Franco Lancia

# STRUMENTI PER L'ANALISI DEI TESTI

## Introduzione all'uso di T-LAB

Strumenti per l'indagine sociale
Collana diretta da A. Claudio Bosio

FrancoAngeli

Ristampa
0 1 2 3 4 5 6 7

Anno
2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015



Stampa: Tipomonza, via Merano 18, Milano.

*ita verum humanum sit, quod homo, dum novit, componit item ac facit;*
*et eo pacto scientia sit cognitio generis, seu modi, quo res fiat,*
*et qua, dum mens cognoscit modum, quia elementa componit, rem faciat;*

(G.B. Vico, *De antiquissima italorum sapientia*, 1710, I, 1)

*la verità umana è quella che l'uomo costruisce nel momento stesso in cui la conosce;*
*quindi la scienza è la conoscenza del modo attraverso il quale la cosa viene fatta,*
*e per mezzo della quale, conoscendo il modo di comporre gli elementi, la mente fa la cosa.*

# Indice

# Introduzione: *leggerezza* ed *esattezza*

*Leggerezza* ed *esattezza* sono due dei *Six memos for the next millennium*[1] che Italo Calvino (1993) ha affidato alla nostra memoria: altrettante chiavi, utili a lui per ripensare la sua esperienza di scrittore e di lettore, utili a noi per ripensare altre pratiche e altre storie.

Questo libro è dedicato a una pratica, l'analisi dei testi, e al suo esercizio tramite un software che propone strumenti per l'esplorazione e l'analisi dei *contenuti*. Il che porta chi scrive e chi legge a collocarsi in un *genere* di letteratura, certamente diverso da quello dei poeti, ma che pure – nella sua costruzione – pone problemi di *peso* (leggerezza) e di *disegno* (esattezza).

«La mia operazione – scriveva Calvino – è stata il più delle volte una sottrazione di peso» (ibid., p. 7). Come lui sono convinto che esiste una *leggerezza della pensosità*. Per questo motivo, tornando indietro di qualche anno, a quando ho iniziato a usare software per l'analisi di testi, a discutere con i loro autori e a frequentare la letteratura sul tema, non posso non chiedermi il perché della "pesante" modalità con cui si era quasi costretti a interloquire con software e persone[2].

Apparentemente sembrava una questione di *esattezza*, relativa alle operazioni da compiere e al cosiddetto rigore scientifico.

Ancora una citazione da Calvino:

> Esattezza vuol dire per me soprattutto tre cose:
> a) un disegno dell'opera ben definito e ben calcolato;
> b) l'evocazione di immagini visuali nitide, incisive, memorabili (…);
> c) un linguaggio il più preciso possibile come lessico e come resa delle sfumature del pensiero e dell'immaginazione (ibid., p. 65).

[1] Il titolo italiano del libro, pubblicato postumo, è *Lezioni americane*.

[2] Mi riferisco al fatto che le interlocuzioni erano mediate da un linguaggio quasi esoterico, certamente non facilitante la diffusione della *computer-aided text analysis* entro un pubblico più ampio.

Si dirà: una cosa è la letteratura, altra è la scienza. Rispondo: discutiamone. Il modo di scrivere di cose tecnico-scientifiche è codificato all'interno di specifiche comunità, entro le quali – come ci ricorda T.S. Kuhn (1962) – *costruiamo le regole* per fare scienza e per raccontarla.

Personalmente, quando i testi scientifici mi risultano di difficile comprensione, faccio due ipotesi:

a) chi scrive ha poco chiaro ciò di cui parla o, per varie ragioni, vuol comunicare che la sua conoscenza è di difficile accesso;

b) chi scrive ha un modo di vedere le cose talmente diverso dal mio che mi riesce *veramente* difficile capirlo.

