamenti degli    0   4   4
corsi di rieducazion   1   0   1
delinquenza giovanil   2   1   3
"
"
"
-----------------------------------
Totals                125 115 240
```

Fig. 4.18 – Rapporto per i codici

All'analisi dei codici possono essere applicati dei filtri; basta selezionare da **Codes**/*Filter* la modalità per la quale si vogliono filtrare i codici, quale per esempio l'autore che li ha generati (utile quando si lavora in un team di ricerca), la data di creazione e così via.

4. 5. LE FAMIGLIE E LE SUPER FAMIGLIE DI CODICI

L'operazione successiva alla codifica del testo è rappresentata dalla creazione di "famiglie di codici". All'interno delle famiglie si raggruppano i codici a seconda della presenza di un comune significato. L'operazione ha la valenza di ridurre la quantità di informazioni trovate.

Dal menu si sceglie **Codes**/*Edit Families*/*Open Family Manager*. Aperto il modulo di gestione delle famiglie occorre creare una nuova famiglia; tale processo avviene mediante *Families*/*New Family*. A questo punto si apre una finestra *Create Code Family* all'interno della quale va registrato il nome della nuova famiglia. In basso a destra sono elencati tutti i codici generati durante la fase di codifica, il ricercatore selezionerà manualmente quelli da includere all'interno di quella famiglia; il numero di codici attribuiti a ciascuna famiglia ne determina la grandezza (*Size*).

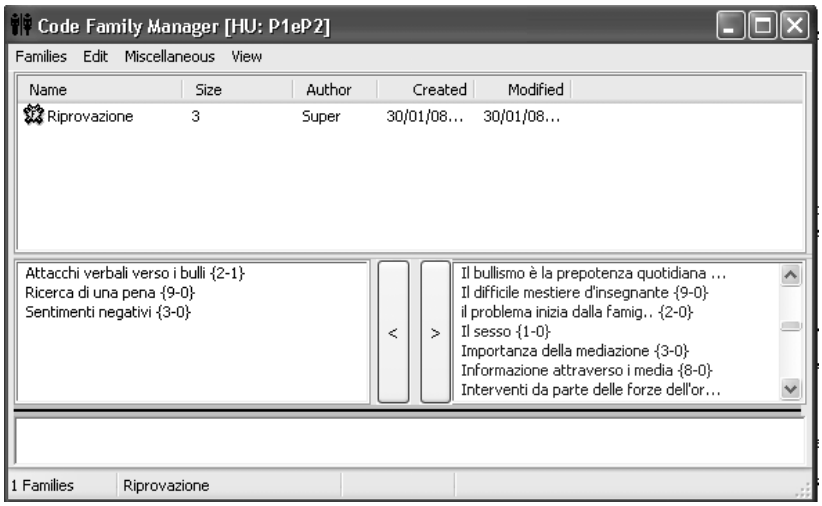


Fig. 4.19 – Modulo di gestione per le famiglie di codici

In basso, nella finestra di dialogo, è riportato il numero delle famiglie: in questo caso è ancora 1 e il nome della famiglia al momento selezionato "Riprovazione". Lo stesso codice può essere attribuito a più famiglie.

All'interno di "Riprovazione" si vogliono far confluire tutti quei codici che si riferiscono a citazioni in cui gli scriventi esprimono un sentimento di disappunto nei confronti del bullismo e delle sue manifestazioni. La creazione di un macro contenitore come "Riprovazione" scaturisce, quindi, da un'operazione di sussunzione di quanto emerso nel testo all'interno di un contenitore concettuale più grande. Questa è un'operazione logico concettuale che il ricercatore opera in proprio. Tuttavia nell'esposizione del proprio lavoro potrebbe essere utile il mantenerne traccia.

Famiglia di codici	Riferimento concettuale
Riprovazione	Si riferisce a quell'insieme di espressioni codificate in: "attacchi verbali verso i bulli", "ricerca di una pena", "sentimenti negativi", in cui affiorano elementi di disappunto, disapprovazione nei confronti dei bulli e dei loro atteggiamenti.

All'interno del modulo *Code Family Manager*, nella parte inferiore della schermata, può essere effettuata l'operazione riportata in tabella. L'appunto mediante il quale si esplicita il legame fra i codici e le famiglie di codici può essere salvato sotto forma di *memo*.

Al modulo di gestione delle famiglie si può accedere anche mentre si sta codificando; in questo caso dal menu **Codes**/*Code Manager*/*Codes*/*Edit Family* si può scegliere se aprire il *Code Family Manager*, poi *Assign Family*.

Nella gestione dei codici da assegnare alle famiglie si può rivelare utile la funzione dei filtri; per esempio se la si imposta su "nessun filtro" (**Codes**/*Filter*/*Families*/*None*) renderà visibili nel modulo *Code Manager* solo i codici non ancora assegnati a nessuna famiglia. Se si de-selezionano le famiglie di codici queste scompaiono automaticamente anche dalla colonna situata alla nostra destra dell'interfaccia dell'unità ermeneutica.

Create le famiglie, l'operazione successiva di sussunzione dell'informazione è rappresentata dalla creazione di "super famiglie", una sorta di ulteriore raggruppamento delle categorie generate. Per accedervi dal menu **Codes** selezionare *Edit Family Manager*/*Open Family Manager*; aperta la finestra *Code Family Manager* dal menu **Families** selezionare *Open Super Family Tool*. Nella colonna a sinistra si visualizzano le famiglie di codici fino a quel momento generate; la parte sulla destra serve invece a visualizzare le relazioni fra due o più famiglie. Per

stabilire i tipi di relazione si utilizzano gli operatori booleani.

OR: nel creare una super famiglia serve per sommarvi all'interno i codici delle due o più famiglie selezionate.

XOR (AND+OR): mantiene all'interno di una super famiglia i codici non comuni alle famiglie che si stanno ponendo in relazione.

AND: nel creare una super famiglia mantiene all'interno di essa solo i codici comuni a entrambe.

NOT: azzera uno degli elementi posti in relazione e ne mantiene solo uno.

Dalla colonna a sinistra si selezionano due o più famiglie con le quali si vuole lavorare; queste vengono visualizzate nella prima finestra in alto. Successivamente si identifica il tipo di operazione selezionando l'operatore booleano corrispondente e si conferma cliccando su *Super Family*, che ci permette non solo di confermare l'operazione ma anche di stabilire il nome della super famiglia. Il tasto *Refresh* svuota le finestre e consente di ricominciare con una nuova operazione. La super famiglia creata andrà a visualizzarsi nella colonna a sinistra con una icona simile ma di colore differente rispetto a quelle delle famiglie. Le super famiglie hanno l'icona di colore rosso, le famiglie giallo.

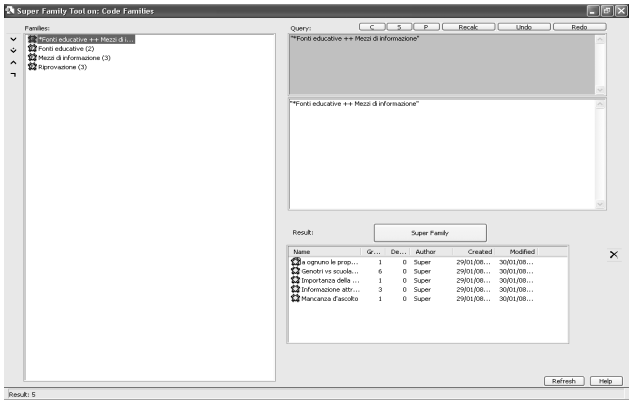


Fig. 4.20 – Modulo di gestione per la creazione di super famiglie di codici

4. 6. LE QUERY

Alcune funzionalità presenti in Atlas.ti5 permettono di verificare le relazioni fra i codici e/o le famiglie. Sono "funzionalità speciali"; fra queste l'*Object Crawler* e il *Query Tool*.

Mediante l'*Object Crawler* è possibile verificare se, in fase di codifica, codici concettualmente vicini e quindi con termini contenuti nelle loro etichette definibili come sinonimi sono stati assegnati a porzioni di testo differenti. L'*Object Crawler* consente di verificare se lemmi uguali sono presenti in più codici, in più citazioni, ecc.

Dal menu **Tools**/*Object Crawler* si accede a tale modulo. L'operazione consiste di tre fasi: definizione della query, selezione degli oggetti all'interno dei quali ricercare, visualizzazione del risultato.

Essendo l'argomento principale della nostra analisi il fenomeno del bullismo definiamo una query volta a rinvenire le citazioni all'interno delle quali si parla di bullo|bulli|bullismo.

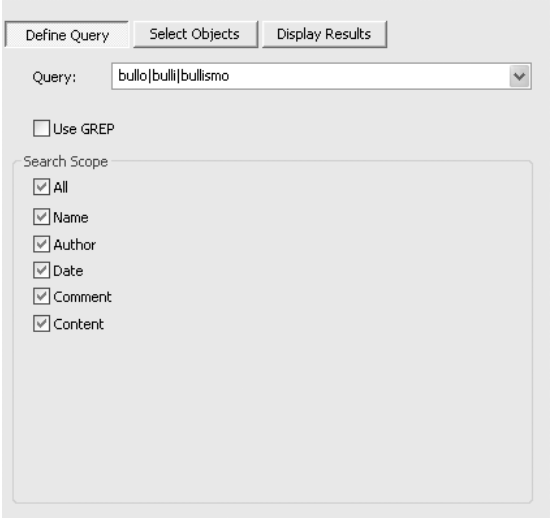


Fig. 4.21 – Finestra per la definizione della query

Poiché l'obiettivo della query è ritrovare tutte le citazioni in cui appaiono i termini della query stessa, da *Select Object* selezioniamo *Quotations* e finalmente è possibile visualizzare il risultato.

Fig. 4.22 – Finestra per la selezione dei campi su cui operare la query

Selezionando la citazione se ne visualizza il contenuto nella finestra posta in basso (fig. 4.23).

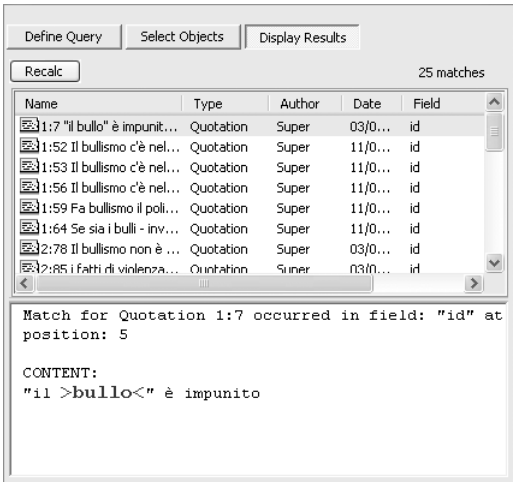


Fig. 4.23 – Finestra per la visualizzazione dei risultati della query

L'*Object Crawler* e il *Query Tool* sono funzioni per la ricerca nel testo sebbene si pongano a livelli di ricerca differenti.

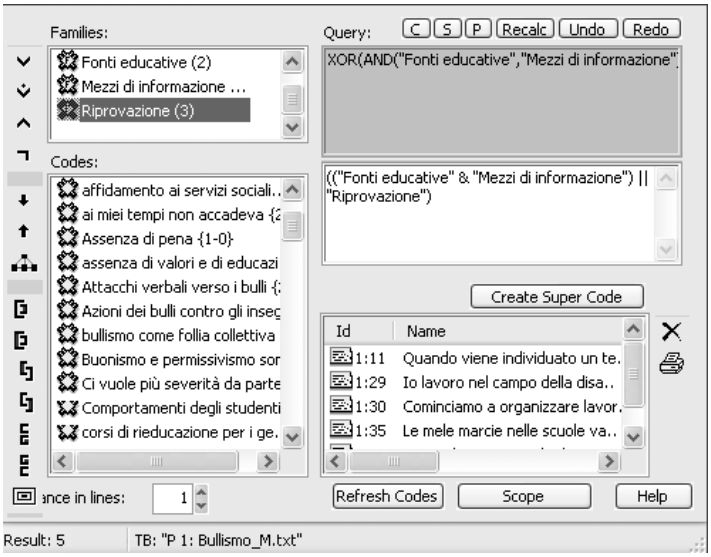


Fig. 4.24 – Finestra per la definizione delle query su codici, famiglie e citazioni

Il *Query Tool* serve per ritrovare citazioni utilizzando i codici che sono stati loro attribuiti durante il processo di codifica; invece l'*Object Crawler* ritrova in una unità di analisi stabilita la parola o la stringa di testo che si è digitato.

Per accedere all'interfaccia delle query si può: partire dal menu **Codes**/*Output*/*Query Tool* oppure dal menu **Tools**/*Query Tool*, senza dimenticare l'accesso veloce dal "binocolo" posto nella barra degli strumenti.

Nell'analisi dei messaggi sul bullismo siamo interessati a conoscere quali citazioni presenti all'interno della famiglia "fonti educative" sono anche presenti in "mezzi di informazione".

Per stabilire questa relazione abbiamo a disposizione il modulo di gestione delle *query*. La colonna esterna a sinistra ci mostra l'insieme di operatori mediante i quali stabilire procedure di relazione. Si tratta di tre gruppi di operatori: gli operatori booleani o logici, gli operatori semantici e gli operatori di prossimità.

OR: si usa per estrarre citazioni in cui sono presenti o uno solo o entrambi i codici selezionati.

XOR: si usa per estrarre citazioni in cui è presente solo uno dei codici selezionati.

AND: si usa per estrarre le citazioni in cui entrambi i termini sono presenti.

NOT: si usa per escludere un codici da un insieme più ampio.

SUB: imposta una ricerca a partire dai livelli superiori verso quelli inferiori.

UP: imposta una ricerca a partire dai livelli inferiori verso quelli superiori.

SIBling: imposta una ricerca per tutte le citazioni connesse al codice selezionato e a ogni altro codice a essa associato.

Entro: recupera le citazioni codificate all'interno del codice A e che allo stesso tempo si trovano anche in B.

Includere: recupera tutte le citazioni codificate in A che contengono citazioni codificate in B.

Sovrapposizione: recupera tutte le citazioni codificate in A che si sovrappongono in B.

Sovrappone: recupera tutte le citazioni codificate in A sovrapposte a B.

Segue: recupera tutte le citazioni codificate in A che continuano in B.

Precede: recupera tutte le citazioni codificate in A che precedono in B.

Co-occorrenze: si usa per combinare i precedenti operatori fra di loro fatta eccezione per "segue" e "precede".

Nella prima finestra in alto sulla sinistra si visualizzano le famiglie o le super famiglie create fino a quel momento. Nell'impostare una relazione è importante considerare l'ordine in cui si inseriscono gli elementi nella finestra in alto a destra. Per esempio, se si vogliono conoscere quali citazioni presenti all'interno della famiglia "fonti educative" sono anche presenti in "mezzi di informazione", occorre prima inserire "fonti educative", poi selezionare il se-

condo termine della relazione "mezzi di informazione" e quindi l'operatore AND per stabilire la relazione. Successivamente poniamo in relazione "riprovazione" mediante l'operatore XOR volendo individuare solo quelle citazioni in cui è presente il codice selezionato. L'operazione posta in essere ci restituisce un nuovo super codice (*Create Super Code*) in cui sono stati selezionati quegli elementi propri delle fonti educative e dei mezzi di informazione che producono riprovazione nei partecipanti al forum.

Stabilite le relazioni, con *Create Super Code* possiamo dare un nome al super codice, per esempio i "motivi della riprovazione". Il nuovo super codice creato diventa esso stesso passibile di relazioni.

Se si volessero estrarre le citazioni afferenti al super codice "motivi della riprovazione" solo per le femmine del documento P2, per farlo bisognerebbe delimitare la ricerca solo a uno dei due documenti. Il tasto *Scope* permette l'avvio di tale procedura. Selezionato il *Primary Doc Families* corrispondente, ovvero "femmine" e il documento P2, dando l'*OK* si ritorna all'interfaccia precedente che ci permette di visualizzare le 8 citazioni all'interno delle quali si verifica la nostra query.

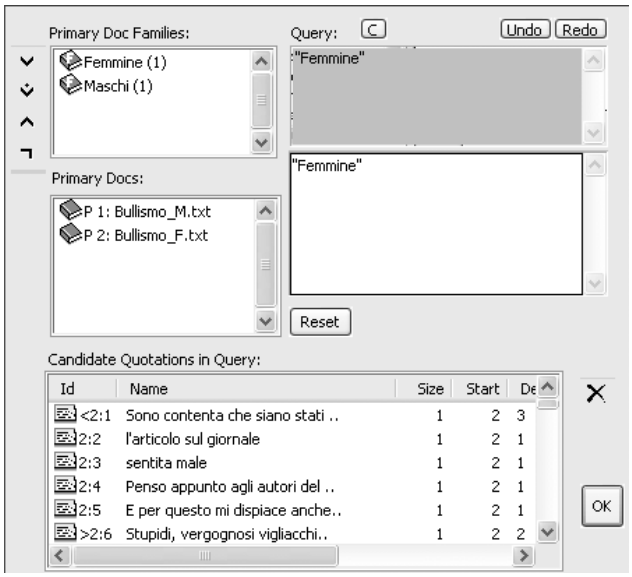


Fig. 4.25 – Finestra per la selezione delle variabili nominali per la query

4.7. I NETWORK – RAPPRESENTAZIONI DI RELAZIONI

Tutte le relazioni generate possono essere visualizzate graficamente all'interno di network grafici. I network possono essere il frutto di una importazione/visualizzazione di nodi creati durante il processo di lavorazione oppure di una nuova generazione.

Tutte le relazione create sono rappresentabili mediante output grafico; è possibile avere:

- *network testuali*: in cui si stabiliscono relazioni per le citazioni;
- *network concettuali di primo livello*: in cui si evidenziano le relazioni per i codici;
- *network concettuali di secondo livello*: per le famiglie e per le super famiglie.

Dal menu principale **Networks** si seleziona *Network View Manager/Network Views/New* per accedere alla finestra di generazione *Name Network View*; si digita il nome del network e si crea l'accesso al primo network. Selezionandolo dalla finestra di dialogo si accede all'area di lavoro.

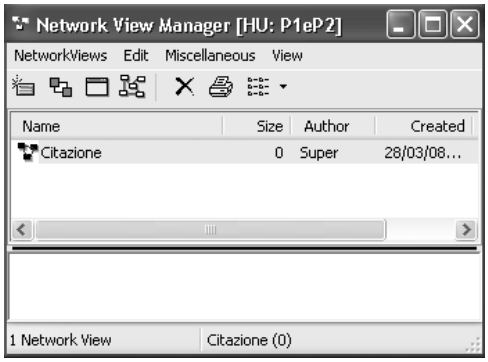


Fig. 4.26 – Modulo per la gestione dei network

Dal sub-menu **Nodes** si sceglie se importare i nodi o generarne di nuovi (*New Node*). Per la prima opzione: **Nodes**/*Import Nodes* si apre, così, la finestra di dialogo in cui occorre selezionare il tipo di nodo (*Node Type*), ovvero scegliere se si vuole lavorare con i codici, le citazioni, le famiglie, le super famiglie e quant'altro si è generato man mano (per esempio i *memos*). Da *Import Nodes*, dopo aver scelto il tipo di nodo, per esempio le citazioni, si selezionano quelle che si vogliono importare, quindi *Import* e si visualizzano i nodi con i tipi di legami/relazioni stabilite dal ricercatore.

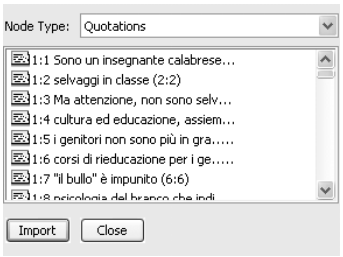


Fig. 4.27 – Finestra per importare i soggetti delle relazioni (citazioni)

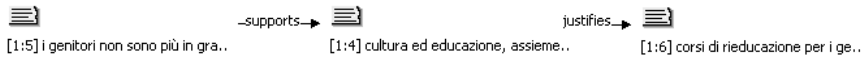


Fig. 4.28 – Esempio di network a catena per le citazioni

In questo caso si è scelto di visualizzare le relazioni poste in essere sottoforma di catena tra "i genitori non sono più in grado di comunicare il rispetto per gli altri, soprattutto per i più deboli", sostenuta (*Supports*) dalla successiva "cultura ed educazione, assieme, sono sparite", che giustifica (*Justifies*) la necessità di "corsi di rieducazione per i genitori".

Qualora, invece, si volesse riflettere sulle possibili relazioni al momento di generare il network e non durante le fasi di lavoro è sempre possibile farlo.

Con la medesima procedura si importano i soggetti del reticolo, per esempio i codici, ma non le relazioni fra gli stessi. I codici scelti appariranno nella finestra senza nessun ordine ma come delle carte da gioco da distribuire.

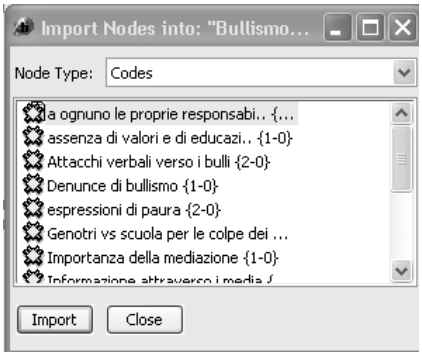


Fig. 4.29 – Finestra per importare i soggetti delle relazioni (codici)

Per attribuire una relazione occorre evidenziare il soggetto della relazione: dal menu **Links**/*Link Nodes* si genererà una freccia che manualmente congiungerà il nodo principale con quello oggetto di relazione; cliccando sul secondo membro della relazione appare una finestra all'interno della quale occorre scegliere il tipo di relazione.

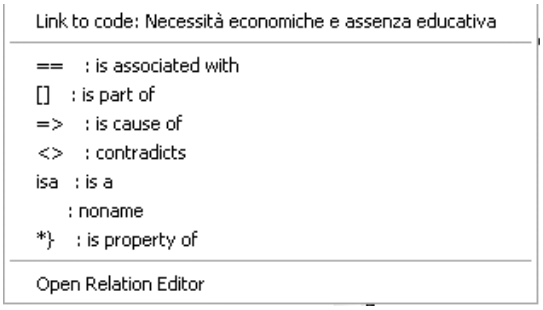


Fig. 4.30 – Finestra per l'attribuzione del tipo di relazione fra i codici

Si viene così a strutturare un reticolo relazionale che, se basato sulle citazioni, può definirsi "testuale", poiché riporta frasi selezionate nel testo, se basato su codici, famiglie o super famiglie, si può definire "concettuale" di primo o di secondo livello.

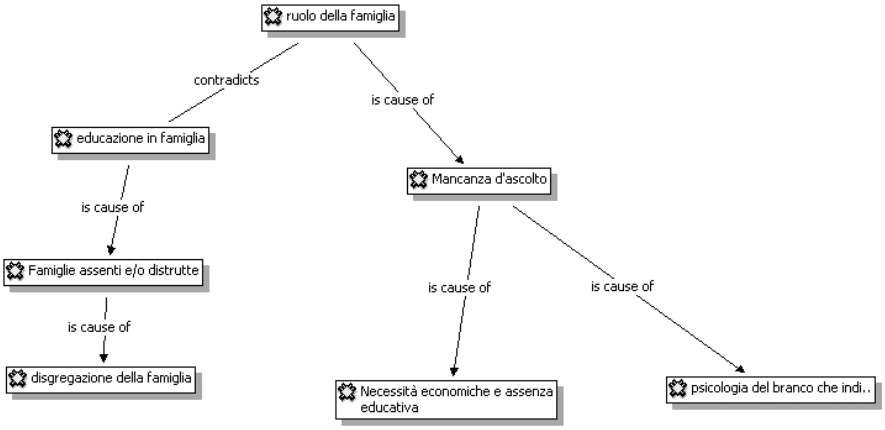


Fig. 4.31 – Netwrok concettuale per il ruolo della famiglia nel bullismo

L'analisi condotta sul testo ha posto in evidenza l'importanza del **ruolo della famiglia** nella generazione di fenomeni di bullismo.

È possibile organizzare la rappresentazione degli elementi di un network mediante un layout semantico (*Semantic Layout*) o topologico (*Topological Layout*). La differenza consiste nell'ordine che verrà attribuito alla disposizione degli elementi nel network. Il layout semantico privilegia una relazione fra gli elementi basata sui significati; il layout topologico stabilisce un ordine negli elementi. Il network generato può essere salvato mediante **Network**/*Save as Graphic File*.

5.
LAVORARE CON NVIVO7: ORGANIZZARE E CODIFICARE IL TESTO

La versione 7 di NVivo rappresenta il superamento di Nudist e delle precedenti versioni di NVivo. Consente lo sviluppo di percorsi di analisi semi-automatica e si rivela estremamente utile nel maneggiare testi qualitativi anche molto ampi. Consente di codificare e organizzare le informazioni, in modo da poterne esplorare il contenuto o costruire e testare teorie sui dati testuali. Infatti, nella codifica è possibile procedere sia con un approccio dall'alto che dal basso. La versione qui utilizzata è disponibile all'indirizzo <http://www.qsrinternational.com/products_nvivo.aspx>.

5. 1. CREARE UN PROGETTO DI LAVORO

Una delle prime operazioni da realizzare con NVivo7 consiste nella creazione di un progetto di lavoro. Aprendo il programma bisogna scegliere *New Project*. Occorre attribuire un titolo al lavoro *Bullismo* e si può dare una breve descrizione del contenuto del file.

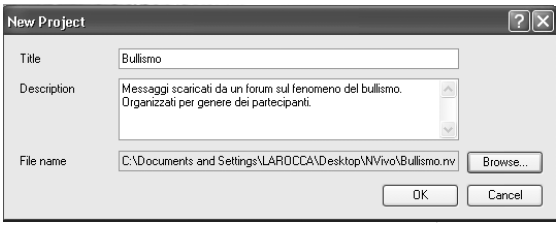


Fig. 5.1 – Finestra per nominare il progetto di lavoro

Si crea in questo modo una sessione di lavoro con estensione *nvp*. Attraverso

Browse scegliamo dove salvare i vari file che si andranno generando. Di default il progetto viene salvato nella cartella *Documenti*.

Selezionato *OK* si accede all'interfaccia di lavoro, per molti aspetti simile a quella di Microsoft Outlook. Nella barra di titolo in alto a sinistra troviamo il nome del nostro progetto di lavoro *Bullismo.nvp*, in basso sotto la barra degli strumenti abbiamo tre piani di lavoro. Nella colonna a sinistra si visualizza il pannello di navigazione; da qui semplicemente con un doppio click sulle icone è possibile accedere alle funzioni principali del software. Al centro si trova il pannello dei documenti; selezionando il singolo documento, questo appare nella finestra di dettaglio sottostante che chiameremo pannello di lavoro.

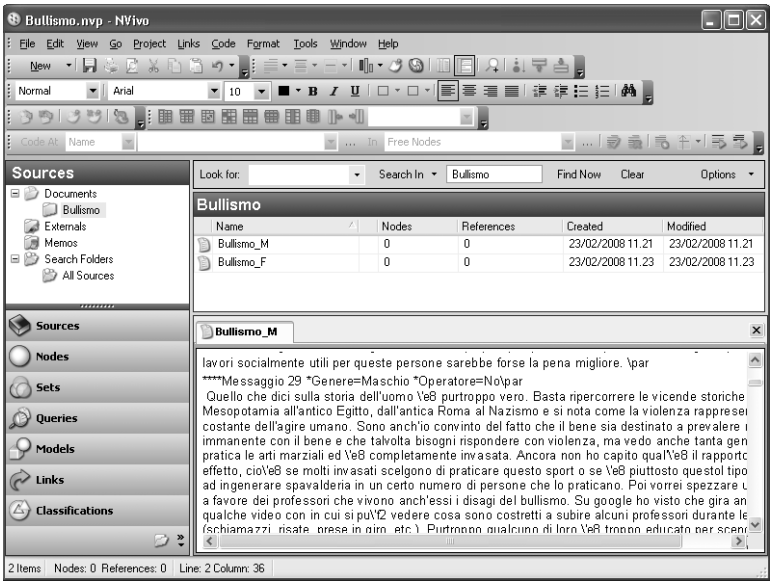


Fig. 5.2 – Interfaccia di NVivo7

Il pannello di navigazione posto sulla sinistra elenca tutti gli elementi con cui è possibile lavorare in NVivo7: **Sources**, **Nodes**, **Sets**, **Queries**, **Models**, **Links**, **Classifications**, **Folders**. Esaminiamo **Sources**: al suo interno si trovano altre sotto cartelle, fra queste:

- *Documents*: contiene la nuova cartella di lavoro all'interno della quale sono stati inseriti i documenti testuali (Bullismo);
- *Memos*: sono gli appunti presi durante il lavoro;
- *Externals*: contiene tutti quei documenti che pur non facendo parte del ma-

teriale oggetto d'analisi possono costituire una fonte di informazione utile per il lavoro, quali bibliografie, e-book e altro.

Il lavoro con NVivo7 può essere condotto utilizzando tre livelli di comando, ovvero si possono impartire i comandi: dalla barra degli strumenti in alto; cliccando con il tasto destro del mouse sul titolo del documento dal pannello di lavoro; oppure accedendo dalla finestra di navigazione posta a sinistra.

La prima operazione consiste nell'importare all'interno del software il documento, che può essere in formato testo (*doc*, *rtf*, *txt*) o anche immagini. Si utilizzano qui i file *Bullismo_M.txt* e *Bullismo_F.txt*; i messaggi dei forum sono stati suddivisi per genere dei partecipanti.

Da **Project**, posto nella barra dei menu in alto, si seleziona *New Folder*; nuovamente ci viene chiesto di assegnare un nome alla cartella e una descrizione.

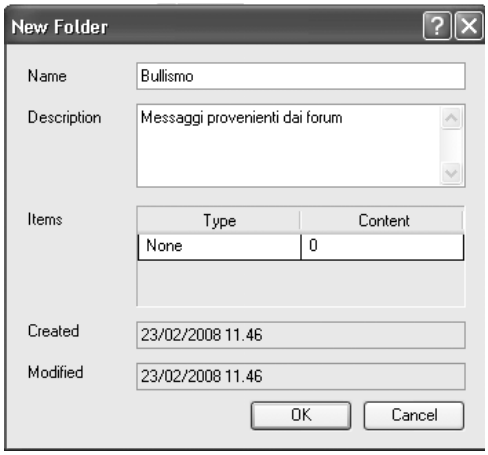


Fig. 5.3 – Finestra per creare una cartella di lavoro

La nuova cartella andrà a visualizzarsi a sinistra nel pannello di navigazione. A questo punto i due file possono essere importati all'interno della cartella *Bullismo*. Sempre dal menu **Project** si seleziona *Import Documents* e si importano uno alla volta i due o più file di lavoro. Il percorso viene visualizzato nella stringa sotto *Import from* e si seleziona da *Browse* (fig. 5.4).

Occorre spuntare anche *Code sources at new cases located under* (i casi possono essere importati da un file Excel o inseriti manualmente).

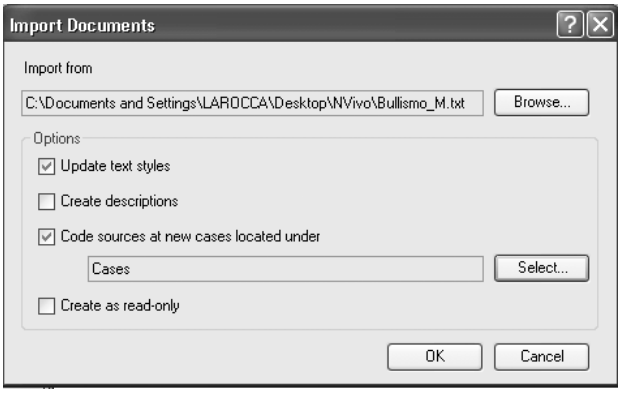


Fig. 5.4 – Finestra per specificare il percorso di importazione del testo

Qui è importante spuntare questa casella perché questa operazione ci permette di abbinare il file di testo con le sue caratteristiche socio-anagrafiche (cfr. § 5.2): i casi sono per noi le persone. In tal senso si sarebbe potuto suddividere ulteriormente il file dei maschi considerando il messaggio di ogni intervistato come un singolo documento, oppure – qualora si individuasse una modalità di pre-trattamento più idonea – raggruppare tutti i messaggi di ogni rispondente come risposte date a domande di un'intervista. Volendo continuare a lavorare solo su due file i nostri casi saranno rappresentati dal "gruppo_maschi" e "gruppo_femmine" (cfr. § 5.2).

L'assegnazione di un primo o di un successivo documento oltre che da **Project** può essere realizzata anche da **New/***Document in This Folder*. Mediante **New** è possibile selezionare non solo un nuovo documento ma anche tutti gli altri possibili elementi. Per creare un nuovo documento è possibile utilizzare anche il tasto destro del mouse cliccando stando posizionati all'interno del pannello dei documenti e scegliendo *New Document* dalla finestra che appare.

Il contenuto del documento viene visualizzato di default in basso nel pannello di lavoro. Da **View**/*Detail View*/*Right* è possibile cambiarvi collocazione e spostarlo alla destra del pannello di lavoro. Sempre da **View**/*Detail View*/*Bottom* si riporta nella posizione precedente. **Window**/*Docked* è invece il percorso per creare una finestra, un foglio a parte, sganciato dal piano di lavoro. Per ripristinare lo stato iniziale selezionare **Window**/*Undocked*.

Fig. 5.5 – Menu del tasto New

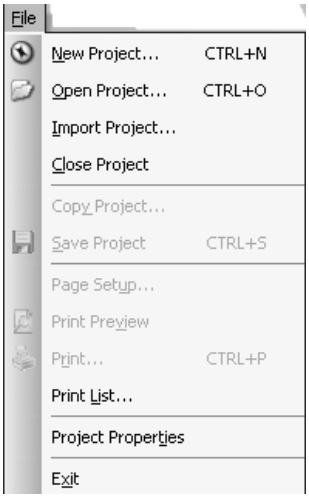


Fig. 5.6 – Menu della funzione File

Per salvare o chiudere il progetto di lavoro si procede da **File**. Mediante *Import Project* è possibile unire più progetti; questa funzione è utile per esempio qualora si lavori in gruppo.

Selezionato il progetto di lavoro con estensione *nvp* si sceglie (fig. 5.7) se unire tutti gli elementi creati (*All item*) o se selezionarne solo alcuni (*Selected item*).

La chiusura del progetto di lavoro, avviene, invece mediante **File**/*Close Project*. Il salvataggio avviene in automatico ogni 15 minuti; tale intervallo può essere modificato da **Tools**/*Options*/*General*. Sempre da **Tools** è possibile cambiare le impostazioni principali del software attinenti ai *fonts*, ai colori e agli altri aspetti legati all'editing.

A volte può rivelarsi utile pulire il progetto di lavoro dagli elementi creati. In questo caso dal menu **Tools** selezioniamo *Options*/*Clear Recent Project List*.

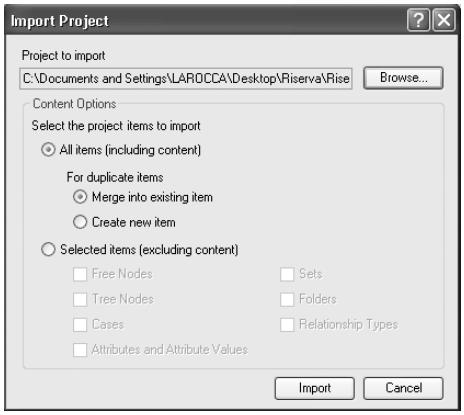


Fig. 5.7 – Finestra per importare un progetto di lavoro

5. 2. L'ORGANIZZAZIONE DEI DATI: I CASI E GLI ATTRIBUTI

Le informazioni inerenti i soggetti intervistati o titolari dei messaggi possono essere organizzate mediante il *casebook*. Il *casebook* è formato da casi e attributi. Prima di poterlo visualizzare è quindi necessario generare i casi e gli attributi. I casi possono corrispondere ai rispondenti e gli attributi alle caratteristiche degli stessi. Ogni attributo si compone di diversi valori (*values*).

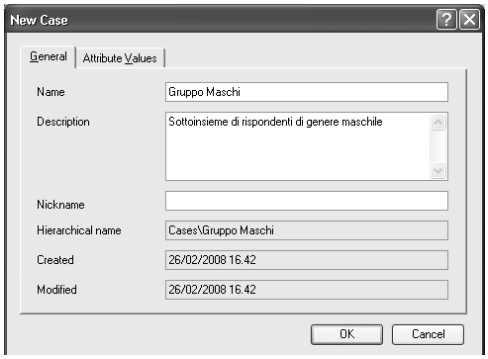


Fig. 5.8 – Finestra per creare i casi

Per creare i casi **Nodes**/*Cases*; poi dal menu **New/***Case in This Folder* si accede alla finestra di dialogo *New Case* e si attribuisce un nome ai casi, per esempio

"gruppo maschi" e "gruppo femmine". Dal pannello di navigazione ora occorre selezionare **Classifications**/*Attributes*, poi **New**/*Attribute in This Folder*; anche in questo caso si accede a una finestra di dialogo *New Attribute* dove bisogna specificare il nome dell'attributo, la descrizione e il tipo che può essere o stringa (*string*), numero (*number*) o data (*date*).

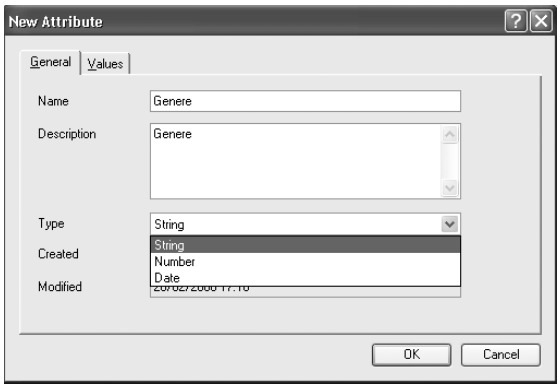


Fig. 5.9 – General - Finestra per creare gli attributi

Iniziamo introducendo l'attributo di genere, che è una stringa. Ci spostiamo al secondo livello della finestra *Values*. Per aggiungere una riga cliccare *Add*; apparsa la riga si può digitare il testo: uomo, donna. Per spostare i valori sopra o sotto bisogna utilizzare *Sort*.

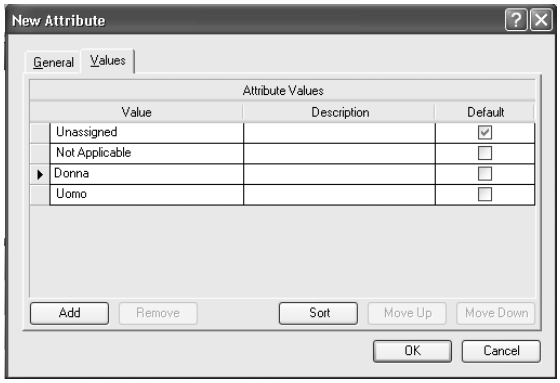


Fig. 5.10 – Values - Finestra per specificare i valori degli attributi

Ora bisogna attribuire il valore al caso. Dal pannello di navigazione **Nodes**/*Cases* evidenziamo uno dei due casi e poi o dal menu **Project** selezioniamo *Case Properties* (fig. 5.11).

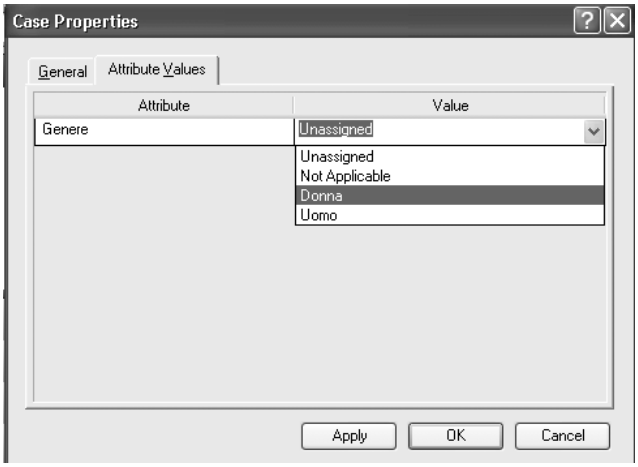


Fig. 5.11 – Attribute Values - Finestra per attribuire i valori ai casi

Il risultato dell'attribuzione dei valori ai casi è visualizzabile mediante il *casebook*. Da **Tools**/*Casebook*/*Open Casebook* il *casebook* viene visualizzato nel pannello di lavoro in basso (fig. 5.12).

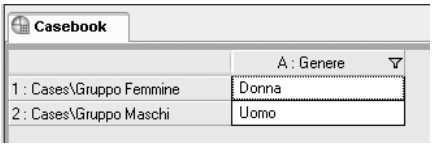


Fig. 5.12 – Dettaglio del *casebook*

Gli attributi possono essere sia importati che esportati con **Tools**/*Casebook*/*Import Casebook* o *Export Casebook*.

Se si sceglie di importarli, il foglio Excel in cui si preparano deve poi essere salvato come un foglio di testo separato da tabulazione.

Quindi con **Tools**/*Casebook*/*Import* si apre la finestra per importare i dati; è necessario specificare il percorso all'interno del quale si trova il file *.txt* e mediante *Format* si selezionano le caratteristiche del testo e del file.

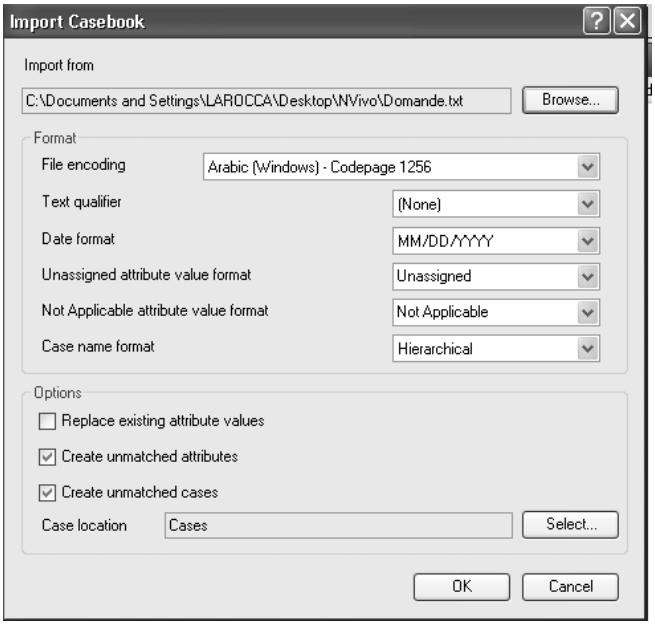


Fig. 5.13 – Finestra per l'importazione del *casebook*

Per i casi è possibile e a volte utile creare dei Set (cfr. paragrafo 5.6). Si pensi per esempio all'utilità di raggruppare i rispondenti secondo alcune specifiche caratteristiche oppure in base ai tipi di risposta forniti.

5. 3. LA BARRA DEGLI STRUMENTI

L'interfaccia di NVivo7 è simile a quella utilizzata in Outlook o Word e molte delle icone presenti sono quelle proprie di un sistema di videoscrittura.

Fig. 5.14 – Dettaglio della barra degli strumenti

Fig. 5.15 – Dettaglio della barra di formattazione

Di particolare interesse è in questa *release* l'introduzione della barra per la ricerca degli elementi.

Fig. 5.16 – Barra di ricerca

L'utilizzo è molto semplice, basta selezionare dal menù a tendina della stringa **Look for** l'oggetto da ricercare e da **Search In** l'elemento all'interno del quale sviluppare la ricerca. A questo punto cliccando su *Find Now* nel pannello sottostante alla barra vengono visualizzati i risultati.

Accanto a *Find Now* troviamo **Options** dalla quale è possibile accedere a due modalità di ricerca avanzata: *Adavanced Find* e *Grouped Find*.

Fig. 5.17 – Opzioni presenti nella barra di ricerca

Sempre mediante **Look for** (fig. 5.18) è possibile selezionare l'elemento all'interno del quale avviare la ricerca.

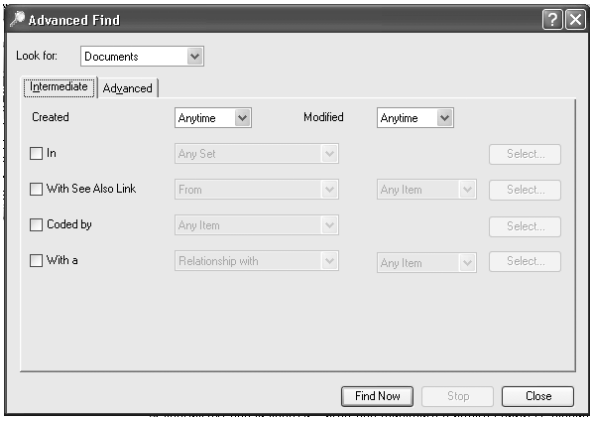


Fig. 5.18 – Finestra per la funzione avanzata di ricerca

Selezionando *Grouped Find* si avvia, invece, la ricerca all'interno dei gruppi creati. Le funzioni di ricerca del *Find Tools* sono semplici funzioni di ritrovamento dei documenti e pertanto sono da ritenersi differenti dalle query che realizzano delle vere e proprie interrogazioni sui dati.

5. 4. LA FORMATTAZIONE DEL TESTO

Le peculiarità di un sistema di videoscrittura consentono al software di avere anche funzionalità dedicate alla formattazione, che si integrano con quelle di analisi.

Molte volte ci è capitato di lavorare a un testo e aver attribuito stili diversi al titolo, sottotitolo, paragrafi e così via. I diversi livelli di stile si ritrovano anche in NVivo7. Evidenziato il testo cui applicare il formato scelto basta semplicemente selezionare il livello di stile dalla barra degli strumenti posta in alto. Tale funzionalità è molto utile qualora si voglia già distinguere mediante la formattazione, per esempio, fra le domande e le risposte contenute all'interno di un unico documento.

Attraverso l'icona **Auto Code** i livelli di stile attribuiti possono ritornare utile quali strumenti di codifica veloce (fig. 5.19).

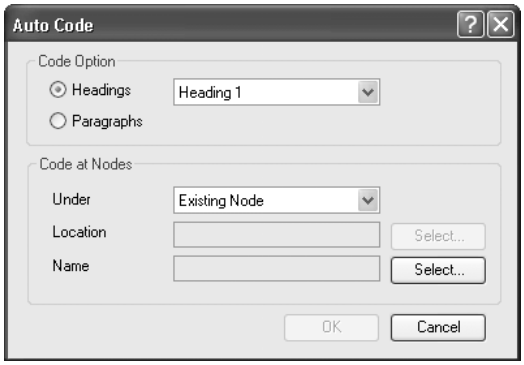


Fig. 5.19 – Finestra per l'autocodifica

Fig. 5.20. – Icone per l'autocodifica

È poi possibile anche mediante **Edit**/*Find* e **Edit**/*Replace* individuare gli elementi formattati.