Nel primo caso, mi annoio subito e – per così dire – rispedisco la lettera al mittente. Nel secondo, in genere, raccolgo la sfida, tanto seriamente che ho accumulato debiti incalcolabili verso autori che mi hanno insegnato a pensare, autori di testi difficili che mi sono diventati familiari quasi come i racconti dei miei nonni.

Detto ciò, se chi legge questo libro ha – per fondate ragioni – intenzione di "rispedire al mittente", può farlo in modo diretto inviando una e-mail al mio indirizzo di posta elettronica[3].

Ma veniamo, o torniamo, all'analisi dei testi. Il fatto che in questi casi i materiali da analizzare vengano definiti *non strutturati* è un buon indizio: sembra che, quasi per necessità, chi se ne occupa rischia di restare in balia dell'*indeterminatezza*.

Domanda: l'indeterminatezza è una caratteristica dei *dati* da analizzare o, piuttosto, deriva dal modo di organizzarli e di interpretarli? Certo, nell'ambito delle cosiddette scienze umane non è facile elaborare modelli e fare inferenze. Più avanti avremo modo di considerare questi aspetti in modo più approfondito. Per ora, mi limito a suggerire qualche ipotesi.

Una delle più importanti rivoluzioni scientifiche del nostro tempo è associata al modello di funzionamento del DNA: un *codice* che, in modo invariante, regola la riproduzione di tutti gli organismi viventi. Come è noto, il modello in questione ha molto a che fare con i modelli elaborati nell'ambito della linguistica e della semiotica, due tra le più *rigorose* scienze umane. Roland Barthes, che in queste scienze era maestro, nel suo saggio sulla fotografia, dopo aver ammesso la sua «disperata resistenza verso ogni sistema riduttivo», scrive:

> Decisi perciò di assumere come punto di partenza della mia ricerca solo poche foto: quelle che ero sicuro esistessero *per me*. Niente a che vedere con un corpus: solamente alcuni corpi. In questa controversia tutto sommato convenzionale tra soggettività e scienza, maturai un'idea bizzarra: perché mai non avrebbe dovuto esserci, in un

[3] franco.lancia@tlab.it.

certo senso, una nuova scienza per ogni soggetto? Una *Mathesis singularis*[4] (e non più *universalis*)?» (1980, tr. it., p. 10).

Il che, tornando alla biologia, significherebbe ipotizzare una scienza per *quella* quercia, una per *quel* pettirosso, e così via.

Il fatto che, nell'ambito delle scienze umane, molti ricercatori siano affezionati all'*unicità* dei loro oggetti di studio non è motivo di scandalo. Più di un secolo fa, W. Windelband (1894) propose di distinguere due modi di far scienza: quello nomotetico e quello idiografico. Il primo orientato alla ricerca di qualche legge (*nomos*) generale, il secondo interessato alla ricostruzione di particolari (*idios*) eventi o "storie". Con una precisazione: i due modi non sono criteri per distinguere tipi di discipline scientifiche, ad esempio per distinguere le scienze biologiche da quelle sociali. Per convincersene, basta leggere o rileggere i testi di K. Lorenz sul comportamento animale. Ad esempio, nel libro dedicato all'etologia dell'oca selvatica (K. Lorenz, 1989), prima di inventariare gli etogrammi, l'autore racconta "storie" che hanno specifici protagonisti: Martina, Mercedes e Florian, Sinda e Blasius, Ada. Tutte oche, naturalmente.

Ebbene, è mia opinione che nell'ambito delle scienze umane esiste un "oggetto" analogo al DNA, in grado di funzionare da catalizzatore per molti programmi di ricerca e in grado di portare a ridisegnare i confini tra le varie discipline. Il suo nome è *rappresentazione*.

I nostri sogni, i nostri miti, le nostre religioni e le nostre culture sono, per così dire, costruite con gli stessi mattoni: le rappresentazioni, appunto[5]. E, si dà il caso, tutte le rappresentazioni si cristallizzano in qualche testo.