5. 5. LA CREAZIONE DI NODI DI CODICI

Il lavoro di analisi sul testo è rappresentato dalla creazione di nodi di codici. Con "nodo", nella terminologia di NVivo7, si può intendere una specifica posizione occupata all'interno del database da una parte dei documenti. Lavorando sul testo la definizione dei nodi diventano le etichette da attribuire ai codici. Nodo, si può infine riferire anche agli aspetti di organizzazione dei dati, per esempio i *case nodes*, o le *relationship nodes*. In questo ultimo caso è evidente come i nodi vengano a indicare sia la chiusura del testo con una etichetta (espresso anche dal codice), ma anche la relazione fra i diversi elementi o i diversi codici attribuiti all'interno degli elementi stessi.

NVivo7 permette di creare nodi e codici secondo due diversi approcci: dall'alto o dal basso. La generazione di nodi e codici dall'alto è indipendente dal testo: i codici in questo caso sono delle categorie che provengono direttamente dal disegno della ricerca secondo il metodo deduttivo. Procedendo invece dal basso si applica il metodo proposto dalla *Grounded Theory*, che vuole che la teoria emerga dal basso.

5. 5. 1. GENERARE CODICI DALL'ALTO

Dal pannello di navigazione si seleziona **Nodes** che permette di visualizzare tutti i tipi di nodi.

Fig. 5.21 – Contenuto della cartella *Nodes* nel pannello di navigazione

- *Free Nodes*: si riferiscono ai nodi non gerarchici.
- *Tree Nodes*: i nodi ad albero sono generati stabilendo una gerarchia, per sempio creando delle categorie o sotto categorie secondo un criterio definito.
- *Cases*: i casi sono rappresentati dalle interviste, dalle persone ecc. Solo i casi hanno gli attributi, dove per attributi si intendono le caratteristiche, per esempio, socio-demografiche dei rispondenti (età, genere ecc.).
- *Relationships*: definiscono il legame creato fra i codici.
- *Matrices*: sono tavole di nodi.
- *All Nodes*: visualizza tutti i nodi creati.

La generazione dei nodi come per tutti gli elementi creabili con NVivo7 può avvenire: dal menu **Project** selezionando *Create a Tree Node*; utilizzando il tasto destro del mouse ponendoci all'interno del pannello di lavoro; oppure dal menu **New** selezionando *Tree Nodes in This Folder*. La finestra di dialogo che appare chiede di specificare il nome del nodo "La responsabilità della famiglia", di darne una descrizione e di indicare un *nickname* "respfam".

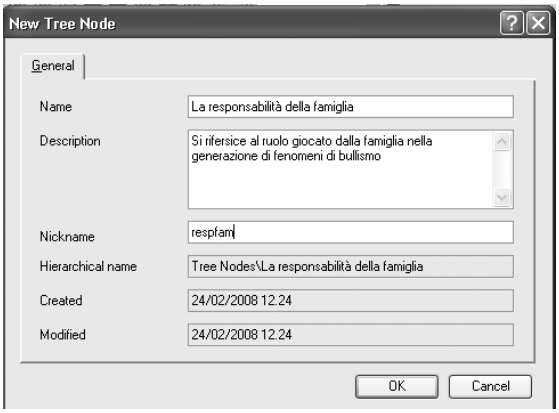


Fig. 5.22 – Finestra per la creazione di un nuovo nodo

Una volta generati, il nodo e il suo codice sono visualizzabili nel pannello di lavoro dei documenti.

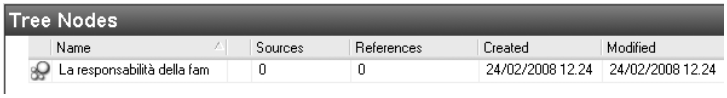


Fig. 5.23 – Presentazione di un nodo nel pannello dei documenti

Per applicare il codice creato bisogna evidenziare la porzione di testo cui si riferisce, quindi con il tasto destro del mouse da **Code** si seleziona *Code Selection at Existing Nodes*.

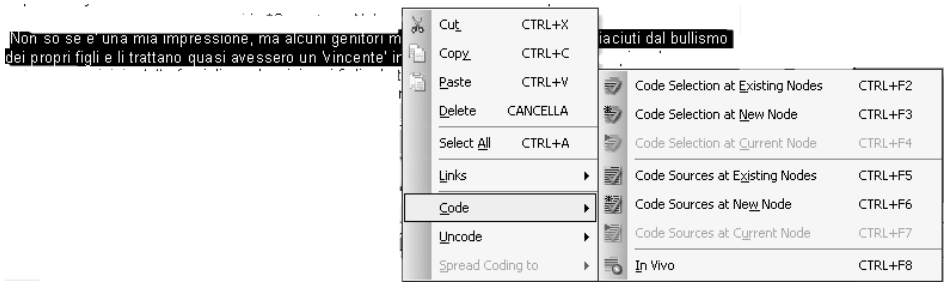


Fig. 5.24 – Procedura per la creazione di un codice

Si accede a una seconda schermata dalla quale bisogna selezionare il tipo di nodo, *Tree Nodes*, e il nome del codice "La responsabilità della famiglia"; il codice può essere anche selezionato mediante la stringa posta in basso *Select item from nickname* "respfam". A questo punto dando l'*OK* il codice viene applicato al testo.

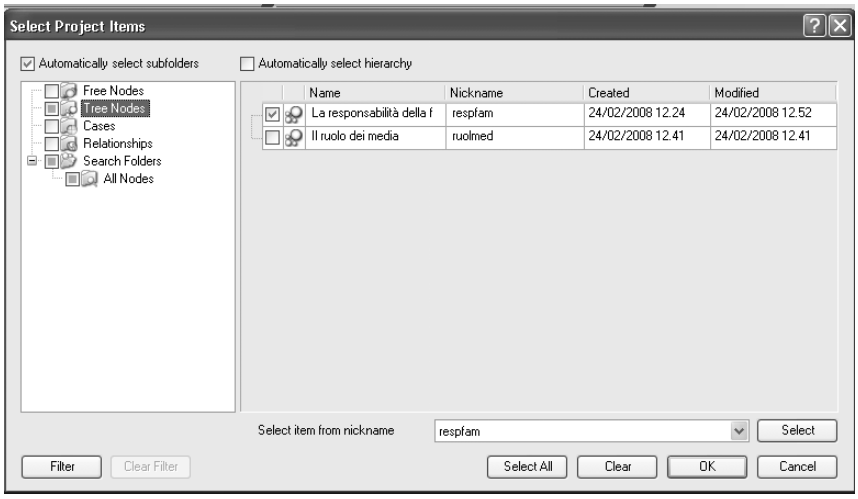


Fig. 5.25 – Finestra per la selezione degli elementi

Il codice può essere applicato anche dalla barra degli strumenti (fig. 5.26): **Code at** permette di visualizzare l'elenco dei codici o per nome (*Name*) o mediante il nickname attribuito. **In** – la terza finestra da sinistra – permette di scegliere il tipo di nodo da applicare.

Fig. 5.26 – Barra di codifica

Selezionati i parametri, l'icona successiva (**Code**; fig. 5.26) realizza l'operazione. Il codice e la porzione di testo cui si riferisce possono essere richiamati selezionando o *Tree Nodes* dal pannello di navigazione o *All Nodes*. In entrambi i casi nel pannello sottostante si visualizza il codice e la porzione di testo cui è stato applicato.

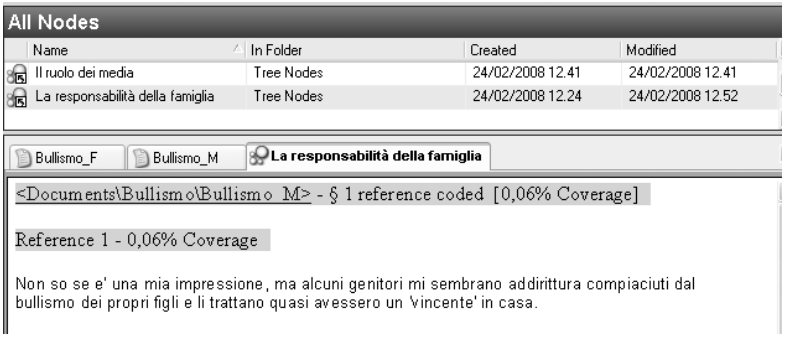


Fig. 5.27 – Presentazione del risultato per il codice "La responsabilità della famiglia"

Il contenuto del nodo può anche essere esportato in formato *doc*.

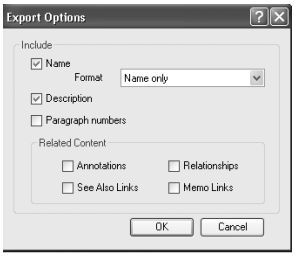
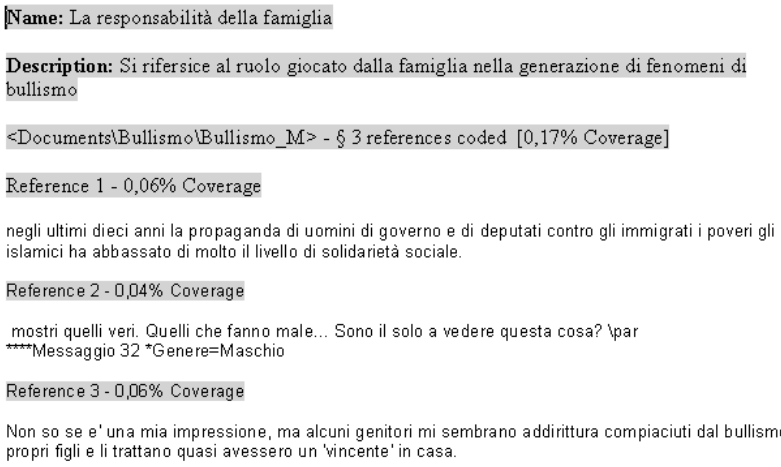


Fig. 5.28 – Opzioni per esportare un codice

Dal pannello di navigazione si seleziona **Nodes,** si sceglie il tipo di nodo e si attiva il codice selezionato "La responsabilità della famiglia"; quindi dal menu **Project** si seleziona *Export Tree Node* dalla finestra *Export Options*, si include nell'output il nome del codice e per esempio la sua descrizione; poi si dà l'*OK*.

Di default l'output viene salvato nella cartella *Documenti*; qualora si volesse salvare in un altro spazio occorre semplicemente selezionare il nuovo percorso. Il risultato dell'output si visualizza come documento Word.



Name: La responsabilità della famiglia

Description: Si rifersice al ruolo giocato dalla famiglia nella generazione di fenomeni di bullismo

<Documents\Bullismo\Bullismo_M> - § 3 references coded [0,17% Coverage]

Reference 1 - 0,06% Coverage

negli ultimi dieci anni la propaganda di uomini di governo e di deputati contro gli immigrati i poveri gli islamici ha abbassato di molto il livello di solidarietà sociale.

Reference 2 - 0,04% Coverage

mostri quelli veri. Quelli che fanno male... Sono il solo a vedere questa cosa? \par
****Messaggio 32 *Genere=Maschio

Reference 3 - 0,06% Coverage

Non so se e' una mia impressione, ma alcuni genitori mi sembrano addirittura compiaciuti dal bullismo dei propri figli e li trattano quasi avessero un 'vincente' in casa.

Fig. 5.29 – Output dell'esportazione di un nodi di codici in un documento Word

A un nodo è possibile attribuire un sotto-nodo. Tale procedura è molto semplice e prevede che si selezioni dal pannello dei documenti il nodo da cui si vuole che discenda il sotto-nodo. Per esempio a "Il ruolo dei media" si vuole aggiungere "L'importanza di Internet", al fine di specificare all'interno dei mezzi di informazione tutti quei riferimenti presenti nel testo e rivolti a sottolineare il ruolo svolto dalla Rete.

Evidenziato il nodo cui si deve agganciare il sotto-nodo, mediante il tasto destro del mouse si seleziona *Create as Node*; a questo punto occorre specificare dove andare a prendere l'elemento cui agganciare il nuovo sotto-nodo (finestra *Select Location*, cfr. fig. 5.33), si seleziona *Tree Nodes* e poi si sceglie fra i *Tree Nodes* "Il ruolo dei media". A questo punto si apre la finestra *New Tree Node* (cfr. fig. 5.22) che ci chiede di nominare e descrivere il nodo; una volta compilati tali appositi spazi, dando l'*OK* il sotto-nodo viene a visualizzarsi all'interno del nodo principale.

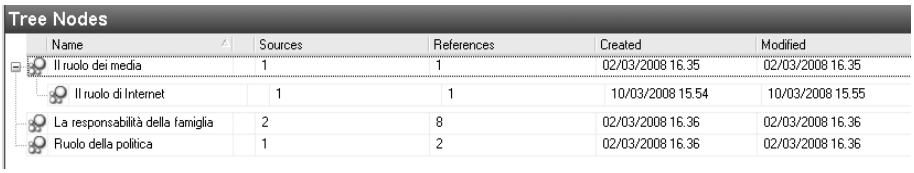


Fig. 5.30 – Nodo e sotto-nodo nel pannello dei documenti

5. 5. 2. GENERARE CODICI DAL BASSO

In un approccio dal basso, i codici vengono generati a partire dal testo: scaturiscono dalla lettura dei messaggi o dei documenti. Occorre quindi aprire i documenti uno alla volta e leggere attentamente quanto in essi contenuto. Si procederà leggendo il testo e attribuendo un codice a ciascuna porzione di testo ritenuta importante.

Fig. 5.31 – Procedura per la creazione di un codice

Per l'attribuzione del codice, evidenziato il testo da codificare si procede mediante il tasto destro del mouse scegliendo fra le diverse possibilità.

- *Code Selection at Existing Nodes*: attribuisce un codice già esistente. La procedura è la medesima di quella visualizzata in fig. 5.25.
- *Code Selection at New Node*: ci dà la possibilità di generare un nuovo codice.
- *Code Selection at Current Node*: attribuisce al testo l'ultimo codice utilizzato.
- *In Vivo*: utilizza la stessa porzione di testo selezionato come stringa di codice.

Per generare un nuovo codice da **Code** (fig. 5.31) si seleziona *Code Selection at*

New Node e si visualizza una finestra che ci chiede di attribuire un nome e una descrizione al codice. La scelta del tipo di nodo all'interno del quale andare a collocare il codice avviene mediante *Location/Select.*

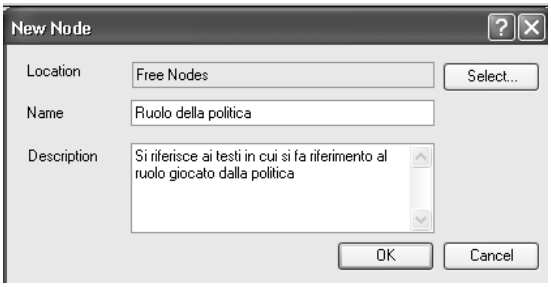


Fig. 5.32 – Finestra per nominare un nodo

In questo caso selezioniamo *Free Nodes* diamo l'*OK* e il nostro nuovo codice è stato generato.

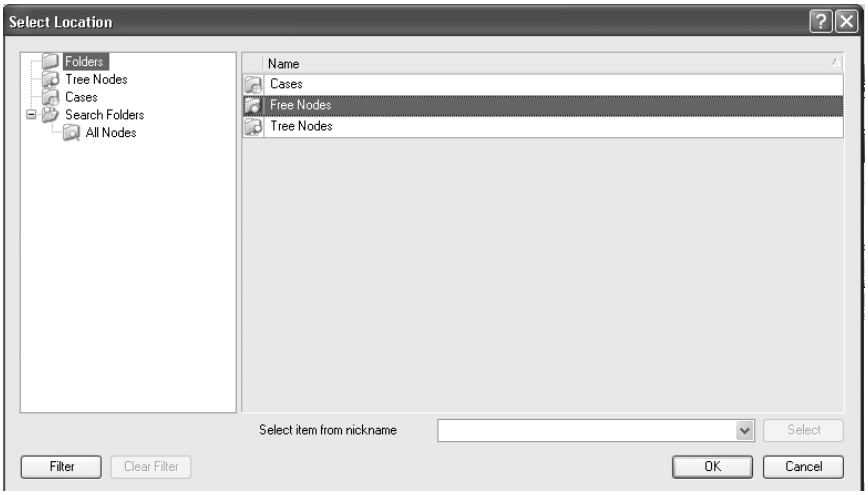


Fig. 5.33 – Finestra per selezionare dove salvare l'elemento

Sempre mediante il tasto destro è possibile eliminare i codici creati (fig. 5.34): **Uncode**/*Uncode Selection at Existing Nodes* oppure *Uncode Selection at current Node*.

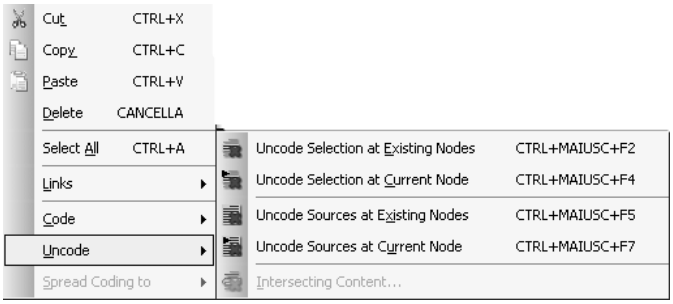


Fig. 5.34 – Finestra per eliminare la codifica da un testo

Come per la generazione dei codici si apre una finestra che ci chiede di specificare il codice da eliminare (fig. 5.35).

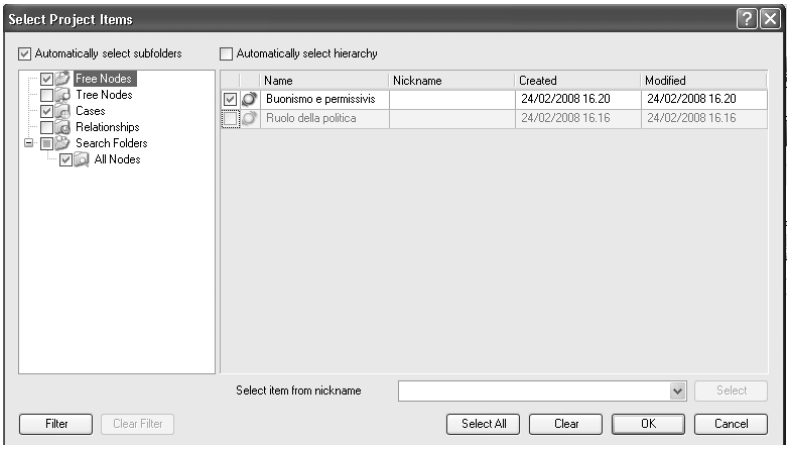


Fig. 5.35 – Finestra per la selezione degli elementi

Durante la fase di codifica può rivelarsi utile il voler controllare i codici e i nodi creati. La funzionalità di NVivo7 che permette tale visualizzazione è il *Coding Stripes*, accessibile dalla barra dei menu da **View**/*Coding Stripes* o direttamente dall'icona **Strisce di codifica** riportata nella barra degli strumenti (fig. 5.36).

Fig. 5.36 – Icona per visualizzare le strisce di codici

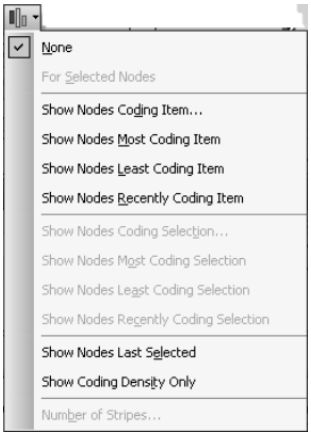


Fig. 5.37 – Menù della funzione strisce di codici

Il passo successivo (fig. 5.37) consiste nel selezionare il tipo di nodo che si vuole visualizzare. Si ricorda che di default l'opzione che appare è *None*.

A questo punto del lavoro visualizziamo quanto fino ad ora fatto sul "gruppo maschi". Il percorso da seguire prevede: **View**/*Coding Stripes*/*Show Nodes Coding Item*: appare la finestra di dialogo *Select Project Items* per selezionare i codici da visualizzare (è la medesima della fig. 5.35) e a destra in basso nel pannello di lavoro visualizziamo i codici "La responsabilità della famiglia" e "Il ruolo della politica" e l'*In vivo code* "Buonismo e permissivismo sono deleteri..." A ciascun codice è associato un colore differente.

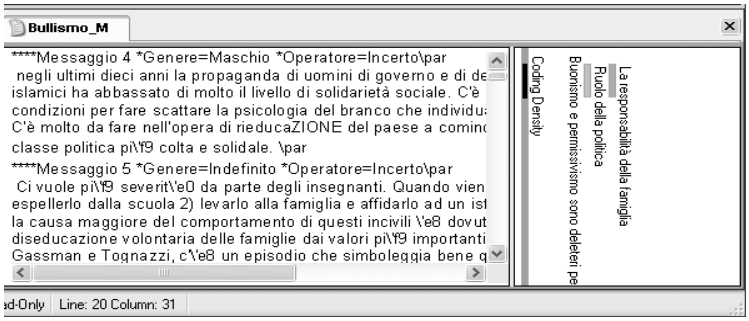


Fig. 5.38 – Dettaglio del pannello di lavoro, sulla destra la visualizzazione delle strisce di codice

Sotto i codici viene visualizzata la *Coding Density*. La densità di codifica è rappresentata attraverso una striscia grigia lunga con picchi di ombre nere. L'intera striscia della densità rappresenta tutto il documento; le parti più scure indicano invece i frammenti del testo che sono stati codificati. Posizionandoci con il cursore del mouse sulla parte scura della striscia di densità essa mostra le codifiche effettuate in quella parte di testo.

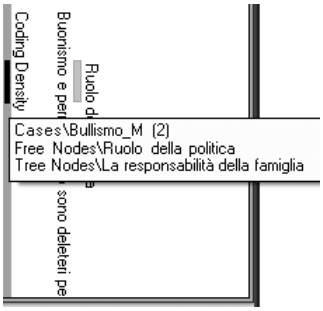


Fig. 5.39 – Dettaglio delle strisce di codice

Qualora si volesse solo visualizzare la striscia di densità escludendo i codici, dal menu **View** selezionare *Coding Stripes/Show Coding Density Only*.

Il numero massimo di strisce per codici che appaiano è di sette. Questo può essere modificato da **View/***Coding Stripes/Number of Stripes* (fig. 5.40). La modifica viene applicata solo al documento corrente.

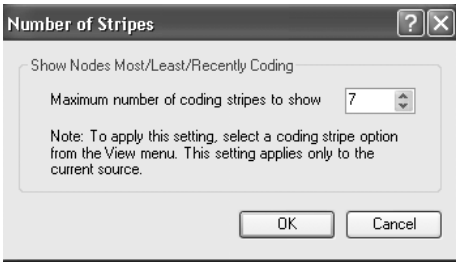


Fig. 5.40 – Finestra per aumentare o diminuire il numero delle strisce di codice visualizzabili

I risultati della codifica e il conseguente output con le strisce di codifica, che possiamo chiamare **tracciato di codifica**, possono essere stampati. Dal menu **File** selezioniamo *Print* e dalla finestra di dialogo *Print Options* (fig. 5.41) spuntiamo *Coding Stripes* e quindi diamo l'*OK*.

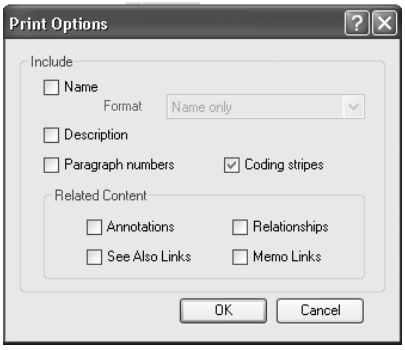


Fig. 5.41 – Finestra per attivare le opzioni di stampa

5. 6. RI-ORGANIZZARE CODICI E NODI

Nel lavoro di analisi ci si potrebbe accorgere a posteriori di aver codificato o raggruppato alcuni elementi all'interno di una risorsa o sotto un codice non pertinente. Nel lavoro di riesame è possibile ovviare a questi inconvenienti spostando, eliminando o raggruppando i codici e i nodi.

Per avere una rapida panoramica del lavoro effettuato dal pannello di navigazione **Nodes**/*All Nodes* si ha la possibilità di controllare il tipo di nodi generati.

Fig. 5.42 – Pannello dei documenti con i nodi generati

Desta curiosità nell'analisi condotta fino a questo momento "La responsabilità della famiglia". Mediante **Nodes**/*Tree Nodes* si visualizzano i nodi presenti in questa cartella e già dal pannello dei documenti ci si accorge che, al momento, all'interno di tale nodo sono stati codificati sei elementi (*References*); sotto **Sources** appare 1 che sta ad indicare che al momento questo nodo è stato attribuito solo a uno dei due documenti inseriti, ovvero al "gruppo_maschi".

Selezionando il codice esso appare con il suo contenuto nel pannello in basso riportando il nome del documento *Bullismo_M*, il numero complessivo di elementi codificati al suo interno (6) e la percentuale di copertura nel testo

(in questo caso pari allo 0,31%). Tale percentuale si può intendere come somma delle singole percentuali attribuite ai paragrafi codificati. Il valore della singola percentuale deriva dall'estensione della porzione di testo codificata: più è lungo il testo codificato più alto sarà il valore della percentuale.

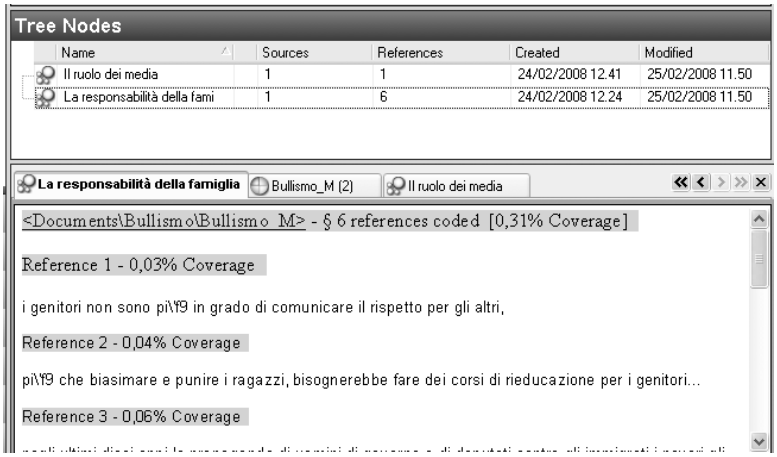


Fig. 5.43 – Pannello dei documenti e di lavoro per i nodi generati

Nella lettura di queste citazioni ci si potrebbe accorgere che alcune dovrebbero essere raggruppate sotto altri nodi. Si procede evidenziando il riferimento da spostare e mediante il tasto destro del mouse con **Code**/*Code Selection at Existing Nodes* si accede alla finestra *Select Project Items* e, selezionando il nuovo codice, si sposta sotto il nodo ritenuto più appropriato.

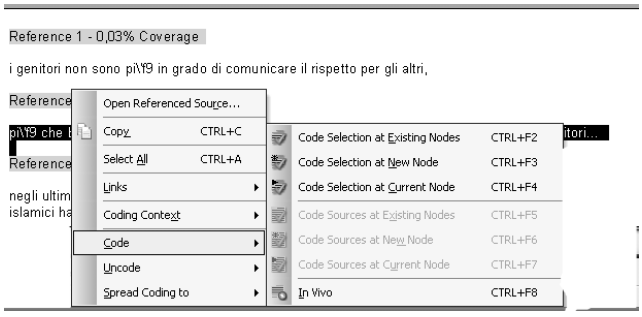


Fig. 5.44 – Dettaglio della procedura di codifica

Nel lavoro di controllo e di riallocazione dei codici occorre prestare attenzione a non selezionare da **Code**/*Code Sources at Existing Nodes* (fig. 5.44) poiché in questo caso si sposterebbe il codice sotto l'altro gruppo o documento.

A volte può anche essere necessario spostare un intero nodo sotto un altro tipo. Per esempio, potrebbe essere utile ai fine dell'analisi ricondurre tutti i nodi sotto il medesimo tipo, quindi spostare i *Tree Nodes* sotto *Free Nodes*. Evidenziamo l'elemento da spostare, in questo caso "Il ruolo dei media", e mediante l'icona "taglia" dalla barra degli strumenti o mediante il tasto destro del mouse, tagliamo l'elemento.

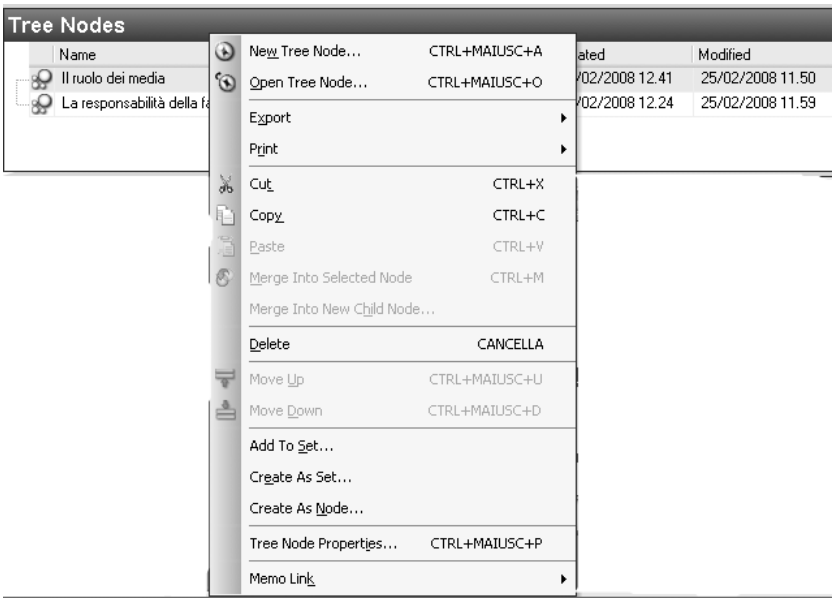


Fig. 5.45 – Finestra attivata mediante il tasto destro del mouse

Da **Nodes** si seleziona *Free Nodes* (fig. 5.45) e sempre mediante tasto destro si seleziona *Paste*; ci viene chiesto di confermare la trasformazione del nodo che viene quindi ora salvato come un *Free Nodes*.

Nel lavoro di ricodifica ci si potrebbe accorgere che nodi con denominazioni simili possono essere uniti in un unico nodo. Tale operazione è realizzabile mediante *Merge Into Selected Node* (fig. 5.45).

Durante la lettura del testo è stato creato un primo *In Vivo code* che aveva come tema il buonismo e il permissivismo delle famiglie dei bulli; a seguito

della lettura del testo si ritiene congrua l'operazione di annessione all'interno del nodo "La responsabilità della famiglia". Entrambi i nodi si trovano già nella cartella *Free Nodes*.

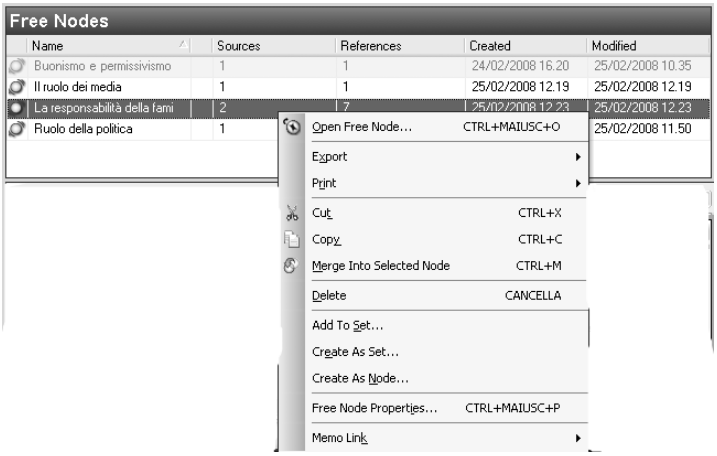


Fig. 5.46 – Finestra attivata mediante il tasto destro del mouse

Dal tasto destro del mouse si evidenzia il nodo da accorpare, si taglia (*Cut*), si seleziona il nodo all'interno del quale andrà introdotto e, sempre mediante il tasto destro, si sceglie *Merge Into Selected Node* (fig. 5.46). La finestra *Merge Into Node* (fig. 5.46) ci consente di specificare tutti gli elementi associati al precedente nodo che si vuole vengano importati assieme al nodo stesso.

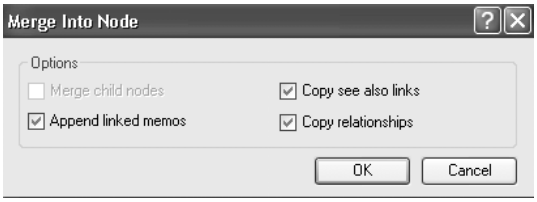


Fig. 5.47 – Finestra per il *merge* dei codici

A seconda dell'estensione dei documenti e del loro contenuto potrebbe rivelarsi utile raggruppare i nodi sotto dei macro elementi. Tale funzionalità è permessa da **Sets,** facilmente accessibile dal pannello di navigazione. I set sono da considerarsi come una selezione o raggruppamento di nodi rilevanti, de-

finiti secondo uno specifico criterio caro all'analista. La loro creazione è importante per le analisi successive, ovvero si rivela utile nel momento in cui si passa alla creazione di modelli. Infatti, essa rappresenta una prima scrematura di elementi utili nel percorso di ricerca.

Centrale nella lettura dei messaggi sul bullismo appare il ruolo giocato dalla famiglia come centro educativo e anti-educativo. Si potrebbe ipotizzare che essa rappresenti un in-group per la responsabilità e che da contorno le facciano gli altri istituti delegati all'educazione, come la scuola, o alla trasmissione di valori, ad esempio la politica.

Può essere utile raggruppare tutti gli elementi esterni alla famiglia e indicati come importanti nell'educazione dei ragazzi in un unico set che chiamiamo "Responsabilità esterne".

Per creare il set dal menu **New** si seleziona *Set in This Folder*, si indica il nome del set e se ne fornisce una descrizione, sempre all'interno della finestra *New Set*.

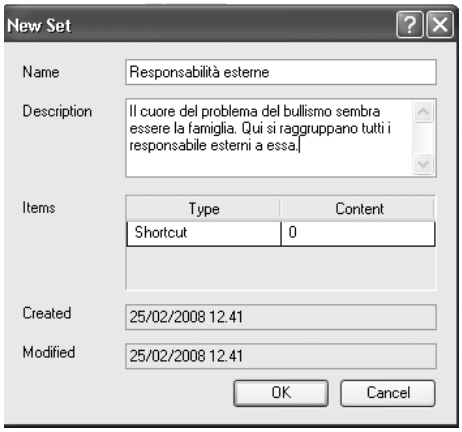


Fig. 5.48 – Finestra per nominare un set

Creato il set occorre scegliere gli elementi da introdurvi. Dal pannello di navigazione **Nodes**/*Free Nodes* i nodi creati vengono visualizzati nel pannello di lavoro; si selezionano quelli da introdurre all'interno del set. Cliccando sul nodo selezionato con il tasto destro del mouse appare il menu a discesa (fig. 5.49). Qui si seleziona *Add To Set* e si apre la finestra di dialogo *Select Set* (fig. 5.50) in cui **si** seleziona il set "Responsabilità esterne"; infine si chiude con *OK*.

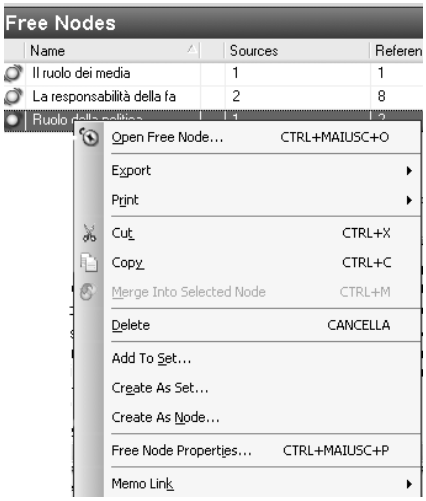


Fig. 5.49 – Finestra attivata dal tasto destro del mouse per creare un nuovo set o aggiungere un nodo a un set già esistente

Fig. 5.50 – Finestra per selezionare i set

Si ripete l'operazione tante volte quanti sono i nodi che si vogliono importare all'interno del set.

Alla fine dell'operazione i nodi sono visualizzabili all'interno del set "Responsabilità esterne" appena creato (fig. 5.41).

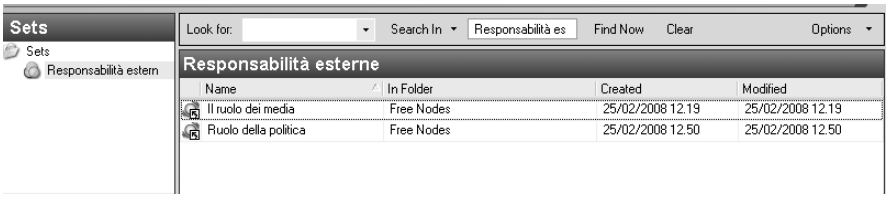


Fig. 5.51 – Dettaglio per il set Responsabilità esterne

5. 7. I RAPPORTI DI LAVORO

Ai risultati del lavoro svolto si accede mediante **Tools**/*Reports* scegliendo tra le varie opzioni i rapporti da visualizzare:

- *Project Summary*: riepiloga tutti gli elementi creati.
- *Source Summary*: fornisce un riepilogo dei documenti contenuti per ogni fonte, qui per fonte si intende il documento. Quindi le nostre fonti saranno *Bullismo_M* e *Bullismo_F*.
- *Node Summary*: riepiloga tutti i nodi e i codici generati.
- *Relationship Summary*: riepiloga le relazioni attribuite specificandone il tipo e gli elementi posti in connessione.
- *Attribute Summary*: riassume le modalità attribuite ai casi, nel nostro esempio sui file del bullismo ne abbiamo solo due e sono uomo e donna.
- *Coding Summary*: riepiloga tutte le operazioni in cui si è lavorato sui codici e sui nodi, si trovano quindi raggruppati per documento i tipi di nodi generati (*Free* o *Tree*), i risultati delle query sui codici e vengono indicati il numero di riferimenti all'interno di essi codificati.
- *Coding Comparison*: fornisce un confronto fra i documenti per i codici creati.

Coding Comparison Report

Project: Bullismo
Generated: 03/03/2008 10.04

Source A	Bullismo_F	Document
Source B	Bullismo_M	Document

	Source A	**Source B**
Coding Coverage	20.22 %	7.78 %
Overlapping References	16	21
Non-Overlapping References	5	122
Difference in Coding	11	101

Fig. 5.52 – Output del rapporto

Ogni file di rapporto è dotato di una sua barra degli strumenti che permette di esportare il testo, di stamparlo o di scorrerlo.

Fig. 5.53 – Barra degli strumenti presente nel file dei rapporti di lavoro

5. 8. CREARE ELEMENTI DI LAVORO AGGIUNTIVI

Il lavoro sul testo può richiedere che vengano introdotti elementi o che si creino, mediante scrittura, documenti specifici; fra questi vi sono anche i *memos*.

Per crearne uno, selezionare da **Sources**/*Memos*; è importante ricordare che tutte le volte che si opta per la creazione di un diverso elemento, questo deve essere prima attivato selezionandolo dal pannello di navigazione posto a sinistra dell'interfaccia di lavoro del progetto. Da **New**/*Memo in This Folder*, anche per i *memos* appare la finestra che ci chiede di attribuire un nome allo stesso (in questo caso "Elementi repressivi") e di darne una descrizione. Dalla lettura dei messaggi postati ci si accorge di come "Accanto a una visione del fenomeno del bullismo che ne attribuisce la responsabilità ora alla scuola ora ai genitori emergono proposte repressive nei confronti di chi manifesta tali atteggiamenti"; per questo motivo è stato creato tale *memo*.

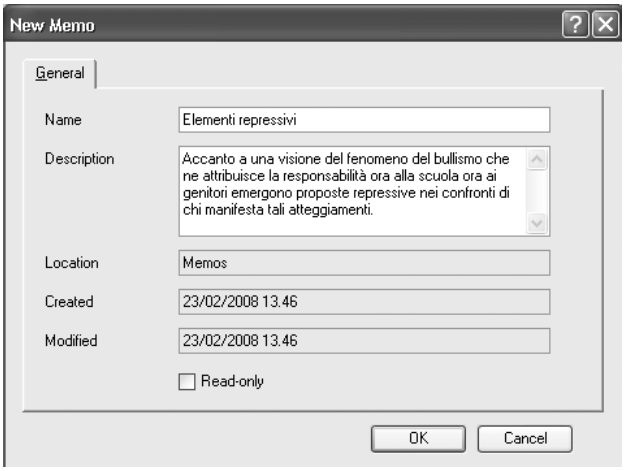


Fig. 5.54 – General - Finestra per nominare un memo

La stessa procedura si può avviare per introdurre delle risorse esterne. Da **Sources** si attiva *External*, quindi **New**/*External in This Folder*. Anche per le risorse esterne appare la medesima finestra che ci chiede di nominare e descrivere la risorsa, ma accanto a questa schermata (*General*) se ne visualizza una seconda (*External*).

La seconda schermata ci chiede di specificare il tipo di risorsa che si intende introdurre; essa può essere: *File Link*, *Web Link*, *Other*. Supponiamo di voler introdurre un manuale sul software, evidentemente non ha relazione con

il bullismo ma potrebbe rivelarsi utile per lavorare sui file. Il tipo del documento sarà *Other*; si importa il percorso da *Browse* e si seleziona il tipo di contenuto a scelta fra: *Audio*, *Image*, *Printed Document*, *Video*.

Sappiamo che tale manuale è diviso in più capitoli; come unità si sceglierà quindi *Chapter* (le altre opzioni sono: *Page*, *Paragraph*, *Section*, *Sentence*, *Verse*). *Start range* sarà dato dal primo capitolo ed *End range* dall'ultimo.

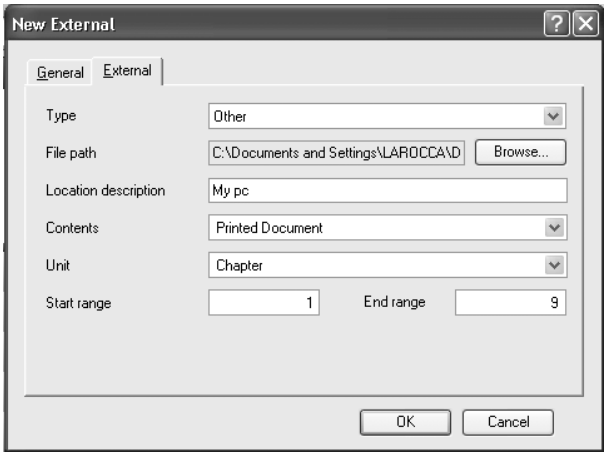


Fig. 5.55 – External - Finestra per specificare le caratteristiche del documento da importare

Dal pannello di lavoro è poi visualizzabile il contenuto della risorsa esterna.

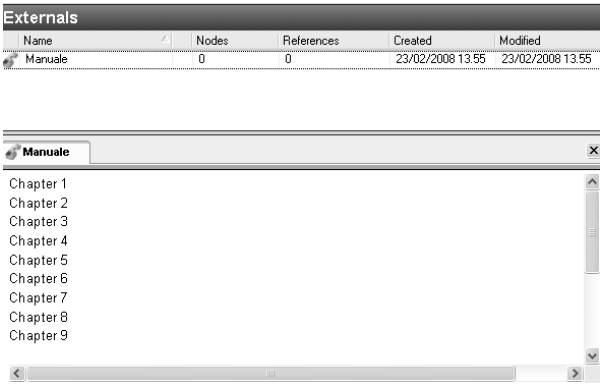


Fig. 5.56 – Dettaglio della risorsa importata

La possibilità di tenere traccia dei propri pensieri è garantita anche durante il lavoro. Evidenziando il testo che suggerisce una riflessione, mediante il tasto destro del mouse si apre una finestra, si seleziona **Links**/*Annotation*/*New Annotation* e appare una finestra all'interno della quale digitare il contenuto della nostra riflessione.

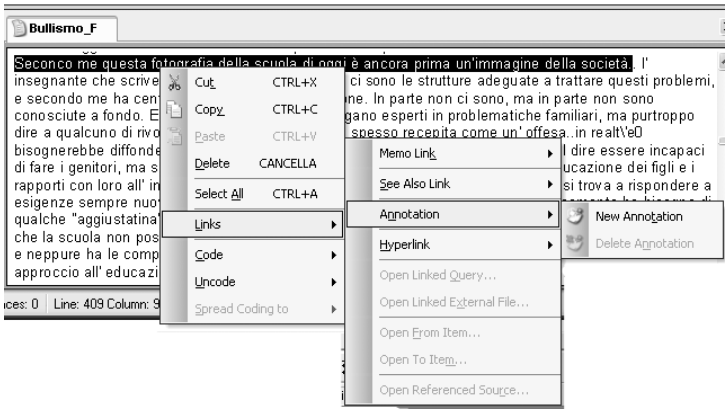


Fig. 5.57 – Percorso per creare un'annotazione

Le annotazioni si usano solitamente per piccole riflessioni; invece si ricorre ai *memos* per appunti di più estese dimensioni.

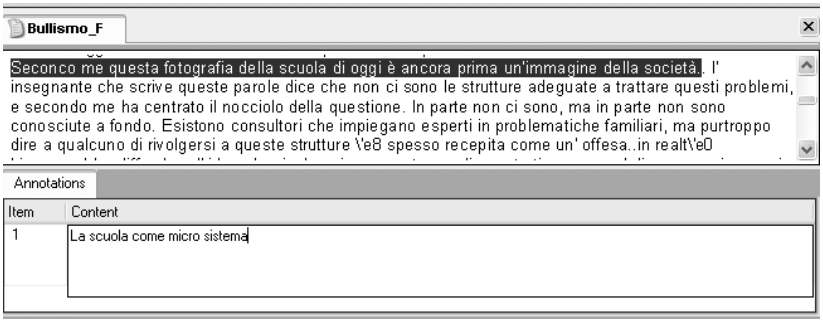


Fig. 5.58 – Finestra per digitare il contenuto dell'annotazione

Quanto scritto viene salvato automaticamente. Per visualizzare tutte le annotazioni basta selezionare **Folders** dal pannello di navigazione, quindi *Annotations*: nel primo pannello si visualizza l'elenco di tutte le annotazione realizzate; sele-

zionandone una dall'elenco appare per primo il riferimento nel testo cui si riferisce e a seguire il commento predisposto.