Per questa ragione l'analisi dei testi è una pratica molto antica e molto diffusa; così come l'ermeneutica, la scienza dell'interpretazione che porta il nome di una divinità greca[6].

La mia esperienza al riguardo: ho iniziato a studiare di linguistica, ermeneutica e filosofia del linguaggio, agli inizi degli anni settanta. A quell'epoca mi sono laureato con una tesi sul pensiero di P. Ricoeur che, non a caso, in quegli anni sosteneva la necessità di integrare strutturalismo ed ermeneutica. Il mio relatore fu P. Filiasi Carcano (1969), allora noto metodologo della

[4] Quella che l'autore, nello stesso libro (ibid., p. 72), chiamerà *scienza impossibile dell'essere unico*.

[5] Per restare nell'analogia con il DNA, le differenze tra i vari tipi di rappresentazioni (mentali e/o sociali, siano esse denominate "immagini", "simboli", "modelli", ecc.) sono analoghe alle differenze esistenti tra le varie specie di esseri viventi. Sempre per analogia, così come gli animali e le piante condividono lo stesso "codice vivente" (il DNA, appunto), allo stesso modo i vari tipi di rappresentazioni possono essere ricondotte a una comune matrice genetica. Come vedremo più avanti (paragrafo 1.2 della prima parte), le rappresentazioni, oltre ad essere "mattoni", sono anche *regole* di costruzione.

[6] Hermes.

scienza, anche lui molto interessato alla integrazione dei vari metodi di ricerca. Successivamente ho iniziato ad occuparmi di psicologia e, percorrendo uno specifico itinerario professionale, a lavorare in un ambito in cui le competenze concernenti l'analisi del linguaggio[7] erano al servizio di una pratica di ricerca-intervento: la consulenza per l'analisi e lo sviluppo delle culture organizzative[8]. In questo contesto, circa quindici anni fa, alcuni progetti mi hanno portato a sviluppare specifiche competenze concernenti metodi e tecniche per l'analisi dei dati, cioè – all'epoca – di "numeri" in cui, in modo prevalente, erano codificate risposte a questionari con domande chiuse[9]. I software che utilizzavo sono ancora sul mercato; magari con una diversa interfaccia, ma la filosofia è restata invariata: importazione di tabelle e possibilità di utilizzare vari strumenti di tipo statistico. Poco più tardi ho incominciato ad usare anche software per l'analisi dei testi: trascrizioni di interviste, risposte a domande aperte, ecc. Nell'arco di un paio di anni mi procurai le licenze di quattro prodotti: ciascuno a suo modo interessante, ognuno in grado di fare alcune operazioni e non altre, quasi tutti non semplici da usare. Conclusione: per completare un progetto di analisi, e prima di passare alla stesura del report, dovevo lavorare vari giorni, spesso più di qualche settimana, con alla fine la falsa impressione che il più era fatto.

Caso ha voluto che, mentre "bisticciavo" con alcuni di questi software, qualche amico informatico di professione "ha visto" le mie difficoltà ed ha osservato: «Queste operazioni possono essere effettuate in modo più *semplice* e *integrato*». Approfondendo il discorso, abbiamo iniziato a costruire un prototipo che poi è diventato T-LAB[10].

Qualche osservazione: i software per analizzare dati numerici sono gli stessi per i ricercatori di molte le discipline (biologi, economisti, psicologi, ecc.); al contrario, quelli per analizzare dati testuali sono usati solo nell'ambito delle scienze umane e sociali. I primi hanno degli standard, i secondi no. I primi sono poco numerosi e hanno molto mercato, i secondi sono innumerevoli e hanno un mercato di *aficionados*. I primi, in genere, offrono molti strumenti integrati tra loro, i secondi no. Le competenze richieste per il loro uso: nel caso dei primi, basta un buon corso di statistica; nel caso dei secondi, quasi per ciascuno di essi, c'è da pagare lo scotto di un qualche processo di iniziazione.

[7] Cfr. F. Lancia, 1984.