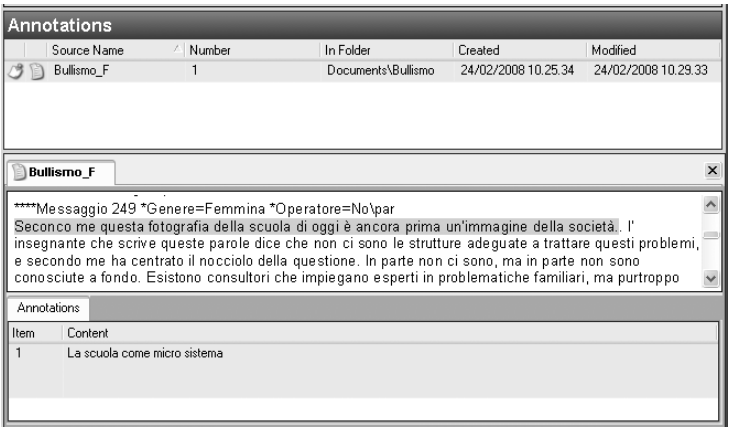


Fig. 5.59 – Visualizzazione del testo cui si riferisce l'annotazione

Per stampare il contenuto delle annotazioni, dalla barra degli strumenti selezionare l'icona con la stampante **Print** quindi scegliere l'elemento da stampare (*Related Content*).

6.
LAVORARE CON NVIVO7: INTERROGARE E RAPPRESENTARE IL TESTO

La codifica del testo rappresenta la prima operazione da realizzare lavorando con NVivo7. Immediatamente dopo si presenta la necessità di lavorare sui codici creati, di raggrupparli e di interrogare il testo mediante il lavoro di codifica svolto. A tale scopo rispondono le query.

Le query sono delle interrogazione rivolte direttamente al testo o ai codici, pertanto differiscono dalla semplice opzione di ricerca attivabile dalla barra degli strumenti. L'esplicitazione del proprio lavoro è affidata ai modelli che rappresentano una mappa dei concetti rinvenuti o del lavoro svolto.

6. 1. LE QUERY

Le query sono uno dei modi disponibili in NVivo7 per interrogare i dati.
Dal pannello di navigazione si seleziona **Queries** poi è possibile o procedere da **Project**/*New Query* o mediante il tasto destro del mouse, scegliendo sempre *New Query* o da **New** selezionando il tipo di query.

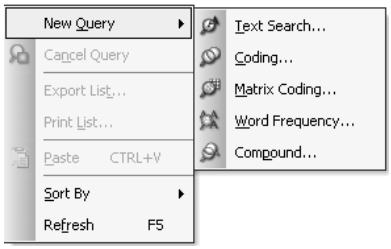


Fig. 6.1 – Creazione della query mediante il tasto destro del mouse

Mediante *New Query* si accede a vari tipi di interrogazioni:
- testuali (*Text Search Query*);
- di codici (*Coding Query*);
- di matrici di codici (*Matrix Coding Query*);
- creazione di vocabolari (*Word Frequency Query*):
- combinazione di query/query multiple (*Compound Query*).

6. 1. 1. LA CREAZIONE DEL VOCABOLARIO

La creazione del vocabolario si ha mediante l'opzione *Word Frequency Query*. Si apre una finestra dedicata (fig. 6.2) che ci chiede di specificare i criteri della query (*Word Frequency Criteria*).

L'estrazione del vocabolario può avvenire sia sui testi che sulle annotazioni prodotte, oppure su entrambi. Creiamo il vocabolario per i due gruppi sui quali si sta lavorando. Da *Search in* scegliamo *Text*, quindi selezioniamo gli elementi sui quali lavorare mediante *Of/Selected Items/Select* e apriamo la finestra di dialogo (fig. 6.3). Qui si spuntano i due testi di lavoro.

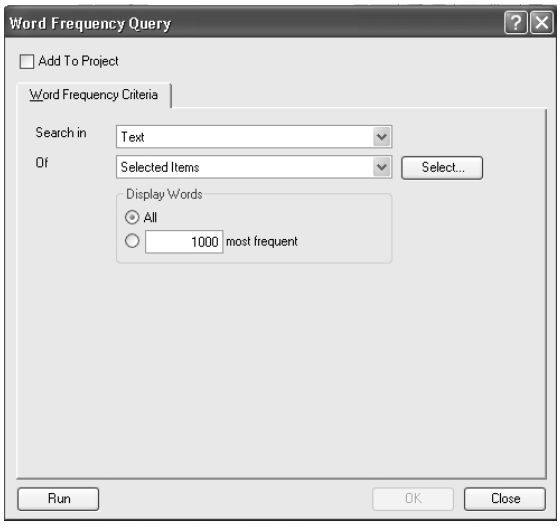


Fig. 6.2 – Finestra per l'estrazione del vocabolario

Dalla finestra *Select Project Items* – che è quella che ci permette di scegliere gli elementi su cui applicare la query – è possibile applicare ulteriori filtri alla ricerca

(fig. 6.3). Da *Filter* si passa alla successiva finestra *Advanced Find* e si specificano gli ulteriori criteri di ricerca, quali per esempio i documenti o la data di creazione degli stessi e quanto può essere utile ad affinare l'interrogazione. Con l'obiettivo di voler creare un dizionario per i due gruppi non occorre specificare ulteriori elementi; chiudiamo quindi la finestra e lanciamo la query con il pulsante *Run* (fig. 6.2).

È importante ricordarsi che per salvare il risultato della query è necessario spuntare *Add To Project* nella cella posta in alto a sinistra (fig. 6.2). Tale operazione è necessaria per ogni tipo di query.

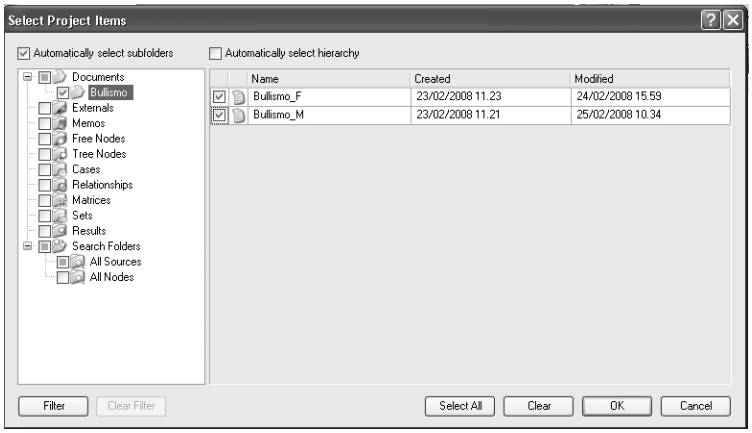


Fig. 6.3 – Finestra per la selezione degli elementi

Accanto a ciascuna parola è indicata la frequenza in valori assoluti (occorrenze).

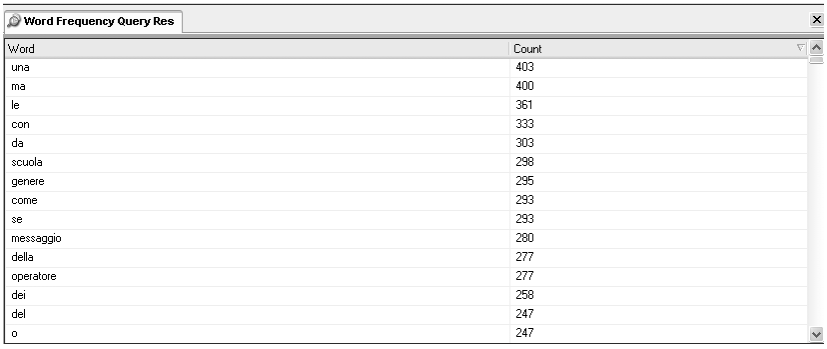


Fig. 6.4 – Visualizzazione del vocabolario nel pannello di lavoro

La lista può essere esportata in un file Excel o di testo, utilizzando il tasto destro del mouse e selezionando *Export List*.

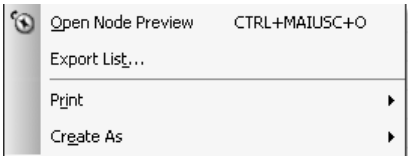


Fig. 6.5 – Finestra per visualizzare i dettagli di una parola selezionata

Dall'elenco di parole visualizzato nel pannello di lavoro non si hanno informazioni sulla presenza delle parole nei due testi selezionati. La scuola sembra essere un elemento presente in entrambi i gruppi. Per capire come questa parola, che in totale compare 298 volte, si distribuisca nei due diversi testi, sempre dal tasto destro del mouse selezioniamo *Open Node Preview*. In questo modo si vedrà che "scuola" compare 62 volte nel "gruppo_femmine" e 236 nel "gruppo uomini".

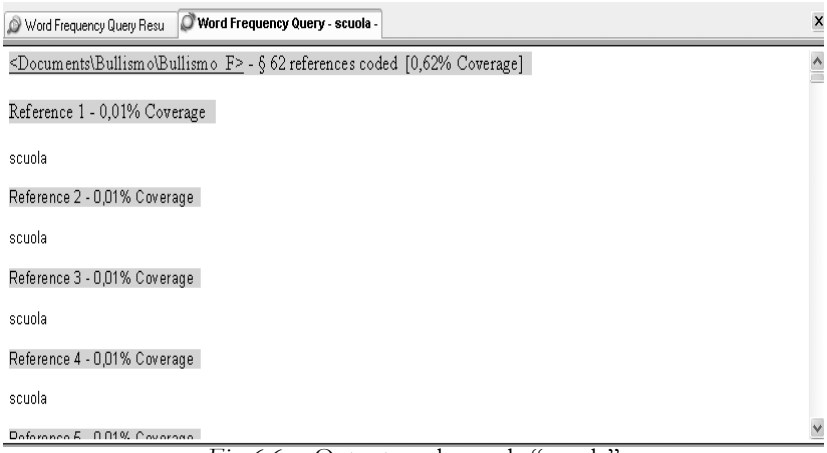


Fig. 6.6 – Output per la parola "scuola"

Se fossimo interessati a visualizzare dove nell'intero testo delle donne si trova la parola "scuola", evidenziando tale parola con il tasto destro del mouse si seleziona *Open Referenced Source* ed appare il testo "gruppo_femmine" con evidenziata la parola "scuola".

Sempre lavorando sulla singola parola (*Word Frequency Query – scuola -*), frammento per frammento è possibile individuare cosa si trova nell'intorno della parola "scuola". Per esempio selezioniamo il riferimento 4: *Coding Context/Number of Words* e specifichiamo il numero di parole da includere prima e dopo il termine della ricerca "scuola" (fig. 6.8).

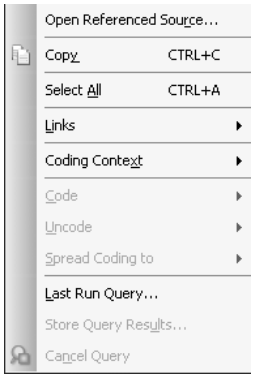


Fig. 6.7 – Finestra attivabile dal tasto destro del mouse per esplorare la parola

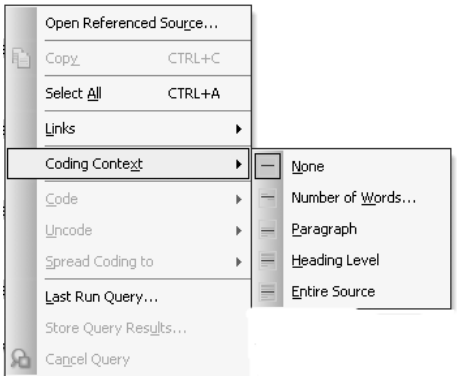


Fig. 6.8 – Dettaglio delle funzioni di *Coding Context*

Da *Coding Content* è possibile accedere a ulteriori funzioni.

- *Coding Context/Paragraph*: recupera per i singoli riferimenti l'intero paragrafo.
- *Coding Context/Heading Level*: recupera, se formattato il testo come indicato nel § 5.4, il livello superiore di codifica.
- *Coding Context/Entire Source*: recupera tutto il documento all'interno del quale la parola selezionata è contenuta.

6. 1. 2. QUERY TESTUALI

La query testuale permette di estrarre porzioni di testo ritenute significative. Si può effettuare per esempio un ricerca per la radice 'bull+' volendo estrarre tutte quelle parole che cominciano con tale parte, quali "bullo/i", "bulla/e", "bullismo" ecc. Si può realizzare anche una ricerca per gruppi di parole che ruotano intorno a un medesimo universo, per esempio: "scuola", "insegnante", "studenti" e così via.

Per realizzare questo tipo di interrogazione si ricorre alla *Text Search*

Query, cui si accede secondo le stesse modalità indicate per la creazione di un vocabolario. Dal pannello di navigazione si seleziona **Queries** e poi, con il tasto destro del mouse dall'interno del pannello dei documenti, si seleziona *New Query/Text Search*.

Dai testi "gruppo_maschi" e "gruppo_femmine" si vogliono estrarre i riferimenti in cui si discute o di scuola o di famiglia. In *Search for* si digita la query "Scuola OR famiglia" e da *Special* si scelgono i criteri della query (fig. 6.9).

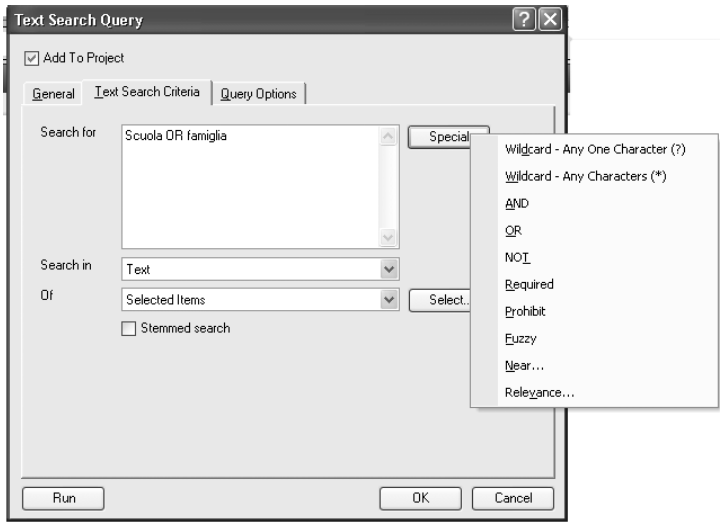


Fig. 6.9 – Text Search Criteria - Finestra per avviare una query testuale

Da *Search In* si seleziona il testo e da *Of /Selected Items/Select* mediante la finestra *Select Project Items* si selezionano i documenti.

Dalla terza schermata disponibile, *Query Options*, si selezionano le modalità in cui si vuol salvare la query (fig. 6.10).

Se si desidera che il risultato della query diventi un nodo: *Option/Create Results as New Node*; salvato nei risultati: *Location/Results* con il suo nome "Scuola OR famiglia".

Sarebbe utile anche visualizzare l'intorno della parola: *Spread to/Surrounding Paragraph*.

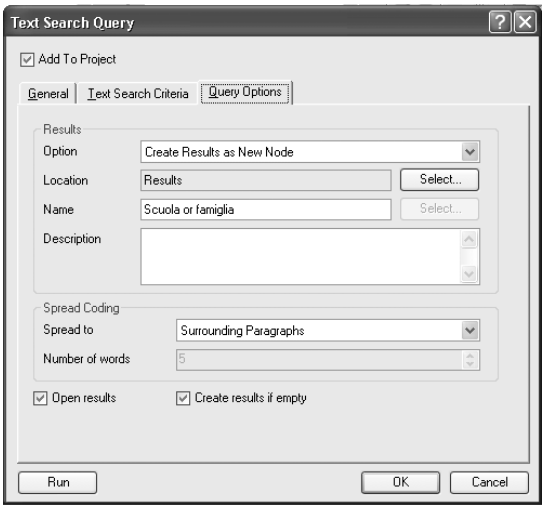
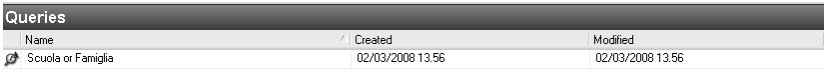


Fig. 6.10 – Query Options - Finestra per specificare le opzioni della query testuale



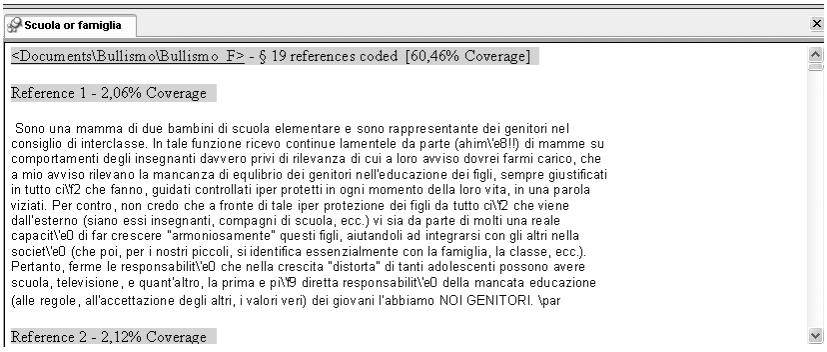


Fig. 6.11 – Risultato della query testuale

È possibile agganciare il risultato della query a un nodo. Nell'opera di codifica è stato creato il nodo "La responsabilità della famiglia". Unire i testi contenenti parole riferite alla scuola o alla famiglia con il nodo potrebbe rivelarsi utile. Da **Queries**/*Results* visualizziamo la query "Scuola OR famiglia"; mediante tasto destro copiamo l'elemento e dal pannello di navigazione attiviamo **Nodes**/*Free Nodes*. Visualizzato l'elenco dei nodi dal pannello dei documenti si seleziona quello scelto e sempre mediante tasto destro selezioniamo *Merge Into Selected Node*.

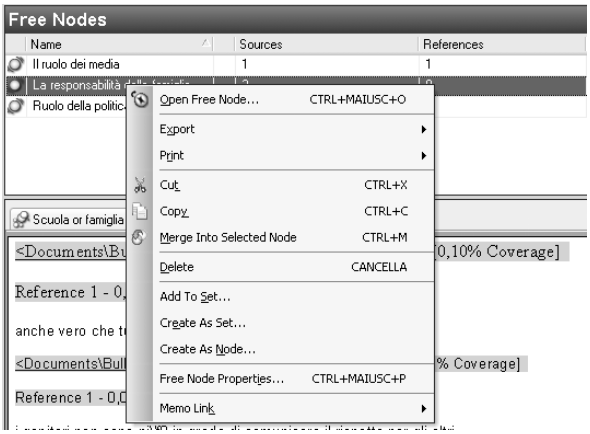


Fig. 6.12 – Finestra per agganciare la query a un nodo

A volte può rivelarsi necessario lanciare la query senza salvarla. Per esempio quando si vuole testare la bontà dell'interrogazione. In tal caso dalla finestra di dialogo *Text Search Query* è possibile selezionare l'opzione *Preview Only* e lanciarla. A questo punto saggiatane la bontà, per rilanciarla e salvarla senza ripetere tutta la procedura da **Tools** si seleziona *Query*/*Last Run Query*.

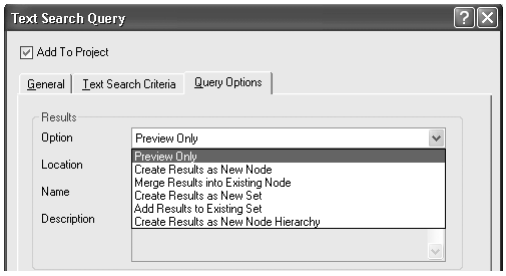


Fig. 6.13 – Dettaglio per lanciare come anteprima il risultato della query

6. 1. 3. QUERY DI CODICI

L'interrogazione mediante codici può essere semplice (*Simple*) o avanzata (*Advanced*). Attivando come modalità di lavoro le **Queries** dal pannello di navigazione da **New**/*Coding Query in This Folder* si accede alla finestra di dialogo.
Ognuna di queste finestre dal momento in cui si spunta *Add to Project* genera una prima schermata *General* in cui chiede di indicare il nome della query e di darne un descrizione.
Da *Simple* si può selezionare o non selezionare un nodo sul quale applicare la query semplice. Per esempio se i nostri i testi non fossero già stati raggruppati per genere si sarebbe potuto scegliere di lanciare la query su tutti i commenti lasciati dagli uomini, ma i nostri sono già raccolti in un unico testo. Sembra quindi più opportuno selezionare un solo nodo. Da **Node**/*Select* si sceglie *Free Nodes*/*La responsabilità della famiglia*; a questo punto da *Attribute*/*Gender* si seleziona "Uomo" e si lancia mediante *Run* questa query semplice.

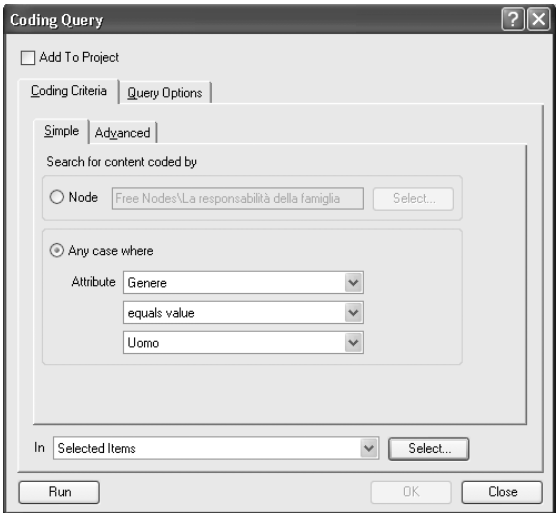


Fig. 6.14 – Finestra per avviare una query di codici

I frammenti di testo estratti si vanno a visualizzare nel pannello di lavoro.

Per lanciare una query avanzata sui codici è necessario spostarsi su *Advanced* e definirne i parametri. Per primo da *Define more criteria* si sceglie *Any Case Where*/*Select* e la modalità dell'attributo desiderata: per noi il genere "maschio".

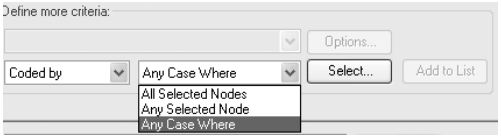


Fig. 6.15 – Procedura per specificare i casi da sottoporre a query

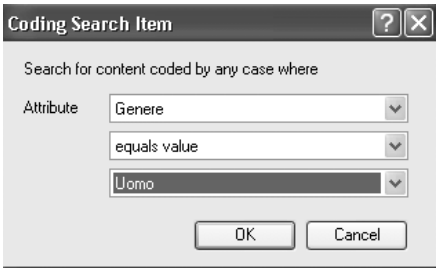


Fig. 6.16 – Procedura di selezione dei casi

Questo primo criterio va memorizzato; *Add to List* (fig. 6.15) è il tasto che ci permette di selezionarlo e andare avanti.

A questo punto scegliamo i nodi su cui lavorare: *All Seleceted Node/Select* e si spuntano nella *Select Project Items* i tre nodi scelti; anche qui *Add to List*. La sintassi della query si viene a visualizzare nel riquadro *Search for content matching these criteria*. Non resta che lanciarla con il tasto *Run* (fig. 6.17).

Fig. 6.17 – Visualizzazione della procedura di query

Il risultato del lavoro svolto si può salvare anche mediante **Tools**/*Query*/*Store Query Results*.

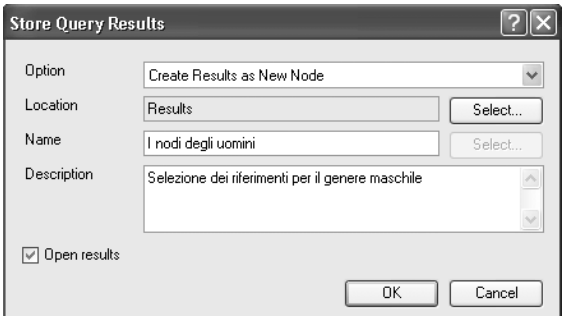


Fig. 6.18 – Finestra per salvare il risultato della query come un nodo

6. 1. 4. QUERY COMPOSTE

Le query composte nascono dall'unione della query testuale e della query di codici. Attiviamo le query e accediamo alla relativa finestra di dialogo. Per la sua creazione: **New**/*Coumponed Query inThis Folder*.

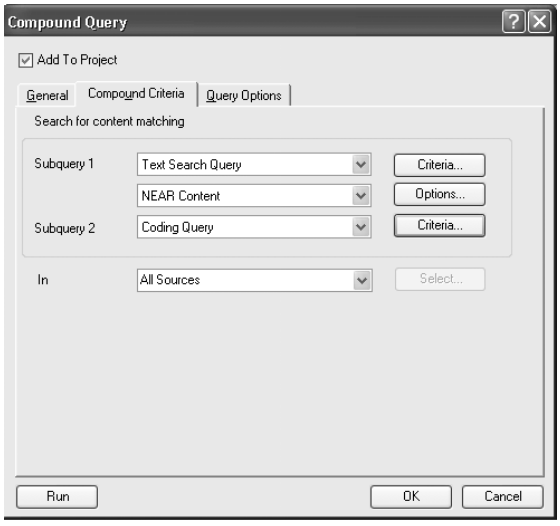


Fig. 6.19 – Finestra per specificare i criteri della query (*Compound Criteria*)

Proviamo a cercare come primo elemento (*Subquery 1*) le parole "pugni" o "botte" nel testo codificato per le donne (*Subquery 2*). Il risultato apparirà nel pannello di lavoro.

6. 1. 5. MATRICI DI QUERY

Nell'interrogazione per matrice quello che occorre definire sono proprio gli elementi che andranno nelle righe e nelle colonne di cui si viene a comporre la matrice di interrogazione stessa.

Nelle righe introdurremo tre distinti nodi. Per selezionarli: *Selected Items/Select* e dalla finestra *Select Project Items* scegliamo i nodi sui quali lavorare. Diamo l'*OK* e tornando alla finestra *Matrix Coding Query/Row* li aggiungiamo alla lista con *Add to List* (fig. 20).

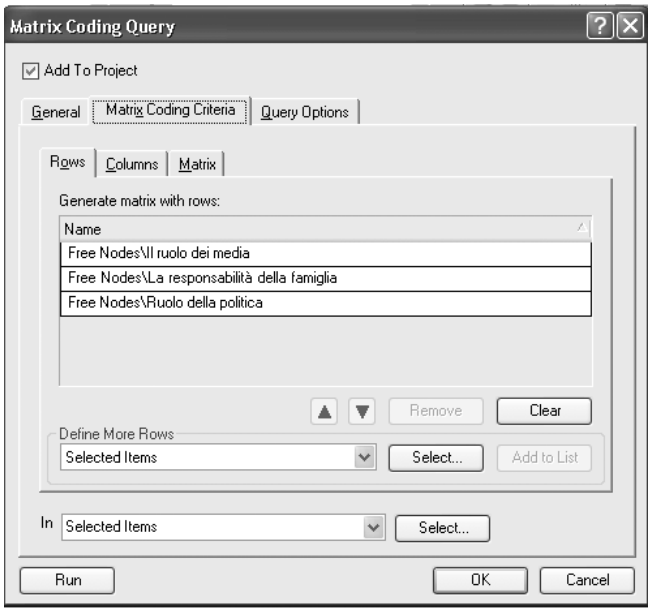


Fig. 6.20 – Visualizzazione dei nodi sui quali applicare la query (*Matrix Coding Criteria/Row)*

Individuati gli elementi da porre in riga occorre selezionare quelli da mettere nelle colonne. In colonna metteremo la variabile di genere nelle sue modalità

"uomo" e "donna"; quindi con *Attribute Condition/Select* aggiungeremo le modalità una alla volta.

Fig. 6.21 – Visualizzazione dei casi sui quali applicare la query (*Matrix Coding Criteria / Columns*)

Anche in questo caso i risultati della query andranno a visualizzarsi nel pannello di lavoro.

6. 2. I MODELLI

Analizzati i testi, i risultati della loro esplorazione possono essere rappresentati mediante i modelli. I modelli esprimono graficamente le relazioni fra gli elementi creati durante il processo di codifica.

Per realizzare un modello, dal pannello di navigazione si attiva **Models** e dal menu **New** si seleziona *Dynamic Model in This Folder*. Si apre la finestra *New Model* (fig. 6.22) e si attribuisce un nome al modello, "Codici e fonti", fornendone una breve descrizione: "relazione tra i testi e i codici creati".

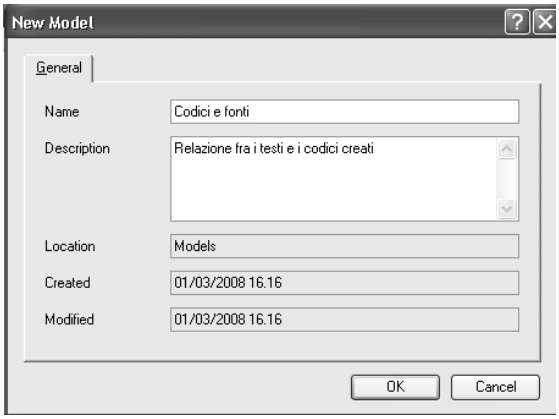


Fig. 6.22 – Finestra per nominare un modello

Mediante questo modello si vogliono visualizzare i legami fra i testi *Bullismo_M* e *Bullismo_F* e i codici fino al momento realizzati.

Il modello "Codici e Fonti" viene a visualizzarsi nel pannello dei documenti. Nel piano di lavoro appare invece un foglio a quadretti (fig. 6.23): alla sua sinistra si visualizzano le forme (*Shapes*) che è possibile utilizzare e a destra *Custome Groups* e *Project Groups*.

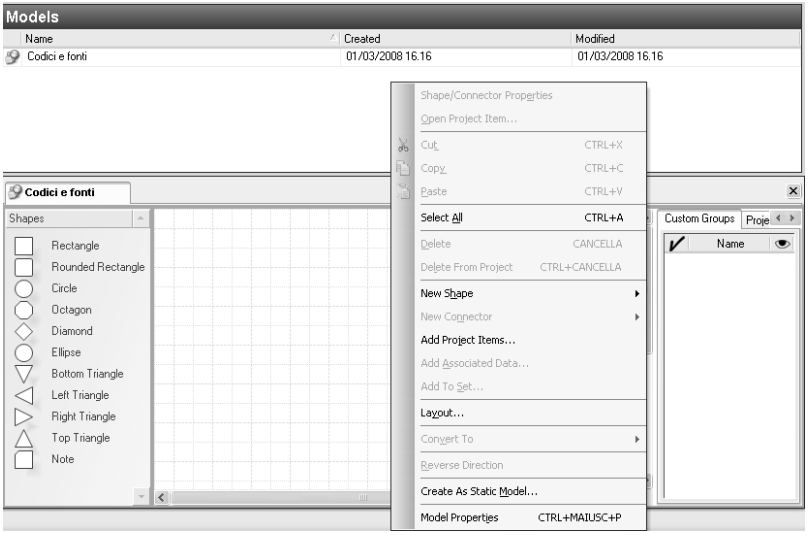


Fig. 6.23 – Procedura per importare gli elementi del modello

Ponendoci sul pannello a quadretti, mediante il tasto destro si seleziona *Add Project Items* e dalla finestra *Select Project Items* (fig. 6.24) si spuntano gli elementi interessati: *Free Nodes* e *Cases*. Dando l'*OK* si apre una successiva finestra di dialogo *Add Associated Data* (fig. 6.26) mediante la quale è possibile associare ulteriori elementi a quelli già impostati.

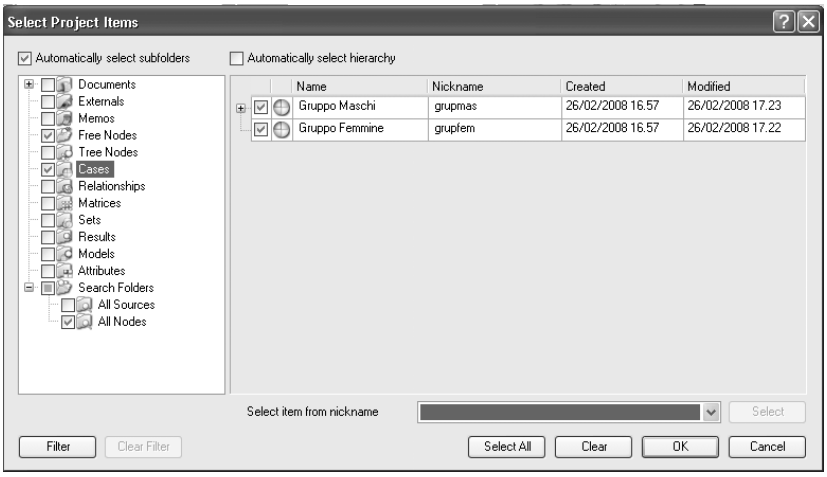


Fig. 6.24 – Finestra per la selezione degli elementi

del progetto di lavoro si seleziona qui *Sources Coded* e *Attribute Values*.

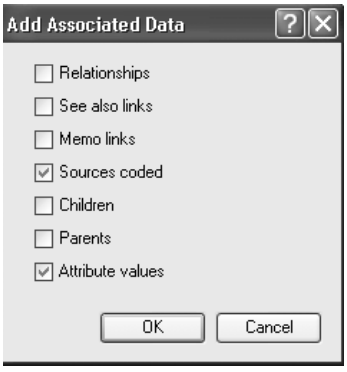


Fig. 6.25 – Finestra per associare informazioni aggiuntive agli elementi selezionati

Nello spazio a destra troviamo le fonti dati, quindi i testi "gruppo_maschi" e "gruppo_femmine" con i rispettivi attributi.

Nello spazio centrale visualizziamo i soggetti del modello ovvero i codici e i testi. È facile notare che mentre "il ruolo dei media" e "il ruolo della politica" sono elementi posti in evidenza soprattutto dagli uomini, "la responsabilità della famiglia" nell'insorgere di fenomeni di bullismo è un elemento posto in luce tanto dagli uomini che dalle donne.

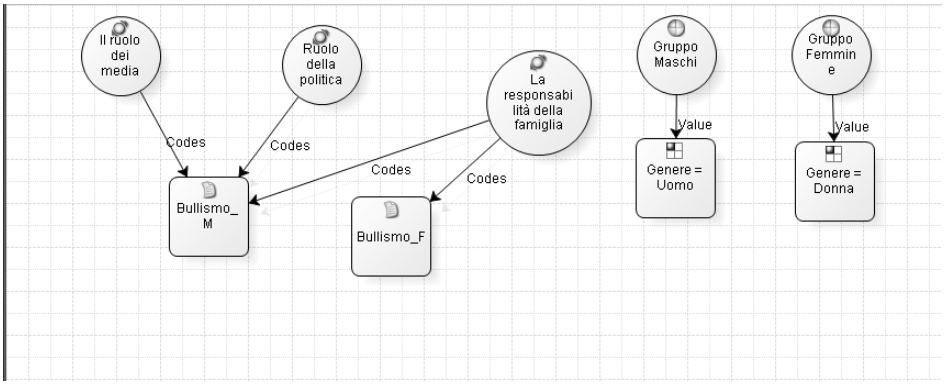


Fig. 6.26 – Modello esplicativo del contenuto del progetto di lavoro

È possibile copiare e incollare secondo le tradizionali procedure il proprio modello. Si possono creare modelli dinamici e statici; la differenza consiste proprio nella libertà di movimento degli elementi presenti nei modelli. Gli elementi di un modello dinamico possono essere sistemati nel foglio a piacimento. Gli elementi di un modello statico sono invece dati una volta per tutte. Colori e stili degli elementi dei modelli dinamici possono essere modificati dalla barra degli strumenti, cambiando il tipo di carattere e la dimensione degli stessi.

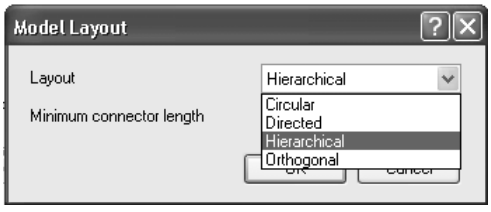


Fig. 6.27 – Dettaglio della finestra di selezione del layout del modello

La disposizione degli elementi può essere cambiata selezionando *Layout/Model Layout* dal tasto destro del mouse (fig. 6.27). Per avere un po' più di spazio nel pannello di lavoro si può invece mediante il menu **View** selezionare *Models Shapes Palette* ed eliminare il riquadro delle forme. Sempre da **View**/*Models Groups* si può far sparire il riquadro posto a sinistra.

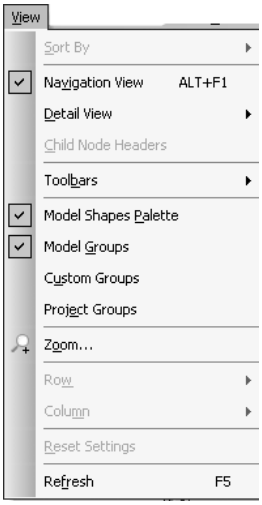


Fig. 6.28 – Menu della funzionalità *View*

6. 3. LE RELAZIONI

Le relazioni servono per stabilire delle connessioni fra più elementi presenti nel progetto di lavoro o semplicemente per porre in interazione degli appunti, dei pensieri propri del ricercatore. A seconda del tipo di freccia utilizzata per indicare la relazione è possibile definire differenti tipi relazioni: associativa, a un senso e simmetrica).

Fig. 6.29 – Tipi di relazioni

Tuttavia questa tipologia dipende esclusivamente dalla direzione delle frecce, poiché – come è proprio a NVivo7 – per ogni elemento creato è possibile indicare il nome e descrivere il contenuto dello stesso.

Dal pannello di navigazione si attiva **Classifications**/*Relationships Types* quindi dalla barra degli strumenti **New**/*Relationship Types in This Folder* e dalla finestra *New Relationship Type* si attribuisce il nome alla relazione (fig. 6.30); è d'aiuto esplicitare nel nome il tipo di connessione che si sta realizzando. Nello specifico, dall'analisi dei documenti emerge una relazione fra l'assenza di un ruolo forte giocato dalla famiglia nell'educazione dei ragazzi e l'eccessiva esposizione degli stessi ai media.

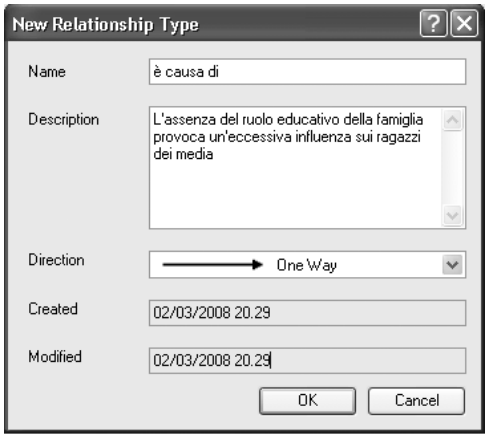


Fig. 6.30 – Finestra per nominare una relazione

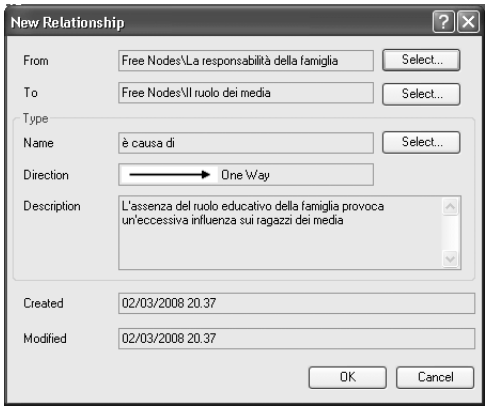


Fig. 6.31 – Modalità di definizione di una relazione

Esplicitata la relazione si dà l'*OK* e questa si viene a visualizzare nel pannello dei documenti. A questo punto dal pannello di navigazione **Nodes** si seleziona *Relationships*, poi dal menu **New** si seleziona *Relationship in This Folder*. Nella finestra *New Relationship* (fig. 6.31) da *From* si clicca sul bottone *Select* e si individua il nodo di partenza denominato "La responsabilità della famiglia", e da *To* si clicca su *Select* "Il ruolo dei media" (fig. 6.31).

Da Name si clicca su *Select* e si recupera la relazione scegliendola dalla finestra *Select Project Items* (fig. 6.32): "è causa di".

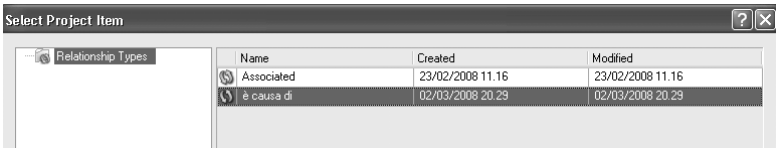


Fig. 6.32 – Procedura di selezione delle relazioni

La relazione si viene a visualizzare nel pannello dei documenti nel quale si sintetizza tutto il percorso realizzato.

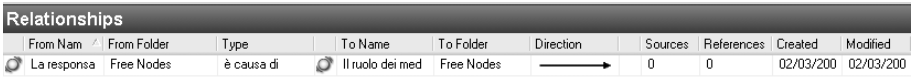


Fig. 6.33 – Visualizzazione della relazione nel pannello dei documenti

La creazione della relazione è poi rappresentabile anche mediante modello grafico.

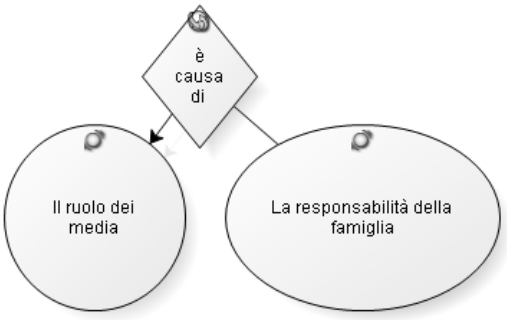


Fig. 6.34 – Modello per la relazione

7.
L'ANALISI QUANTITATIVA DEL LESSICO

L'analisi quantitativa del linguaggio naturale rappresenta una sfida straordinaria per la statistica e per la metodologia della ricerca nelle scienze sociali. Lo sviluppo dell'informatica dagli anni '50 in poi ha esercitato un grande fascino sui linguisti e, in generale, sugli studiosi di discipline affini come la sociolinguistica e la psicolinguistica, dando vita al settore disciplinare autonomo, sebbene difficilmente isolabile e poco omogeneo, della linguistica computazionale (Chiari, 2007). Di fatto l'incontro tra informatica e linguistica, fortemente intrecciato – come vedremo – con la statistica e la matematica – avviene su un terreno trans-disciplinare in cui tutte le competenze delle discipline del linguaggio, anche le più specialistiche, trovano un momento di sintesi applicativa che ha ricadute importanti sia per la ricerca di base che per le applicazioni tecnologiche, in particolare nei settori della traduzione automatica, del riconoscimento e sintesi digitale del parlato e della gestione dei grandi sistemi informativi, un tema tipico del *Knowledge Management*.

Una storia della linguistica quantitativa non è ancora stata scritta, ma proviamo a seguirne le tappe di sviluppo a partire dalle linee generali tracciate da Sergio Bolasco, uno dei primi studiosi italiani ad occuparsi di analisi statistica dei dati testuali, in apertura della Giornata di studio su "Applicazioni di analisi testuale" tenutasi alla Sapienza – Università di Roma il 16 dicembre 2003, integrando le sue indicazioni con altre notizie raccolte da contributi disciplinari diversi (Bolasco, 2004).

7. 1. I PIONIERI DELLA LINGUISTICA QUANTITATIVA

L'approccio quantitativo allo studio del linguaggio trae origine dalla psicologia sperimentale positivista e dalla necessità di tracciare una linea di demarcazione tra questa nuova scienza e la filosofia. Benjamin Bourdon nel 1888, allora professore all'Università di Rennes, in una ricerca sull'espressione delle emozioni e delle intenzioni attraverso il linguaggio (Bourdon, 1892) prese come riferimento un capitolo della Bibbia, l'Esodo, ne individuò le parole essenziali (trascurando le cosiddette "parole vuote") e fece una calcolo delle frequenze di queste parole riordinandole in una classificazione tematica.

La statistica linguistica e l'analisi del contenuto nascono pertanto insieme in un terreno di confine tra la critica letteraria, la linguistica vera e propria e la psicologia con l'intento di trovare un metodo in grado di garantire, attraverso la quantificazione, criteri di oggettività e di controllo delle ipotesi. Questo approccio era già stato annunciato nel 1848 dal matematico russo Victor Buniakowski in un articolo sulla rivista *Sovremennik* che trattava della applicazione delle misure di confidenza ai risultati delle scienze empiriche, in cui traccia un abbozzo di "descrizione aritmetica" delle caratteristiche del linguaggio che tenesse conto della frequenza delle parole, della loro lunghezza, delle lettere iniziali, finali, ecc., al fine di poter applicare la matematica e la statistica in campi di applicazione nuovi, come gli studi grammaticali, etimologici e la filologia comparata. Buniakowski è convinto che probabilmente le sue proposte sarebbero state giudicate come una perdita di tempo non sufficientemente compensata dalla maggior precisione nella conoscenza dei fenomeni linguistici, ma nonostante ciò ritiene che la statistica del linguaggio sia un "tentativo industrioso" da perseguire su un terreno comune di incontro tra matematici e specialisti di filologia (Gryzbeck, 2003).

Tuttavia, il vero e proprio interesse per l'analisi statistica del lessico (lessicometria) si deve soprattutto a stenografi come F. W. Kaeding, che nel 1898 coordinò una ricerca sulle frequenze dei grafemi, delle sillabe e delle parole nella lingua tedesca, e J. B. Estoup, uno dei maggiori esponenti della stenografia francese, il quale in uno studio pubblicato per la prima volta nel 1907 definì la nozione di rango come posizione occupata da una parola in una lista di parole ordinate secondo frequenze decrescenti, mettendone in evidenza l'importanza per l'acquisizione della velocità nella scrittura (Estoup, 1916).

Successivamente è la psicolinguistica a imprimere un'accelerazione nell'analisi statistica del linguaggio. Adolf Busemann in un suo studio pilota sul linguaggio dei bambini (1925), dimostrò che vi erano due stili linguistici ben distinti, ma correlati: uno attivo, rappresentato da un uso preponderante di

verbi nel discorso, e uno qualitativo, con una prevalenza di aggettivi e avverbi. I bambini emotivamente instabili presentavano un "quoziente d'azione" più elevato. Da questi studi, David P. Boder, negli anni '40, ricavò la nota misura grammaticale detta "quoziente aggettivo/verbo" (*adjective-verb ratio*) utilizzando differenti generi letterari e scoprendo che nel dramma, presumibilmente il più vicino al linguaggio parlato, i verbi erano cinque volte di più degli aggettivi nell'85% dei casi.

Il lavoro di Jean-Baptiste Estoup è ripreso da George K. Zipf , docente di linguistica ad Harvard, che già nella sua prima pubblicazione enuncia i termini generali e su base fonetica il "principio di frequenza relativa" che diverrà poi la "legge di Zipf", secondo la quale in un testo sufficientemente lungo, classificando le parole per ranghi decrescenti, il rango moltiplicato la frequenza, rimane approssimativamente costante (Zipf, 1929; 1935). La legge di Zipf è una legge empirica, della quale si può solo dare una dimostrazione induttiva, ma che non ha alcun sostegno teorico-esplicativo, analogamente al cosiddetto principio di Pareto, formulato dal grande economista nel 1896 studiando la distribuzione dei redditi, secondo il quale solo pochi individui posseggono la maggior parte dei redditi e la loro concentrazione intorno alla media è sempre approssimativamente costante in tutti i tempi e in tutti i sistemi politici (Pareto, 1896).

Zipf era animato dalla forte convinzione che fosse possibile studiare il linguaggio come un fenomeno naturale e oggettivo, come se vi fosse una qualche forza operante nella natura che potesse determinarne gli effetti e le regolarità rilevabili statisticamente. La legge di Zipf è stata successivamente generalizzata ad altri fenomeni naturali e sociali, dalla distribuzione del diametro dei crateri lunari, fino alla distribuzione della popolazione nei centri abitati, alle frequenze di accesso alle pagine web o dei messaggi inviati nei forum e newsgroup, ma nessuno è stato in grado di spiegare perché queste regolarità si verificano. Il contributo di Zipf alla linguistica, sebbene geniale, è controverso. Tuttavia, sul piano dell'analisi quantitativa, la sua visione scientifica del linguaggio ha dato un forte impulso ad altri contributi teorici ed empirici che solo in tempi recenti – con lo sviluppo del calcolo automatico – hanno trovato più estesi campi di applicazione.

Nel 1939 è lo statistico e matematico inglese George Udney Yule a occuparsi dell'analisi del vocabolario "letterario" mettendolo in relazione con il problema dello stile e aprendo la strada alla vera e propria statistica lessicale (Yule, 1944) con un primo studio sul caso di dubbia attribuzione dell'*Imitazione di Cristo* a Tommaso da Kempis o Giovanni Gerson, più noto come San Giovanni della Croce (Yule, 1939). Sempre su un tema religioso, c'è poi il lavoro

pionieristico del gesuita Roberto Busa sull'*Index Tomisticus*, la costruzione di un lessico di frequenza sulle opere di San Tommaso D'Aquino, iniziato in proprio nel 1946 e poi realizzato nel corso di trent'anni di lavoro con l'aiuto dell'IBM e con la pubblicazione di 56 volumi (Busa, 1974-1980), dando vita a quella che poi sarà l'informatica linguistica.

Il tema della frequenza delle parole nel testo, già trattato da Zipf e Yule, ritorna nel lavoro del linguista Pierre Guiraud, che identifica delle relazioni semplici e costanti tra lunghezza del testo ed estensione del vocabolario. Per esempio, vi sono pochissime parole (in gran parte "parole grammaticali" o *mots-outils*) che coprono più della metà delle occorrenze di parole che compongono la maggioranza dei testi. Pertanto Guiraud definisce il lessico secondo la sua estensione (varietà delle parole che lo compongono) e la sua struttura (la frequenza relativa di ciascuna parola). Nella struttura del lessico, Guiraud sottolinea la rilevanza della concentrazione delle parole più frequenti (parole tematiche) e della dispersione delle parole meno frequenti, che rappresentano però una misura della ricchezza del vocabolario dal quale sono tratte le parole stesse. La concentrazione delle parole è legata all'argomento del testo e alla motivazione che lo origina (parole chiave), mentre la dispersione è legata alla sua caratterizzazione (originalità o eccentricità del testo). L'analisi statistica della distribuzione delle parole nel testo è quindi fondata su un certo numero di costanti che definiscono e misurano queste proprietà (Guiraud, 1954, pp. 68-69).

La misura della ricchezza del vocabolario di un testo, inteso come un campione del lessico di riferimento, trova una formulazione matematica da parte di Gustav Herdan (1956), che propone l'utilizzazione di una misura fondata sulla funzione esponenziale.

Verso la fine degli anni '50, il Centro studi della lingua francese di Besançon inizia uno spoglio meccanografico delle opere di Corneille, analogo a quello condotta da Busa su Tommaso d'Aquino, che porta nel 1957 a dar vita al progetto denominato *Trésor de la Langue Française*, all'interno del quale Charles Muller sviluppa le sue analisi lessicometriche (Muller, 1967; 1977). L'idea di fondo è che i testi sono un campione rappresentativo della lingua e che attraverso lo studio dei corpora sia possibile risalire alle informazioni sul linguaggio. Da qui nascono gli studi sulla specificità del vocabolario, intendendo così mettere a confronto le frequenze relative delle parole in un testo con le frequenze teoriche attese nell'insieme del corpus e di valutarne la sovra o sottoutilizzazione. La costituzione di un corpus di riferimento per la lingua francese (*Frantext*) avrebbe permesso così di utilizzare un modello probabilistico per l'individuazione di un linguaggio peculiare o tipico. È da qui che hanno origine indici e misurazioni divenute poi di uso comune nella statistica linguistica e

lessicale (Tournier, 1980 e Lafon, 1984) e che hanno portato allo sviluppo del software *Hyperbase* di Etienne Brunet (1993).

7. 2. LA COSTRUZIONE DEI LESSICI DI FREQUENZA

Negli anni '60, parallelamente agli studi sul lessico degli autori classici, il rapido diffondersi della *Information Technology* e della ricerca computazionale diedero un forte impulso alla costituzione dei lessici di frequenza, il cui obiettivo, molto più vasto e impegnativo, non è soltanto di descrivere la lingua di un'opera specifica (la *Divina Commendia*) o di un autore (Corneille o Shakespeare) quanto piuttosto di descrivere la lingua scritta (o anche parlata) da un'intera comunità linguistica attraverso un campione rappresentativo dei diversi generi testuali: lettere, articoli di giornali, teatro, letteratura, ecc.

Da qui nasce la **linguistica dei corpora**, un metodo di indagine sul linguaggio che assume come riferimento un insieme finito di "esecuzioni linguistiche" che possono essere tratte dal parlato e poi trascritte oppure direttamente costituite da testi scritti. Questo approccio alla linguistica non è condiviso da tutti perché una lingua è qualcosa di vivo e quindi, di per sé, virtualmente infinito. Un corpus, comunque costituito, non può rappresentare che in modo incompleto tutti gli enunciati possibili; pertanto questo modo di procedere dell'indagine è stato spesso criticato e ostacolato, come per esempio dal maggior linguista teorico del XX secolo, Noam Chomsky, secondo il quale l'oggetto di studio della linguistica non è costituito da un insieme, inevitabilmente parziale, di esecuzioni (*perfomances*) ma dalle strutture implicite e inconsce del linguaggio naturale del quale è necessario ricostruire le regole (grammatica generativo-trasformazionale). Questa polemica ricorda molto da vicino la contrapposizione tra metodo quantitativo e metodo qualitativo, tra spiegazione e interpretazione. La linguistica di un corpus è utile se il corpus è costituito da un campione sufficientemente rappresentativo della lingua di riferimento, sebbene non possa pretendere di esaurire la complessità dell'analisi linguistica, perché può fornire informazioni attendibili, generalizzabili e controllabili.

Il primo esperimento in questo senso è il *Dictionnaire Fondamental de la langue française* di Georges Gougenheim del 1958, composto di 3.550 parole e basato su una raccolta di 163 registrazioni al magnetofono; mentre per la lingua inglese è stata fondamentale la pubblicazione del *Computational Analysis of Present-Day American English* del 1967, di Henry Kučera e Nelson Francis.

In Italia la linguistica quantitativa è rappresentata soprattutto dai lavori di Antonio Zampolli, al quale risale il primo *Lessico di frequenza della lingua italiana*

contemporanea (LIF) del 1971 (con U. Bortolini e G. Tagliavini) e di Tullio de Mauro, al quale si deve il più ampio progetto del *Vocabolario elettronico della lingua italiana* (VELI) del 1989 e il *Lessico di frequenza dell'italiano parlato* (LIP) del 1993 (con F. Mancini, M. Vedovelli e M. Voghera).

7. 3. LA SCUOLA FRANCESE DELLA STATISTICA TESTUALE

Negli anni '60 un brillante matematico e statistico francese, Jean-Paul Benzécri, sulla base dei suoi studi di linguistica quantitativa e con un orientamento del tutto opposto a quello di Noam Chomsky, acquisisce una posizione di rilievo nel panorama della statistica multidimensionale, conquistando uno spazio che, a fronte della crescente richiesta di analisi automatica dei dati, era rimasto libero da parte dell'approccio razionalista della statistica inferenziale, soprattutto nell'ambito dell'analisi dei dati categoriali e della classificazione automatica (*cluster analysis*). Dagli studi e dal successo ottenuto da Benzécri con l'*Analyse des Correspondances* ha origine la scuola francese dell'*Analyse des Données*, all'interno della quale, negli anni '80, un'équipe diretta da Ludovic Lebart e Alain Morineau (1985) sviluppano il software applicativo SPAD (*Système Portable pour l'Analyse des Données*).

In questo ambito, sempre sotto la direzione di Lebart e con la collaborazione di André Salem, Mónica Bécue e altri, si viene a definire un settore specifico della statistica testuale basata sulla forma grafica delle parole e un software dedicato all'analisi dei dati testuali (SPAD-T), dal quale poi troveranno origine Lexico (al quale è dedicato il cap. 8 di questo volume) e DTM (*Data Text Mining*), un ottimo software di analisi multidimensionale sviluppato dallo stesso Ludovic Lebart (http://ses.enst.fr/lebart/) per uso gratuito di studenti e ricercatori. L'approccio all'analisi testuale sulla base della forma grafica delle parole introduce un elemento di novità sostanziale perché si tratta di un approccio formale indipendente dalla lingua, in cui l'analisi (automatica) è basata sul significante (la forma scritta della parola) e l'obiettivo è di arrivare al **senso** degli enunciati concreti, cioè a un insieme di significati esplicitati dal contesto. Il significato di una parola, pertanto, è a sua volta distinto in una parte puramente formale (l'appartenenza a una classe sintattica e grammaticale) e da una parte sostanziale, il vero e proprio contenuto (l'appartenenza a una classe semantica). Dal punto di vista statistico, la ricostruzione del senso avviene lungo l'**asse sintagmatico**, in cui le parole di combinano tra loro in sequenze, e lungo l'**asse paradigmatico** che è determinato dalla sostituibilità ed equivalenza delle parole nel sintagma, senza che per questo ne risulti una modifica

nel contenuto dell'enunciato (Bolasco, 2004, p. 10).

Questo approccio lessicometrico si struttura in tre momenti principali (Jenny, 1997):

- inventario e conteggio di tutte le forma grafiche "grezze" che costituiscono il corpus e ordinamento di esse in ordine alfabetico e in ordine di occorrenze decrescenti;

- costruzione della tabella lessicale del corpus, composta di tante righe quante sono le forme, classificate in ranghi di occorrenze decrescenti, e di tante colonne quante sono le parti in cui è stato suddiviso precedentemente il corpus in base a una certa proprietà caratteristica;

- calcolo delle frequenze relative secondo le parti di cui è composto il corpus e comparazione dei profili lessicali.

Su questa base, utilizzando i metodi e le tecniche di analisi multimensionale (soprattutto l'analisi delle corrispondenze e la *cluster analysis*) si producono rappresentazioni grafiche delle lessie (semplici e complesse) che permettono di visualizzare modelli di senso apprezzando vicinanze, somiglianze, distanze e contrapposizioni tra insiemi di forme grafiche all'interno del corpus in esame.

A questo approccio si sono ispirati i principali software attualmente disponibili tra i quali, oltre ai precedenti già citati, Alceste di Max Reinert, TaLTaC^2 di Sergio Bolasco e coll. (cui sono dedicati i capitoli 9, 10 e 11 di questo manuale), T-LAB di Franco Lancia e WordMapper di Jean François Grimmer.

7. 4. ESTRAZIONE DELL'INFORMAZIONE E TECNOLOGIE DI TEXT MINING

Dall'inizio degli anni '90 la crescita delle informazioni memorizzate su supporti informatici ha reso potenzialmente disponibili enormi risorse di dati e testi creando l'illusione che questo avrebbe contribuito a una maggiore efficienza e capacità decisionale. Fino al 1999 l'unità a misura più consueta utilizzata per indicare la quantità di informazione memorizzabile su hard disk era il megabyte e i testi che si analizzavano andavano da 500.000 a 4 milioni di occorrenze . Oggi un computer portatile di fascia media ha un hard disk di 120 gigabyte ed è possibile anche trattare testi di oltre 250 milioni di occorrenze.

Il vero problema nella gestione delle informazioni, oggi, non è più l'immagazzinamento ma la sua interpretazione e pertanto la selezione delle informazioni importanti rispetto a quelle superflue o, sotto determinati punti di vista, irrilevanti.

Ecco allora che l'estrazione di informazione utile, quindi dotata di valore

economico e trasformabile in conoscenza, diventa un obiettivo fondamentale. In un primo tempo, l'interesse si concentra sulla gestione di banche dati e di sistemi informativi strutturati per i quali si sviluppano tecniche e software di *Data Mining*. Poi, con il fenomeno crescente della comunicazione mediata dal computer in tutte le transazioni, commerciali e personali, e grazie ai successi ottenuti nelle procedure di analisi del linguaggio naturale e della maggiore disponibilità di risorse linguistiche, emerge la necessità di trattamento delle informazioni non strutturate, che risiedono in file di testo, come e-mail, pagine web, forum, relazioni di focus group, notizie, verbali di riunioni, contratti, cartelle cliniche, interviste e risposte a domande aperte nei questionari. Già nel 1995, un rapporto di Forrester Research, stimava che l'80% delle informazioni utili in un'azienda è nascosto all'interno di documenti non strutturati e pertanto non disponibili per la stessa azienda, che non sa nemmeno di possedere queste informazioni e quindi non è in grado di valorizzarle inserendole nel processo produttivo (Forrester Research, 1995). Le tecnologie di *Knowledge Discovery in Text* (KDT) e di *Text Mining* (TM) permettono di fronteggiare questi problemi e di cercare delle soluzioni. I produttori di software che più hanno lavorato in questo senso sono SAS (Text Miner), SPSS (Lexi Quest Mine e Text mining), TEMIS (Online Miner) e Synthema (Lexical Studio).

Il TM è un settore complesso nel quale confluiscono diverse tecnologie che vanno dal recupero delle informazioni alla archiviazione dei documenti, dalla progettazione dei sistemi di "intelligenza artificiale" alla traduzione automatica dei testi, dalla costruzione di meta informazioni tratte dai dati testuali alla costruzione di programmi automatici di apprendimento in grado attivare processi di scoperta a partire da un set predefinito di conoscenze. Si tratta di temi che toccano solo in piccola parte gli argomenti di questo manuale, ma che, tuttavia, non possono prescindere dalla acquisizione di metodi, strumenti e tecniche di analisi automatica dei testi.

7. 5. GLI ELEMENTI COSTITUTIVI DEL TESTO: LE PAROLE

Il corpus quindi è costituito di testi e ogni testo rappresenta una delle possibili partizioni di un corpus. Ogni testo si può suddividere a sua volta in **frammenti**. Senza alcuna pretesa di voler affrontare argomenti che esulano dagli scopi di questo manuale, come la semiologia, la linguistica, la retorica e la stilistica, ci sono però delle conoscenze fondamentali nello studio della lingua che devono essere ben comprese per affrontare lo studio quantitativo dei testi.

La **frase** è l'unità massima in cui vigono relazioni di costruzione sintatti-

ca (Renzi, 2001, p. 37 e sgg.): "Va via domani" è una frase.

"Va via sono occupato" è un enunciato costituito da due frasi: "va via" e "sono occupato". L'**enunciato** non è una unità grammaticale ma un pensiero completo. Noi parliamo per enunciati.

La **proposizione** è l'unità sintattica con cui si indica una frase elementare minima (soggetto e predicato; per i verbi meteorologici basta il predicato verbale: "piove", "nevica").

Più proposizioni possono far parte di una frase complessa.

Proposizioni subordinate e proposizioni coordinate costituiscono dei **periodi**. Ogni periodo grammaticale di solito termina con un segno di interpunzione forte (.!?).

I principali elementi costitutivi di un testo sono le **parole**. Di ogni parola dobbiamo conoscere la pronunzia, la grafia e il significato. Gli elementi fondamentali di questa conoscenza costituiscono il **lessico**. Gli oggetti del lessico sono raccolti in opere che descrivono le parole all'interno dei settori in cui esse vengono utilizzate: i **vocabolari**.

Le regole che descrivono e spiegano le strutture della lingua costituiscono la **grammatica**. La **morfologia** è quella parte della grammatica che si occupa delle forme (flessione, declinazione e coniugazione) che le parole assumono. La **sintassi** detta invece le regole per una disposizione ordinata e corretta delle parole all'interno del **discorso**.

Senza entrare nel dettaglio della nomenclatura grammaticale, ci sono alcune definizioni che è bene rammentare perché sono rilevanti ai fini di una corretta utilizzazione dell'analisi automatica dei testi.

Le parti del discorso nella sintassi:

- *soggetto*: ciò di cui si parla e a cui si riferisce l'azione (compiuta o subita);
- *verbo* o *predicato verbale*: l'azione riferita al soggetto, il modo di essere, ecc.;
- *oggetto*: lo scopo dell'azione;
- *complementi*: parti della proposizione che ne completano il significato esprimendo relazioni e circostanze in cui l'azione viene svolta.

I termini grammaticali:

- *aggettivo*, *attributo*: si aggiunge al sostantivo per determinarlo e qualificarlo;
- *articolo* (determinativo o indeterminativo): particella premessa al sostantivo che precisa il genere e il numero del nome che la segue;
- *avverbio*: si pone vicino al verbo o all'aggettivo per modificarlo;
- *congiunzione*: serve a legare parti di una frase tra loro;
- *interiezione*: esprime un'esclamazione di tipo emotivo (dolore, gioia, sorpresa, ecc.);
- *numero*: definisce la singolarità o pluralità di sostantivi e verbi;
- *preposizione*: serve a legare una parola (nome, aggettivi o verbi) all'altra;

- *pronome*: sostituisce il nome senza specificarlo;
- *sostantivo*, *nome*: indica esseri animati, inanimati, oggetti concreti o astratti;
- *verbo*: indica l'azione o il modo di essere di una entità, la relazione tra entità, l'appartenenza ad una categoria esistenziale.

Di fatto la distinzione tra grammatica e lessico oggi passa attraverso la nozione di regolarità della lingua. La grammatica è il regno della sistematicità, mentre il lessico è il regno dell'irregolarità, si sottrae alle generalizzazioni (Lepschy, 1979, p. 134).

Le parole costituiscono le proposizioni (le frasi). Le frasi messe in sequenza costituiscono un discorso (o un testo, quando il discorso è trascritto). Il testo però non è una somma di frasi: è qualcosa di più che contribuisce a costruire il senso. Il testo permette di restituire il significato delle parole anche quando queste, prese isolatamente, sono **ambigue**. La parola con grafia <rosa> può essere riferita al nome del fiore, ad un colore, al participio passato del verbo *rodere* o all'aggettivo corrispondente, espresso nel genere femminile. La pronunzia è solo in parte sufficiente a disambiguare la parola (se si parla in modo corretto): *róso*, nel senso di *corróso* ha l'accento acuto come *fióre*; mentre *ròsa* (il fiore) ha l'accento grave. Tuttavia *ròsa* (il colore, il gruppo, la serie) è indistinguibile da *ròsa*. Solo il contesto della frase permette di rendere univoco il significato della parola. Così intere frasi possono essere ambigue e solo il testo permette di comprenderne il senso. "Quel cane del tenore ulula da un'ora" assume un senso completamente diverso secondo il testo in cui la frase è inserita: potrebbe riferirsi ad una povera bestia (il cane) che aspetta il ritorno del suo padrone (il tenore); oppure essere riferito al tenore (un cane) che canta in modo sgradevole per il pubblico (l'esempio è tratto da Marchese, 1978, p. 67).

La parola scritta ha quindi tante forme diverse che possono derivare da significati diversi (**polisemia**) con pronunzia identica (parole **omofone**) e da significati diversi con grafia identica (parole **omografe**). D'altra parte ci possono essere significati che hanno una stessa forma grafica ma che appartengono a due parole diverse (**omonimia**); oppure due forme grafiche diverse che esprimono lo stesso significato (**sinonimia**). Le omonimie sono frequenti nell'inglese, in cui la pronunzia e la grafia sono in un rapporto complesso.

Le parole possono anche essere raccolte in lemmi, cioè essere classificate in forme che hanno fra di loro qualche attributo comune, per esempio il fatto di essere coniugazioni dello stesso verbo (*venire* come lemma di *vengo*, *vieni*, *viene*, ecc.). Le parole così come sono classificate nei vocabolari costituiscono dei lemmi. Tuttavia la classificazione delle forme grafiche non è un'operazione univoca né semplice da definire. Si possono classificare in un unico "ceppo" parole che hanno una funzione grammaticale diversa (*termine*, *termini*, *terminare*,

terminante) oppure parole che appartengono a uno stesso ambito semantico perché legate a un'origine comune (che hanno lo stesso etimo) come la "famiglia di parole" classificabile sotto la voce "fare" che comprende il verbo *fare* con tutte le sue coniugazioni, ma anche parole come *faccenda*, *faccendiere*, *facile*, *facinoroso*, *facoltà*, *fazione* e *fazioso* (Gianni, 1988).

Le parole sono costituite, come si è detto, da un significante e da un significato. Il significante è la rappresentazione grafica della fonetica della parola e consiste di una stringa di lettere ordinate linearmente. L'inventario dei significanti è limitato, si presta a vari tipi di ordinamento (l'ordine alfabetico è quello solitamente utilizzato nei vocabolari) ed è analizzabile con i sistemi di trattamento automatico dell'informazione. I significati fanno riferimento a un campo dai confini sfumati, non sono ordinabili e non sono mai stati inventariati in modo esaustivo in nessuna lingua. Noi cerchiamo nei vocabolari il significato di una parola a partire da un significante noto. Non possiamo fare l'inverso e non c'è nessun vocabolario che permetta di farlo, salvo alcuni prodotti culturali interessanti come il *Dizionario analogico della Lingua italiana* (UTET, Torino, 1991) o il *Dizionario alla rovescia* (Selezione del Reader's Digest, Milano, 1992).

Tra le varie categorie di vocabolari utilizzati nella linguistica, per noi sono particolarmente interessanti i **vocabolari (o lessici) di frequenza**, in cui le parole sono ordinate per ordine di frequenza. Il problema naturalmente è quello della rappresentatività dei testi sui quali è stato effettuato lo spoglio. Fino a che punto possiamo dire che un determinato numero di frasi, tratte da campi diversi del discorso, possono essere considerati un campione rappresentativo della lingua italiana?

La domanda che si pone il lessicografo è: quante parole ha una lingua? Di solito un vocabolario contiene da cinquantamila a cinquecentomila voci per i più estesi. Un individuo di media cultura conosce normalmente da 3.000 a un massimo di 20.000 parole. Gli scrittori utilizzano mediamente da 5.000 a 15.000 parole (Lepschy, 1979, p. 146).

La frammentazione minima di un testo è la parola, una sequenza di caratteri alfabetici delimitata da due separatori. L'insieme dei separatori deve essere convenzionalmente definito come un insieme di caratteri che non appartengono all'alfabeto. Questo non è un problema banale perché in un alfabeto possono esserci dei caratteri che in alcune circostanze possono essere considerati dei separatori (per esempio i numeri). Sono quasi sempre da considerare come separatori: lo spazio bianco (*blank*), la punteggiatura, le virgolette, i "trattini" e le parentesi. Per i caratteri speciali (# @ $ £ § ° % & ^ * < >) e per i numeri si dovrà decidere caso per caso secondo gli obiettivi dell'analisi.

Una sequenza di caratteri delimitata da due separatori definisce una **forma grafica**. Le forme grafiche intese come unità di conto vengono definite **occorrenze** (*words token*). L'analisi automatica di un testo fornisce come primo risultato un conteggio delle forme grafiche. Ad ogni forma grafica diversa viene associato un codice numerico, pertanto è possibile costruire un indice delle forme grafiche che sarà rappresentato dalle **forme grafiche distinte** (*words type*).

Non tutte le parole di un testo possono essere considerate come equivalenti dal punto di vista semantico. Vi sono delle difficoltà nel considerare la parola una unità di base elementare della semantica. Pensiamo a parole come *gatto*, *finestra*, *mela*, *amare*, *odiare*. Ognuna di queste parole presente in una frase rinvia a un contenuto e noi possiamo scegliere se utilizzare l'una o l'altra secondo il concetto che vogliamo esprimere. Parole come *il*, *e*, *che*, *di* sono presenti in una frase in rapporto con altre parole e non possono essere sostituite con parole equivalenti sebbene con un altro significato. "Amare il gatto" può diventare "odiare il gatto", ma l'articolo *il* è insostituibile senza generare una frase grammaticalmente diversa. Uno studioso inglese del XIX secolo, Henry Sweet, propose di distinguere tra **parole piene** (*full words*) e **parole forma** (*form words*). Le parole forma hanno un significato esclusivamente grammaticale che può essere stabilito solo in relazione con le altre parole.

Nell'analisi automatica del linguaggio si distingue più correntemente tra parole piene e **parole vuote**. La distinzione non è, come nel caso di Sweet, di carattere funzionale ma è strumentale ai fini della ricerca. Le parole vuote vengono definite di volta in volta come parole che non esprimono un contenuto interessante ai fini dell'analisi (e spesso sono parole grammaticali o di semplice legame nella frase), mentre le parole piene sono quelle che contribuiscono significativamente all'interpretazione del testo. In alcune analisi – come si vedrà in seguito – le parole grammaticali, di solito considerate come parole vuote, possono diventare molto importanti per l'interpretazione automatica del testo. Un testo che presenta un numero di occorrenze sopra la media nelle proposizioni di *in* e *di* potrebbe essere un testo descrittivo; mentre la prevalenza di *ma* e *se* indicherebbe la presenza di elementi di incertezza (Bolasco, 1999, p. 193).

La frammentazione del testo è un problema di analisi che richiede delle decisioni meditate. Il problema non ha una soluzione univoca. Non si possono fornire indicazioni valide per tutti i casi. Una **frammentazione sintattica** (frasi delimitate da una punteggiatura che ha una rilevanza nella ricostruzione del senso), adeguata per un testo in prosa, può essere sostituita da una **frammentazione stilistica** (il verso di testo poetico), da una **frammentazione di senso** (un enunciato di senso compiuto ricavato in modo strumentale da una suddivisione del testo), oppure da una **frammentazione di comodo**

(una riga di testo). Ogni corpus necessita di una frammentazione del testo che sia adeguata alle ipotesi di lavoro sulla base delle quali è stato costruito.

APPROFONDIMENTI TEMATICI

Sulla storia della linguistica quantitativa c'è un articolo abbastanza elaborato su *Glottopedia* <http://urts120.uni-trier.de/glottopedia/index.php/Main_Page> cliccando dalla home page su *Quantitative linguistics* e poi selezionando "history of quantitative linguistics" (3/3/2008). Con uno spiccato carattere introduttivo sono da segnalare anche Isabella Chiari, *Introduzione alla linguistica computazionale*, Bari, Laterza, 2007; Mauro La Torre, *Le parole che contano. Proposte di analisi testuale automatizzata*, Milnao, FrancoAngeli, 2005; Franco Lancia, *Strumenti per l'analisi dei testi. Introduzione all'uso di T-LAB*, FrancoAngeli, 2004.

Per i più recenti sviluppi del *Text Mining* vi sono due riferimenti che offrono una panoramica ampia e aggiornata dello "stato dell'arte: S. Dulli, P. Polpettini, M. Trotta (a cura di), *Text Mining: teoria e applicazioni*, Milano, FrancoAngeli, 2004; e S. Bolasco, A. Canzonetti, F. Capo (a cura di), *Text Mining. Uno strumento per imprese e istituzioni*, Roma, CISU, 2005.

Numerosi sono i siti web di riferimento sulla linguistica quantitativa e si possono facilmente rintracciare con l'inserimento della opportune parole chiave nei motori di ricerca che, tra l'altro, sono già di per sé strumenti di reperimento delle informazioni nell'immenso *text warehouse* rappresentato da Internet. Per l'analisi dei testi, il punto di partenza ideale è *Text Analysis Info Page* < http://www.textanalysis.info/> curato da molti anni da Harald Klein, cui si deve uno dei primi tentativi di rendere conto in modo sistematico degli sviluppi della ricerca in questo campo, già dalle prime applicazioni in MS-DOS del 1987-1990. *Lexicometrica* è una rivista online curata da André Salem e Serge Fleury <http://www.cavi.univ-paris3.fr/lexicometrica/>, che offre invece una panoramica della ricerca accademica con un taglio fortemente interdisciplinare e pubblica tutti gli atti delle JADT (*Journées internazionales d'Analyse Statistique des Données Textuelles)*, il principale appuntamento internazionale, con cadenza biennale, su questo tema. Per il *Text Mining*, tra i siti più interessanti si possono segnalare: *The National Centre for Text Mining* <http://www.nactem.ac.uk/> un punto di riferimento fondamentale per la ricerca e lo sviluppo di questa tecnologia; il blog di Matthew Horst, ricercatore del Microsoft Liv's Lab, che presenta curiose elaborazioni grafiche dell'infospazio: *Data Mining: Text Mining, Visualisation e Social Media* <http://datamining.typepad.com/data_mining/>, e *Temis*, una società di consulenza nata nel 2000 come spin-off della IBM, con sedi in Francia, Italia e Germania <http://www.temis.com/>. Tra le società italiane: *Èulogos*, fondata nel 1994, molto attiva soprattutto nella Pubblica Amministrazione <http://www.eulogos.net/>; *Celi*, fondata nel 1999, specializzata in tecnologie di trattamento automatico dei testi, con uno specifico software di TM, Sophia <http://www.celi.it/>; e *Synthema*, fondata nel 1994 da specialisti dell'IBM <http://www.synthema.it/>.

RIFERIMENTI BIBLIOGRAFICI

BODER D. P. (1940) "The adjective-verb quotient: A contribution to the psychology of language", Psychological Record, 3, pp. 310-343.

BOLASCO S. (1999) *Analisi multimensionale del dati. Metodi, strategie e criteri d'interpretazione*, Roma, Carocci (II ed. 2004).

BOLASCO S. (2004) "L'analisi statistica dei dati testuali: intrecci problematici e prospettive, in E. Aureli Cutillo e S. Bolasco (a cura di), *Applicazioni di analisi statistica dei dati testuali*, Roma, Casa Editrice La Sapienza, pp. 9-26.

BOURDON B. (1892) *L'expression des émotions et des tendence dans le langage*, Paris, Alcan.

BRUNET É. (1993) "Un hypertexte statistique: Hyperbase", in S. J. Anastex (eds.), (1993), *JADT 1993. Actes des secondes Journées internationales d'Analyse Statistique des Données Textuelles*, Paris, ENST, pp. 1-14.

BUSA R. (1974-1980) *Index Thomisticus: Sancti Thomas Operum Omnium Indices et Concordantiae*, Stuttgart, 56 voll. (consultabile online a cura di Enrique Alarcón, Università di Navarra <http://www.corpusthomisticum.org/> - 3/3/2008).

BUSEMANN A. (1925) *Die Sprache der Jugend als Ausdruck des Entwicklungsrhythmus*, Jena, Fischer.

CHIARI I. (2007) *Introduzione alla linguistica computazionale*, Bari, Laterza.

ESTOUP J. B. (1916) *Gammes sténographiques*. Paris, Institut sténographique de France.

FORRESTER RESEARCH (1995) "Coping with Complex Data", *The Forrester Report*, April.

GIANNI A. , a cura di (1988) *Dizionario italiano ragionato*, Firenze, G. D'Anna – Sintesi.

GUIRAUD P. (1954) *Les caractères statistiques du vocabulaire*, Paris, P.U.F.

GRYZBECK P. (2003). "History of quantitative linguistic", *Glottometrics*, 6, 103-106.

HERDAN G. (1956) *Language as choices and chance*, Groningen, Noordhoff.

JENNY J. (1997) "Méthodes et pratique formalisées d'analyse de contenu et de discours dans la recherche sociologique contemporaine. État des lieux et assai de classification", *Bulletin de Méthodologie Sociologique*, 54, pp. 64-122.

LAFON P. (1984) *Dépouillements et statistique en lexicometrie*, Genève-Paris, Ed. Slatkine et Champion.

LEBART L., MORINEAU A. (1985) *Systeme portable pour l'analyse des données* (SPAD), Paris, Cesia.

LEPSCHY G. C. (1979) "Lessico", in *Enciclopedia*, vol. VIII, Torino, Einaudi, pp. 129-151.

MARCHESE A. (1978) *Dizionario di retorica e stilistica*, Milano, Mondadori.

MULLER CH. (1967) *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse.

MULLER CH. (1977) *Principes et méthodes de statistique lexicale*, Paris, Larousse.

PARETO V. (1896) *Cours d'economie politique*, Genève, Droz.

RENZI L. (2001) "La frase semplice", in L. Renzi, G. Salvi, A. Cardinaletti, *Grande grammatica italiana di consultazione*, vol I, Bologna, il Mulino, pp. 37-127.

TOURNIER M. (1980) "D'où vienne les frequences de vocabulaire", *Mots*, 1, pp. 189-209.

YULE G. U. (1939) "On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship", *Biometrika*, 30 (3-4), pp. 363-390.

YULE G. U. (1944) *Statistical Study of Literary Vocabulary*, Cambridge, Cambridge University Press.
ZIPF G. K. (1929) "Relative frequency as a determinant of phonetic change", *Harvard Studies in classical Philology*, 40, pp. 1-95.
ZIPF, G. K. (1935) *The psychobiology of language : An introduction to dynamic philology*, Boston, Mass., Houghton-Mifflin.

8.
LAVORARE CON LEXICO3

Lexico3 è un software adatto per il trattamento lessicometrico di corpora molto grandi, che possono arrivare a centinaia di migliaia di occorrenze. È stato sviluppato da André Salem e poi ampliato dal gruppo di lavoro SYLED-CLA2T dell'Università Sorbonne Nouvelle 3, di cui fanno parte C. Lamalle, W. Martinez, S. Flaury.

L'interfaccia del programma è in francese, ma non ha limitazioni linguistiche per quanto riguarda i testi da analizzare. La sua impostazione – fondata sul conteggio delle occorrenze a partire dalla forma grafica delle parole – lo rende del tutto indipendente dalla lingua.

I testi sono elaborati così come si presentano, con pochissime preparazioni preliminari. Questo è un vantaggio, soprattutto in una fase di esplorazione del corpus, per compiere alcune osservazioni prima di procedere a decisioni e modifiche per le quali è preferibile e necessario ricorrere ad altri software. Ovviamente non mancano i limiti; primo tra tutti il fatto che il testo è sottoposto ad analisi senza alcuna procedura di "normalizzazione". Vedremo meglio in seguito cosa significa questo; per il momento teniamo presente che accenti, apostrofi e caratteri speciali tendono ad accrescere la variabilità del testo e, in certi casi – se non si introducono delle "correzioni" – possono portare ad alcuni errori nell'apprezzamento delle sue caratteristiche.

Lexico3 è oggetto di distribuzione commerciale; la copia di download può essere utilizzata gratuitamente per limitati scopi personali e didattici. Le informazioni dettagliate sul programma e sulla licenza si possono consultare sul sito: <http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/>.

8. 1. PREPARAZIONE DEL CORPUS

Il primo passo da fare è la preparazione del corpus in modo che possa essere "interpretato" correttamente dal software. Il testo deve essere in formato ASCII e pertanto deve essere stato salvato come "solo testo" e non con un formato *docx*, *doc*, *rtf* ecc. Tutti i software di elaborazione dei testi inseriscono delle informazioni di formattazione (attribuzione di una specifica veste grafica al documento, come il tipo di carattere, la dimensione, l'interlinea, ecc.) che non sono utili per l'analisi statistica delle forme grafiche. La base dei dati testuali non deve contenere informazioni che non siano utili all'analisi delle parole e al conteggio delle occorrenze. Ogni informazione non necessaria (e non "controllata" in termini sperimentali) può essere fonte di errore o di rumore. Per questo motivo nel capitolo 2 (§ 2.5) si è suggerito di utilizzare un editor di testi come TextPad che permette eventualmente di compiere le modifiche necessarie tenendo "sotto controllo" i codici dei caratteri.

La fase di preparazione del corpus è una fase complessa e di assoluta importanza nell'analisi testuale. Non vi sono regole generali che possono essere applicate meccanicamente. Ogni intervento di preparazione di un corpus comporta inevitabilmente delle decisioni che devono essere consapevoli e adeguate allo scopo che ci si prefigge. Per il momento ci siamo posti uno scopo esplorativo, pertanto decidiamo di lasciare il testo il più possibile inalterato, facendo solo attenzione a due aspetti formali:

- che non ci siano ambiguità nell'uso degli accenti e degli apostrofi;
- che ogni forma grafica sia codificata dal programma in modo univoco.

Nella tabella dei caratteri ASCII vi sono cinque codici utilizzati per identificare l'apostrofo:

- codice 39 '
- codice 96 `
- codice 145 ‘
- codice 146 ’
- codice 180 ´

Prima di proseguire occorre controllare che nel corpus gli accenti siano tutti nella forma del codice 39 e procedere alle sostituzioni necessarie.

Nei testi scaricati da WWW, da newsgroup o da e-mail accade spesso che l'apostrofo sia utilizzato come "accento" (*e'*, *perche'*, *puo'*, *piu'*). Se il corpus fosse lasciato in questa forma il programma non potrebbe riconoscere *e'* come la terza persona dell'indicativo presente del verbo <essere> e classificherebbe la forma *e'* come congiunzione *e* seguita da apostrofo.

Inoltre occorre fare molta attenzione alla ortografia degli accenti. In lin-

gua italiana l'accento grafico può essere **acuto** (per le vocali chiuse come in *perché*, *affichè*) oppure **grave** (per le vocali aperte come in *cioè*, *caffè*). A noi interessano in particolare gli accenti alla fine di alcune parole di uso comune. Di norma l'accento grafico in queste parole è sempre grave, salvo che per la vocale *e*.

Si scrivono con accento grave le parole: *cioè*, *è*.

Si scrivono con accento acuto le parole: *affinché*, *benché*, *cosicché*, *finché*, *perché*, *poiché*, *purché*, *sé* (pronome, ma non in *se stesso*), *né* (nella negazione).

Nella preparazione del corpus è bene uniformare tutto il testo secondo queste norme ortografiche con le opportune operazioni di "cerca e sostituisci" nell'editor dei testi.

Un altro problema è generato dalle lettere con carattere **maiuscolo**. Il caso è più complesso del precedente. Una forma con l'iniziale in maiuscolo sarà interpretata dal software come diversa rispetto alla forma identica in minuscolo: *Il* e *il*, *Hai* e *hai*, *Io* e *io*. Lo stesso per forme come *Attenzione*, *attenzione*, *ATTENZIONE*. L'ideale sarebbe poter cambiare in minuscolo tutte le lettere situate dopo il "punto" o all'inizio di un paragrafo e lasciare inalterate le altre. Questo è quanto sarà possibile fare con una procedura di "normalizzazione" (come vedremo in seguito). Per ora, una soluzione drastica e sofferta, ma necessaria per un software come Lexico3, è di abbassare tutte le lettere in minuscolo. In questo modo si perde qualche informazione sui nomi propri (*Mario*, *Rosa*, *Roma*, *Campania*) e sulle sigle (*ONU*, *UE*, *USA*) ma si riduce la variabilità del testo per ragioni esclusivamente grafiche. Se, per qualche ragione, si ritenesse necessario conservare l'informazione relativa alle parole con carattere maiuscolo, dopo aver controllato che nel testo non vi sia il carattere *, si può procedere (prima del loro abbassamento automatico) a una sostituzione (che sarà inevitabilmente non automatica) dell'iniziale maiuscola con *iniziale minuscola (**mario*, **rosa*, **roma*, **onu*, **ue*, **usa*).

Infine, un ultimo controllo dovrà verificare l'assenza nel testo delle "virgolette basse singole" (dette anche "minore" e "maggiore": < >) che serviranno esclusivamente per indicare le chiavi di partizione del corpus.

Il software deve essere messo in condizione di riconoscere le forme grafiche del corpus, pertanto oltre allo spazio ci saranno altri separatori come la punteggiatura, l'enfasi e l'apostrofo che non devono essere considerati come parte dell'alfabeto. Lexico3, di default, ne ha 20, ma la lista può essere modificata secondo le necessità:

```
. , : ; ! ? / _ – \ " ' ( ) [ ] { } § $
```

In alcuni casi questi marcatori (in particolare il "punto" e il "paragrafo" §) potranno essere utilizzati per delimitare una frammentazione del corpus che sarà utile per la sua esplorazione.

A questo punto potremmo procedere alla creazione del vocabolario, perché le chiavi di partizione non sono strettamente necessarie in questa prima fase. La loro assenza non impedisce al programma di individuare le forme grafiche e di calcolare le occorrenze. Certamente le informazioni che se ne ricavano sono riferite solo all'insieme del corpus senza alcuna modalità di descrizione. L'inserimento di una o più chiavi di partizione permette invece di effettuare confronti e analisi sulla distribuzione delle forme grafiche tra le diverse modalità di caratterizzazione dei testi all'interno del corpus.

8. 2. LE CHIAVI DI PARTIZIONE DEL CORPUS

La preparazione del corpus richiede ancora l'inserimento di qualche "chiave" o marcatore che permetta la suddivisione del corpus in parti e costituisca almeno una variabile con due modalità in base alla quale poter confrontare le occorrenze e quindi la frequenza delle parole secondo le parti prestabilite. Utilizzando le chiavi di partizione è possibile introdurre diversi "descrittori" dei testi (tematici o cronologici) che permettono di esaminarne la struttura. Per il corpus che stiamo esaminando i descrittori sono stati individuati in:

- *genere*, con le modalità "femmina", "maschio" e "indefinito";
- *operatore scolastico*, "sì", "no", "incerto".

L'inserimento delle chiavi avviene tra i segni di "virgolette basse singole" (< >): con il formato "nome della variabile" = "modalità", facendo attenzione che non ci siano spazi vuoti tra un carattere e l'altro. Le chiavi, di norma, sono inserite all'inizio di ogni testo:

Codifica di una chiave

Una chiave (ad esempio: <genere=femmina>) è composta di cinque elementi:

1	<	virgoletta bassa aperta
2	genere	il nome della chiave
3	=	il carattere "uguale"
4	femmina	il contenuto della chiave
5	>	virgoletta bassa chiusa

Può risultare utile inserire anche un "marcatore di paragrafo" che, in questo

caso, permette di individuare il messaggio. Il corpus *Bullismo.txt* si presenta come segue:

```
§ <genere=maschio> <operatore=sì>
sono un insegnante calabrese, e ho a che fare ogni giorno con
dei selvaggi in classe. ma attenzione, non sono selvaggi per-
ché risentono di una difficile situazione ambientale, ma sem-
plicemente perché seguono i modelli proposti dalla tv, da
questa vergognosa tv
§ <genere=indefinito> <operatore=incerto>
....cultura ed educazione, assieme, assenti in troppe fami-
glie... i ragazzi sono naturalmente crudeli, spesso inconsa-
pevolmente... i genitori non sono più in grado di comunicare
il rispetto per gli altri, soprattutto per i più
.............................................................
§ <genere=indefinito> <operatore=incerto>
chi vive nella scuola sa bene che "il bullo" è impunito. non
sarebbe il caso di sanzionare subito certi comportamenti? se
si continua a fare della sociologia spicciola "il bullo" si
sente autorizzato a fare ciò che vuole.
.............................................................
```

Il seguente è un corpus di poeti degli inizi del Novecento. Le chiavi di partizione sono: autore, regione di nascita, anno di pubblicazione dell'opera. Il marcatore di paragrafo delimita il verso.

```
§ <Autore=GOVONI> <Regione=EMILIA-ROMAGNA> <Anno=1911>
La bufera è lontana.
§ Sull'aia allegra cantano i galletti.
§ ........................
§ <Autore=PALAZZESCHI> <Regione=TOSCANA> <Anno=1909>
Clof, clop, cloch
§ cloffete,
§ cloppete
§ chchch...
§ E' giù nel
§ cortile,
§ ........
§ <Autore=SOFFICI> <Regione=TOSCANA> <Anno=1915>
La luce non è un mazzolino di fiori più sottili
§ Un ronzio di mosche d'oro e verdi il cielo
§ ..............................................
```

Le chiavi di partizione devono essere bilanciate; a ogni marcatore < deve seguire un marcatore > che chiude la chiave; come si è detto non devono esserci spazi nel contenuto delle chiavi e il nome deve essere separato dal contenuto con un segno di uguale. Gli errori sono segnalati dal software nel file di controllo *atrace.txt* che viene generato automaticamente nella cartella di lavoro.

8. 3. LA BARRA DEGLI STRUMENTI

A questo punto diamo una prima descrizione della barra degli strumenti. Sulla base di questa descrizione, che illustra anche le fasi fondamentali per una esplorazione del corpus, introduciamo alcuni concetti nuovi che saranno utilissimi in seguito e sui quali torneremo con maggiore attenzione.

Nuova base (segmentazione, frammentazione o *parsing*): è l'operazione di avvio dell'elaborazione del corpus e la sua trasposizione in una base di dati. Questo comporta il riconoscimento delle forme e dei separatori, la loro codifica (numerizzazione = ogni forma grafica viene fatta corrispondere ad un codice numerico che la identifica) e quindi la suddivisione dei testi in unità minime dotate di significato: le parole.
Il risultato di questa prima operazione di frammentazione sarà un file *Bullismo.dic* (vocabolario) che conterrà nella prima colonna le occorrenze, nella seconda il codice numerico e nella terza la forma grafica riconosciuta.

Apertura di una base preesistente: permette di aprire una sessione di lavoro già avviata in precedenza. Occorre selezionare il file (*Bullismo.par*) che contiene i parametri della sessione, cioè le informazioni relative al percorso e alle elaborazioni già compiute.

Concordanze: permette di definire il "contesto locale" in cui si trova una determinata parola che funge da polo (*pivot*). Il programma fornisce un elenco delle parole che precedono o seguono la forma pivot.

Segmenti ripetuti: il segmento in questo caso identifica un frammento di testo composto da tutte le disposizioni a 2, 3, ..., *q* parole che si ripetono nel testo.