[8] Nel 1985 sono stato tra i fondatori di SPS (Studio di Psicosociologia) e per circa quindici anni ho condiviso preziose esperienze di lavoro con i colleghi di questo studio professionale.

[9] Quelle del tipo "In che misura lei è soddisfatto del nostro servizio? (Molto / Abbastanza / Poco / Per niente)".

[10] In particolare ringrazio Marco Silvestri: per il suo paziente lavoro e per avermi aiutato a padroneggiare la logica dei linguaggi di programmazione.

Forse pecco di *leggerezza*, ma salta agli occhi che, nell'ambito delle scienze sociali, letteratura e software sono due facce della stessa medaglia: da una parte le pratiche *standard*, dall'altra quelle *non-standard*[11]. Il punto è che i testi, come vedremo più avanti, possono essere trasformati in tabelle dati, affidabili quanto quelle derivate dal censimento della popolazione italiana. Tutto sta a vedere cosa viene misurato e come. Quanto poi all'analisi delle tabelle, se i dati sono in forma di numeri, le tecniche statistiche offrono *pari opportunità* di trattamento.

Ciascuno di noi sa che come individuo il suo nome è presente in innumerevoli banche dati, associato a caratteristiche che vanno dalla data di nascita, alla residenza, al codice fiscale, al reddito, ai consumi, al traffico telefonico e a tante altre cose ancora. La stessa cosa vale per le parole e per i testi. In termini generali, parole, frasi ed altre unità di analisi sono *individui* con determinate caratteristiche; quindi ogni testo, così come ogni tipo di popolazione, può essere rappresentato in un database.

E, a partire dal momento in cui uno o più testi sono rappresentati in un database, l'*esattezza*[12] dipende degli algoritmi di calcolo che, in funzione di specifici obiettivi, si decide di applicare.

Come avremo modo di vedere, l'architettura di T-LAB, pur aperta ad ulteriori sviluppi, prevede l'archiviazione automatica di ogni corpus in un database ed offre una serie di strumenti sia per consultare e/o modificare il database, sia per effettuare analisi che producono vari tipi di grafici e tabelle. Il tutto in un ambiente integrato, in qualche modo all'insegna della *leggerezza* e dell'*esattezza*.

I memo di Calvino portano ancora a considerare il tipo di *esperienza* che accompagna la lettura di un buon libro: sfogliare una pagina dopo l'altra, immergersi nel mondo costruito dall'autore, ricostruirlo con la propria immaginazione, esplorare altri tempi e altri luoghi, ricordare altre storie, ascoltare il ritmo del linguaggio, emozionarsi, essere tentati di leggere ad alta voce, tornare a rileggere gli stessi brani, e altre cose ancora.

Verrebbe da dire: ahimè, questo libro non è un romanzo!

Il mio augurio è che possa risultare semplicemente utile, e non solo come introduzione all'uso di T-LAB. Giocando con le parole, potrei dire che il di-

[11] Mi riferisco alla distinzione proposta da A. Marradi (1966). Secondo questo autore, l'"insieme" *standard* include le pratiche scientifiche che, attraverso l'uso di variabili e di matrici dati, realizzano esperimenti o analisi di covariazioni; mentre l'insieme *non-standard* include le pratiche di ricerca che «producono asserti privi di ragionevoli pretese di impersonalità» (*ibid*. 172).

[12] A scanso di equivoci: uso "esattezza" nell'accezione proposta da Italo Calvino. Quanto alla "precisione", nell'analisi dei testi essa è funzione del tipo di operazioni che si stanno mettendo in atto. Ad esempio, nel riconoscimento e nella classificazione delle parole c'è sempre da scontare e da gestire la considerevole rilevanza dei casi di ambiguità.

scorso sul software è anche un pretesto teso a promuovere una cultura dell'analisi dei testi, e non solo tra chi, per professione, si occupa di ricerca.