Caratteristiche lessicometriche: permette di generare delle tabelle e di compiere alcune operazioni che descrivono le misure compiute sul lessico del corpus e sulle rispettive partizioni.

Statistiche delle partizioni: permette di selezionare la partizione del corpus rispetto alla quale visualizzare sui grafici le variazioni nell'uso delle parole ed effettuare determinati calcoli statistici in vista delle analisi successive.

Cartografia dei paragrafi: permette di esaminare graficamente la frammentazione del corpus selezionando i separatori e di visualizzare la distribuzioni delle forme (parole, segmenti o gruppi).

Gruppo di forme grafiche: permette di creare un "gruppo di forme" (per esempio un lemma) e di esaminarne la distribuzione anche dal punto di vista grafico secondo le partizioni del corpus e la frammentazione.

Portaparole: permette di memorizzare un oggetto (forma grafica, segmento, gruppi di forme) per un suo uso successivo (concordanze, rappresentazioni grafiche, ecc.). L'oggetto è immagazzinato trascinandolo con il mouse nel "cubetto" rosso e recuperato nello stesso modo per essere utilizzato in un altra funzione.

Mosaico: permette di riorganizzare le diverse applicazioni su uno stesso foglio di lavoro.

Sposta verso un altro foglio: permette di muoversi tra i fogli di lavoro.

Nuovo foglio: permette di creare un nuovo foglio di lavoro che si aggiunge al precedente con una "linguetta" alla destra del finestra principale.

Aggiungi al rapporto finale: tutte i grafici e le tabelle prodotte nel corso dell'esplorazione ed elaborazione dei dati testuali possono essere raccolte in un file in formato HTML. Basta cliccare su questa icona per aggiungere il foglio di lavoro al rapporto.
Il rapporto viene salvato in una cartella (nome del file di lavoro, data e ora) dalla finestra principale cliccando sulla scheda *Rapport* e poi su

Enrégistrer. Fino a quando non si è fatta pratica sufficiente con questa procedura è consigliabile salvare il rapporto nella cartella principale del programma (di solito con il percorso "C: Programmi/Lexico3/Rapport".

Opzioni: permette di intervenire sulla dimensionamento di grandi corpora da analizzare.

Modifica: permette di aprire i file del Rapporto o altri documenti, di visualizzarli e modificarli.

Esci: bisogna ricordarsi di salvare i risultati nel Rapporto prima di uscire dal programma.

Finestre aperte: visualizza un elenco delle finestre di lavoro dal quale selezionare la finestra da portare in primo piano.

8. 4. FRAMMENTAZIONE DEL CORPUS E FORMAZIONE DEL VOCABOLARIO

Ora possiamo iniziare l'esplorazione del corpus *Bullismo.txt* che si sarà copiato in una "cartella di lavoro" creata appositamente per contenere i file generati da Lexico3 durante l'esecuzione. Cliccando sull'icona **Nuova base** si apre nel modo consueto il file *Bullismo.txt* e ci viene chiesta conferma dei separatori (che volendo possiamo modificare). Cliccando su *OK* si esegue il programma.

Lexico3, prima di tutto, verifica se ci sono errori nei marcatori e nella codifica delle chiavi. Se c'è qualcosa che non va (spazi bianchi nelle chiavi, virgolette basse in apertura e non in chiusura o viceversa, ecc.) ci chiede di esaminare il file *atrace.txt* (generato nella cartella di lavoro) in cui vengono rilevati gli errori per permetterne la correzione.

Se non ci sono errori viene generato un output nella finestra *Dictionnaire* (Vocabolario; fig. 8.1). Nella prima colonna sono elencate le forme in ordine di frequenza. Nella seconda colonna sono elencate le rispettive frequenze.

Apprendiamo immediatamente che nel nostro corpus vi sono 8.344 forme. La forma più frequente è *di* con 1.751 occorrenze.

Fig. 8.1 – Vocabolario del corpus *Bullismo*

Le forme grafiche, oltre i numeri e i caratteri speciali, comprendono anche i 14 separatori individuati e posti in fondo all'elenco delle forme:

```
. , : ; ! ? / _ – " ' ( ) §
```

Nella finestra *Dictionnaire* possiamo cliccare su *Formes (ordre léxicometrique)* e otterremo un ordinamento delle forme grafiche secondo l'ordine alfabetico: *Formes (ordre léxicographique)*, e viceversa. Convenzionalmente i numeri e i caratteri speciali, in un ordine alfabetico, sono collocati prima delle lettere.

8. 5. ANALISI DELLE PARTIZIONI DEL CORPUS

Per continuare con l'esplorazione del corpus, a questo punto possiamo esaminare le occorrenze secondo le partizioni che lo compongono e ottenere molte altre informazioni. Cliccando sull'icona **Statistiche delle partizioni** possiamo selezionare la partizione "operatore" e aprire (*Créer*) una finestra *Graphique de ventilation pour la partition: data* (Grafico di distribuzione per la partizione: operatore). Vedremo successivamente come utilizzare questa finestra. Ora cliccando sull'icona **Caratteristiche lessicometriche** selezioniamo la partizione "operatore" e generiamo una tabella con alcune misure riferite a questa variabile.

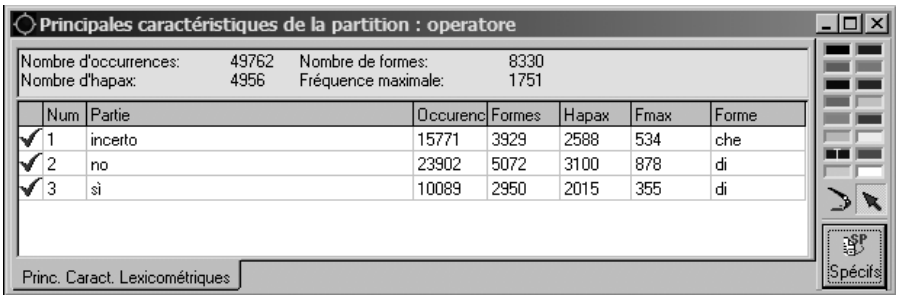


Fig. 8.2 – Caratteristiche lessicometriche della partizione: "operatore"

Vediamo in dettaglio che il numero di **occorrenze delle parole** (N) conteggiate è pari a 49.792. Le **parole distinte** (V) sono le forme grafiche diverse (8.330), ognuna delle quali appartiene a una classe *ì-esima* di occorrenze. Ad esempio *scuola* (fig. 8.1) appare 298 volte e appartiene alla 298-esima classe di occorrenze. La forma *mi* appare 181 volte e appartiene alla 181-esima classe di occorrenze insieme alla forma *solo* che appare ugualmente con 181 occorrenze.

Le parole distinte, come si è detto, sono meno delle forme (8.344) indicate sul margine inferiore della finestra, perché queste ultime comprendono i 14 separatori posti in fondo all'elenco. Pertanto da Bolasco (1999, p. 187) prendiamo la seguente definizione:

> Più in generale si indichi con V_i il numero di parole diverse che appaiono (o ricorrono) *i* volte in un vocabolario. V_1 rappresenta quindi l'insieme delle parole che appaiono una sola volta, ossia l'insieme degli hapax di un testo, V_2 quelle che ricorrono due volte ecc. Vale la relazione seguente:
>
> $$V_1 + V_2 + V_3 + ... + V_{fmax} = V$$
>
> dove f_{max} esprime il valore delle occorrenze della parola con il maggior numero di occorrenze del vocabolario.

Gli **hapax** (*hapax-legomena*) sono le parole che compaiono una sola volta nel corpus; mentre le parole che compaiono due volte sono dette *dis-legomena* (Tuzzi, 2003, p. 73). La classe f_{max}, come le altre immediatamente successive nella tabella di frequenza delle parole, è formata da una sola parola.

Come si può osservare il rapporto $(V/N)\times100$ può rappresentare un misura della **estensione lessicale**, cioè quante parole distinte vi sono in un corpus rispetto al totale delle occorrenze. In questo caso questa misura è pari al 16,74%.

Il rapporto $(V_1/V)\times100$ rappresenta una misura della **ricercatezza del linguaggio**, cioè quante parole compaiono una sola volta rispetto al totale delle parole distinte. In questo caso la "ricercatezza" è del 59,49%.

Entrambe queste misure sono importanti, come vedremo, per valutare la validità di un corpus ai fini di un'analisi statistica. Un linguaggio che si presenta con una grande estensione lessicale, ed è quindi dotato di un vocabolario molto ricco, necessita di un numero altrettanto grande di occorrenze. In altre parole il corpus deve essere un "campione" abbastanza rappresentativo del linguaggio affinché si possano applicare adeguatamente strumenti di analisi quantitativa.

Nello stesso tempo una presenza eccessiva di hapax rende del tutto inutile l'applicazione di tecniche che sono fondate sulla ricorrenza delle parole e sulle **co-occorrenze** delle stesse (e cioè sul loro apparire per un certo numero di volte insieme in uno stesso contesto).

Questi criteri di valutazione, in gran parte di carattere empirico, dovranno essere introdotti in seguito perché, per il momento, la fase esplorativa del corpus è ancora troppo "grossolana" per tali approfondimenti.

Se ci spostiamo sulla scheda *Rapport* cliccando sull'icona **Aggiungi al**

rapporto finale della barra degli strumenti e poi su *Enregistrer* (in basso a sinistra) possiamo inserire nel rapporto finale le principali caratteristiche lessicometriche del corpus e salvare il rapporto nella cartella *Rapport* (il percorso di solito è "C: Programmi/Lexico3/Rapport"). Il file viene salvato come *index.html* in una cartella che porta il nome del file di lavoro, la data e l'ora. Il file è leggibile da un qualsiasi browser di navigazione come Explorer o equivalente. Chi ha una maggiore confidenza con la gestione delle cartelle di Windows può agevolmente navigare alla ricerca della cartella di lavoro in cui è collocato il file *Bullismo.txt* (spesso, per comodità, si tratta del desktop) e salvarvi direttamente questo rapporto e quelli che saranno generati successivamente.

8. 6. GRAFICO DI DISTRIBUZIONE PER LA PARTIZIONE

Con l'operazione **Statistiche delle partizioni** era stata generata una finestra *Graphique de ventilation pour la partition: operatore* (Grafico di distribuzione per la partizione: operatore). Cliccando sull'icona **Finestre aperte** della barra degli strumenti possiamo selezionare la finestra del grafico e portarla in primo piano. Ora possiamo selezionare la parola *scuola* tra le forme del vocabolario, a sinistra del monitor, e trascinarla nella finestra del grafico ottenendo così la rappresentazione di essa nelle tre parti (o testi) in cui è suddiviso il corpus secondo la variabile "operatore". La rappresentazione è espressa in termini di frequenze relative o meglio di occorrenze normalizzate su base 10.000 in modo da poter confrontare tra di loro i testi come se avessero la stessa dimensione (10.000 occorrenze). Nel corpus la frequenza normalizzata di *scuola* è $(298/49.762)\times10.000 = 59{,}88$. Come possiamo osservare, nella parte del corpus che raccoglie i messaggi degli operatori scolastici la frequenza normalizzata è di 90 (fig. 8.3). Le frequenze assolute si possono esaminare marcando l'indicatore *absolues* alla base del grafico.

Il grafico (e in generale tutte le finestre di lavoro che appaiono a destra del vocabolario) può essere aggiunto al rapporto cliccando sulla barra di trascinamento (per renderlo attivo) e poi sull'icona **Aggiungi al rapporto finale** della barra strumenti. Aprendo la scheda *Rapport* si può verificare la presenza del grafico e salvarlo con *Enregistrer*.

Per ciascuna partizione prevista nella descrizione del corpus ("genere", "operatore"), cliccando sull'icona **Statistiche delle partizioni** e selezionando la partizione di interesse, si apre una finestra con il grafico corrispondente.

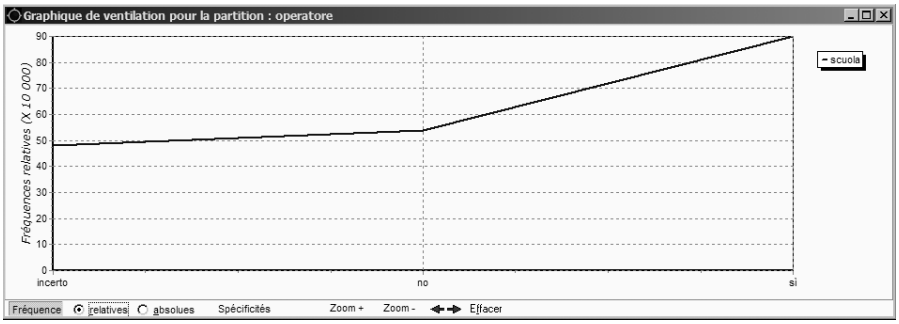


Fig. 8.3 – Grafico di distribuzione per la partizione: "operatore"

8.7. ANALISI DELLE SPECIFICITÀ

La specificità consiste nella estrazione delle forme "anormalmente" frequenti in una parte del corpus. Un test fondato sulla legge ipergeometrica permette di selezionare le parole la cui frequenza in una parte è sensibilmente superiore (o inferiore per le parole "anti-caratteristiche") alla frequenza media nel corpus (Lebart e Salem, 1994). Da **Statistiche delle partizioni** selezioniamo la partizione "genere" ottenendo la finestra corrispondente alle principali caratteristiche della partizione (fig. 8.4).

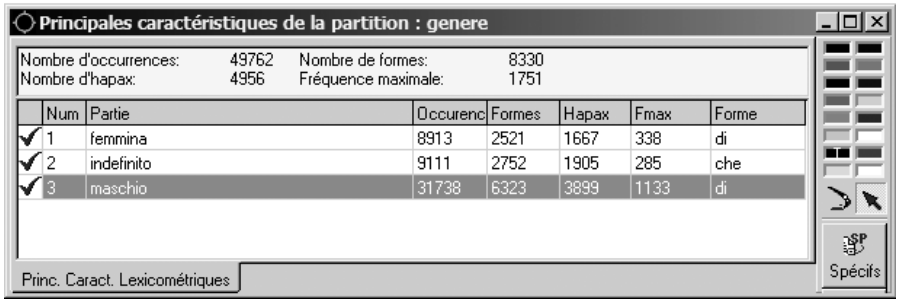


Fig. 8.4 – Caratteristiche lessicometriche della partizione: genere

Per calcolare la specificità delle parole nei messaggi imputabili ai maschi rispetto agli altri messaggi, si seleziona la parte 3 e poi si clicca sull'icona **Spécifs** per richiamare la finestra con i parametri per il calcolo. Di default i parametri

sono prefissati a una soglia di probabilità del 5% per le forme con frequenza superiore a 10 nel corpus.

Il risultato (fig. 8.5) appare nella finestra a sinistra come specificità positiva (o negativa).

Navigation | Rapport

Dictionnaire | Spécifs - Part : genere

Corpus de référence : indefinito, maschio,
Parties sélectionnées : maschio,
Spécificités (•) positives ( ) négatives

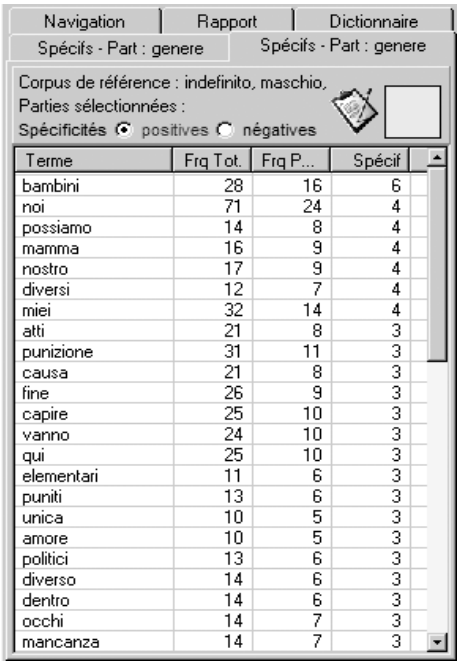
Terme	Frq Tot.	Frq P...	Spécif
prof	39	35	5
stato	62	54	5
video	40	35	4
un	722	498	4
dove	70	56	4
lo	131	99	4
anche	243	178	4
professori	56	44	3
nei	51	40	3
faccia	22	19	3
istituzioni	14	13	3
posto	18	16	3
ragazzo	50	41	3
senso	37	31	3
figlio	36	29	3
disabile	15	14	3
ciò	46	38	3
ti	33	28	3
mentre	14	13	3
solidarietà	10	10	3
ed	129	95	3
nel	94	69	3
è	781	531	3
verso	23	19	2
attenzione	19	16	2
ha	152	108	2
amici	12	11	2
mia	61	46	2
li	53	40	2
atti	21	9	-2
sotto	17	7	-2

Fig. 8.5 – Analisi di specificità: "maschi"

La forma grafica *prof* è presente con 39 occorrenze nel corpus, di cui 35 sono concentrate nei messaggi imputabili ai maschi. L'indice di specificità, segnalato da un valore 5 indica che la probabilità di ottenere uno scarto (tra la frequenza attesa della parola e la effettiva frequenza della parola nel testo 3: maschi) su-

periore o uguale a quello osservato è inferiore a un valore pari a 10^{-5} (cioè 0,00001). Si direbbe pertanto che i maschi nei loro messaggi tendano a mettere in evidenza il ruolo svolto dai *professori* e dalle *istituzioni* nella vicenda, i *video* incriminati, gli *studenti* e le forme: *ragazzo*, *figlio*, *disabile*.

Per calcolare la specificità delle parole nei messaggi imputabili alle femmine, si seleziona la parte 1 e poi si clicca sull'icona **Spécifs** proseguendo nello stesso modo. Il risultato (fig. 8.6) indica che le femmine nei loro messaggi tendono a mettere in evidenza *bambini*, *mamma*, *punizione*, *puniti* e *amore*, con una presenza significativa del pronome personale *noi* e degli aggettivi possessivi *nostro* e *mio*.


Navigation | Rapport | Dictionnaire

Spécifs - Part : genere | Spécifs - Part : genere

Corpus de référence : indefinito, maschio,
Parties sélectionnées :
Spécificités (•) positives ( ) négatives

Terme	Frq Tot.	Frq P...	Spécif
bambini	28	16	6
noi	71	24	4
possiamo	14	8	4
mamma	16	9	4
nostro	17	9	4
diversi	12	7	4
miei	32	14	4
atti	21	8	3
punizione	31	11	3
causa	21	8	3
fine	26	9	3
capire	25	10	3
vanno	24	10	3
qui	25	10	3
elementari	11	6	3
puniti	13	6	3
unica	10	5	3
amore	10	5	3
politici	13	6	3
diverso	14	6	3
dentro	14	6	3
occhi	14	7	3
mancanza	14	7	3

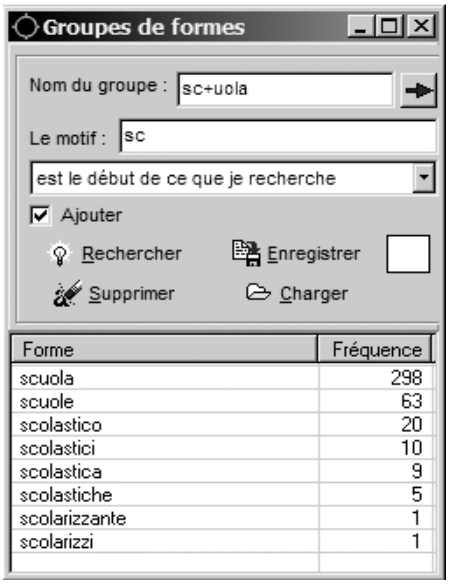
Fig. 8.6 – Analisi di specificità: "femmine"

Il marcatore di "spunta" nella prima colonna della figura 8.4 indica che la parte è inclusa nel conteggio delle occorrenze e pertanto nell'analisi delle specificità. Se si esclude dal calcolo la parte 2 il confronto viene effettuato tra maschi e femmine escludendo i messaggi dai quali non è stato possibile rilevare il genere. Il risultato di ciascuna analisi delle specificità può essere aggiunto al rapporto cliccando sull'icona *Rapport* (figg. 8.5 e 8.6, in alto a destra).

8. 8. RAGGRUPPAMENTI DI FORME GRAFICHE

Lo strumento **Gruppo di forme** (*Types généralisés – Tgens*) permette di raggruppare forme grafiche che hanno qualche proprietà in comune. La costruzione del gruppo può essere fondata su vari criteri. I più comuni sono la classificazione tematica (che tiene conto delle proprietà semantiche delle parole) e la classificazione grammaticale (in base alle loro proprietà morfologiche e sintattiche). In quest'ultimo caso si parla più propriamente di **lemmi** e di **lemmatizzazione**, una procedura essenziale e complessa che ha come obiettivo l'individuazione delle unità minime significanti (**lessie**). Ad esempio, le coniugazioni dei verbi producono forme flesse che possono essere classificate in uno stesso lemma, come *capisco*, *capisci*, *capisce* in <capire>.

Una categoria tematica invece raggruppa le forme che sono riconducibili a un contenuto comune per poterne apprezzare la presenza nelle diverse parti del corpus. Ad esempio possiamo raggruppare nella forma unica *sc+uola* tutte le parole che iniziano con *sc* e che sono riferite all'ambiente scolastico (fig. 8.7). Le liste generate possono essere salvate in file separati e ricaricati al momento opportuno.



Forme	Fréquence
scuola	298
scuole	63
scolastico	20
scolastici	10
scolastica	9
scolastiche	5
scolarizzante	1
scolarizzi	1

Fig. 8.7 – Gruppi di forme

Per procedere, cliccare sull'icona **Gruppo di forme grafiche** della barra strumenti. La finestra di dialogo ci offre diverse opzioni. Dopo aver selezionato il criterio (in questo caso *est le début de ce que je recherche*, cioè "è l'inizio di ciò che sto cercando"), nel campo *Le motif* inseriamo la radice *sc*. Cliccando su *Rechercher* nella finestra sottostante compaiono le forme richieste. Le forme individuate dal criterio di ricerca e che si vogliono eliminare dal gruppo, perché non coerenti con il criterio di raggruppamento (ad es. *scritto*, *schifo*, *scrive*, *scatole*, ecc.), si cancellano dal risultato selezionandole e cliccando su *Supprimer*.

Il gruppo può essere trascinato con il mouse dalla freccia rossa, a sinistra del nome, sull'icona **Portaparole** (cubo rosso) della barra degli strumenti oppure l'intero elenco può essere salvato nella cartella di lavoro, cliccando su *Enregistrer*, per essere recuperato successivamente cliccando su *Charger*.

Successivamente il gruppo può essere ripreso dallo strumento **Portaparole** e trascinato su una finestra *Graphique de ventilation* per esaminarne la distribuzione secondo la partizione del corpus prescelta (fig. 8.8).

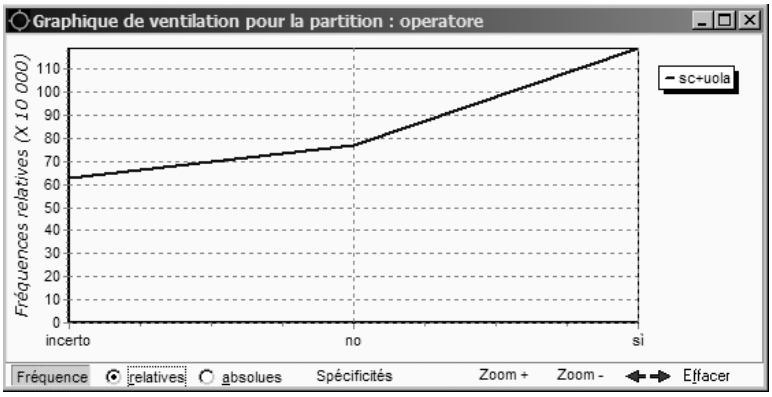


Fig. 8.8 – Grafico di distribuzione del gruppo *sc+uola* per la partizione: operatore

Sullo stesso grafico si può apprezzare anche la specificità di questo gruppo di forme per la parte: operatori scolastici (sì).

I gruppi possono essere formati anche da parole più "eterogenee". Con la marcatura della casella *Ajuter* le eventuali ricerche successive si aggiungeranno all'elenco precedente.

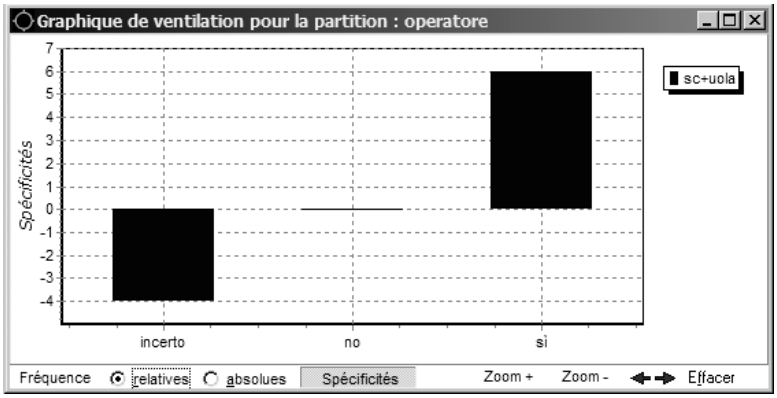


Fig. 8.9 – Grafico di distribuzione del gruppo *sc+uola* per la partizione: "operatore"

8. 9. CONCORDANZE

L'analisi delle concordanze è stato probabilmente uno dei primi strumenti utilizzati già dai filologi medievali per ricostruire il senso delle parole all'interno del contesto in cui la parola è situata e, di conseguenza, per essere di aiuto alla interpretazione del testo stesso. La concordanza più antica di cui abbiamo testimonianza risale al 1247, ed è stata realizzata nel convento di San Giacomo a Parigi sulla Vulgata, la Bibbia tradotta in latino da San Girolamo nel V secolo (Lana, 2004, pag. 212).

La concordanza di una parola è data da un elenco per forme in cui è visualizzata una certa porzione di testo prima o dopo la forma scelta come perno (o pivot). Questa, ad esempio, è la concordanza completa della parola *sonno* nella Divina Commedia di Dante Alighieri (fonte: htt://www.intratext.com):

```
 1 In, 1, 11 | tant'era pien di sonno a quel punto~
 2 In, 3, 136| e caddi come l'uom cui sonno piglia.~ ~
 3 In, 4, 1 | Ruppemi l'alto sonno ne la testa~
 4 In, 4, 68 | di qua dal sonno, quand'io vidi un foco~
 5 In, 25, 90 | pur come sonno o febbre l'assalisse.~ ~
 6 In, 33, 26 | già, quand'io feci 'l mal sonno~
 7 In, 33, 38 | pianger senti' fra 'l sonno i miei figliuoli~
 8 Pu, 9, 11 | vinto dal sonno, in su l'erba inchinai~
 9 Pu, 9, 33 | che convenne che 'l sonno si rompesse.~ ~
10 Pu, 9, 41 | mi fuggì 'l sonno, e diventa' ismorto,~
11 Pu, 9, 63 | poi ella e 'l sonno ad una se n'andaro».~ ~
12 Pu, 15, 119| far sì com'om che dal sonno si slega,~
13 Pu, 15, 123| a guisa di cui vino o sonno piega?».~ ~
14 Pu, 17, 40 | Come si frange il sonno ove di butto~
```

```
15 Pu, 27, 92 | mi prese il sonno; il sonno che sovente,~
16 Pu, 27, 92 | mi prese il sonno; il sonno che sovente,~
17 Pu, 27, 113| e 'l sonno mio con esse; ond'io leva'
18 Pu, 30, 104| sì che notte né sonno a voi non fura~
19 Pu, 32, 72 | del sonno e un chiamar: «Surgi: che
20 Pa, 12, 65 | vide nel sonno il mirabile frutto~
```

Il presupposto sul quale si fonda la concordanza è che il testo sia costruito sulla base di un principio unitario e che utilizzi le parole in modo coerente, sebbene all'interno delle consuete ambiguità sintattiche e lessicali. La parola *sonno* è portatrice sempre dello stesso significato. Ma questo non accade con *scorte* che nella riga 1 è una forma flessa del verbo *scorgere*, mentre in 2 e 4 stanno per *guide* e in 3 per *guidate:*

```
1 In, 1, 9 | l'altre cose ch'i' v'ho scorte.~ ~
2 Pu, 16, 45| tue parole fier le nostre scorte».~ ~
3 Pu, 21, 21| ha per la sua scala tanto scorte?».~ ~
4 Pu, 27, 19| Volsersi verso me le buone scorte;~
```

Senza l'uso delle concordanze sarebbe stato impossibile ricostruire i tre diversi significati di *scorte*, che sono parole omografe.

Ecco quindi che attraverso le concordanze possiamo disambiguare le forme grafiche, sia dal punto di vista semantico (significati diversi con forme omografe) che sintattico-grammaticale (categorie grammaticali non classificabili automaticamente, come nel caso di *quando* che è avverbio, congiunzione e nome) oppure in forme miste come *stato* (participio passato di *essere*/*stare* e nome per *Stato)* oppure *sole* (forma femminile dell'aggettivo solo; forma plurale del nome sola; e nome della stella più vicina alla Terra).

Con lo strumento **Concordanze** è possibile visualizzare le occorrenze di una forma o di un gruppo di forme (*Tgens*) all'interno di un contesto rappresentato dalle parole adiacenti a un termine scelto come parola perno (*pivot*). Lexico3 permette di selezionare le concordanze sulla base di un certo numero di caratteri prima e/o dopo la forma pivot che viene scritta nella finestra di dialogo *Forme* (per avere il risultato premere "invio" sulla tastiera).

Le occorrenze sono visualizzate secondo l'ordine di apparizione della forma pivot nel testo (*Aucun*), secondo l'ordine alfabetico della forma grafica che precede la forma pivot (*Avant*) o secondo l'ordine alfabetico della forma che la segue (*Après*). Il numero di caratteri prima e dopo la forma pivot può essere modificato (*Largeur*) e la visualizzazione viene aggiornata cliccando sull'icona verde *Réfraichir* sull'estrema destra. La forma pivot può anche essere trascinata con il cursore dalla finestra *Vocabolario* alla finestra del risultato con un effetto più immediato. Un gruppo di forme trascinato con il cursore nello strumento **Portaparole** può essere a sua volta trascinato dal cubo rosso nella

finestra delle concordanze come forma pivot.

Per una visualizzazione delle concordanze secondo la partizione del corpus, selezionare la partizione di interesse (data, genere, operatore) nella finestra di dialogo *Regroupement* (fig. 8.10). Le concordanze risultanti possono essere salvate nel rapporto per analisi successive.

Fig. 8.10 – Concordanza della forma pivot *punizione*

Le concordanze sono di grande utilità sia nell'analisi lessicale che nell'analisi del contenuto. Evidentemente l'interesse per le concordanze comporta uno slittamento del metodo verso l'analisi semi-automatica dei co-testi, prestando attenzione alle relazioni che le singole parole hanno sia con la lingua cui appartengono che con il contesto specifico in cui compaiono. Le relazioni possono essere di due tipi: sintagmatiche o paradigmatiche (Violi, 1997, p. 36).

La **relazione sintagmatica** prende in esame le relazioni che la forma grafica intrattiene con le altre forme che la precedono e la seguono. La relazione sintagmatica si manifesta nella lingua e ha un carattere, per così dire, strutturale, che prescinde da qualsiasi considerazione di carattere psicologico da parte di chi scrive. I rapporti sintagmatici dipendono dalle regole grammaticali

e offrono un contributo fondamentale per individuare le omografie (forme che si scrivono nello stesso modo ma hanno significati diversi: *stato*, "participio passato del verbo essere", e *stato*, "persona giuridica territoriale sovrana"; *calcio*, "gioco del pallone", e *calcio*, "elemento chimico"). Esaminando le forme che precedono *stato*, ad esempio, è agevole disambiguarne il significato: il sostantivo è spesso preceduto dall'articolo (*lo stato*), mentre il participio passato è necessariamente preceduto da un verbo (*è stato*, *sono stato*).

La **relazione paradigmatica** (o associativa) è riferita a un insieme di parole che possono essere commutate conservando il senso generale del contesto in cui sono inserite. Per esempio, nella figura 8.10 possiamo prendere in esame la parola *punizione* e chiederci se i co-testi cambierebbero di senso sostituendo *punizione* con altre forme lessicali che appartengono allo stesso modello (paradigma), come *pena*, *espiazione*, *penitenza*, *intervento disciplinare*, *castigo*, *sberla.* Il significato di queste parole non è lo stesso, ma riportano al campo semantico della "punizione". Con questa tecnica si possono individuare delle sinonimie di fatto che possono avere una valenza tematica. Ad esempio, nel corpus *Bullismo* sono presenti *pena* (16), *sberle* (3), *botte* (4).

I rapporti paradigmatici possono dare vita a famiglie associative molto complesse e personalizzate che si basano su affinità **foniche** (dann*osa*, dolor*osa*, paur*osa*,) **morfologiche** (ben*issimo*, bell*issimo*, bull*issimo*) o **semantiche** (*docente*, *maestra*, *insegnante*, *professore*, *prof*). Attraverso le concordanze si identificano gli usi "figurati" delle parole, per esempio il ricorso alla metafora (*un povero ragazzo che ha perso la testa*, *io ci scommetto la testa)* rispetto a un uso non traslato della parola come in *la testa presa di mira con una sedia* oppure *pistole giocattolo puntate alla testa dei professori.*

8. 10. CARTOGRAFIA DEI PARAGRAFI

Questa procedura permette di visualizzare la dispersione delle forme grafiche all'interno dei paragrafi in cui è stato frammentato il corpus. Cliccando sull'icona **Cartografia dei paragrafi** sulla barra degli strumenti si richiama una finestra di dialogo con la tabella dei separatori di sezione rinvenuti nel testo. Tra i separatori, oltre ai segni di punteggiatura, troviamo anche il marcatore di paragrafo § che abbiamo inserito volontariamente all'inizio di messaggio proprio per questo scopo. L'inserimento del marcatore di paragrafo è del tutto strumentale. In altre situazioni può essere utile per tenere traccia di altre frammentazioni del corpus. Selezionando il marcatore e cliccando sul bottone *Créer* viene aperta una finestra in cui sono visualizzati graficamente i 277 mes-

saggi in cui è suddiviso il corpus.
Selezionando la forma *bullismo* tra le forme della finestra *Vocabolario* e trascinandola nella finestra del grafico otteniamo una rappresentazione della posizione che la forma occupa nel corpus e dei messaggi che trattano l'argomento in modo esplicito. Come possiamo notare (fig. 8.11) la forma selezionata è più frequente tra il messaggio 51 e il messaggio 100. Cliccando su uno dei quadratini colorati si visualizza, nella finestra in basso, il testo corrispondente.

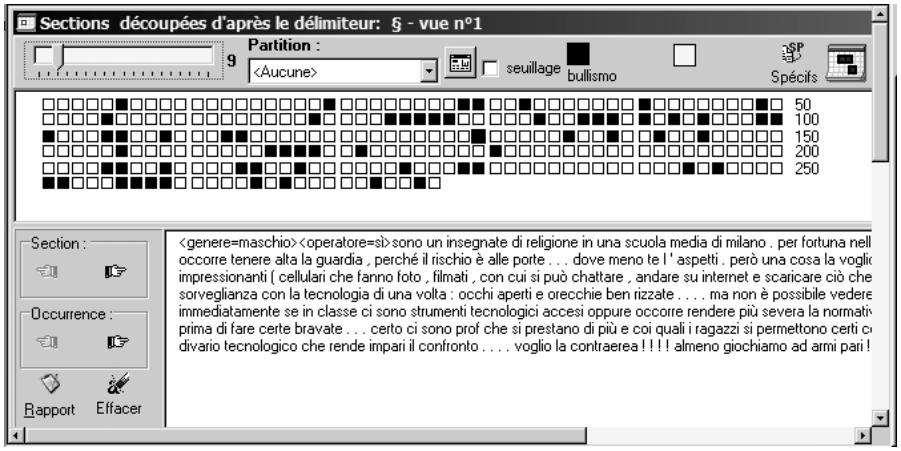


Fig. 8.11 – Cartografia dei messaggi per la forma: *bullismo*

8. 11. ALTRE FUNZIONI E SALVATAGGIO DEL RAPPORTO

Il software Lexico3 permette di compiere altre esplorazioni del testo come l'estrazione dei **segmenti ripetuti**, stringhe di testo che si ripetono un certo numero di volte nel corpus (vedi § 10.2), e analisi statistiche come il calcolo del grado di accrescimento del vocabolario e l'analisi delle corrispondenze. Quest'ultima è un'analisi di statica multivariata che non è trattata in questo manuale ed è svolta in modo più efficiente e completo da altri software (ad esempio SPAD).

Lexico3 offre l'ineguagliabile vantaggio, tra le altre cose, di un rapido accesso ai testi in vista di modifiche preliminari per la loro preparazione. La sua facilità d'uso e le funzioni lessicometriche principali, unite alla semplicità di accesso al testo per gli approfondimenti semi-automatici, lo rendono particolar-

mente adatto per integrare osservazioni già compiute con i software CAQDAS. La lettura del manuale, disponibile con la procedura di installazione, permetterà di conoscere in modo più dettagliato e puntuale le varie funzioni del programma.

La scheda *Navigation*, nella finestra *Vocabolario* a sinistra del monitor, permette di navigare agilmente tra le fasi dell'analisi svolta, mentre l'icona **Mosaico** sulla barra degli strumenti permette di riorganizzare e visualizzare le finestre di lavoro solo temporaneamente chiuse. Nel corso della utilizzazione del programma si possono generare una quantità considerevole di finestre e si potrebbe avere una sensazione di smarrimento. L'icona **Finestre aperte** sulla barra degli strumenti permette, in qualsiasi momento, di visualizza l'elenco delle finestre di lavoro e di selezionare la finestra di nostro interesse.

Va ricordato che in genere le finestre di lavoro attive (con la barra del titolo accesa con un clic del mouse) sono salvabili nel Rapporto cliccando sull'icona **Aggiungi al rapporto finale**. Dalla scheda *Rapport* della finestra *Vocabolario* il rapporto completo può essere salvato in un file formato HTML nella cartella di lavoro cliccando su *Enregistrer* e successivamente aperto e modificato con un qualsiasi browser come Internet Explorer, Netscape Navigator, Opera, Mozilla.

RIFERIMENTI BIBLIOGRAFICI

BOLASCO S. (1999) *Analisi multimensionale del dati. Metodi, strategie e criteri d'interpretazione*, Roma, Carocci (II ed. 2004).

LANA M. (2004) *Il testo nel computer. Dal web all'analisi dei testi*, Torino, Bollati Boringheri.

LEBART L., SALEM A. (1994) *Statistique textuelle*, Paris, Dunod (disp. integralmente on line: http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/.

TUZZI A. (2003*) L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*, Roma, Carocci.

VIOLI P. (1997) *Significato ed esperienza*, Milano, Bompiani.

9.
LAVORARE CON TALTAC²: IL TRATTAMENTO DEL TESTO

TaLTaC (Trattamento Automatico Lessico-Testuale per l'Analisi del Contenuto) è un software per l'analisi testuale sviluppato da Sergio Bolasco, Francesco Baiocchi e Adolfo Morrone (http://www.taltac.it).

Seguendo le fasi principali di trattamento suggerite dal software è possibile delineare, a livello introduttivo, alcune strategie di base utili per qualsiasi analisi automatica dei dati testuali. Alcune funzioni possono risultare utili anche per l'analisi semi-automatica. TaLTaC integra risorse e tecniche che fanno riferimento sia alla statistica che alla linguistica. A sua volta è predisposto per ricevere dati testuali che provengono da altri software, sia per esportare dati testuali disposti in matrici per l'analisi multidimensionale.

La versione di riferimento per questo manuale è la versione 2: TaLTaC². Tutte le informazioni qui contenute sono da ritenersi come puramente introduttive. TaLTaC² è un software complesso dotato di moltissime funzioni e strumenti che qui non possono essere affrontati in modo esaustivo e che sono oggetto invece della Guida ufficiale che accompagna la distribuzione del programma.

9. 1. LA BARRA DEGLI STRUMENTI

Le icone sulla barra degli strumenti permettono di identificare le fasi principali dell'analisi.

Esplora il corpus: permette di leggere il corpus, dopo averlo caricato, con tutte le sue caratteristiche descrittive (frammentazione, variabili e sezioni).

Normalizzazione: fase di pre-trattamento in cui si cerca di ridurre la variabilità del testo con l'applicazione di alcune procedure standard.

Misure lessicometriche: fase di analisi del Vocabolario generato durante la fase di normalizzazione.

Segmentazione: fase di estrazione dei segmenti ripetuti.

Lessicalizzazione: fase di identificazione delle sequenze di forme (segmenti) definite dall'utente e di trasformazione di esse in forme grafiche semplici.

Tagging grammaticale: fase di riconoscimento delle forme del Vocabolario e applicazione delle categorie grammaticali.

Tagging semantico: fase di associazione di un'etichetta tematica alle forme del Vocabolario ritenute rilevanti per un certo oggetto di studio.

Confronto lessici: fase di confronto con i lessici di frequenza (che fanno parte delle risorse statistico-linguistiche esterne) per la estrazione del linguaggio peculiare.

Concordanze: permette di eseguire l'analisi delle concordanze semplici e per categorie.

Estrazione d'informazione (IE): modulo di ricerca nel corpus con espressioni regolari definite dall'utente.

Accesso ai Database: modulo in cui sono raccolte le tabelle/liste del Database di TaLTaC^2 e quelle generate durante la sessione di lavoro.

Azzera evidenziatore: permette di annullare le forme evidenziate nel corpus in seguito a una query.

Text-Data Mining: modulo di gestione e ricerca sulle liste selezionate.

Ordine crescente: permette di ordinare il contenuto di una colonna selezionata nella tabella attiva secondi valori numerici crescenti o l'ordine alfabetico dal primo all'ultimo grafema.

Ordine decrescente: permette di ordinare il contenuto di una colonna selezionata nella tabella attiva secondi valori numerici decrescenti o l'ordine alfabetico dall'ultimo al primo grafema.

Ricostruzione corpus: permette di generare un file in formato ASCII annotato con informazioni di tipo grammaticale o semantico eseguite nelle fasi di *tagging*.

9. 2. PREPARAZIONE DEL CORPUS

Il primo passo da compiere è di preparare il corpus per la lettura e acquisizione da parte di TaLTaC². Il corpus può essere acquisito in diverse modalità:

- in un singolo file di testo con l'inserimento dei marcatori richiesti;
- in un file di testo strutturato in campi;
- in una collezione di file in formato *txt*, *doc* o *rtf*.

Qui non sarà esaminata quest'ultima modalità, ma solo le prime due.

Il corpus in un singolo file di testo si presenta con questa sintassi:

Codifica tipo per un corpus composto di un singolo file di testo

```
****Descrizione frammento 1 *NomeVariabile=ValoreVariabile
++++Sezione1
Testo
++++Sezione2
Testo
```

"Descrizione del frammento", preceduto da quattro asterischi (****), identifica il frammento; i successivi "NomeVariabile=ValoreVariabile", preceduti da un asterisco (*), identificano le caratteristiche descrittive del frammento se-

condo le categorie di analisi individuate. L'asterisco che precede il nome della variabile indica al software che ciò che segue deve considerarsi una variabile. Il nome della variabile "NomeVariabile" non deve contenere spazi, punti, virgole, apici o asterischi, né i simboli di operatore algebrico +, -, / (addizione, sottrazione, divisione), mentre il valore/modalità (numerico o alfanumerico) della variabile può contenere qualsiasi carattere (spazi inclusi) eccetto l'asterisco.

Le "Sezioni", precedute da (++++), sono opzionali e identificano ulteriori frammentazioni del testo che possono risultare utili: paragrafi, parti specifiche di un articolo a stampa (titolo, occhiello, sottotitolo e corpo dell'articolo), diverse risposte a domande aperte di un questionario.

Pertanto il corpus *Bullismo* composto da un singolo file di testo con l'inserimento dei marcatori ha il seguente formato:

```
****Messaggio 1 *Genere=Maschio *Operatore=Sì
Sono un insegnante calabrese, e ho a che fare ogni giorno con dei
selvaggi in classe. Ma attenzione, non sono selvaggi perché risentono
di una difficile situazione ambientale, ma semplicemente perché
seguono i modelli proposti dalla tv, da questa vergognosa tv
****Messaggio 2 *Genere=Indefinito *Operatore=Incerto
.... cultura ed educazione, assieme, assenti in troppe famiglie ....
i ragazzi sono naturalmente crudeli, spesso inconsapevolmente .... i
genitori non sono più in grado di comunicare il rispetto per gli
altri, soprattutto per i più deboli ... naturalmente mi riferisco ai
genitori di cotanti figli, che confondono la liberalità con la
licenza .... più che biasimare e punire i ragazzi, bisognerebbe fare
dei corsi di rieducazione per i genitori ..........................
.................................................................
****Messaggio 277 *Genere=Indefinito *Operatore=Sì
A seguito di un grave episodio di violenza verificatosi tra 5 nostri
allievi (4 di questi si sono recati a casa di un compagno per
dirimere una stupida questione che in realtà è finita con
.................................................................
```

Il corpus in un file di testo strutturato in campi si presenta con questa sintassi:

Codifica tipo per un corpus composto di un file strutturato in campi

Desc. frammento	NomeVariabile	Sezione1	*Sezione2*
Frammento 1	ValoreVariabile	Testo sezione 1	*Testo sezione 2*
....	....	....	....
Frammento 2	*ValoreVariabile*	*Testo sezione 1*	*Testo sezione 2*

In questo caso il corpus *Bullismo.txt* strutturato in campi avrà una provenienza da un file Excel, dal quale sarà stato salvato in formato *txt* delimitato da tabulazione:

Messaggio	Genere	Operatore	Corpo del messaggio
Messaggio 1	Maschio	Sì	Sono un insegnante cala-brese, e ho a che fare ogni giorno ....
Messaggio 2	Indefinito	Incerto	....cultura ed educazio-ne, assieme, assenti in troppe .....
.....	.....	.....	....
Messaggio 262	Femmina	Sì	A seguito di un grave episodio di violenza verificatosi .....

Il file *Bullismo_TT2.txt* verrà copiato in una cartella di sessione che denominiamo *Bullismo_TT2* e che sarà destinata ad accogliere tutti gli output della sessione di lavoro. Spesso può risultare comodo aprire la cartella sul desktop, ma per facilitare l'accesso durante i primi passi di utilizzo del software è bene creare la cartella all'interno della cartella *Sessioni* creata da TaLTaC² all'atto della installazione. TaLTaC² si posiziona di default in questa cartella e ciò facilita le operazioni nella fase successiva.

9. 3. CREAZIONE DI UNA SESSIONE DI LAVORO

Dal menu **File** si sceglie *Nuova/Apri Sessione*, si apre la maschera, si spunta l'opzione *Nuova Sessione* e si preme *OK*. TaLTaC² apre una finestra di dialogo che permette di assegnare il nome *Bullismo* alla sessione e ai Database corrispondenti e di indicare la cartella di lavoro *Bullismo_TT2* in cui verranno salvati tutti i file della sessione. Questa operazione costruisce un ambiente di lavoro che riporta i riferimenti di percorso necessari nel Registro di Configurazione. In questo modo ogni volta che si vorrà ritornare alla sessione definita da quel nome basterà selezionarla nell'elenco delle Sessioni aperte, spuntando l'opzione *Apri sessione esistente*.

Dopo aver assegnato il nome, premendo su *Salva* si genera il Database di sessione: *Bullismo.tsdb2* e una cartella *DB_di_Sessione.tswf* che conterrà i file generati automaticamente da TaLTaC². Il contenuto di questa cartella non deve essere modificato dall'utente.

Per aprire una sessione sulla quale si è lavorato in precedenza è sufficiente spuntare l'opzione che richiama l'ambiente di lavoro già precostituito. A questo pun-

to, se si tratta di una nuova sessione, sarà necessario caricare il file *Bullismo_TT2.txt* che contiene il corpus preparato con i marcatori o strutturato in campi.

Dal menu **File** si sceglie *Corpus/Seleziona/Assembla/Struttura*, si apre la finestra *Selezione del Corp*us, ci si posiziona sulla scheda *File singolo* o *File strutturato in campi* (a seconda del modello di preparazione che si sarà seguito) e poi - tramite il pulsante *Sfoglia* - si accede alla cartella di lavoro per la selezione del file da caricare nel Database della sessione.

Nel nostro caso il corpus non è suddiviso in sezioni, pertanto cliccando su *OK* il programma prosegue chiedendo: "Vuoi eseguire il parsing?"; rispondendo "Sì" verrà completato il caricamento del corpus all'interno della sessione con la visualizzazione della maschera di *Definizione dei caratteri* che permette di identificare i caratteri alfabetici e i separatori (fig. 9.1).

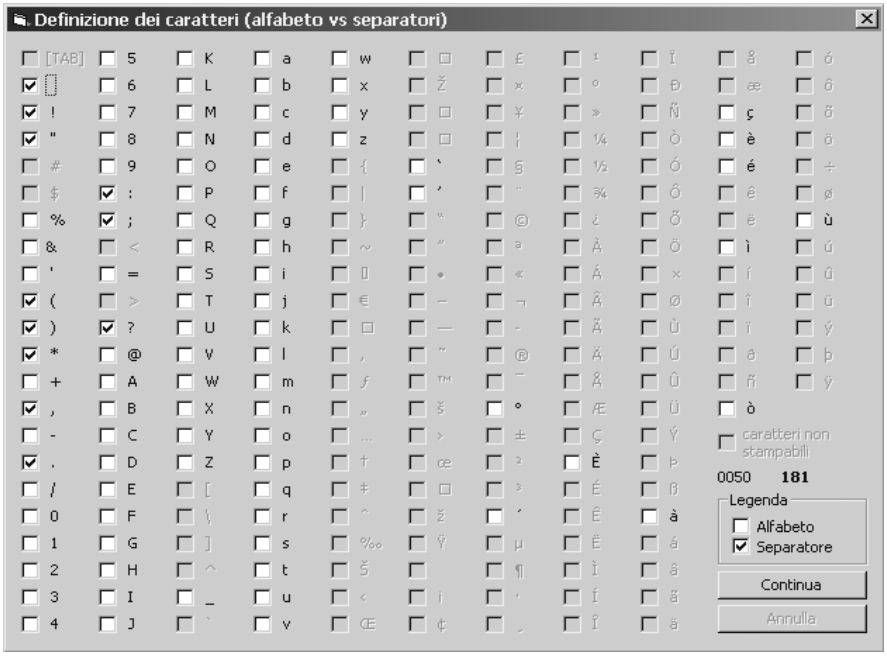


Fig. 9.1 – Finestra di definizione dei caratteri alfabetici e dei separatori

Il programma considera convenzionalmente come caratteri alfabetici tutti i caratteri che non vengono definiti come separatori. I caratteri separatori di default sono i seguenti:

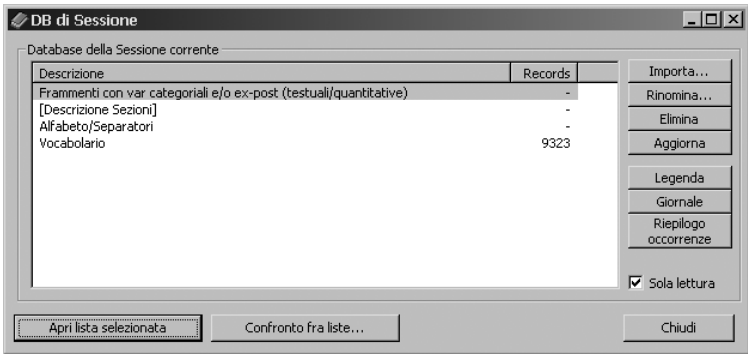
```
[TAB] [spazio] , . ; : ? ! ( ) [ ] { } \ / |
```

Altri caratteri (o simboli) possono essere definiti come separatori dall'utente stesso. La maschera, posizionando il cursore su un singolo carattere, fornisce l'occorrenza e il corrispondente codice ASCII. TaLTaC² decodifica i codici carattere ANSI di Windows pertanto, per i file che provengono da altri sistemi operativi o da Internet, si possono presentare gli stessi problemi di transcodifica già evidenziati per Lexico3 nel paragrafo 8.1. Tuttavia TaLTaC² è dotato di una procedura di trasformazione degli "apici difformi" (dei quali viene chiesta la conversione nell'apostrofo - codice 39 – al termine del parsing). La sostituzione degli accenti non viene eseguita in questa fase perché farà parte della procedura di normalizzazione.

Al termine del parsing il corpus è caricato nel Database di sessione. Cliccando su **Accesso ai Database** dalla barra degli strumenti si visualizza la finestra *Database della sessione corrente* (fig. 9.2) che contiene:

- *Vocabolario*: lista delle forme grafiche presenti nel corpus (9.323 prima della normalizzazione).
- *Alfabeto/Separatori*: caratteri presenti nel corpus con la occorrenze corrispondenti.
- *Frammenti con variabili categoriali e/o ex-post (testuali/qualitative)*: informazioni sulle variabili categoriali, se esistenti.
- *[Descrizione Sezioni]*: informazioni sulle sezioni, se esistenti.

Fig. 9.2 – Database di sessione



È possibile esaminare ciascuna di queste liste con l'opzione *Apri la lista selezionata*. Una prima esplorazione del *Vocabolario* (fig. 9.3) permetterà di osservare che il parsing ha dato come risultato 9.323 forme grafiche distinte, delle quali sono elencate le frequenze e la lunghezza in caratteri. Come vedremo, le for-

me sono destinate a ridursi in seguito alla procedura di normalizzazione. Selezionando la colonna CAT e cliccando su **Ordine decrescente** sulla barra degli strumenti, potremo renderci conto che, già in questa prima fase di pretrattamento del corpus, TaLTaC² ha etichettato i numeri con la categoria NUM. In seguito molte altre informazioni si aggiungeranno in questa colonna.

Fig. 9.3 – Vocabolario di base prima della normalizzazione

9. 4. FASE DI PRE-TRATTAMENTO: NORMALIZZAZIONE

Ora possiamo avviare la procedura di normalizzazione del corpus. La normalizzazione agisce sui caratteri alfabetici e consente di:

- trasformare gli apostrofi in accenti (*perche'* in *perchè*) e regolarizzare gli accenti gravi e acuti della grafia italiana (*perchè* in *perchê*);
- eliminare le fonti più frequenti di sdoppiamento del dato (presenza di maiuscole non significative in principio di periodo e non sempre necessariamente precedute da punto fermo, punto interrogativo o punto esclamativo);
- uniformare la grafia dei nomi propri, celebrità, sigle e altre entità.

Premendo il bottone **Normalizzazione** della barra degli strumenti, apriamo la finestra di dialogo *Normalizzazione del testo* (fig. 9.4). Da questa finestra è possibile eseguire una serie di operazioni che modificano la struttura del vocabolario. Non è agevole fornire una regola generale valida per tutte le analisi. Le scelte che si compiono in questa fase di normalizzazione sono determinate dagli obiettivi dell'utente che, con molta approssimazione possiamo distinguere

tra obiettivi "lessico-testuali" e di "analisi del contenuto". L'utente interessato ad analizzare il corpus da un punto di vista lessicale potrebbe escludere, in un primo tempo, dalla procedura di normalizzazione le locuzioni grammaticali, i gruppi nominali e le polirematiche (tab. 9.1). Mentre un utente interessato all'analisi del contenuto potrebbe procedere immediatamente alla applicazione di tutte le opzioni indicate nelle normalizzazioni basate su liste.

La nostra scelta in questo manuale è di fornire le indicazioni per una normalizzazione di base per l'analisi del contenuto. Ogni altra opzione deve essere eseguita consultando direttamente la *Guida* di TaLTaC².

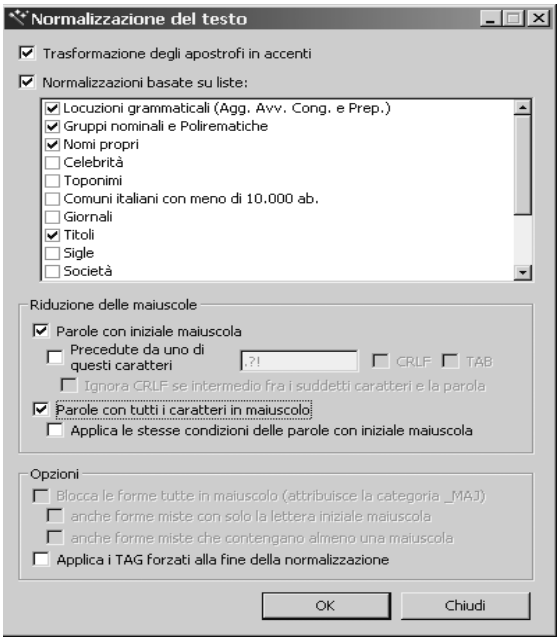


Fig. 9.4 – Normalizzazione del testo

La prima opzione *Trasformazione degli apostrofi in accenti* esegue alcune operazioni di sostituzione che avevamo dovuto eseguire manualmente nella preparazione del testo per Lexico3 (vedi paragrafo 8.1):

- *a'*, *A'* – sostituzione in *à*, *À* (a meno che non si tratti della parola *ca'*);
- *e'* – sostituzione con *è* (accento grave) o *é* (accento acuto) a seconda dei casi (a meno che non si tratti di *de'*);
- *E'* – sostituzione con *È*;

- *i'*, *I'* – sostituzione con *ì*, *Ì*;
- *o'* – sostituzione con *ò* (a meno che non si tratti della preposizione *co'*, oppure delle parole *po'* e *mo'*;
- *O'* – sostituzione con *Ò*;
- *u'*, *U'* – sostituzione con *ù* e *Ù*.

La *Normalizzazione basata su liste* si pone come obiettivo di categorizzare, già in questa fase di pre-trattamento, le forme delle quali si vuole conservare la specificità. La forma grafica *La Repubblica* è ben diversa da *la repubblica*; *Bossi* (cognome del segretario della Lega Nord Umberto Bossi) non può essere confuso con *bossi* (plurale di *bosso*, arbusto perenne sempreverde); *ONU* deve essere uniformato con *Onu*; e così via. Le informazioni sulle liste sono inserite nel Database di normalizzazione di TaLTaC[2]. Ovviamente non possono essere esaustive di tutte le fonti di ambiguità e variabilità del lessico. È sempre necessario che l'utente controlli attentamente la tabella vocabolario secondo le proprie esigenze e secondo le caratteristiche specifiche del corpus da analizzare.

Una delle funzioni principali in questa fase è il riconoscimento dei poliformi, locuzioni grammaticali (aggettivi, avverbi, congiunzioni e preposizioni), gruppi nominali e poliremathiche (tab. 9.1).

Tab. 9.1 – Esempi di poliformi (Bolasco, 1999, p. 195)

1) locuzioni grammaticali con funzioni di:	
- *avverbi:*	di più, non solo, per esempio, di nuovo, in realtà, più o meno, di fatto, del resto
	(luogo) a casa, in chiesa, al di là
	(tempo) di sera, un anno fa, al più presto
	(modo) in particolare, d'accordo, in piedi
- preposizioni:	fino a, da parte di, prima di, rispetto a, in modo da, per quanto riguarda
- *aggettivi*:*	in punto, di oggi, del genere, in crisi, di cotone, in fiamme, alla mano
- *congiunzioni:*	il fatto che, dal momento che, prima che, nel senso che, a patto che
- *interiezioni:*	va bene!, grazie a Dio, mamma mia!, hai voglia!, punto e basta
2) idiomi e modi di dire:	io penso che, è vero che, non è che, per così dire, questo è tutto, non c'è niente da fare, è un peccato
3) gruppi nominali poliremathici:	buona fede, lavoro nero, mercato unico, punto di vista, cassa integrazione
4) verbi supporto e idiomatici:	si tratta di, tener conto, portare avanti, far fronte, far parte, prendere atto, dare vita, dare luogo, mettere a punto, venirne fuori, rendersi conto
* Alcuni aggettivi si possono anche trovare con funzione di avverbi	

I **poliformi**, unità di senso (**lessie**) da considerare come unità minime del discorso non scomponibili, possono essere locuzioni grammaticali (*di_nuovo*, *di_fatto*, *del_resto*, *fino_a*, *in_modo_da*, *in_punto*, *il_fatto_che* ecc.); oppure modi di dire (*io_penso_che*, *è_vero_che* ecc.).

I poliformi possono dare luogo a **poliremathiche**, lessie che hanno complessivamente un significato diverso dalle parole che le compongono (Bolasco, 1999, p. 196) e che quindi il software di analisi automatica dei testi deve saper riconoscere e conteggiare come occorrenze: *Camera_dei_Deputati*, *punto_di_riferimento*, *tessuto_sociale*, *posti_di_lavoro*. Alcune poliremathiche possono essere valide solo all'interno di lessici specifici: *in_vigore*, per esempio, può essere trattato come una poliremathica nel lessico giuridico, ma non nel lessico medico.

In generale il riconoscimento dei poliformi, e delle poliremathiche in particolare, offre un contributo essenziale alla **disambiguazione** delle forme grafiche riducendo la **polisemia** delle parole per conseguire un livello accettabile di **monosemia**, che rimane comunque una meta molto difficile da conseguire se non attraverso linguaggi artificiali e simbolici (per esempio il linguaggio della matematica e della logica formale).

La terza e la quarta parte riguardano lo spinoso problema della *Riduzione delle maiuscole*. Le opzioni riguardano esclusivamente le forme che non sono state trattate nella sequenza precedente (*Rosa* sarà già stato etichettato come nome proprio e pertanto non sarà più modificato). L'opzione che permette di ridurre le maiuscole in modo relativamente standard è quella indicata nella figura 9.4. Per una applicazione delle altre opzioni è bene seguire attentamente le indicazioni della Guida.

L'opzione *Blocca le forme tutte in maiuscolo* e *anche forme miste con sola lettera iniziale maiuscola* attribuisce una categoria fittizia MAJ nel campo CAT a forme con iniziale maiuscola nel testo, ma che non sono presenti nelle liste del Database, non seguono un inizio di frase e non sono state ridotte a minuscolo nella fase precedente. In questo modo forme grafiche come *Verdi* e *Rossi* nelle fasi successive non saranno interpretati dal programma come aggettivi e ricondotti al lemma <verde> e <rosso>.

L'opzione *Applica i TAG forzati alla fine della normalizzazione* opera, già in questa fase, una forzatura nel tagging grammaticale classificando le parole più frequenti all'interno della loro categoria quasi esclusiva (oltre il 99%) riducendo le ambiguità grammaticali, soprattutto quando l'interesse dell'analista è rivolto al contenuto piuttosto che alla espressione linguistica. Con questa opzione, ad esempio, la forma *sono*, che è teoricamente anche un sostantivo (inteso come "suono") verrà classificato come verbo e segnalato nel campo Tagger della tabella Vocabolario con il codice 30.

9. 5. ANALI DEL VOCABOLARIO

A questo punto possiamo passare a una prima analisi del vocabolario selezionando la voce *Database di sessione* dal menu **Visualizza**, oppure cliccando sul bottone **Accesso ai Database** dalla barra degli strumenti. Nella finestra selezioniamo *Vocabolario* e con *Apri lista selezionata* visualizziamo la tabella (fig. 9.5).

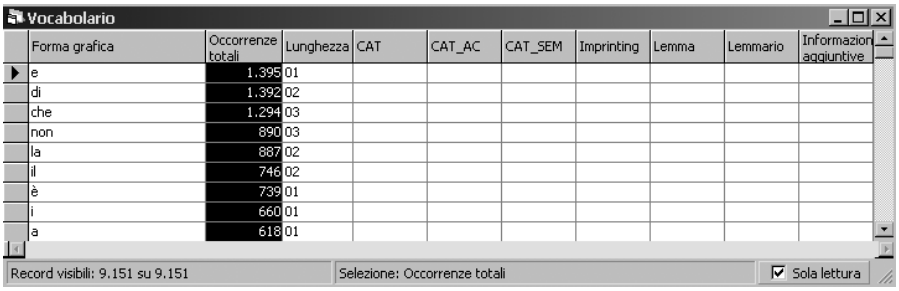


Fig. 9.5 – Vocabolario della sessione

Dalla lettura della tabella ("Record visibili", in basso a sinistra) apprendiamo subito le che forme grafiche dopo la normalizzazione sono 9.151).

La colonna 1 riporta l'elenco delle forme, la colonna 2 le occorrenze in ordine decrescente, la colonna 3 la lunghezza di ciascuna forma, la colonna 4 le categorie grammaticali delle parole già definite in fase di normalizzazione, e in particolare:

- i nomi propri (NM), i numerali (NUM) e alcuni sostantivi (N);
- le forme idiomatiche (FORM);
- le locuzioni grammaticali riconosciute nel pre-trattamento (AGG, AVV, CONG, PREP e PRON).

Queste informazioni sono visualizzabili selezionando la colonna "CAT" e cliccando sul pulsante **Z→A (Ordine decrescente)** della barra degli strumenti. Le altre colonne sono vuote: si riempiranno nel corso dell'analisi.

Su questa tabella *Vocabolario* possiamo eseguire le principali misure lessicometriche. Sulla barra degli strumenti, cliccando sul bottone **Misure lessicometriche**, visualizziamo la finestra di interrogazione (fig. 9.6).

Il **totale delle occorrenze** o **dimensione del corpus** (N), come si è visto, è il totale delle forme grafiche (parole, lessie, grafie) intese come "unità di conto" (*word token*).

Il **totale delle forme grafiche** o **ampiezza del vocabolario** (V) è il to-

tale delle forme grafiche conteggiate come forme grafiche distinte (*word type*).

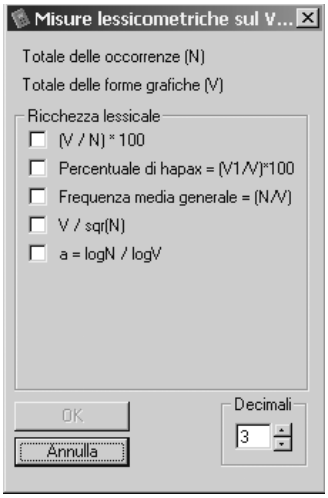


Fig. 9.6 – Misure lessicometriche sul Vocabolario

La **ricchezza lessicale** viene misurata con alcuni indicatori in uso nella statistica linguistica.

Estensione lessicale (*type/token ratio*):

$$\frac{V}{N} \times 100$$

Percentuale di *hapax*:

$$\frac{V_1}{V} \times 100$$

Frequenza media generale:

$$\frac{N}{V}$$

Coefficiente G (di Guiraud):

$$G = \frac{V}{\sqrt{N}}$$

Le misure di ricchezza del vocabolario sono sempre influenzate dal numero delle occorrenze. Ad esempio, la frequenza media generale è tanto più alta quanto più è esteso il corpus perché il totale delle occorrenze (per effetto delle alte frequenze delle parole forma) tende a crescere più rapidamente del totale delle parole distinte e, in ogni caso, le parole tendono a ripetersi con l'aumentare delle dimensione del corpus.

Il discorso inverso vale per la *type/token ratio* (che infatti è l'inverso della media generale delle parole): quanto più il corpus è grande tanto più il valore del rapporto è piccolo.

Alcuni autori hanno proposto delle misure indipendenti dall'ampiezza del vocabolario come la ***K* di Yule**. La caratteristica di Yule nei testi più grandi è meno influenzata dalla presenza degli hapax (nella formula di *K* la *i* rappresenta la classe di frequenza e V_i il numero di forme grafiche appartenenti alla classe di frequenza *i*).

$$K = \frac{10^4(\sum_{i=1}^{\infty} i^2 V_i - N)}{N^2}$$

Prima di proseguire occorre soffermarsi brevemente sulla relazione riscontrata tra rango e frequenza da un geniale linguista: George Kingsley Zipf (1902-1950). Se si ordinano secondo il rango le parole di un testo sufficientemente esteso (Zipf aveva preso come riferimento l'*Ulisse* di James Joyce, con 260.000 occorrenze), partendo dal rango più elevato, la frequenza delle parole è ovviamente in relazione inversa (tab. 9.2).

Tab. 9.2 – Rapporto tra rango e frequenza delle parole nell'Ulisse di J. Joyce.

rango (*r*)	frequenza (*f*)	*f* × *r* = *c*
la parola 10ª	è usata 2.653 volte	26.530
la parola 100ª	265	25.500
la parola 1.000ª	26	26.000
la parola 10.000ª	2	20.000
la parola 29.000ª	1	29.000

Questo non sorprende perché siamo stati noi ad assegnare il rango 1 alla parola con frequenza maggiore e così di seguito. Quello che sorprende è osservare che il prodotto della frequenza per il rango è approssimativamente costante.

Questa osservazione della **legge di Zipf** richiede naturalmente che a un certo rango (ad esempio, al rango 10) si assuma come valore di frequenza il valore medio delle occorrenze delle parole intorno al rango considerato (ad esempio da 1 a 20); infatti nel vocabolario di un corpus non vi sono tutte le classi di frequenza possibili (per classi di frequenza si intende un insieme di parole che hanno lo stesso numero di occorrenze; Bolasco, 1999, p. 186 sg.). In seguito a un'ampia discussione sulla validità di questa legge, i linguisti hanno ritenuto più opportuno esprimere l'equazione come:

$$f \times r^{a} = c$$

che in scala logaritmica si può esprimere anche come equazione della retta di regressione (fig. 9.7):

$$\log f = c + a \times \log r$$

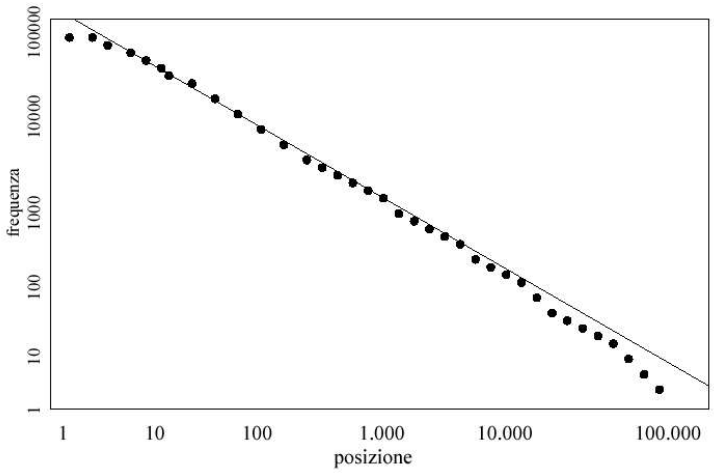


Fig. 9.7 – Legge di Zipf

per cui il coefficiente *a* indica l'angolo della retta (da intendersi con segno negativo) su un grafico a coordinate logaritmiche, in cui sull'asse *x* si riporta il

logaritmo del rango e sull'asse *y* il logaritmo della frequenza; *c* è il punto in cui la retta interseca l'ordinata. Il coefficiente *a* è approssimato dal rapporto:

$$\frac{\log N}{\log V}$$

e definisce, come si è detto, la pendenza della retta di regressione; in testi con un numero abbastanza elevato di occorrenze (50.000) il suo valore (con segno negativo) dovrebbe essere intorno a 1,15. Valori più elevati di 1,3 indicano che il vocabolario utilizzato non è particolarmente ricco (Tuzzi, 2003, p. 127). La G di Guiraud, per testi delle stesse dimensioni, assume un valore intorno a 22. Valori più grandi indicano una maggiore ricchezza lessicale, ma occorre tenere conto delle particolari tipologie di testi che vengono sottoposti a trattamento[10].

Nella finestra di dialogo (fig. 9.6) spuntiamo tutte le opzioni disponibili e clicchiamo su *OK*. Alla richiesta di salvare il file nella cartella di lavoro rispondiamo *Salva* e otteniamo la schermata di passaggio della figura 9.8. Durante l'esecuzione di questa procedura si apre una ulteriore finestra che ci chiede di indicare l'ordine di grandezza delle frequenze normalizzate con un valore che approssima meglio il totale delle occorrenze (N) del corpus (nel nostro esempio con 47.056 occorrenze dovremo battere 10.000).

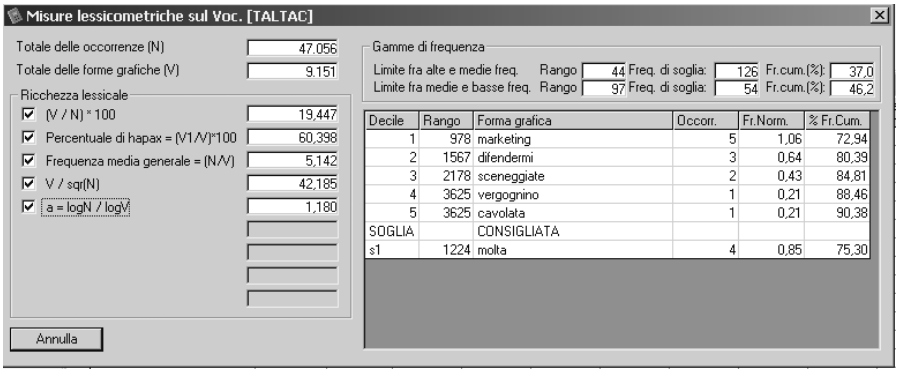


Fig. 9.8 – Misure lessicometriche sul vocabolario

[10] Per approfondire l'interpretazione delle misure della ricchezza lessicale, molto complessa e molto discussa nella statistica linguistica, conviene riferirsi ai lavori di Cossette (1994) e di Labbé (1995).

Un confronto tra diversi corpora di dimensione crescente permette di apprezzare la validità e la sensibilità di alcuni indicatori di ricchezza lessicale (tab. 9.3). Aumentando le dimensioni del corpus la *type/token ratio* diminuisce. La frequenza media generale aumenta con il crescere delle dimensioni del corpus. La G di Guiraud è sensibilmente più bassa nel corpus *Lex* in cui vi sono testi di carattere legislativo rispetto ai testi letterari del corpus *Mongai* e del corpus *AnticoT*. La G è abbastanza alta anche nel corpus *Bullismo*, ma si tratta testi con una grande variabilità di forme grafiche anche per errori di ortografici, come accade tipicamente per i testi tratti da Internet. Tuttavia questi confronti tra testi non omogenei devono essere condotti con cautela: in generale i confronti di ricchezza lessicale si dovrebbero fare tra testi dello stesso genere (tra testi giornalistici, tra testi di autori letterari ecc.).

Tab. 9.3 – Misure lessicometriche per diversi corpora secondo la dimensione

Corpora	**Lex**	**Bullismo**	**Mongai**	**AnticoT**
Occorrenze N	14.733	47.056	142.485	627.325
Forme grafiche distinte	3.145	9.151	19.024	33.368
***type/token r.* = (V/N)*100**	21,35	19,44	13,32	5,32
% di hapax = (V1/V)*100	55.87	60,39	53,04	44,75
Frequenza media gen.= N/V	4.63	5,14	7,51	18,80
G di Guiraud	25,91	42,18	50,33	42,13
coefficiente *a*	1,19	1,18	1,20	1,28

Descrizione dei corpora:
Lex: corpus costituito dalle testi "costituzionali": *Statuto del Regno di Sardegna* (1848); *Costituzione della Repubblica romana* (1849); *Costituzione della Repubblica italiana* (1948); *Dichiarazione Universale dei Diritti dell'Uomo* (1948).
Bullismo: corpus costituito da 277 messaggi del Forum dedicato al bullismo sul sito www.repubblica.it dal 16 novembre 2006 al 29 aprile 2007.
Mongai: corpus costituito da due romanzi di fantascienza di Massimo Mongai (*Il gioco degli immortali*, 1999 e *Memorie di un cuoco d'astronave*, 1997) pubblicati da Mondadori.
AnticoT: corpus costituito dai testi della *Bibbia*, *Antico Testamento* (Pentateuco, Libri storici, Sapienziali, Profetici).
I testi che costituiscono i corpora *Lex*, *Mongai* e *AnticoT* sono sono stati scaricati dal sito *Liber Liber* del progetto Manuzio per la costituzione di una biblioteca telematica ad accesso gratuito <http://www.liberliber.it>.

La schermata delle misure lessicometriche riporta altre informazioni che però necessitano di alcune spiegazioni preliminari sulla tabella *Vocabolario* che, a seguito di queste operazioni di misura, si è popolata di altre informazioni. Nella figura 9.9 (e nelle figure seguenti) alcuni campi, per comodità di rappresenta-

zione, sono stati nascosti con il comando *Nascondi campi selezionati* del menu **Formato**. In qualsiasi momento i campi nascosti possono essere visualizzati con il comando *Scopri campi...* nello stesso menu.

Nella colonna 7 (fig. 9.9), accanto al rango viene indicata la gamma (**fascia**) di frequenza e nella colonna 8, la frequenza relativa su base 10.000.

Vocabolario

Forma grafica	Occorrenze totali	Lunghezza	CAT	Frequenza cumulata	Rango	Fasce	Frequenza relativa
scuola	215	06		31,2	27	Alta	45,69
del	204	03		31,6	28	Alta	43,35
loro	191	04		32,1	29	Alta	40,59
gli	184	03		32,4	30	Alta	39,10
ci	183	02		32,8	31	Alta	38,89
mi	180	02		33,2	32	Alta	38,25
ragazzi	174	07		33,6	33	Alta	36,98
più	165	03		33,9	34	Alta	35,06
perché	164	06	J	34,3	35	Alta	34,85
ha	152	02		34,6	36	Alta	32,30
ho	150	02		34,9	37	Alta	31,88
solo	148	04		35,2	38	Alta	31,45
genitori	147	08		35,6	39	Alta	31,24
tutti	142	05		35,9	40	Alta	30,18
essere	141	06		36,2	41	Alta	29,96
ed	129	02		36,4	42	Alta	27,41
questo	128	06		36,7	43	Alta	27,20
fare	126	04		37,0	44	Media	26,78
chi	126	03		37,2	44	Media	26,78
questi	123	06		37,5	46	Media	26,14
lo	121	02		37,8	47	Media	25,71
io	120	02		38,0	48	Media	25,50

Record visibili: 9.151 su 9.151 | Nessuna colonna selezionata | Sola lettura

Fig. 9.9 – Vocabolario di base *Bullismo:* alte e medie frequenze

Le forme grafiche in ordine lessicometrico (ordinamento secondo la frequenza) si presentano anche in un ordinamento per **ranghi crescenti** (fig. 9.9, col. 6). Il rango è il posto occupato da una forma grafica nella graduatoria. La forma *e* (congiunzione) occupa il primo posto (rango 1) e appartiene alla **classe di occorrenze** *i*=1.395 come unica forma (nel senso che nessun'altra forma conta 1.395 occorrenze). La forma *scuola* occupa il rango 27 ed è la prima parola piena, cioè una parola che ci rivela qualche cosa della struttura semantica del corpus.

Al rango 44 troviamo due forme: *fare* e *chi*. Entrambe appartengono alla classe di occorrenze *i*=126 e definiscono il passaggio tra la fascia delle **alte frequenze** e la fascia delle **medie frequenze** (fig. 9.9, col. 7). La fascia delle medie frequenze inizia con la prima coppia di parole che hanno uno stesso

numero di occorrenze.

Le parole che appartengono alla fascia delle alte frequenze sono in massima parte parole vuote. Per la statistica testuale è rilevante osservare che nei primi 25 ranghi decrescenti troviamo solo parole grammaticali e queste rappresentano il 30,3% delle occorrenze. Tuttavia nella fascia delle alte frequenze (fig. 9.9) si incontrano anche alcune **parole-chiave** che possono descrivere i temi principali dei testi in esame: *scuola* (215), *ragazzi* (174) e *genitori* (147).

Scorrendo rapidamente i ranghi decrescenti dal fondo della lista delle parole, partendo quindi dagli hapax per risalire verso l'alto, incontriamo classi di occorrenze crescenti consecutive: 1, 2, 3, ..., *i*, ... fino al rango 54 (fig. 9.10), in corrispondenza della preposizione *dalla*, cui segue (risalendo verso l'alto) una lacuna nelle classi di occorrenze crescenti (*i*=55). Dal rango 54 inizia quindi la fascia delle **basse frequenze**. Nella fascia delle basse frequenze, con classi di frequenze decrescenti fino a 1, si trova sempre la grandissima parte delle parole distinte del vocabolario, in genere le parole principali. In questo caso sono 9.055 parole distinte (pari al 98,95% del totale).

Vocabolario

Forma grafica	Occorrenze totali	Lunghezza	CAT	Frequenza cumulata	Rango	Fasce	Frequenza relativa
mia	60	03		45,5	89	Media	12,75
a scuola	58	08	AVV	45,6	92	Media	12,33
un'	58	03		45,7	92	Media	12,33
fatto	56	05		45,9	94	Media	11,90
nelle	56	05		46,0	94	Media	11,90
professori	56	10		46,1	94	Media	11,90
dalla	54	05		46,2	97	Bassa	11,48
educazione	54	10		46,3	97	Bassa	11,48
spesso	54	06		46,5	97	Bassa	11,48
li	53	02		46,6	100	Bassa	11,26
classe	52	06		46,7	101	Bassa	11,05
forse	52	05		46,8	101	Bassa	11,05
quelli	51	06		46,9	103	Bassa	10,84
deve	51	04		47,0	103	Bassa	10,84
molto	51	05		47,1	103	Bassa	10,84
colpa	51	05		47,2	103	Bassa	10,84
ragazzo	50	07		47,3	107	Bassa	10,63
persone	50	07		47,4	107	Bassa	10,63

Record visibili: 9.151 su 9.151 | Nessuna colonna selezionata | Sola lettura

Fig. 9.10 – Vocabolario di base *Bullismo:* medie e basse frequenze

L'output delle misure lessicometriche presenta altre informazioni interessanti sulle gamme di frequenza (fig. 9.8, che si riporta in primo piano sul monitor cliccando sul segno "–" in alto a destra della finestra di Windows).

Le informazioni della figura 9.8 sono essenziali per compiere alcune valutazioni sulle **dimensioni minime** del corpus per una analisi automatica e

sulla copertura del testo in funzione della scelta di determinate soglie alle quali collocare la selezione delle forme da analizzare. Si tratta di scelte che non possono essere di natura esclusivamente quantitativa, ma che devono essere compiute con la massima attenzione tenendo conto anche di alcuni aspetti che riguardano le misure effettuate sul testo.

Quali sono le dimensioni minime che un corpus deve avere affinché sia adeguato per un'analisi statistica? Un criterio empirico suggerito dagli analisti (Bolasco, 1999, p. 203) è di osservare la *type/token ratio*: quando le parole distinte superano il 20% delle occorrenze il corpus non si può considerare sufficientemente esteso per un'analisi quantitativa. Tra i corpora della tab. 9.3 possiamo notare come il corpus *Bullismo* sia appena adeguato (19,44%). Come indicazione generale, un corpus di 15.000 occorrenze si può considerare di "piccola dimensione"; un corpus di 50.000-100.000 occorrenze di "media dimensione" e un corpus maggiore di 200.000 occorrenze di "grande dimensione". Un corpus sufficientemente grande (oltre le 500.000 occorrenze) può costituire una base per la costruzione di un lessico di frequenza rappresentativo di un linguaggio purché i testi siano abbastanza rappresentativi della sua eterogeneità. Le unità di un lessico di frequenza sono espresse in lemmi.

L'analisi testuale di un corpus, soprattutto nelle applicazioni della statistica multidimensionale, non può prendere in esame l'intero vocabolario. Pertanto la copertura del testo non potrà mai essere del 100%. In genere, in tutte le indagini statistiche, l'obiettivo del ricercatore è di selezionare un piccolo numero di variabili che siano sufficientemente rappresentative dei caratteri essenziali del fenomeno oggetto di studio. Nell'analisi testuale questo obiettivo si consegue attraverso la scelta di una soglia di frequenza al di sotto della quale le parole possono essere abbandonate senza una significativa perdita di informazione.

Il **tasso di copertura del testo** è pari alla percentuale di occorrenze che derivano dalle parole $V_{(s)}$ al di sopra della soglia *s* sul totale N di tutte le occorrenze del corpus.

Dalle misure lessicometriche sul vocabolario del corpus *Bullismo* (fig. 9.8) apprendiamo che sul primo decile delle basse frequenze (il 10% delle parole distinte di bassa frequenza che sono al di sopra della soglia 5 indicata al rango 978 con la forma grafica *marketing*) si ottiene una copertura del testo pari al 72,94%. Sulla mediana (5° decile) delle basse frequenze in corrispondenza della forma *cavolata* si ottiene il 90,38% di copertura del testo. La soglia consigliata da TaLTaC² per l'analisi multimensionale è 4 che offre una copertura del testo del 75,30%. La copertura del testo migliora quando il corpus è più grande. Con il corpus *AnticoT* la soglia consigliata è 13 con una copertura dell'87,45%.

9. 6. IL RICONOSCIMENTO DELLE FORME GRAMMATICALI

La procedura di attribuzione di ciascuna forma a una categoria grammaticale è fondamentale per la disambiguazione delle parole.

Tab. 9.4 – Categorie grammaticali del database di TaLTaC²

N	sostantivo	PREP	preposizione	FORM	forma idiom.
A	aggettivo	CONG	congiunzione	NM	nome proprio
V	verbo	PRON	pronome	DAT	data
AVV	avverbio	ESC	interiezione	NUM	numerale
DET	determinante	J	ambigua	O	stranierismo

La categoria "ambigua" (J) identifica le parole compatibili con più categorie. Ad esempio, *legge* (sostantivo femminile) e *legge* (terza persona singolare dell'indicativo presente del verbo *leggere*). Le forme non riconosciute (parole rare, parole straniere non presenti nel database, parole di un lessico specialistico, errori di ortografia ecc.) non sono attribuite ad alcuna categoria (il campo di attribuzione rimane vuoto).

Dalla barra degli strumenti, cliccando sul bottone **Tagging grammaticale**, visualizziamo una finestra di dialogo che ci offre la scelta delle categorie grammaticali da inserire nell'analisi.

Salvo esigenze particolari, è preferibile procedere al tagging grammaticale completo, comprese le opzioni basate su criteri morfologici che classificano anche i numeri scritti in lettere; le enclitiche verbali (*mangiarlo*, *dammelo*, *daglielo*, *prendine*); i derivati di sostantivi e aggettivi (*azionismo*, *assistenziale*, *buonista*, *ventenne*) e alterati con suffissi diminutivi, vezzeggiativi, peggiorativi (*casina*, *casona*, *casaccia*); le forme con i prefissi più comuni (microbomba, dopobomba) e altre forme più complesse come prefisso+base+suffisso (*neo+colonial+ismo*).

Premendo su *OK* si esegue la procedura e si ottiene la tabella *Vocabolario* arricchita con le informazioni sulle categorie grammaticali (fig. 9.11). La colonna "CAT" contiene per ciascuna forma una etichetta grammaticale di classificazione. Se clicchiamo con il tasto destro del mouse su una forma otteniamo tutte le informazioni contenute nel database DIZTALTAC, il lessico di riferimento di TaLTaC² che contiene circa 74.000 lemmi e oltre 500.000 forme flesse (fig. 9.11).

Per la forma *sono* troviamo tre lessie:

- *sono*, sostantivo maschile per una forma letteraria di *suono*;
- *sono*, indicativo presente, prima persona singolare e terza persona plurale del verbo *essere*;
- *sono*, indicativo presente, prima persona singolare del verbo *sonare*.

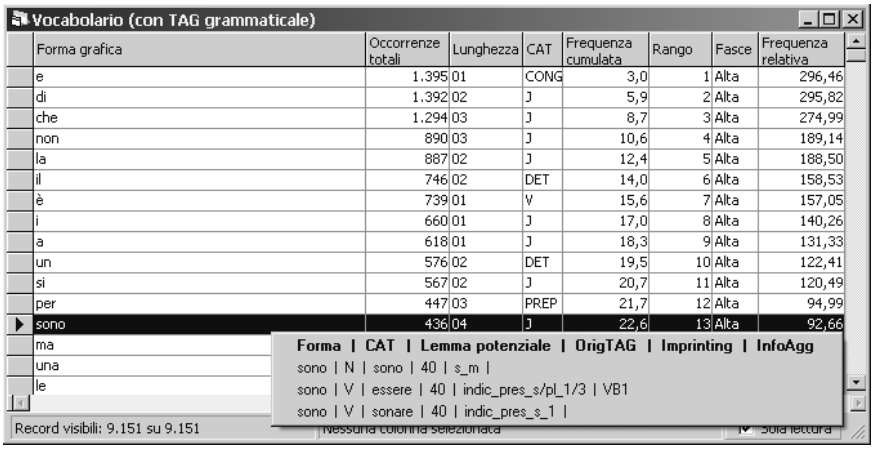


Fig. 9.11 – Dettaglio del tagging grammaticale

Nella fase di tagging grammaticale alcune forme grafiche non sono riconosciute dal software. Le possiamo evidenziare nella tabella del vocabolario selezionando la colonna "CAT" e cliccando sul bottone **A→Z – Ordine crescente** della barra degli strumenti. I primi record della tabella sono record con il "campo vuoto". Ad esempio, la forma *puffo* non è stata identificata correttamente perché non è presente nel Database di TaLTaC²; la forma *bowman@repubblica* è parte di un indirizzo e-mail (*.it* è stato separato dal corpo dell'indirizzo perché dopo il punto il software inserisce automaticamente uno spazio). La maggior parte delle forme non riconosciute sono dovute a forme composte non prevedibili (*bulli/somari*, *bullismo/vandalismo*), a errori di ortografia (*presentanto*, *aggravente*, *socialè*); a errori di trascrizione (*-cioè*, *violenza-*); a forme dialettali (*ragà*); a grafie tipiche della scrittura veloce (*kiedo*); a parole inconsuete o nuove (*giocologia*); a nomi propri non presenti nel Database (*erasmus*, *stakanov*, *damocle*); a forme tipiche di "rumore ambientale" dovuto alla comunicazione elettronica: transcodifiche errate da parte del server o sequenze alfanumeriche (*v=UlB2BiiI91E&NR*) che entrano nel corpo del messaggio per effetto di citazioni automatiche (*quoting*) o operazioni di copia e incolla da parte dell'utente. Di solito il "rumore di fondo", presente in quantità rilevante nei newsgroup, nei forum e nelle e-mail, può essere trascurato nell'analisi automatica del testo (Giuliano, 2004) soprattutto quando il trattamento è applicato a corpora di grandi dimensioni. In un corpus piccolo come quello utilizzato in questo esempio, l'analista potrebbe ritenere utile procedere a una correzione

manuale (riducendo gli hapax dal 60,39% al 55,87%) in modo da migliorare la qualità del dato testuale. Queste sono decisioni che riguardano momenti tattici della ricerca e che devono essere valutate caso per caso secondo gli obiettivi e le risorse disponibili.

Dal menu **Formato**, selezionando *Scopri campi*, possiamo marcare la colonna "CAT-AC" e smarcare le colonne "Lunghezza", "Frequenza cumulata", "Rango" e "Fasce" che ora sono meno interessanti) e osservare (fig. 9.14) come in CAT-AC siano contenute le attribuzioni multiple che caratterizzano le forme grafiche classificate come ambigue (J).

Vocabolario (con TAG grammaticale)

Forma grafica	Occorrenze totali	CAT	CAT_AC	Frequenza relativa
non	890	J	A+AVV	189,14
la	887	J	DET+N+PRON	188,50
il	746	DET	DET	158,53
è	739	V	V	157,05
i	660	J	DET+N	140,26
a	618	J	N+PREP	131,33
un	576	DET	DET	122,41
si	567	J	N+PRON	120,49
per	447	PREP	PREP	94,99
sono	436	J	N+V	92,66
ma	400	J	CONG+ESC	85,01
una	376	J	DET+NUM+PRON	79,90
le	360	J	DET+PRON	76,50
in	352	J	A+PREP	74,80

Record visibili: 9.151 su 9.151 | Nessuna colonna selezionata | Sola lettura

Fig. 9.12 – Vocabolario *Bullismo*: colonna delle categorie grammaticali ambigue

9. 7. LA LEMMATIZZAZIONE

Nella fase di tagging grammaticale il software ha riconosciuto i lemmi, pertanto è possibile generare una lista di forme grafiche classificate secondo il lemma corrispondente. Il **lemma** è la forma canonica della parola così come appare in un vocabolario della lingua italiana: il verbo all'infinito, il sostantivo e l'aggettivi al maschile, ecc. Dopo aver richiamato la tabella *Vocabolario*, dal menu **Formato** selezioniamo *Scopri campi* e le voci che ci interessano: *Lemma* e *Lemmario* (fig. 9.15). Nella colonna 5, quando la lessia è univoca, le forme sono contrassegnate dal lemma. Ad esempio, le forme *applica* e *applicarsi* sono classificate nel verbo *applicare*; *anni* e *anno* nel sostantivo *anno*. Nella tabella vediamo come la forma *è* sia classificata in *essere*; tuttavia *sono* rimane inalterata nella forma grafica originale perché è impossibile al programma (al di fuori del con-

testo in cui è collocata la parola) decidere a quale delle "entrate" del lemmario attribuire le occorrenze (colonna 6 in cui sono conteggiate 3 entrate; fig. 9.13).

La lemmatizzazione è parzialmente indipendente dal tagging. Una forma ambigua può essere ricondotta a un lemma senza modificare la sua classificazione come ambigua. Ad esempio, *altra* e *altre*, pur essendo forme ambigue (J) sono classificate nel lemma *altro* poiché la forma canonica delle differenti CAT di queste forme è sempre la stessa. Così pure la forma *che* può essere attribuita a 4 categorie grammaticali e ha 4 entrate nel lemmario, ma la forma canonica è sempre la stessa.