D'altro canto, c'è ricerca ogni qual volta qualcuno si interroga sul senso di qualcosa e, usando gli strumenti di cui dispone, prova a trovare risposte alle sue domande.

All'inizio di un bel libro scritto qualche millennio fa si legge:

> Ogni arte e ogni ricerca, e similmente ogni azione e ogni proposito sembrano mirare a qualche bene; perciò a ragione definirono il bene: ciò a cui ogni cosa tende.
>
> Tuttavia sembra esservi una differenza tra i fini: talora infatti essi sono attività, talora invece, oltre ad esse, opere definite. Quando vi sono dei fini definiti nelle azioni, allora le opere sono più importanti delle attività (Aristotele, *Etica Nicomachea*, 1094a, 1)[13].

In base alla *differenza tra i fini*, Aristotele distingueva due tipi di attività: la *poiésis*, in cui il fine è costituito dalla realizzazione di un prodotto (*èrgon*), e la *prâxis*, in cui il fine si realizza nelle attività (*enérgheia*) stesse; le attività artigianali e tecnologiche da una parte, le attività quali il viaggiare e il riflettere dall'altra.

Qualche secolo più tardi alcuni studiosi (R. Harré e P.F. Record, 1972) hanno proposto un diverso criterio di classificazione: per sapere che tipo di azione viene fatta, bisogna "chiederlo" alle persone che ne sono attori o spettatori. Ciò in base al fatto che il tipo di azione è definito dal tipo di resoconto che se ne può fornire.

In quest'ottica questo libro vuol essere uno strumento soprattutto per chi è interessato a resocontare attività di ricerca. Quindi, uno strumento non solo per apprendere *come* fare ma anche per riflettere sul *cosa* e sul *perché*. Tuttavia, poiché T-LAB può essere usato in vari modi e con differenti obiettivi, non ultimo per *giocare* con i testi e per sperimentare nuove tecniche di lettura, questo libro potrebbe anche accompagnare la pratica di attività "fini a sé stesse".

La prima parte è suddivisa in due capitoli:

[13] In sequenza, i due capoversi evocano una visione del mondo e un metodo di analisi. Il primo, proponendo la concezione del bene come "fine ultimo" di ogni azione, ci giunge come un messaggio da un altro mondo e da un'altra cultura. I secondo, mostrando all'opera la costruzione logica di categorie (definizione per "genere" e per "differenza specifica": genere = fine; specie = attività e opere) ci ricorda che il pensiero di Aristotele è "anche" attuale. Ad esempio, molte *ontologie* implementate in sistemi di intelligenza artificiale, sia per rappresentare insiemi di *oggetti* che per identificare *concetti* all'interno dei testi, sono costruite secondo i principi della logica aristotelica (Vedi: http://ontology.buffalo.edu/smith/articles/ontologies.htm e http://www.loa-cnr.it/Publications.html).

- *un contesto di riferimento*, in cui sono presentati concetti, modelli e problemi concernenti la metodologia della ricerca applicata all'analisi dei testi;
- *la logica analitica di T-LAB*, in cui viene proposta un'accurata e accessibile descrizione dei processi che regolano il funzionamento del software, con particolare attenzione alle logiche dei trattamenti linguistici e di quelli statistici.

Nella seconda parte, un capitolo introduttivo (*Strategie e percorsi di analisi*) include 18 schede, con altrettante testimonianze di ricercatori qualificati, che danno modo di capire quali sono le potenzialità di T-LAB e i suoi possibili campi di applicazione.

Seguono tre capitoli dedicati ad esempi narrati "passo dopo passo" che, in forma di *esercizi*, mostrano all'opera varie strategie di analisi.