Vocabolario (con TAG grammaticale)

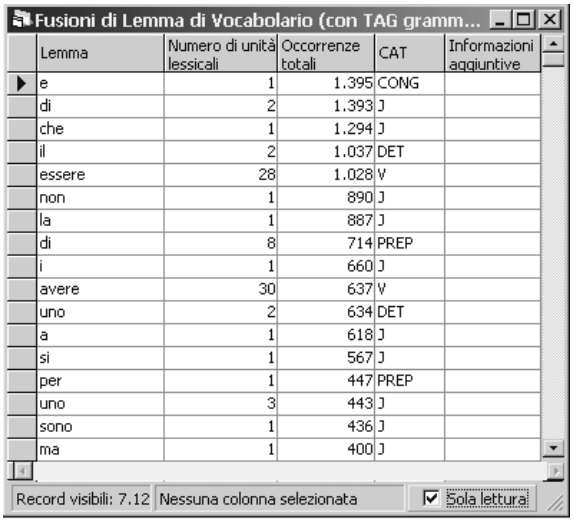
Forma grafica	Occorrenze totali	CAT	CAT_AC	Lemma	Lemmario
e	1.395	CONG	CONG	e	1
di	1.392	J	N+PREP	di	2
che	1.294	J	A+CONG+N+PRON	che	4
non	890	J	A+AVV	non	2
la	887	J	DET+N+PRON	la	3
il	746	DET	DET	il	1
è	739	V	V	essere	1
i	660	J	DET+N	i	2
a	618	J	N+PREP	a	2
un	576	DET	DET	uno	1
si	567	J	N+PRON	si	2
per	447	PREP	PREP	per	1
sono	436	J	N+V	sono	3

Fig. 9.13 – Vocabolario per la lemmatizzazione

Per eseguire la lemmatizzazione selezioniamo la colonna "Lemma"; dal menu **Calcola**, selezioniamo la voce *Fusioni*: una finestra di dialogo ci chiede di inserire una descrizione per la tabella che sarà generata. Il nome assegnato di default è: *Fusioni di Lemma di Vocabolario (con TAG grammaticale)* con il quale la tabella sarà inserita nel Database della sessione (fig. 9.14).

La tabella presenta 7.123 record con tutte le forme grafiche (comprese quelle non disambiguate J e le forme non riconosciute) classificate (fuse) nel lemma corrispondente. Ad esempio, dove è stato possibile, tutte le coniugazioni del verbo <essere> sono state classificate nel lemma corrispondente (V, 1.028 occorrenze). La colonna 2 "Numero di unità lessicali" indica quante forme flesse del verbo sono state conteggiate all'interno del lemma (28).

Tuttavia questa lemmatizzazione "grezza", effettuata su tutto l'insieme delle categorie grammaticali, non è attendibile: il 36,67% dei lemmi è classificato come ambiguo e 474 forme non sono riconosciute (5,18% del totale).



Lemma	Numero di unità lessicali	Occorrenze totali	CAT	Informazioni aggiuntive
e	1	1.395	CONG	
di	2	1.393	J	
che	1	1.294	J	
il	2	1.037	DET	
essere	28	1.028	V	
non	1	890	J	
la	1	887	J	
di	8	714	PREP	
i	1	660	J	
avere	30	637	V	
uno	2	634	DET	
a	1	618	J	
si	1	567	J	
per	1	447	PREP	
uno	3	443	J	
sono	1	436	J	
ma	1	400	J	

Fig. 9.14 – Fusioni di lemma

Le classificazioni ambigue sono eccessive e per ridurle sarebbero necessari degli interventi puntuali di disambiguazione che possono essere effettuati solo con un attento esame delle concordanze. In generale, la procedura di lemmatizzazione, essendo sostanzialmente un'operazione di misura tramite classificazione, va condotta con cura e con la piena consapevolezza del ricercatore rispetto agli scopi che intende perseguire. L'attenzione alla qualità del dato deve sempre prevalere rispetto alla quantità.

RIFERIMENTI BIBLIOGRAFICI

BOLASCO S. (1999) *Analisi multimensionale del dati. Metodi, strategie e criteri d'interpretazione*, Roma, Carocci (II ed. 2004).

COSSETTE A. (1994) *La richesse lexicale et sa mesure*, Paris, Honoré Champion.

GIULIANO L. (2004) "L'analisi automatica dei testi ad alta componente di rumore", in in E. Aureli Cutillo e S. Bolasco (a cura di), *Applicazioni di analisi statistica dei dati testuali*, Roma, Casa Editrice La Sapienza, pp. 41-54.

LABBE D. (1995) "La structure du vocabulaire du Général De Gaulle", in S. Bolasco, L. Lebart, A. Salem, *JADT 1995. III Giornate Internazionali di Analisi statistica dei dati testuali*, Roma, CISU, pp. 165-176.

TUZZI A. (2003*) L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*, Roma, Carocci.

10.
LAVORARE CON TALTAC²: L'ANALISI LESSICALE

Le fasi di normalizzazione, di costruzione del vocabolario e di tagging grammaticale sono fondamentali per un esame completo del corpus in vista di qualsiasi strategia di analisi. Quando ci troviamo di fronte a vocabolari molto ampi, composti di oltre 30.000 forme grafiche distinte, l'interpretazione automatica dei testi richiede necessariamente la selezione di un sottoinsieme di parole con un alto contenuto di informazione che sia rappresentativo del contenuto del corpus. L'estrazione dei segmenti ripetuti, la lessicalizzazione e l'individuazione delle forme peculiari rappresentano momenti significativi in vista di questo obiettivo.

10.1. TEXT/DATA MINING ED ESPLORAZIONE DELLE TABELLE

Prima di proseguire con le altre analisi è necessario apprendere l'uso di uno strumento fondamentale per la gestione dei dati testuali e delle tabelle di TaLTaC². Lo strumento **Text/Data Mining (TDM)** è di grande utilità, non solo per l'esplorazione del corpus attraverso le liste generate dal programma, ma anche per selezionare le matrici da esportare per il loro utilizzo in altri software di analisi testuale. Negli esempi che seguono prenderemo come riferimento il *Vocabolario* della sessione, ma le stesse operazioni possono essere compiute sulle altre tabelle e liste generate dal software e salvate del Database di sessione.

Con la lista *Vocabolario* aperta nella finestra di lavoro, selezioniamo il campo sul quale desideriamo operare; per esempio selezioniamo il campo "Forma grafica" ponendo il cursore sull'intestazione della colonna stessa: l'inversione di colore indica che il campo è "attivo".

Ora clicchiamo sull'icona **TDM** della barra degli strumenti (oppure selezioniamo il comando *Text/Data Mining* dal menu **Record**) e apriamo la finestra di dialogo:

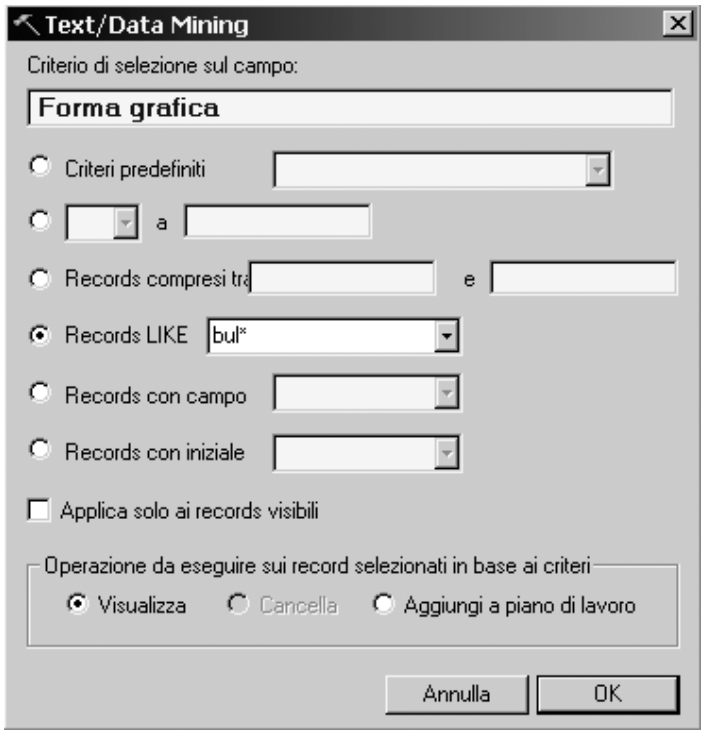


Fig. 10.1 – Finestra di dialogo del Text/Data Mining

Come si può osservare (fig. 10.1) il "criterio di selezione sul campo" indicato nella casella è quello da noi selezionato: "Forma grafica". I parametri successivi ci permettono di compiere delle operazioni logiche e delle operazioni di selezione tali da visualizzare esclusivamente i record di nostro interesse. Tali record "visibili" possono essere salvati in matrici da esportare e da utilizzare con altri software. Per esempio rispetto al campo "Forma grafica", per il criterio *Records LIKE* scriviamo "bul*". L'asterisco * sta per "qualsiasi carattere". Così facendo otterremo l'elenco delle forme grafiche che iniziano per "bul" (fig. 10.2). Dall'elenco apprendiamo quali sono le forme lessicali riconducili al bullismo e troviamo dei neologismi (*bullizzando* e *bullissimo*) o degli errori ortogra-

fici che dovranno essere corretti per una analisi semantica attendibile del corpus (*bullismò*, *bullì*, *bulismo*, *bullismo-*).

	Forma grafica	Occorrenze totali	CAT	CAT_AC	Frequenza relativa
▶	bullismo	77	N	N	16,36
	bulli	41	J	A+N	8,71
	bullo	21	J	A+N	4,46
	bulletti	16	V	V	3,40
	bulletto	2	V	V	0,43
	bulli/somari	1			0,21
	bullismò	1			0,21
	bullì	1			0,21
	bulle	1	A	A	0,21
	bulismo	1	N	N	0,21
	bullismo/vandalismo	1			0,21
	bullismo-	1			0,21
	bulli-leader	1			0,21
	bullissimo	1	A	A	0,21
	bullizzando	1			0,21

Fig. 10.2 – Selezione delle forme grafiche che iniziano con "bul".

Oppure, selezionando il campo "CAT", per il criterio *Records LIKE* scriviamo "V" ottenendo solo le forme grafiche della categoria "Verbi" (fig. 10.3).

Ciascuna di queste matrici è dotata di tutte le informazioni contenute nelle colonne della tabella Vocabolario, compresi i campi "nascosti", sempre visualizzabili dal menu **Formato**, voce *Scopri campi*, e poi selezionando i campi da scoprire. La matrice attiva nella finestra di lavoro è esportabile in un file di testo (successivamente importabile, per esempio, in Excel) dal menu **File**, comando *Esporta in un file di testo*, voce *Solo i record visibili*.

Le operazioni di selezione più interessanti che si possono compiere dalla finestra di dialogo dello strumento **TDM** sono:

1) La query predefinita di selezione dei **nomi astratti** (non direttamente percepibili dai sensi) come *libertà*, *proprietà*, *famiglia*, *violenza*, *sicurezza*, *dignità*, *società* ecc. Campo attivo: "Forma grafica"; selezione *Criteri predefiniti*: "N_astratto". Nel corpus *Bullismo* sono 362. I più frequenti sono *violenza* (69), *società* (64), *problema* (42), *responsabilità* (38), *famiglia* (34). Il risultato della query viene scritto nella tabella *Vocabolario* nel campo "Informazioni aggiuntive" con l'etichetta: "N_astratto".
2) La selezione secondo il numero delle occorrenze con gli operatori di > < e =. Campo attivo: "Occorrenze totali"; selezione >5, per esempio, visualizza una matrice di 977 record (forme grafiche distinte) che contiene solo le

forme con occorrenze maggiori o uguali a 6.

Vocabolario (con TAG grammaticale)

Forma grafica	Occorrenze totali	CAT	CAT_AC	Frequenza relativa
è	739	V	V	157,05
ha	152	V	V	32,30
ho	150	V	V	31,88
hanno	118	V	V	25,08
può	89	V	V	18,91
siamo	67	V	V	14,24
deve	51	V	V	10,84
sarebbe	42	V	V	8,93
va	41	V	V	8,71
erano	38	V	V	8,08
abbiamo	36	V	V	7,65
fanno	35	V	V	7,44
viene	34	V	V	7,23
vedere	32	V	V	6,80

Record visibili: 2.343 su 9.151 | Nessuna colonna selezionata | Sola lettura

Fig. 10.3 – Selezione delle forme grafiche: verbi

3) La selezione *Record compresi tra* che permette di indicare un range di validità della selezione. Ad esempio le forme comprese tre 10 e 100 occorrenze.
4) La selezione *Records LIKE* che, come si è visto, permette di individuare le forme grafiche, le categorie grammaticali e di operare, in genere, su una selezione di stringhe di testo.
5) La selezione dei *Records con campo vuoto/non vuoto*. Il campo vuoto è un campo in cui non vi è alcun carattere. Le forme non riconosciute dal tagging grammaticale hanno il campo vuoto.
6) La selezione dei *Records con iniziale* "MAIUSCOLA" o "minuscola".

La casella *Applica solo i records visibili* permette le operazioni di affinamento delle selezioni: spuntando la casella, la selezione successiva avviene solo sulla lista selezionata attiva.

L'opzione *Aggiungi a piano di lavoro* permette di conservare il criterio di selezione scelto (o meglio la combinazione di diversi criteri di selezione anche su campi differenti) in modo da eseguire query complesse che possono essere salvate e utilizzate successivamente secondo le necessità specifiche della ricerca. Scegliendo questa opzione, TaLTaC[2] aprirà un nuova finestra di dialogo con cui eseguire la procedura richiesta (consultare la *Guida in linea* del programma).

Lo strumento **TDM** permette di eseguire molte altre operazioni utili per la

costruzione di matrici di lavoro. La consultazione del manuale, l'esperienza dell'utente, la sua fantasia e i problemi che si presentano in ogni analisi in modo originale e imprevisto, possono generare nuove soluzioni e nuovi percorsi.

10. 2. ESTRAZIONE DEI SEGMENTI RIPETUTI E LESSICALIZZAZIONE

I segmenti ripetuti sono sequenze di forme grafiche formate da tutte le disposizioni a 2,3, ..., *q* forme che si ripetono per un certo numero di volte nel corpus (Bolasco, 1999, p. 194).

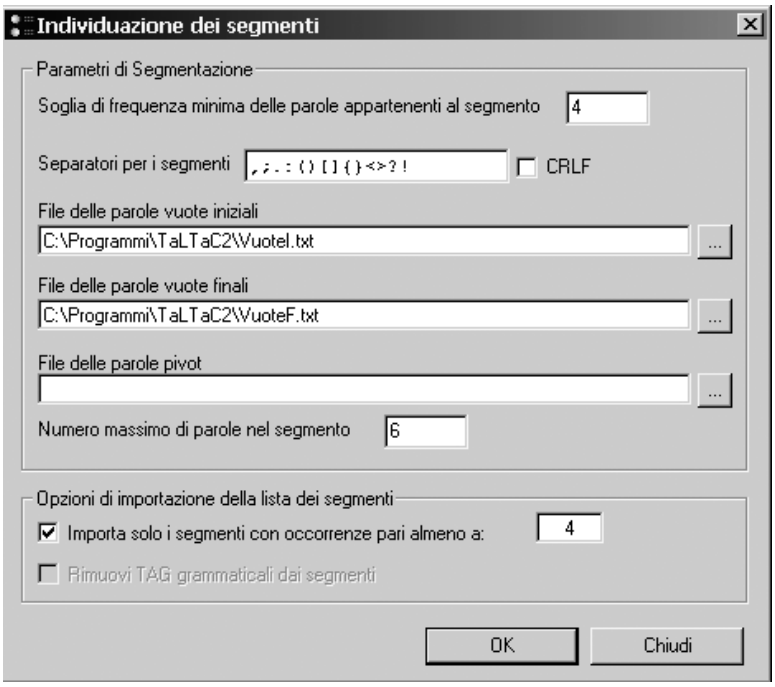


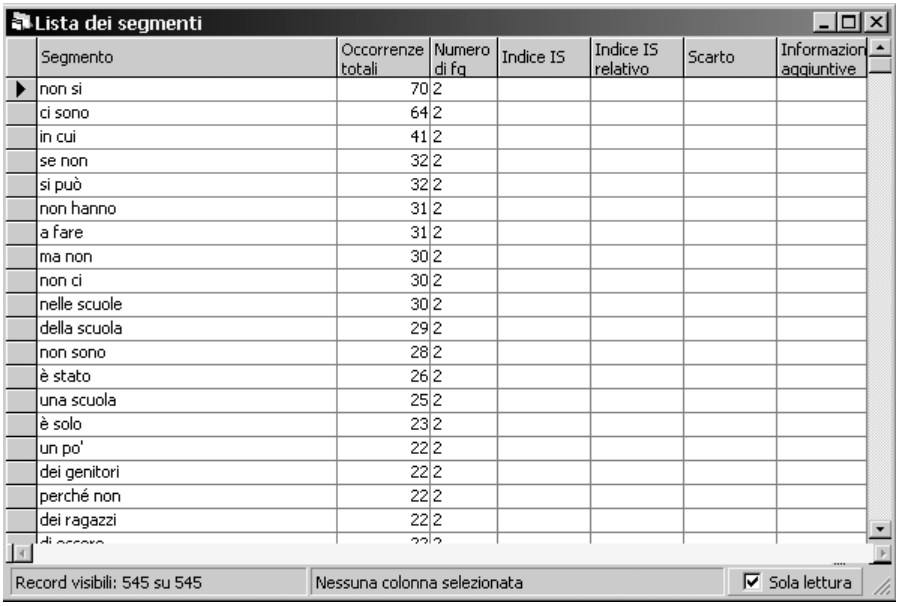
Fig. 10.4 – Parametri di segmentazione del corpus

TaLTaC² permette di eliminare i segmenti ridondanti e i segmenti vuoti utilizzando, durante la procedura di estrazione, due file (adeguatamente modificabili dall'utente) di parole vuote all'inizio (*VuoteI.txt*) e alla fine (*VuoteF.txt*) del segmento.

Dal barra degli strumenti, cliccando sul **Individua segmenti** (o dal me-

nu **Analisi** – *Analisi Lessicale*, selezionando *Analisi dei Segmenti* e poi la voce *Individuazione dei segmenti*) si accede alla finestra di dialogo per l'esecuzione della procedura (fig. 10.4).

L'individuazione dei segmenti ripetuti, soprattutto nei file molto grandi, può essere una procedura *time expensive*. È sempre opportuno scegliere una soglia minima di occorrenze per i segmenti da individuare e importare nella lista. I segmenti ripetuti vengono individuati come tutte le disposizioni a 2, 3, ..., *q* parole (numero massimo di parole nel segmento) con classe di occorrenze *i* (soglia minima di occorrenze delle parole appartenenti al segmento) che si ripetono *n* volte nel corpus (importa solo i segmenti con occorrenze pari almeno a...); i segmenti individuati dipendono da *i* , ma i segmenti visualizzati nella lista dipendono dalla soglia *n*.


Lista dei segmenti

Segmento	Occorrenze totali	Numero di fq	Indice IS	Indice IS relativo	Scarto	Informazioni aggiuntive
non si	70	2				
ci sono	64	2				
in cui	41	2				
se non	32	2				
si può	32	2				
non hanno	31	2				
a fare	31	2				
ma non	30	2				
non ci	30	2				
nelle scuole	30	2				
della scuola	29	2				
non sono	28	2				
è stato	26	2				
una scuola	25	2				
è solo	23	2				
un po'	22	2				
dei genitori	22	2				
perché non	22	2				
dei ragazzi	22	2				

Record visibili: 545 su 545 | Nessuna colonna selezionata | Sola lettura

Fig. 10.5 – Segmenti ripetuti

È evidente che se si vogliono visualizzare i segmenti che si ripetono almeno *n* volte è del tutto inutile fissare $i < n$ (un segmento costituito da almeno una parola che ha 3 occorrenze non può presentarsi con 4 occorrenze). Nel nostro esempio scegliamo di selezionare 4 come soglia minima di occorrenze delle parole e 4 come soglia minima di occorrenze dei segmenti da importare nella lista. Con questi parametri vengono individuati 545 segmenti (fig. 10.5).

Come di consueto, i segmenti possono essere ordinati per occorrenze o per lunghezza utilizzando i bottoni di ordinamento sulla barra degli strumenti.

Sulla selezione dei segmenti ripetuti è possibile calcolare un **indice di significatività** (IS) dei segmenti per valutare la loro rilevanza nel corpus. Dal menu **Analisi** – *Analisi Lessicale*, selezionare *Analisi dei Segmenti* e poi la voce *Calcolo indice IS su – Lista dei segmenti [TALTAC]*. Alla richiesta di calcolare l'indice IS anche sugli hapax rispondiamo "No". L'output della lista dei segmenti (visualizzabile dal bottone **Accesso ai Database** sulla barra degli strumenti) si arricchisce di due colonne: l'indice IS e l'indice IS relativo (fig. 10.6).

Lista dei segmenti (con indice IS)

Segmento	Occorrenze totali	Numero di fq	Indice IS	Indice IS relativo	Scarto	Informazio aggiuntive
arti marziali	6	2	4,00	1,00		
aria fritta	5	2	3,67	0,92		
diversamente abili	4	2	3,60	0,90		
di arti marziali	4	3	2,67	0,30		
assistenti sociali	4	2	2,62	0,65		
certezza della pena	6	3	2,41	0,27		
sta accadendo	4	2	2,32	0,58		
35 anni	4	2	2,08	0,52		
11 anni	4	2	2,08	0,52		
è diventata	7	2	2,02	0,50		
nuove generazioni	5	2	1,91	0,48		
maggior parte	4	2	1,88	0,47		
quando andavo a scuola	5	3	1,70	0,19		

Record visibili: 545 su 545 | Nessuna colonna selezionata | Sola lettura

Fig. 10.6 – Segmenti ripetuti: indice di significatività (IS)

Come si interpreta l'indice assoluto IS? L'indice mostra il grado di assorbimento del segmento ripetuto rispetto alle parole che lo costituiscono (Morrone, 1993; Bolasco, 1999, p. 221).

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

Per ciascuna delle forme (*L*) che compongono il segmento si considera il rapporto tra le occorrenze del segmento (f_{segm}) sulla forma grafica che ne fa parte (f_{fg}). La somma da 1 a *L* di questi rapporti viene moltiplicata per le parole piene (*P*) che costituiscono il segmento.

Il segmento *certezza della pena* (3) è composto da *certezza* (8), *della* (249) e

pena (14), pertanto:

$$IS = \left[\frac{6}{8} + \frac{6}{249} + \frac{6}{14}\right] \times 2 = 2{,}41$$

L'**indice IS relativo** viene rapportato al suo massimo (L^2) e quindi varia tra 0 e 1. I due indicatori offrono informazioni diverse. L'indice IS assoluto è fortemente condizionato dal numero di parole piene che costituiscono il segmento, pertanto mette in evidenza i segmenti più lunghi, costituiti da un maggior numero di parole, ma anche meno frequenti. L'indice IS relativo mette ai primi ranghi i segmenti più corti che spesso rappresentano i termini specialistici del lessico. Infatti chiedendo l'ordinamento dei segmenti secondo il valore decrescente dell'indice IS relativo troveremo nei primi ranghi *arti marziali* (1,00), *aria fritta* (0,92), *diversamente abili* (0,90), *assistenti sociali* (0,65). I segmenti ripetuti con un grado di assorbimento più elevato sono evidentemente dei poliformi che conviene trattare come una sola forma grafica (un'unica parola formata da una lessia complessa) piuttosto che attraverso le forme grafiche che li compongono.

Il trattamento dei poliformi avviene attraverso la procedura di **lessicalizzazione**, attraverso la quale il software viene istruito a riconoscere i segmenti che vogliamo trasformare in lessie complesse fino a modificare il vocabolario di base per le operazioni successive. Il segmento *arti marziali*, ad esempio, con la lessicalizzazione viene modificato in *arti_marziali*.

La procedura di lessicalizzazione inizia con la marcatura dei segmenti nella colonna delle "Informazioni aggiuntive" scrivendo nel campo in corrispondenza del segmento scelto un codice a scelta, ad esempio "s". Per rendere possibile la scrittura nei campi della lista dei segmenti è necessario smarcare la casella *Sola lettura* in basso a destra. Questa operazione deve essere eseguita con molta attenzione utilizzando il cursore e il mouse per passare da una riga all'altra. Si eviti invece di cliccare sulla barra di scorrimento a sinistra, perché questo provocherà un errore (sebbene senza perdita di dati) e la necessità di riavviare il programma.

Ultimata la fase di marcatura dei segmenti prescelti:

1) Selezionare l'intestazione di colonna "Informazioni aggiuntive".
2) Cliccare sull'icona **TDM** sulla barra degli strumenti per aprire la corrispondente finestra di dialogo.
3) Marcare l'opzione di ricerca *Records LIKE* inserendo nel campo il codice prescelto (per es. "s").
4) Cliccare su *OK*.

A questo punto la lista selezionata conterrà esclusivamente i segmenti da lessicalizzare. Pertanto la lista dovrà essere salvata dal menu **File** selezionando *E-*

sporta - In un file di testo e poi la voce *Lista di lessicalizzazione/tematizzazione*. La lista verrà salvata nella cartella di lavoro con il nome *Lista di lessicalizzazione.txt*.

La procedura prosegue con la lessicalizzazione dei segmenti selezionati e la successiva rinumerizzazione che riporta i segmenti selezionati nel vocabolario della sessione.

Dal menu **Analisi** – *Analisi Lessicale*, selezionando *Analisi dei Segmenti* e poi la voce *Lessicalizzazione* viene richiamata la finestra di dialogo *Lessicalizzazione di poliformi* con la richiesta di indicare la lista dei segmenti da unire nel testo. Cliccando sul bottone a destra si può selezionare dalla cartella di lavoro il file *Lista di lessicalizzazione.txt* precedentemente salvato (o un altro file preparato appositamente dall'utente). Cliccando su *OK* la procedura sarà eseguita con una avvertenza finale: "la lessicalizzazione/tematizzazione ha prodotto dei cambiamenti nelle occorrenze delle forme grafiche. È necessario procedere all'aggiornamento della Lista dei Segmenti". Ora verranno aperte le finestre per eseguire i calcoli necessari. Cliccando su *OK* apparirà una nuova finestra di dialogo del calcolo dei segmenti per l'inserimento dei parametri adeguati. Ad operazione conclusa le modifiche saranno riportate sia nella tabella dei segmenti che nella lista *Vocabolario (con TAG grammaticale)*, se quest'ultimo è già stato eseguito. Ovviamente nella tabella dei segmenti l'indice IS e l'indice IS relativo sono azzerati.

Dalla tabella *Vocabolario*, possiamo visualizzare le forme lessicalizzate, selezionando la colonna "CAT_SEM", utilizzando lo strumento **TDM**, con l'opzione di ricerca *Records LIKE*, e scrivendo "*FLESS*" nel campo corrispondente.

10. 3. ESTRAZIONE DELLE FORME SPECIFICHE

Per un approfondimento dell'analisi del corpus possiamo individuare quali sono le forme caratteristiche per ciascuna delle partizioni (nel nostro caso, per ciascuna variabile scelta per caratterizzare i messaggi: "genere" e "operatore").

L'analisi di specificità è sempre basata sulla sovra o sotto-utilizzazione delle forme rispetto a un modello di riferimento (Bolasco, 1999, p. 223). I modelli di riferimento possono essere i lessici di frequenza (come vedremo nel capitolo successivo per l'estrazione del linguaggio peculiare) oppure l'intero corpus rispetto a una sua parte, come nel nostro caso. La misura di specificità è data dal seguente scarto standardizzato della frequenza relativa,

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

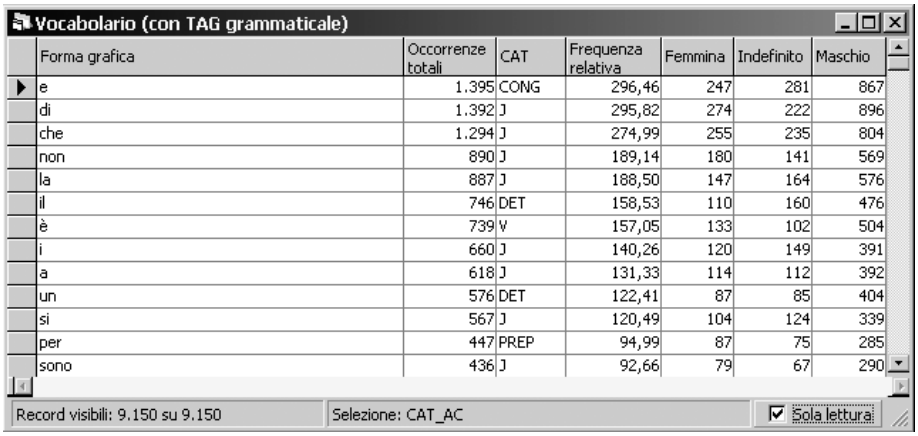
dove f_i è il numero delle occorrenze normalizzate della *i*-esima forma grafica nella parte in esame e f_i^* è il valore corrispondente nel modello di riferimento (corpus o lessico di frequenza). Il valore al denominatore è lo scarto quadratico medio della frequenza relativa; poiché la frequenza relativa di una parola, nell'analisi di un corpus, è sempre bassissima, di fatto lo s.q.m. equivale alla radice quadrata della frequenza teorica (Bolasco, 1999, p. 227). Come già si è visto per quanto riguarda il software Lexico3, la stima della significatività dello scarto standardizzato viene calcolata in base alla legge di distribuzione ipergeometrica che è la legge della variabile *N* = "numero di oggetti del tipo V che trovo raccogliendo a caso *n* oggetti tra una quantità *q* di oggetti", dato come noto il numero *v* degli oggetti di tipo V. Per l'analisi delle forme grafiche si assume che V siano le forme grafiche distinte, *v* siano le rispettive occorrenze, *n* siano le occorrenze delle parole che costituiscono il corpus e *q* siano le occorrenze delle parole che costituiscono il lessico (Tuzzi, 2003, pp. 131-134).

Prima di procedere con l'analisi è necessario preparare la tabella *Vocabolario* con le sub-occorrenze delle partizioni. Dal menu **Analisi** selezioniamo il comando *Pre-trattamento – Calcolo sub-occorrenze*. La finestra di dialogo ci chiede di scegliere una o più variabili di descrizione del corpus; scegliamo "Genere" e confermiamo con *OK*. Nella tabella *Vocabolario* si aggiungono tre nuove colonne (fig. 10.7) che descrivono la distribuzione delle forme secondo il genere degli autori dei messaggi: "Femmina", "Maschio" e "Indefinito" per gli autori che non è stato possibile classificare con certezza.

Ora possiamo procedere con l'analisi. Dal menu **Analisi**, – *Analisi Lessicale*, selezionando *Analisi delle specificità* e poi marcando la variabile "Genere" nella finestra di scelta delle partizioni, cliccando su *OK* otteniamo una finestra di dialogo che ci chiede di indicare il valore di *alfa*, e cioè la soglia di probabilità al di sotto della quale possiamo ritenere che le forme siano caratteristiche.
La soglia di default è 0,025. Abbassandolo il numero di forme diminuisce e si alza il livello di pertinenza delle forme estratte. Con testi piccoli e medi un eventuale abbassamento della soglia può ridurre eccessivamente il campo semantico di interpretazione del risultato.

Cliccando su *OK* ci appare un'altra finestra in cui ci viene chiesto di indicare la soglia di occorrenze delle forme per la partizione. La soglia minima potrebbe essere almeno pari alle modalità della variabile (in questo caso 3) ma, salvo casi particolari, è bene lasciare il valore indicato (10). La procedura di

calcolo termina con la richiesta di salvataggio nella cartella di lavoro del file delle forme specifiche estratte.



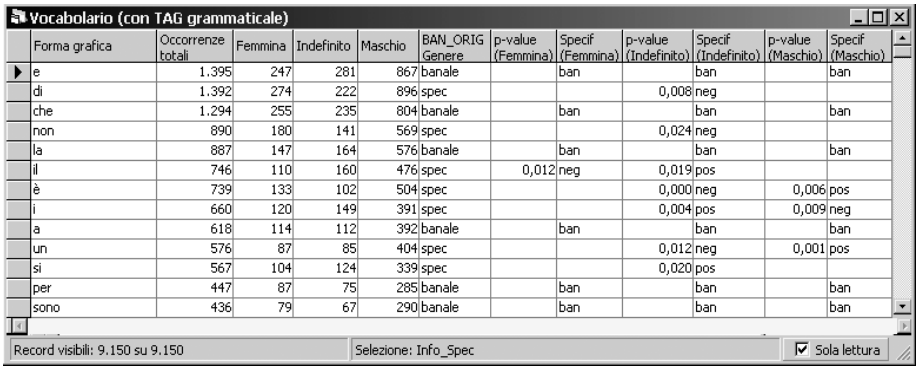
Vocabolario (con TAG grammaticale)

Forma grafica	Occorrenze totali	CAT	Frequenza relativa	Femmina	Indefinito	Maschio
e	1.395	CONG	296,46	247	281	867
di	1.392	J	295,82	274	222	896
che	1.294	J	274,99	255	235	804
non	890	J	189,14	180	141	569
la	887	J	188,50	147	164	576
il	746	DET	158,53	110	160	476
è	739	V	157,05	133	102	504
i	660	J	140,26	120	149	391
a	618	J	131,33	114	112	392
un	576	DET	122,41	87	85	404
si	567	J	120,49	104	124	339
per	447	PREP	94,99	87	75	285
sono	436	J	92,66	79	67	290

Record visibili: 9.150 su 9.150 | Selezione: CAT_AC | Sola lettura

Fig. 10.7 – Vocabolario con calcolo delle sub-occorrenze della partizione "Genere".

La tabella *Vocabolario* si arricchisce di altre informazioni (fig. 10.8): nella colonna 6 (BAN_ORIG Genere) sono indicate le forme che hanno una specificità (positiva o negativa) nella partizione. Le forme con occorrenze inferiori alla soglia richiesta (10) presentano il campo vuoto. Le forme con 10 occorrenze concentrate in una sola parte sono indicate con "spec_orig", evidenziabili con un ordinamento decrescente (*pubblica* e *solidarietà* per la modalità "Maschio").



Vocabolario (con TAG grammaticale)

Forma grafica	Occorrenze totali	Femmina	Indefinito	Maschio	BAN_ORIG Genere	p-value (Femmina)	Specif (Femmina)	p-value (Indefinito)	Specif (Indefinito)	p-value (Maschio)	Specif (Maschio)
e	1.395	247	281	867	banale		ban		ban		ban
di	1.392	274	222	896	spec			0,008	neg		
che	1.294	255	235	804	banale		ban		ban		ban
non	890	180	141	569	spec			0,024	neg		
la	887	147	164	576	banale		ban		ban		ban
il	746	110	160	476	spec	0,012	neg	0,019	pos		
è	739	133	102	504	spec			0,000	neg	0,006	pos
i	660	120	149	391	spec			0,004	pos	0,009	neg
a	618	114	112	392	banale		ban		ban		ban
un	576	87	85	404	spec			0,012	neg	0,001	pos
si	567	104	124	339	spec			0,020	pos		
per	447	87	75	285	banale		ban		ban		ban
sono	436	79	67	290	banale		ban		ban		ban

Record visibili: 9.150 su 9.150 | Selezione: Info_Spec | Sola lettura

Fig. 10.8 - Specificità

Nelle colonne successive per ciascuna forma grafica è indicato il valore di probabilità *alfa* ≤ 0,025 ("p-value") e la sua specificità positiva o negativa; in altre parole se la sua presenza nella parte è sovra o sottodimensionata rispetto alle attese senza che questo risultato sia dovuto al puro effetto del caso.

La lettura della specificità diventa più agevole se utilizziamo lo strumento **TDM** per creare delle tabelle di partizione in cui evidenziare le forme con specificità positiva per ciascuna modalità.

Selezioniamo la colonna 7, "p-value (Femmina)", e con lo strumento **TDM** inseriamo il criterio <= 0.025[11]. Sulla tabella risultante selezioniamo la colonna 8, "Specif (Femmina)", e nel campo *Records LIKE* del **TDM** scriviamo "pos*" ricordandoci di spuntare la casella *Applica solo ai record visibili*. In questo modo otterremo una tabella con le forme con specificità positiva della modalità "Femmina" (fig. 10.9) ordinabili (**ordine crescente A→Z**) a partire dal valore di probabilità più basso (e quindi più caratteristiche).

Vocabolario (con TAG grammaticale)

Forma grafica	Occorrenze totali	Femmina	Indefinito	Maschio	BAN_ORIG Genere	p-value (Femmina)	Specif (Femmina)	p-value (Indefinito)	Specif (Indefinito)	p-value (Maschio)	Specif (Maschio)
bambini	28	16	1	11	spec	0,000	pos	0,023	neg	0,007	neg
miei	25	13	0	12	spec	0,000	pos				
noi	68	24	17	27	spec	0,000	pos			0,000	neg
mamma	16	9	4	3	spec	0,001	pos			0,000	neg
possiamo	14	8	0	6	spec	0,001	pos				
nostro	17	9	3	5	spec	0,001	pos			0,004	neg
diversi	12	7	2	3	spec	0,002	pos			0,008	neg
insegnanti	90	27	15	48	spec	0,003	pos				
due	35	13	4	18	spec	0,006	pos				
vanno	24	10	4	10	spec	0,006	pos			0,023	neg
elementari	11	6	0	5	spec	0,007	pos				
loro	191	48	35	108	spec	0,008	pos			0,024	neg
capire	25	10	2	13	spec	0,008	pos				
qui	25	10	5	10	spec	0,008	pos			0,014	neg
punizione	30	11	4	15	spec	0,012	pos				
vita	48	15	11	22	spec	0,018	pos			0,009	neg
puniti	13	6	3	4	spec	0,018	pos			0,017	neg

Record visibili: 25 su 9.150 | Selezione: p-value (Femmina) | Sola lettura

Fig. 10.9 – Parole specifiche della modalità "Femmina"

Bambini, *miei*, *noi*, *mamma*, *possiamo* sono le parole più caratteristiche nei messaggi con autore di genere femminile. La stessa operazione condotta per il genere maschile metterebbe in evidenze parole come: *stato*, *video*, *ragazzo*, *prof*. Le parole specifiche della modalità "operatore scolastico" sono: *scuola*, *lezioni*, *insegnanti*, *noi*, *ragazzi*; quelle della modalità "non operatore" sono: *ti*, *ho*, *mai*, *addirittura*.

[11] Attenzione! Nello strumento **TDM** i valori numerici devono essere scritti con la notazione inglese 0.025 perché la virgola è interpretata come parte di un stringa di testo.

10. 4. ESTRAZIONE DELLE FORME PECULIARI

L'estrazione delle forme più caratteristiche di un corpus può essere effettuata anche per confronto con un lessico di riferimento. In questo caso si tratta di estrarre le forme peculiari del corpus, indipendentemente dalle partizioni. Vediamo, prima di tutto, quali sono i lessici presenti nelle risorse di TaLTaC². Dal menu **Visualizza** selezioniamo la voce *Risorse statistico-linguistiche*:

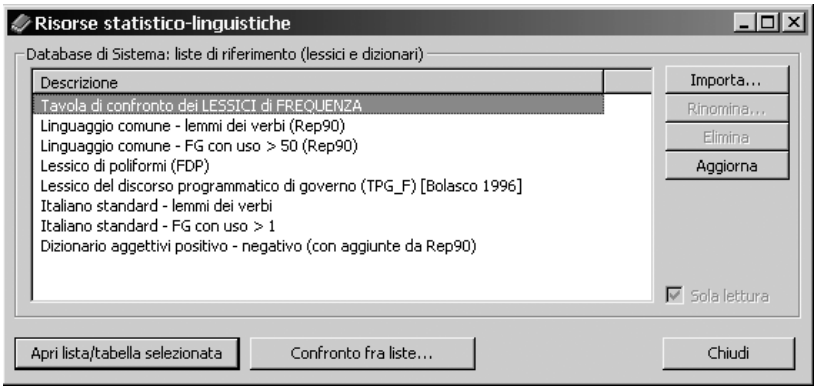


Fig. 10.10 – Risorse statistico-linguistiche

- La *Tavola di confronto dei lessici di frequenza* elenca 14.943 lemmi con le rispettive misure di frequenza tratte da:
  - *Vocabolario di Base della lingua italiana (VdB)* di Thornton, Iacobini e Burani (1997): per questi lemmi si conosce solo la fascia di frequenza.
  - *Vocabolario fondamentale (Vfond)* costituito di 2.739 lemmi di massima disponibilità tratti dal LIF.
  - *Lessico Italiano di Frequenza (LIF) dell'italiano scritto*, di Bortolini e altri (1972) costituito di 5.360 lemmi.
  - *Lessico di frequenza dell'italiano parlato (LIP)* di De Mauro, Vedovelli, Voghera, Mancini (1993).
  - *Vocabolario elettronico della lingua italiana (VELI, 1989)*, costituito da 9.994 lemmi.
  - *Lessico dei bambini (LE)*, 2029 lemmi.
  - *Lessico del discorso politico* (TPG), 2999 lemmi.
- *Linguaggio comune – lemmi dei verbi (Rep90)* è una lista costituita da 4.907 lemmi di verbi tratti da forme non ambigue.
- *Linguaggio comune – FG con uso >50 (Rep90)* è una lista costituita da 60.489

forme grafiche, con indice d'uso > 50 tratte da una raccolta di 270 milioni di occorrenze di 10 annate del quotidiano "La Repubblica" (1990-1999).

- Il *Lessico di Poliformi (FDP)* è stato costruito sulla base di un campione di oltre 4 milioni di occorrenze (121.786 forme grafiche diverse) del linguaggio contemporaneo (scritto e parlato) ed elenca 3.925 poliformi con le rispettive frequenze.
- *Lessico del discorso programmatico di governo (TPG_F)* è il lessico costituito da 3.000 lemmi tratti da un'analisi dei discorsi programmatici dei Presidenti del Consiglio della Prima Repubblica dal 1948 al 1994 (Bolasco, 1996).
- *Italiano standard – lemmi dei verbi* elenca 2.605 lemmi di verbi tratti da forme grafiche non ambigue.
- *Italiano standard – FG>1* è un lessico basato sul campione del lessico dei Poliformi (stampa, discorsi parlamentari, documenti ufficiali, saggistica, biografie, interviste, dialoghi, composizioni scolastiche) costituito di forme grafiche con indice d'uso >1 (50.464 forme grafiche distinte).
- *Dizionario aggettivi positivo - negativo (con aggiunte da Rep90)* è un dizionario di forme flesse di 6.000 aggettivi per il tagging semantico realizzato a partire dal *General Inquirer* di P.J. Stone (1966) con integrazioni dal lessico del quotidiano *La Repubblica*.

Possiamo effettuare un confronto tra il vocabolario del corpus *Bullismo* con tag grammaticale e *Italiano standard* che riporta le forme grafiche del lessico contemporaneo.

Dal menu **Analisi** – *Analisi lessicale*, selezioniamo il comando *Linguaggio peculiare - Confronto con un lessico di frequenza*. La finestra di dialogo (fig. 10.11) ci offre alcune opzioni. Questo primo confronto è effettuato sulla lista di **intersezione**: ciò che vogliamo ottenere è un tabella con le forme grafiche comuni che ci permetta di identificare le forme sovra e sottoutilizzate nel nostro lessico (*Bullismo*) rispetto al lessico di confronto. Per le due liste, opportunamente selezionate dalle liste della sessione (per la lista da confrontare: *Vocabolario con Tag Grammaticale*) e dalle risorse interne di TaLTaC2 (per la lista "modello": *Italiano standard – FG con uso > 1*), deve essere indicato il campo sul quale effettuare il confronto. In questo caso si tratta del campo "Forma grafica". I campi da inserire per la visualizzazione dell'output dipendono dagli scopi del confronto. Il campo su cui calcolare lo scarto standardizzato è quello delle occorrenze totali in entrambe le liste (le frequenze d'uso sono le occorrenze ponderate con una misura di dispersione delle forme nel testo).

La marcatura della casella *Maiuscole/minuscole* comporta, come al solito, il tener conto della presenza delle maiuscole (*case sensitive*). Con la marcatura della casella *Scrivi scarto nel vocabolario* i valori dello scarto standardizzato saranno ag-

giunti alla tabella *Vocabolario (con Tag grammaticale)* per utilizzazioni successive che permettono di combinare linguaggio specifico e linguaggio peculiare. Cliccando su *OK* ci viene chiesto di inserire una descrizione della lista di intersezione che si aggiungerà alle liste della sessione.