# Parte prima

# 1. Un contesto di riferimento

## 1.1. Introduzione

Nei testi di metodologia della ricerca normalmente si fa distinzione tra teorie, metodi, tecniche e strumenti: quattro tipi di "oggetti" che, in varie combinazioni, vengono utilizzati per definire i confini tra discipline e tra programmi (o tradizioni) di ricerca. Ad esempio, per distinguere la sociologia dalla psicologia o dall'antropologia culturale; ma anche per distinguere la psicoanalisi dal comportamentismo, la scuola sociologica di Chicago da quella di Francoforte, l'antropologia strutturale dall'antropologia interpretativa, e così via. Gli stessi oggetti e le loro possibili combinatorie vengono inoltre utilizzati per distinguere vari tipi di analisi dei testi: analisi di contenuto *qualitative* e *quantitative*, analisi del discorso, analisi delle conversazioni, analisi di storie di vita, analisi tematiche, ecc.

Mentre evochiamo queste diverse specie che abitano l'ecosistema delle scienze umane e le classificazioni che ne sono state proposte, la *leggerezza della pensosità* e un po' di autoironia fanno venire alla mente quella strana enciclopedia cinese citata da Jorge Luis Borges (1952), dove si legge:

> gli animali si dividono in: a) appartenenti all'Imperatore, b) imbalsamati, c) addomesticati, d) maialini da latte, e) sirene, f) favolosi, g) cani in libertà, h) inclusi nella presente classificazione, i) che si agitano follemente, j) innumerevoli, k) disegnati con un pennello finissimo di peli di cammello, l) *et caetera*, m) che fanno l'amore, n) che da lontano sembrano mosche (tr. it., p. 1005)[1].

Contemporaneamente, poiché il testo di Borges ha ispirato la ricerca di M.

[1] La citazione è tratta da un breve scritto (*El idioma analitíco de John Wilkins*) in cui Borges "inventa" l'enciclopedia cinese per ironizzare sulla logica delle classificazioni. L'autore citato da Borges, John Wilkins (1614-1672), è autore di un saggio sulla tassonomia universale (*An Essay Towards a Real Character and a Philosophical Language,* 1668).

Foucault (1966) sull'archeologia delle scienze umane, ricordiamo la sua lezione: le somiglianze e le differenze tra le discipline scientifiche, così come sono codificate nelle tassonomie del sapere, sono fenomeni di superficie e vanno interrogate a partire da un'analisi delle condizioni che le hanno rese possibili. Ma, la *nuova analitica* proposta da questo autore, quella che sonda le radici della nostra *epistème*[2], non può farci dimenticare che, nell'orizzonte finito della nostra esperienza, classificazioni e distinzioni possono essere usate come *strumenti* di ricerca, al pari delle teorie e delle discipline che esse consentono di articolare.

Le teorie, scriveva K. R. Popper (1959, tr. it., p. 43), sono *reti* gettate per catturare quello che noi chiamiamo "il mondo"; per razionalizzarlo e per spiegarlo. Ma la metafora del ricercatore-pescatore merita qualche commento. Le teorie-reti sono sì strumenti per pescare, ma – ahimè – i ricercatori sono allo stesso tempo pescatori e pesci. Fuor di metafora: le teorie sono rappresentazioni del mondo che, oltre a consentire l'organizzazione della ricerca e della comunicazione scientifica, organizzano sistemi di appartenenza. L'una cosa non va senza l'altra e tutti, in qualche modo, dobbiamo farci i conti.

Notare: il fatto che le teorie sono **rappresentazioni**, cioè strutture di *contenuti* in forma di ipotesi, concetti e modelli, significa che esse possono cristallizzarsi in *testi*; e, in quanto testi, possono essere analizzate dagli stessi ricercatori che le producono[3]. La ricorsività del linguaggio, il fatto cioè che «le descrizioni possono essere fatte considerando altre descrizioni come se fossero oggetti» (H. Maturana e F. Varela, 1980, tr. it., p. 173) è all'origine della nostra capacità di pensare.