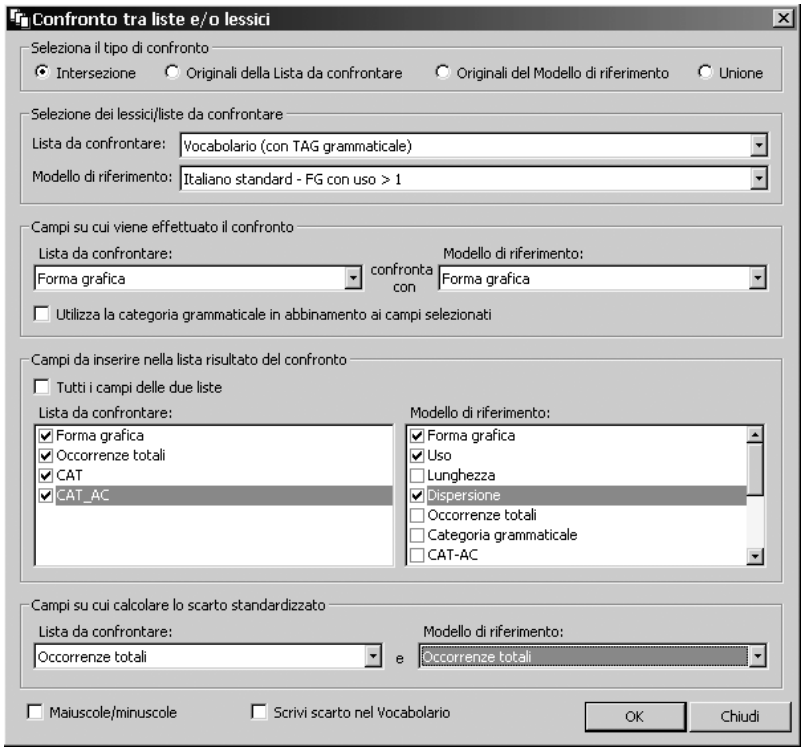
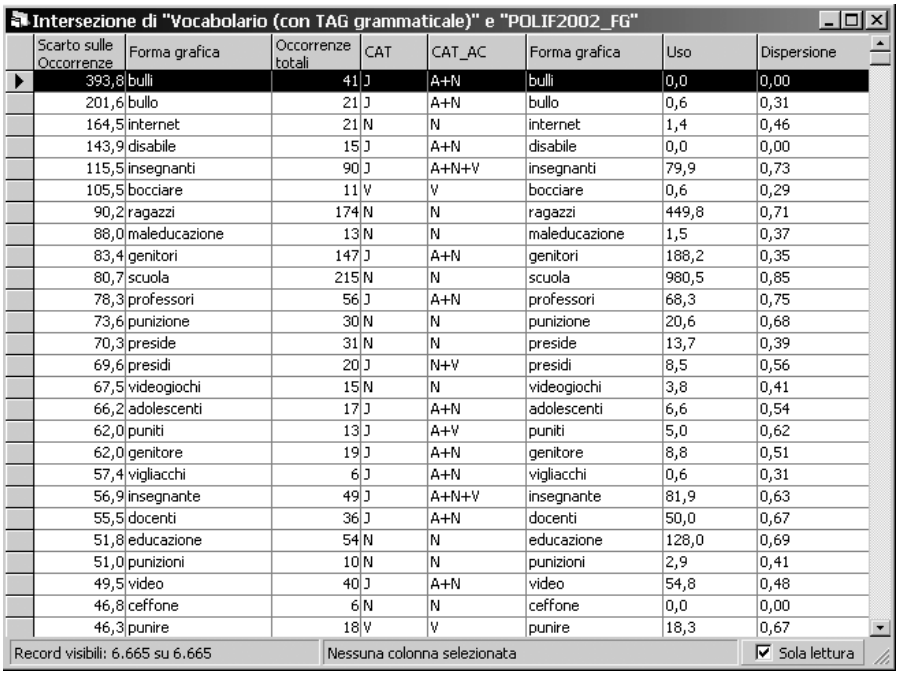


Fig. 10.11 – Finestra di dialogo del confronto con un lessico di frequenza

La tabella risultato del confronto si presenta con un ordinamento delle forme per valori decrescenti dello scarto standardizzato (fig. 10.12). Sono significative, cioè sovra-utilizzate rispetto a quanto lo sono nel linguaggio di riferimento, le forme con uno scarto maggiore di 3,84 che è il valore del χ^2 con 1 grado di libertà e p-value=0,05.

Nella prima colonna sono riportati i valori dello scarto standardizzato sulle occorrenze normalizzate dai quali possiamo trarre informazioni sulle forme peculiari di questo corpus rispetto al lessico standard.



Intersezione di "Vocabolario (con TAG grammaticale)" e "POLIF2002_FG"

Scarto sulle Occorrenze	Forma grafica	Occorrenze totali	CAT	CAT_AC	Forma grafica	Uso	Dispersione
393,8	bulli	41	J	A+N	bulli	0,0	0,00
201,6	bullo	21	J	A+N	bullo	0,6	0,31
164,5	internet	21	N	N	internet	1,4	0,46
143,9	disabile	15	J	A+N	disabile	0,0	0,00
115,5	insegnanti	90	J	A+N+V	insegnanti	79,9	0,73
105,5	bocciare	11	V	V	bocciare	0,6	0,29
90,2	ragazzi	174	N	N	ragazzi	449,8	0,71
88,0	maleducazione	13	N	N	maleducazione	1,5	0,37
83,4	genitori	147	J	A+N	genitori	188,2	0,35
80,7	scuola	215	N	N	scuola	980,5	0,85
78,3	professori	56	J	A+N	professori	68,3	0,75
73,6	punizione	30	N	N	punizione	20,6	0,68
70,3	preside	31	N	N	preside	13,7	0,39
69,6	presidi	20	J	N+V	presidi	8,5	0,56
67,5	videogiochi	15	N	N	videogiochi	3,8	0,41
66,2	adolescenti	17	J	A+N	adolescenti	6,6	0,54
62,0	puniti	13	J	A+V	puniti	5,0	0,62
62,0	genitore	19	J	A+N	genitore	8,8	0,51
57,4	vigliacchi	6	J	A+N	vigliacchi	0,6	0,31
56,9	insegnante	49	J	A+N+V	insegnante	81,9	0,63
55,5	docenti	36	J	A+N	docenti	50,0	0,67
51,8	educazione	54	N	N	educazione	128,0	0,69
51,0	punizioni	10	N	N	punizioni	2,9	0,41
49,5	video	40	J	A+N	video	54,8	0,48
46,8	ceffone	6	N	N	ceffone	0,0	0,00
46,3	punire	18	V	V	punire	18,3	0,67

Record visibili: 6.665 su 6.665 — Nessuna colonna selezionata — Sola lettura

Fig. 10.12 – Intersezione fra il Vocabolario *Bullismo* e il lessico dell'*Italiano Standard*

Le parole con il valore più alto dello scarto sono le più caratterizzanti: *bulli*, *bullo*, *internet*, *disabile*, *insegnanti*. La selezione di queste forme permette di estrarre l'informazione che risponde meglio agli obiettivi di approfondimento del contenuto. Sul complesso delle forme grafiche distinte presenti nel corpus (9.150 dopo la lessicalizzazione) le forme peculiari (con uno scarto > 3,84) sono 2.213 (24,18%) e, di queste solo 447 superano la soglia delle 4 occorrenze (il 4,88% del totale).

Nella colonna 8 troviamo un indice di dispersione delle parole nei testi che sono serviti di base per la costruzione del lessico di confronto (*Italiano Standard*). La dispersione è una misura della diffusione di una parola nelle parti che costituiscono il corpus. Di solito le parole con dispersione più alta nel lessico di confronto sono le parole grammaticali, mentre quelle con dispersione più bassa sono le vere e proprie parole dense di contenuto.

Le "frequenze d'uso" della colonna 7 del lessico di confronto rappresentano una ponderazione delle occorrenze sulla base della dispersione: una forma con un elevato valore di dispersione e una alta occorrenza presenta un in-

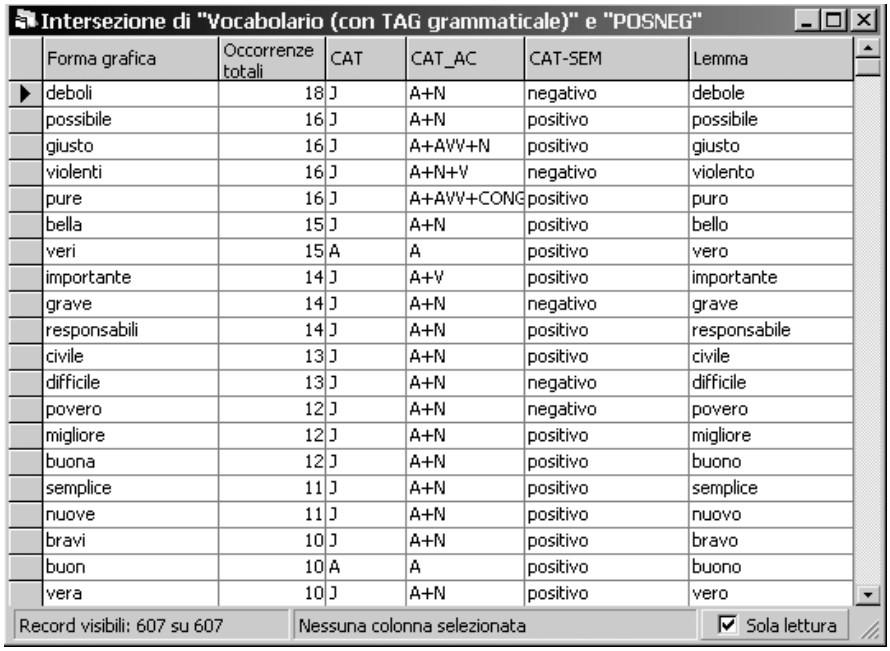
dice d'uso elevato (per es. la congiunzione *e*). Questi indici possono essere calcolati anche per il *Vocabolario* del corpus *Bullismo* dal menu **Calcola** con il comando *Dispersione e Uso*.

La seconda opzione della finestra di dialogo per l'estrazione del linguaggio peculiare (fig. 10.11), *Originali della lista da confrontare*, permette di estrarre le forme grafiche peculiari che sono presenti nel file di confronto ma non nel file di modello. In questo caso, evidentemente, non sarà necessario calcolare lo scarto in quanto le forme visualizzate sono forme "uniche", assolutamente originali. Nel nostro esempio l'output con questa opzione visualizza 2.485 forme, tra le quali: *bullismo*, *tv*, *in classe*, *a casa*, ecc.).

10. 5. CONFRONTO CON UN DIZIONARIO TEMATICO: AGGETTIVI POSITIVI E NEGATIVI

La presenza del dizionario degli aggettivi positivi e negativi tra le risorse statistico-linguistiche di TaLTaC² permette di etichettare la terminologia positiva e negativa di un corpus e di compiere una valutazione d'insieme sul "tono" dei testi sottoposti ad analisi. Si tratta di un approccio tipico all'interno della *Content Analysis* già presente nei primi lavori di Willey (1926) e Woodward (1934) sui giornali quotidiani e poi nell'analisi del linguaggio politico e della propaganda di Lasswell e Leites (1949). La valutazione di una comunicazione o di uno stimolo in termini di contrapposizione positivo-negativo è stata teorizzata da Osgood nei suoi studi sul differenziale semantico (Osgood, Tannenbaum e Suci, 1957) all'interno delle tre dimensioni che definiscono lo spazio del significato (valutazione, potenza e attività). Il tema è stato poi ripreso da Stone e dai suoi collaboratori (1966) per il *General Inquirer*, un programma di categorizzazione automatica dei documenti che utilizza diversi lessici di riferimento oltre alla classificazione psico-cognitiva di Osgood. Nel 1966, Boucher e Osgood formularono la "Pollyanna hypothesis"[12], poi confermata da numerosi studi successivi, secondo la quale c'è una asimmetria tra le qualificazioni positive e negative nel linguaggio perché gli esseri umani tendono a descrivere la realtà più positivamente di quanto non sia (Boucher e Osgood. 1969).

[12] Pollyanna Whittier è una bambina dotata di un inguaribile ottimismo, protagonista di due romanzi per ragazzi di Eleanor H. Porter, *Pollyanna* (1913) e *Pollyanna grows up* (1915).



Intersezione di "Vocabolario (con TAG grammaticale)" e "POSNEG"

Forma grafica	Occorrenze totali	CAT	CAT_AC	CAT-SEM	Lemma
deboli	18	J	A+N	negativo	debole
possibile	16	J	A+N	positivo	possibile
giusto	16	J	A+AVV+N	positivo	giusto
violenti	16	J	A+N+V	negativo	violento
pure	16	J	A+AVV+CONG	positivo	puro
bella	15	J	A+N	positivo	bello
veri	15	A	A	positivo	vero
importante	14	J	A+V	positivo	importante
grave	14	J	A+N	negativo	grave
responsabili	14	J	A+N	positivo	responsabile
civile	13	J	A+N	positivo	civile
difficile	13	J	A+N	negativo	difficile
povero	12	J	A+N	negativo	povero
migliore	12	J	A+N	positivo	migliore
buona	12	J	A+N	positivo	buono
semplice	11	J	A+N	positivo	semplice
nuove	11	J	A+N	positivo	nuovo
bravi	10	J	A+N	positivo	bravo
buon	10	A	A	positivo	buono
vera	10	J	A+N	positivo	vero

Record visibili: 607 su 607 | Nessuna colonna selezionata | Sola lettura

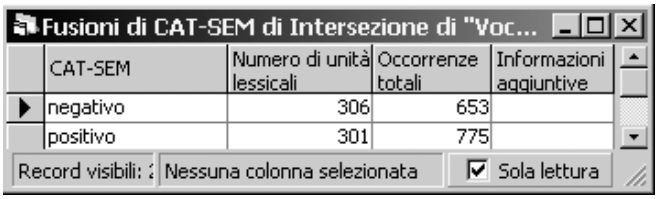
Fig. 10.13 – Intersezione fra il Vocabolario *Bullismo* e il *Dizionario aggettivi*

La procedura per etichettare il corpus utilizzando il dizionario degli aggettivi positivi e negativi è la seguente.

Dal menu **Analisi** – *Analisi lessicale*, selezioniamo il comando *Linguaggio peculiare - Confronto con un lessico di frequenza*. Con l'opzione *Intersezione* mettiamo a confronto il *Vocabolario (con TAG grammaticale)* con il modello di riferimento: *Dizionario aggettivi positivo-negativo*. Il campo sul quale effettuare il confronto è "Forma grafica" sia nella lista da confrontare che nel modello di riferimento Nei campi da inserire nella lista risultato selezionare "Forma grafica", "Occorrenze", "CAT" e "CAT-AC" nella finestra della *Lista da confrontare*; "CAT-SEM" e "Lemma" nella finestra del *Modello di riferimento*. È preferibile non utilizzare la categoria grammaticale di abbinamento per non escludere dal confronto le forme che il tagging grammaticale ha classificato come ambigue (J).

Per conteggiare il totale delle occorrenze e delle unità lessicali degli aggettivi positivi e negativi presenti nella tabella si deve procedere con una lemmatizzazione (paragrafo 9.7): selezionare la colonna "CAT-SEM" e, dal menu **Calcola**, il comando *Fusioni*. Verrà visualizzata una tabella riassuntiva con le due categorie semantiche presenti nella colonna selezionata, il numero di unità les-

sicali (forme flesse degli aggettivi) con cui la categoria è stata formata e il totale delle occorrenze per ciascuna categoria (fig. 10.14). Un indice di valutazione della negatività del testo è costituito da rapporto tra il totale delle occorrenze negative e il totale delle occorrenze positive (Occ Neg/Occ Pos): da una serie di prove effettuate su liste di riferimento risulta che un valore superiore a 0,40 indica un testo con una connotazione negativa (Bolasco e Della Ratta 2004). Nel nostro caso il valore dell'indice di negatività è di 0,84.


Fusioni di CAT-SEM di Intersezione di "Voc...

	CAT-SEM	Numero di unità lessicali	Occorrenze totali	Informazioni aggiuntive
▶	negativo	306	653	
	positivo	301	775	

Record visibili: 2 | Nessuna colonna selezionata | Sola lettura

Fig. 10.14 – Fusione della categorie semantiche Positivo - Negativo

Lo step successivo consiste nella valutazione più raffinata della positività o negatività delle partizioni in cui suddiviso il corpus. Per questo è necessario riportare l'informazione contenuta nella tabella di intersezione della fig. 10.13 nel *Vocabolario* del corpus *Bullismo* che contiene tutte le informazioni necessarie.


Vocabolario

	Forma grafica	Occorrenze totali	CAT	CAT_AC	CAT_SEM	Informazioni aggiuntive	Femmin	Indefini	Maschic	Incerto	No	Si
▶	deboli	18	J	A+N	,negativo,	,Intersez,	2	2	14	9	6	3
	giusto	16	J	A+AVV+N	,positivo,	,Intersez,	3	1	12	6	9	1
	pure	16	J	A+AVV+CONG	,positivo,	,Intersez,	3	2	11	8	6	2
	violenti	16	J	A+N+V	,negativo,	,Intersez,	0	7	9	9	7	0
	possibile	16	J	A+N	,positivo,	,Intersez,	2	0	14	8	4	4
	veri	15	A	A	,positivo,	,Intersez,	4	3	8	6	4	5
	bella	15	J	A+N	,positivo,	,Intersez,	1	4	10	3	10	2
	responsabili	14	J	A+N	,positivo,	,Intersez,	2	4	8	5	6	3
	importante	14	J	A+V	,positivo,	,Intersez,	5	3	6	6	8	0
	grave	14	J	A+N	,negativo,	,Intersez,	2	3	9	5	6	3
	civile	13	J	A+N	,positivo,	,Intersez,	2	3	8	5	6	2
	difficile	13	J	A+N	,negativo,	,Intersez,	3	3	7	6	1	6
	povero	12	J	A+N	,negativo,	,Intersez,	2	2	8	7	4	1
	migliore	12	J	A+N	,positivo,	,Intersez,	4	2	6	4	5	3
	buona	12	J	A+N	,positivo,	,Intersez,	0	3	9	6	6	0
	semplice	11	J	A+N	,positivo,	,Intersez,	2	4	5	5	5	1
	nuove	11	J	A+N	,positivo,	,Intersez,	5	1	5	2	6	3

Record visibili: 602 su 9.150 | Nessuna colonna selezionata | Sola lettura

Fig. 10.15 – Vocabolario *Bullismo* con tagging semantico positivo-negativo

Dal menu **Analisi** selezionare il comando *Analisi Lessicale* – *Tagging Semantico* e poi *Vocabolario [TALTAC]*. Nella finestra *Tagging Semantico* selezionare la scheda *Liste generate dai DB* e marcare l'opzione *Includi le tabelle dal DB della ses-*

sione. In *Seleziona una tabella* inserire: *Intersezione di Vocabolario (con TAG grammaticale)* e *POSNEG*; in *Seleziona il campo contenente la forma da cercare* inserire: *Forma grafica*; in *Seleziona il campo contenente la codifica da attribuire a CAT-SEM* inserire: "CAT-SEM". La tabella *Vocabolario della sessione* ora mostra gli aggettivi etichettati con le modalità negativo-positivo (fig. 10.15).

Sulla tabella *Vocabolario*, che contiene tutte le informazioni sulle suboccorrenze delle partizioni, è possibile calcolare l'indice di valutazione della negatività del testo per ciascuna modalità. Con la consueta procedura di fusione delle forme, selezioniamo la colonna "CAT-SEM" nella tabella *Vocabolario* e dal menu **Calcola** scegliamo il comando *Fusioni*. Alla richiesta del file da salvare inseriamo come descrizione *Fusioni di POS-NEG di Vocabolario*. Cliccando su *OK* otteniamo la tabella di fig. 10.16.

Fusioni di POS-NEG di Vocabolario

CAT_SEM	Numero di unità lessicali	Occorrenze totali	Informazioni aggiuntive	Femmin	Indefini	Maschio	Incerto	No	Si
,negativo,	303	650		113	123	414	233	291	126
,positivo,	299	773		128	135	510	251	372	150

Record visibili: 2 su 2 | Nessuna colonna selezionata | Sola lettura

Fig. 10.16 – Fusioni di *POS-NEG* secondo le partizioni

L'indice di negatività (pari a 0,87 nel corpus, successivamente a questa operazione) è marcatamente più basso nei messaggi inviati dai Maschi (0,81) rispetto a quelli inviati dalle Femmine (0,88). Il divario è ancora più ampio tra i messaggi inviati da coloro i quali non sono operatori scolastici, il cui indice di negatività è pari a 0,78 rispetto ai messaggi di quelli che lavorano nella scuola (0,84).

RIFERIMENTI BIBLIOGRAFICI

BOLASCO S. (1996) "Il lessico del discorso programmatico di governo", in M. Villone, A. Zuliani (a cura di), *L'attività dei governi della repubblica italiana (1947-1994)*, Bologna, il Mulino, pp. 163-349.

BOLASCO S. (1999) *Analisi multimensionale del dati. Metodi, strategie e criteri d'interpretazione*, Roma, Carocci (II ed. 2004).

BOLASCO S., DELLA RATTA-RINALDI F. (2004) "Experiments on Semantic Categorisation of Texts: Analysis of Positive and Negative Dimension", in G. Purnelle, C. Fairon, A. Dister (*eds*) *Les pois des mots. Actes des 7es Journées internationales d'Analyse statistiques des Données Textuelles*, Louvain, Presse Universitarie de Laouvain, pp. 202-210.

BOUCHER T., OSGOOD C. E. (1969) "The Pollyanna hypothesis", *Journal of Verbal*

Learning and Verbal Behavior, 8, pp. 1-8.

LASSWELL H. D., LEITES N. (*EDS*) (1949) *The Language of Politics. Studies in Quantitative Semantics*, New York, Stewart (tr. it. *Il linguaggio della politica. Studi di semantica quantitativa*, Torino, ERI, 1979).

MORRONE A. (1993) "Alcuni criteri di valutazione della significatività dei segmenti ripetuti", in AA.VV. *Secondes Journées internationales d'Analyse statistiques des Données Textuelles* (Montpellier), Paris, Télécom (École Nationale Supérieure des Télécommunications) pp. 445-453.

OSGOOD C. E., SUCI G. J., TANNENBAUM P. H. (1957) *The Measurement of Meaning*, Urbana (IL), University of Illinois Press.

STONE P. J., DUNPHY D. C., SMITH M. S. OGILVY, D. M. (1966) *The General Inquirer: A Computer Approach to Content Analysis*, Cambridge (MA), MIT Press.

TUZZI A. (2003*) L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*, Roma, Carocci.

WILLEY M. M. (1926). *The Country Newspaper: A Study of Socialization and Newspaper Content*, Chapel Hill, University of North Carolina Press.

WOODWARD J. L. (1934) "Quantitative Newspaper Analysis as a Technique of Opinion Research", *Social Forces*, 12, 526-537.

11.
LAVORARE CON TALTAC²: L'ANALISI DEL CONTENUTO

Il trattamento del corpus con analisi statistica del lessico prende in esame i testi nella loro frammentazione minima: la forma grafica. In alcuni casi si tratta di parole o di poliformi, in altri casi di segmenti più o meno estesi. Le forme vanno a costituire un vocabolario che si arricchisce man mano di informazioni. Nel vocabolario le forme grafiche distinte possono essere etichettate e conteggiate, classificate in categorie, confrontate tra di loro all'interno delle partizioni del corpus o rispetto a lessici di tipo diverso. Così facendo si mettono in rilievo diverse caratteristiche linguistiche e semantiche dei testi e questo ci permette di estrarre numerose informazioni in modo automatico. Analizzando corpora molto grandi, costituiti di milioni di occorrenze, l'estrazione del linguaggio peculiare e del linguaggio specifico, con l'applicazione dei relativi test di significatività sul sovra o sotto-uso delle parole, consente la costruzione di liste di forme grafiche con un alto contenuto informativo sebbene rappresentino solo il 10-20% del vocabolario complessivo.

Tuttavia ciò che l'analisi statistica del lessico non ci permette di fare è di tenere conto, in qualche modo, del contesto in cui appare la parola. Per sottoporre a misura il testo siamo costretti a spezzettarlo e questo ci fa perdere di vista, inevitabilmente, il contenuto della comunicazione in senso proprio: la sua natura di "tessuto linguistico".

Gli strumenti di analisi testuale integrati in TaLTaC² ci permettono di ritornare al testo, senza rinunciare all'approccio automatico, e di recuperare qualche modalità di indagine più propriamente ermeneutica e qualitativa.

11. 1. IL RECUPERO DI INFORMAZIONE: LE CONCORDANZE

Il tema delle concordanze e la loro importanza per situare le parole nel contesto è già stato trattato nel paragrafo 8.9 a proposito della procedura analoga in Lexico3. Qui ci limitiamo a fornire le indicazioni generali per la loro esecuzione in TaLTaC[2] .

Le concordanze semplici e complesse si possono eseguire direttamente dalla barra degli strumenti cliccando sul bottone corrispondente. La finestra di interrogazione (fig. 11.1) permette di selezionare la forma pivot sul vocabolario a sinistra del monitor e di visualizzarne immediatamente il risultato. La concordanza è descritta dall'identificatore del frammento (in questo caso, il numero del messaggio). L'ampiezza dell'intorno delle parole prima e dopo la forma pivot è definibile a piacere (di default è fissata a 50 caratteri, con completamento della parola).

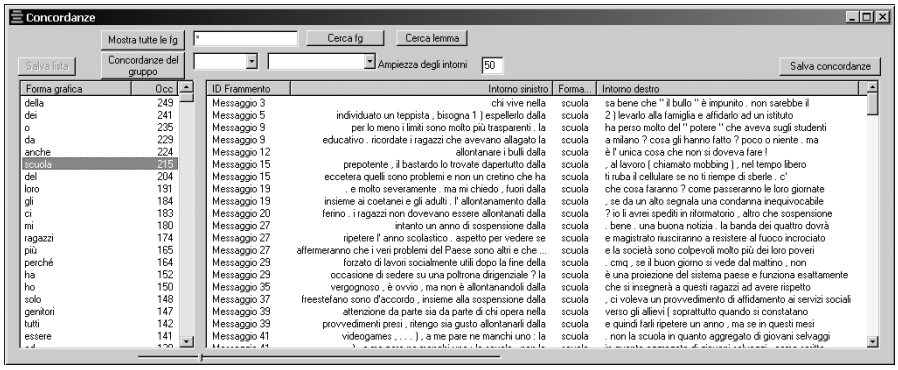


Fig. 11.1 – Concordanza della forma pivot *scuola*

La ricerca delle concordanze può essere effettuata anche per lemmi, scrivendo la forma nell'apposito spazio e cliccando su *Cerca lemma*; oppure per categoria grammaticale (*Concordanze del gruppo*) o per categoria semantica (in questo caso "positivo" o "negativo", ma potrebbero esserci altre categorie inserite dall'utente).

Dal menu **Analisi**, selezionando *Analisi Testuale – Recupero informazioni: Concordanze* e poi *Concordanze di segmenti...* si visualizza una analoga finestra di interrogazione per i segmenti ripetuti. Le concordanze sono esportabili come file di testo cliccando sul bottone *Salva concordanze*.

11. 2. L'ESTRAZIONE DI INFORMAZIONE PER PAROLE CHIAVE

L'indice IFTDF (*Term Frequency*, *Inverse Document Frequency*) è stato sviluppato all'interno di un approccio al Text Mining che si chiama *Information Retrieval* (IR). L'obiettivo è quello di recuperare l'informazione ritenuta significativa o utile da parte dell'utente in un determinato corpus di documenti. L'indice permette di ordinare i documenti secondo la loro rilevanza o pertinenza ed è costruito sulla base della frequenza delle "parole chiave" (*keywords*) e della loro distribuzione all'interno dei documenti che costituiscono il corpus.

Il fattore *tf* (*term frequency*) rappresenta la frequenza della generica *keyword* k_i all'interno del documento d_j, mentre il fattore *idf* (*inverse document frequency*) è dato dall'inverso della frazione di documenti che contengono k_i rispetto al totale dei documenti che compongono il corpus. In altre parole, una *keyword* è utile per rintracciare un documento ritenuto rilevante, e quindi ha un forte peso discriminante, se presenta una frequenza elevata ma, nel contempo, ricorre in un numero ridotto di documenti. La formula di calcolo dell'indice TFIDF è la seguente:

$$\text{TFIDF} = tf \times \log \frac{N}{n}$$

Dove il primo fattore *tf* misura il numero di occorrenze di una *keyword*, mentre il secondo fattore è il logaritmo del rapporto tra il numero (*N*) dei documenti che costituiscono il corpus e il numero (*n*) di documenti che contengono la *keyword*.

In TaLTaC² l'indice TFIDF è normalizzato rispetto alla lunghezza dei documenti per evitare che i documenti più lunghi acquisiscano un peso maggiore rispetto a quelli più corti. L'indice TFIDF può essere calcolato anche rispetto a tutte le forma grafiche del Vocabolario, scegliendo una soglia minima di rilevanza (che di default è pari a 5). In questo caso l'obiettivo diventa quello di estrarre il linguaggio rilevante non solo in base alla frequenza delle forme (che comunque è determinante) ma anche della distribuzione che le parole hanno all'interno dei documenti in cui il corpus è frammentato.

Tuttavia la particolare utilità dell'indice si presenta quando si ha l'obiettivo di ordinare i documenti secondo la pertinenza rispetto a una o più parole chiave scelte dall'utente secondo un certo criterio.

Nel nostro caso abbiamo scelto di preparare una lista di parole (con frequenza maggiore di 4) che riteniamo adeguate per descrivere in modo negativo il bullismo (la lista è stata salvata nella cartella di lavoro in un file denominato: *Vocabolario per TFIDF.txt*).

```
branco
bulletti
bulli
bullo
cattivo
colpevoli
delinquenti
idioti
ignoranti
ignoranza
maleducati
maleducazione
mostri
prepotenti
stupidi
stupidità
vigliacchi
violenti
violento
violenza
```

Per eseguire il calcolo dell'indice, dal menu **Analisi** selezionare il comando *Analisi Testuale – Estrazione di informazione: per parole chiave – Calcolo TFIDF*. Si aprirà la finestra di interrogazione *Calcolo del TFIDF* (fig. 11.2) nella quale (trascurando la soglia di occorrenze) si potrà selezionare il file della lista da utilizzare: *Vocabolario per TFIDF.txt*.

Premendo *OK* il calcolo viene eseguito con la richiesta di indicare il nome della colonna che si andrà ad aggiungere al Vocabolario con il valore dell'indice per ciascuna parola chiave utilizzata.

Ora dal menu **Analisi**, selezionando *Analisi Testuale – Estrazione di informazione: per parole chiave – Frammenti rilevanti*, si potranno visualizzare i messaggi che sono più pertinenti rispetto alle parole scelte per definire i comportamenti negativi del bullismo (fig. 11.2).

Frammenti con var categoriali e/o ex-post (testuali/quantitative)

ID Frammento	Etichetta del Frammento	TFIDF Bullismo	Operatore	Genere
182	Messaggio 182	2,27821	Si	Maschio
262	Messaggio 262	2,27789	Si	Femmina
72	Messaggio 72	2,17852	No	Maschio
214	Messaggio 214	2,17626	No	Femmina
104	Messaggio 104	2,17376	No	Maschio
128	Messaggio 128	2,16887	No	Maschio
47	Messaggio 47	2,16689	Si	Maschio
24	Messaggio 24	2,16575	No	Maschio

Record visibili: 277 su 277 | Nessuna colonna selezionata | Sola lettura

Fig. 11.2 – Messaggi rilevanti rispetto alla definizione dei tratti negativi del bullismo

11. 3. CATEGORIZZAZIONE DEL CORPUS DA REGOLE

Una funzione particolarmente interessante di TaLTaC² è la ricerca per entità che permette di ricercare parole o frasi nel corpus utilizzando interrogazioni complesse con espressioni regolari. Si tratta di una estrazione di informazione che consente di individuare frammenti che presentano certe regolarità nelle sequenze di parole o in parole che appaiono entro una determinata distanza l'una dall'altra (quasi-sequenze). Per una spiegazione completa di questa funzione, piuttosto articolata, è bene consultare direttamente il manuale del programma.

Qui possiamo indicare la procedura per una applicazione particolare, ma molto utile, della funzione di *Ricerca Entità* che a partire dal recupero dell'informazione permette di categorizzare i frammenti con una nuova variabile, tratta direttamente dalla "lettura" automatica del corpus, che si aggiunge alle altre variabili descrittive.

Dal menu **Analisi**, selezionando *Analisi Testuale – Estrazione di informazione: per concetti – Categorizzazione da regole*, apriamo la finestra di interrogazione *Ricerca Entità (RE)* (fig. 11.3).

Nella casella *Espressione regolare da cercare (entità)* scriviamo "bull*". In questo modo diamo istruzione al programma di cercare tutti i messaggi che contengono almeno una delle parole con radice *bull*+: *bulle*, *bulli*, *bulletto*, *bulletti*, *bullismo* ecc.

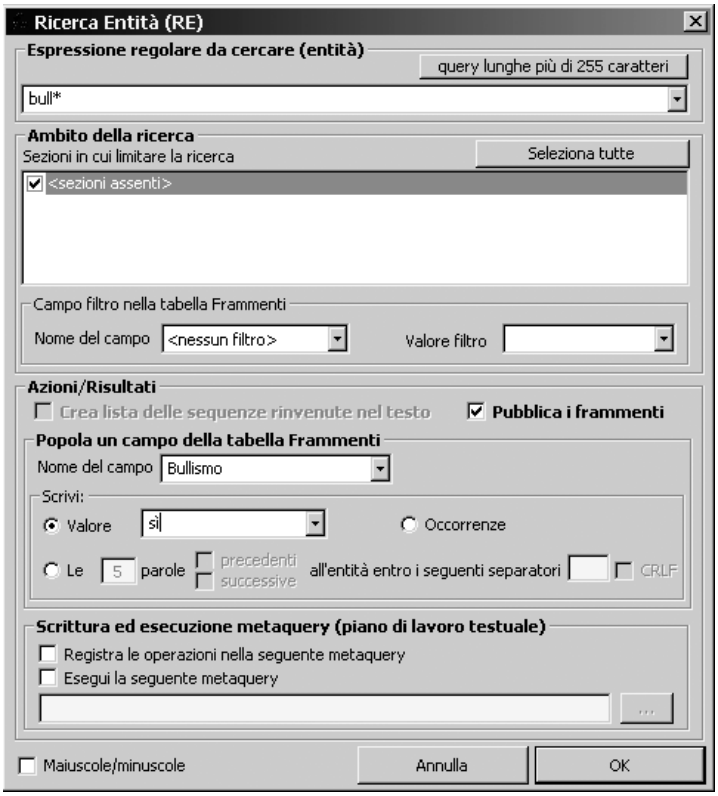


Fig. 11.3 – Finestra per la Ricerca Entità

Selezioniamo tutte le sezioni in cui limitare la ricerca (che in questo caso, in assenza di sezioni, significa estendere la ricerca a tutto il corpus); non selezioniamo alcun *Nome del campo* come filtro e scriviamo "Bullismo" nella casella del campo da aggiungere nella tabella frammenti. Scriviamo "si" in corrispondenza del *Valore* e poi clicchiamo su *OK*.

In questo modo i messaggi saranno arricchiti di una variabile di descrizione composta da due modalità: "si" (messaggi con riferimenti diretti al "bullismo" e "nul" (messaggi senza riferimenti diretti al "bullismo"). Con questa nuova variabile possiamo ricalcolare le sub-occorrenze e aggiungere due nuove colonne nella tabella *Vocabolario*.

11. 4. ESPORTAZIONE DI TABELLE E RICOSTRUZIONE DEL CORPUS

Tutte le tabelle e le liste generate dal programma possono essere salvate nel Database di sessione oppure esportate in formato *txt* per essere elaborate da altri programmi di analisi testuale o semplicemente per essere gestite più comodamente in Excel. L'operazione di salvataggio/esportazione avviene dal menu **File** selezionando *Salva nel DB della Sessione* oppure *Esporta – In un file di testo*. In entrambi i casi è possibile salvare *Tutti i record* oppure soltanto i record selezionati con lo strumento **TDM** e visualizzati nella tabella (*Solo i record visibili*).

Nell'analisi automatica dei testi l'operazione di salvataggio delle liste e delle tabelle è la forma più consueta con la quale si può procedere a una integrazione fra TaLTaC² e altri programmi che permettono l'esplorazione e l'analisi multidimensionale come SPAD, DTM, SPSS, SAS, Statistica e anche R, il più recente e potente prodotto *open source* per l'analisi statistica delle informazioni.
Il alcuni casi può essere utile procedere a una ricostruzione del corpus che sfrutta le procedure di etichettamento grammaticale e semantico che si sono generate nel corso dell'analisi. Ad esempio, dopo aver eseguito la fase di tagging grammaticale si vuole esportare il corpus associando a ogni forma grafica riconosciuta la rispettiva categoria grammaticale.

Dal menu **File** selezioniamo il comando *Esporta* e poi *Ricostruzione del Corpus - Testo annotato con...* La finestra di interrogazione (fig. 11.4) ci presenta alcune opzioni sul tipo di ricostruzione da effettuare:

- *Ricostruzione semplice con annotazioni*: il testo viene ricostruito con la forma grafica originale (con le forme eventualmente lessicalizzate), oppure sostituendo la forma grafica con il lemma corrispondente. Con l'opzione categoria grammaticale o semantica alla forma grafica o al lemma si aggiunge l'etichetta grammaticale del campo "CAT" e/o l'etichetta semantica inserita nel campo "CAT-SEM"; con la scelta della categoria senza la forma grafica, il corpus sarà ricostruito sostituendo le forme grafiche con le sole etichette (grammaticali e/o semantiche).
- *Ricostruzione per forme miste*: con questa opzione, solo per le categorie selezionate, verrà scritto il lemma in sostituzione della forma grafica; marcando le caselle *Categoria grammaticale* e/o *Categoria semantica* per le categorie selezionate saranno aggiunte le etichette corrispondenti.
- *Ricostruzione per pulizia del testo*: con questa funzione, che esclude tutte le altre, è possibile attivare le modalità di "correzione" e "pulizia" del testo, purché si sia precedentemente scritta nella colonna "CAT", in corrispondenza delle forme da correggere il codice ERR o DEL. Nel primo caso, inserendo nella colonna "Lemma" la forma corretta, questa verrà sostituita alla

precedente; nel secondo caso la forma verrà semplicemente cancellata dal testo.

- *Ricostruzione per Treetagger*: con questa opzione il corpus è ricostruito con l'annotazione grammaticale e il lemma utilizzati da TreeTagger, un lemmatizzatore plurilingue (francese, inglese, tedesco, italiano, spagnolo, bulgaro, russo, greco e portoghese) disponibile gratuitamente per il download in Internet, sviluppato dall'Institute for Computational Linguistics of the University of Stuttgart.

Le opzioni *Formato* permettono di esportate il file in diversi formati: *Testo libero* (così come è stato preparato il file *Bullismo_TT2.txt* nel nostro esempio; *Corpus strutturato in campi* o in *Collezione di files* (altre possibili opzioni di caricamento del corpus in TaLTaC[2]; oppure con *Una unità lessicale per riga*, cioè con una forma grafica per ciascuna riga.

Una particolar attenzione va posta alla prima delle *Opzioni generali*. Lasciando inalterata la marcatura di default su *Sostituisci blank con underscore nelle forme lessicalizzate*, le eventuali lessicalizzazioni introdotte durante il trattamento del testo andranno perdute (il poliforme *diversamente_abili* è ricostruito in due forme distinte: *diversamente* e *abili*). Le altre due opzioni hanno usi particolari e, in questo manuale, possono essere trascurate.

Come esempio di applicazione, procediamo a una ricostruzione del corpus con le categorie grammaticali.

Fig. 11.4 – Finestra per la ricostruzione del corpus

L'output, salvato automaticamente nella cartella di lavoro con il nome *Corpus ricostruito (FG_CAT).txt*, ci fornisce nuovamente il corpus originale normalizzato (notare le maiuscole abbassate dopo il punto), con le forme grafiche etichettate dal Tagging grammaticale completo.

```
****Messaggio 1 *Genere=Maschio *Operatore=Si
sono_J un_DET insegnante_J calabrese_J, e_CONG ho_V a che fare_AVV
ogni giorno_AVV con_PREP dei_J selvaggi_J in classe_AVV . ma_J atten-
zione_N, non_J sono_J selvaggi_J perché_J risentono_V di_J una_J dif-
ficile_J situazione_N ambientale_A, ma_J semplicemente_AVV perché_J
seguono_V io_J modelli_J proposti_J dalla_J tv_N, da_PREP questa_J
vergognosa_A tv_N
****Messaggio 2 *Genere=Indefinito *Operatore=Incerto
.... cultura_N ed_CONG educazione_N, assieme_J, assenti_J in_J trop-
pe_J famiglie_N .... i_J ragazzi_N sono_J naturalmente_AVV crudeli_A,
spesso_J inconsapelvolmente_.... i_J genitori_J non_J sono_J più_J in
grado di_PREP comunicare_V il_DET rispetto_J per_PREP gli_J altri_J,
soprattutto_AVV per_PREP i_J più_J deboli_J ... naturalmente_AVV mi_J
riferisco_V ai_J genitori_J di_J cotanti_J figli_J, che_J confondo-
no_V la_J liberalità_N con_PREP la_J licenza_N .... più_J che_J bia-
simare_V e_CONG punire_V i_J ragazzi_N, bisognerebbe_V fare_J dei_J
corsi_J di_J rieducazione_N per_PREP i_J genitori_J ...
```

La ricostruzione del corpus con le forme lemmatizzate, salvato con il nome *Corpus ricostruito (Lemma).txt* , produrrà invece l'output seguente:

```
****Messaggio 1 *Genere=Maschio *Operatore=Si
sono uno insegnante calabrese, e avere a che fare ogni giorno con dei
selvaggio in classe . ma attenzione, non sono selvaggio perché risen-
tire di uno difficile situazione ambientale, ma semplicemente perché
seguire io modelli proposti dalla tv, da questo vergognoso tv
****Messaggio 2 *Genere=Indefinito *Operatore=Incerto
.... cultura ed educazione, assieme, assenti in troppo famiglia.... i
ragazzo sono naturalmente crudele, spesso inconsapelvolmente.... i
genitore non sono più in grado di comunicare il rispetto per gli al-
tro, soprattutto per i più debole... naturalmente mi riferire ai ge-
nitore di cotanto figli, che confondere la liberalità con la licen-
za.... più che biasimare e punire i ragazzo, bisognare fare dei corsi
di rieducazione per i genitore...
```

Il corpus è stato ricostruito sostituendo, dove è possibile, alle forme grafiche il lemma corrispondente. Le forme ambigue o non riconosciute restano inalterate. Come si è detto nel paragrafo 9.7, questo tipo di lemmatizzazione automatica effettuata per tutte le categorie grammaticali non dà risultati attendibili. La

procedura, per essere davvero efficiente, richiede degli interventi manuali di disambiguazione da parte dell'utente.

TaLTaC2 è un programma versatile e complesso. Le opzioni sono numerose e possono essere combinate tra loro in modo da rispondere alle scelte strategiche definite dall'utente. In questo manuale abbiamo esaminato alcuni percorsi di base che permettono all'utente di fare pratica nell'analisi automatica dei testi e di risolvere la maggior parte dei problemi che gli si presentano. La sua funzione è di facilitare l'approccio all'uso del programma e non di sostituire il manuale, nel quale l'utente preparato potrà trovare soluzioni e indicazioni per problemi di analisi specifici di ciascun corpus e tali da rispondere alle sue ipotesi della ricerca.

ESEMPI DI RICERCA

AAVV. (2005) "Le parole per parlare dei media", in Censis - U.C.S.I. *I media che vorrei. Quarto Rapporto sulla comunicazione in Italia*, Milano, FrancoAngeli, pp. 113-157.

BOLASCO S. (1996) "Il lessico del discorso programmatico di governo", in M. Villone, A. Zuliani (a cura di), *L'attività dei governi della repubblica italiana (1947-1994)*, Bologna, il Mulino, pp. 163-349.

BOLASCO S. (2001) "Statistiche sulla partecipazione nel sito FO e analisi testuale dei messaggi", in M. Radiciotti (a cura di) *La formazione on-line dei docenti Funzioni Obiettivo. Indagine qualitativa sugli esiti dei forum attivati dalla Biblioteca di Documentazione Pedagogica*, Milano, FrancoAngeli, pp. 19-78.

BOLASCO S., D'AVINO E., PAVONE P. (2005) *Analisi lessicale dei diari giornalieri con strumenti di statistica testuale e text mining.* Relazione invitata al convegno sul tema "I tempi della vita quotidiana", Istat, Roma, 20 dicembre 2005.

BOLASCO S., DELLA RATTA-RINALDI F. (2004) "Experiments on Semantic Categorisation of Texts: Analysis of Positive and Negative Dimension", in G. Purnelle, C. Fairon, A. Dister (*eds*) *Les pois des mots. Actes des 7es Journées internationales d'Analyse statistiques des Données Textuelles*, Louvain, Presse Universitarie de Laouvain, pp. 202-210.

BOLASCO S., GALLI DE' PARATESI N., GIULIANO L. (2006) *Parole in libertà. Un'analisi statistica e linguistica dei discorsi di Berlusconi*, Roma, ManifestoLibri, 2006.

BOLASCO S., GIOVANNINI D. (2002) "Il Trattamento Automatico Lessico-Testuale per l'Analisi del Contenuto (TALTAC): un'applicazione allo studio delle immagini della formazione professionale trentina", in B. Mazzara (a cura di) *Metodi qualitativi in Psicologia Sociale. Prospettive teoriche e strumenti operativi*, Roma, Carocci Ed., pp. 343-361.

BOLASCO S., PAVONE P. (2008) "Multi-class categorization based on cluster analysis and TFIDF", in *JADT-2008 - Actes des 9èmes Journées Internatial d'Analyse Statistiques*

des Données Textuelles, Lyon. 12-14 mars, Lyon, Presse Universitaire de Lyon, vol. 2, pp. 209-217.

CANZONETTI A. (2008) "Information retrieval e analisi delle cooccorrenze per l'estrazione di informazione specifica da documentazione giuridica", in *JADT-2008 - Actes des 9èmes Journées Internatial d'Analyse Statistiques des Données Textuelles*, Lyon. 12-14 mars, Lyon, Presse Universitaire de Lyon, vol. 2, pp. 277-284.

CORTELLAZZO M.A., TUZZI A., a cura di (2008) *Messaggi dal Colle. I discorsi di fine anno dei presidenti della Repubblica*, Venezia, Marsilio, 2008.

DELLA RATTA-RINALDI F. (2005) "L'interpretazione sistematica del materiale derivante da focus group attraverso l'analisi testuale", *Sociologia e ricerca sociale*, XXVI, 76-77, pp. 91-104.

GALLI DE' PARATESI N., GIULIANO L. (2007). "Pronoun morphology and the semantics of political communication in a face to face Prodi-Berlusconi debate (2006 general election campaign)". *Le discours de campagne – 6èmes journées de la SELP* (Société d'Etude des Langages du Politique), Nice, 29-30 novembre 2007.

GIULIANO L. (2005) "La comunicazione politica nei newsgroup: il caso della guerra in Iraq", in G. Sensales (a cura di) *Rappresentazioni della 'politica'. Ricerche in psicologia sociale della politica*. Milano, FrancoAngeli, pp. 104-122.

GIULIANO L. (2008) "Parole e politica nei «faccia a faccia» della campagna elettorale del 2006", in G. Sensales e M. Bonaiuto (a cura di), *La politica mediatizzata. Forme della comunicazione politica nel confronto elettorale del 2006*, Milano, FrancoAngeli, pp. 163-189.

GIULIANO L., LA ROCCA G. (2005) "The process of sensemaking on the telework virtual community using text mining", in *Data Mining VI. Data Mining, Text Mining and their Business Application*. Sixth International Conference on Data Mining (Wessex Institute of Technology), Southampton, Boston, WIT Press, vol. 35, pp. 143-151.

LA NOCE M., BOLASCO S., ALLEGRA E., RUOCCO V., CAPO F.M. (2006) "Merger control in Italy 1995-2003: a statistical study of the enforcement practice by mining the text of Authority resolutions", in *International Journal of the Economics of Business*, 13, 2, pp. 307-334.

MARICCHIOLO F, GIULIANO L., BONAIUTO M. (2008) "Parole e gesti nell'analisi automatica del testo: il caso dei «faccia a faccia» tra Berlusconi e Prodi (2006), in *JADT-2008 - Actes des 9èmes Journées Internatial d'Analyse Statistiques des Données Textuelles*, Lyon. 12-14 mars, Lyon, Presse Universitaire de Lyon, vol. 2, pp. 787-798.