Già nel 1868, C. S. Peirce, scriveva:

> Dalla proposizione che ogni pensiero è un segno consegue che ogni pensiero deve indirizzarsi a un qualche altro pensiero, deve determinarne qualche altro... Perché ci sia pensiero deve esservi stato un altro pensiero (1868, tr. it., p. 50).

Ancora una nota a proposito del ricercatore-pescatore: le teorie non *sono*

[2] Nella concezione di M. Foucault, l'*episteme* «è l'insieme delle relazioni che per una data epoca si possono scoprire tra le scienze quando si analizzano al livello delle regolarità discorsive….è ciò che si può sapere in una data epoca, tenendo conto delle insufficienze tecniche, delle abitudini mentali e dei limiti posti dalla tradizione; è ciò che, nella positività delle pratiche discorsive, rende possibile l'esistenza delle figure epistemologiche e della scienza» (1969, tr. it. p. 251).

[3] Secondo M. Foucault, questa "duplicazione" è il carattere distintivo delle scienze umane e deriva dal fatto che, per esse, la rappresentazione non è soltanto un "oggetto", ma costituisce "il basamento generale" a partire dal quale la loro esistenza è resa possibile. In effetti, la specificità delle scienze umane non è data dal fatto che i loro "contenuti" parlano dell'uomo (anche l'economia e la biologia fanno qualcosa di simile), ma piuttosto dal "carattere puramente formale" e metodologico della loro "duplicazione" (1966, tr. it., pp. 379-389).

strumenti. La *caratteristica* "strumento" non è una proprietà che appartiene a questa o quella cosa in virtù della sua propria "natura": ogni cosa può *diventare* uno strumento in funzione di uno o più possibili *usi*. Quindi dire, come vorrei dire, che «le teorie sono i più importanti strumenti di ricerca» è un'affermazione impropria, e non solo perché lo sviluppo delle teorie può anche essere legittimamente assunto come un *fine*.

Anni fa, insieme ad alcuni colleghi[4], ho proposto la *teoria della tecnica* come uno strumento utile per ripensare alcune pratiche psicologiche. *Mutatis mutandis*, nelle pratiche di ricerca la teoria della tecnica porta a considerare due aspetti fondamentali:

- il **contesto** in cui si opera;
- il tipo di **prodotto** che, nel modo più adeguato, risponde alla domanda del **cliente**.

Quanto al contesto, vale la pena di precisare che esso non è *dato* e che non corrisponde a "luoghi" fisici o organizzativi; bensì è il risultato di un pensiero, cioè di un'attività che stabilisce un qualche tipo di relazione tra qualcosa e qualcosa d'altro[5]. In effetti, nei dizionari etimologici *contesto* è rubricato come un sostantivo che deriva da una voce verbale: il participio passato del verbo *contexere*, ovvero «ciò che è tessuto insieme (intrecciato) a qualcosa d'altro». Il che induce a trasformare interrogativi del tipo «a quale contesto ti riferisci?» in interrogativi del tipo «quale contesto stai costruendo?».

In questo caso, la mia risposta è nelle relazioni stabilite tra i vari box della Fig. 1.

Qualche chiave di lettura:

- la voce **dati testuali** è stata scelta per ancorare il discorso al tema di questo libro. In altri contesti, più semplicemente, si potrebbe parlare di *dati*;
- la voce **tecniche e strumenti**, ovviamente, include anche i software;
- la voce **ricercatore**, al singolare, sta anche ad indicare uno o più gruppi di ricerca;
- la voce **cliente** non deve trarre in inganno. Una pratica di ricerca ha sempre un cliente: fosse anche costituito da una comunità scientifica di «pochi intimi»;

[4] Cfr. R. Carli, R.M. Paniccia e F. Lancia, 1988; F. Lancia, 1990.

[5] In genere, quando parliamo di "contesto", la relazione tra il "qualcosa" e il "qualcosa d'altro" ha la forma di una relazione tra una "parte" e un "tutto" (ad es. «La parola nel contesto della frase», «Pierino nel contesto scolastico», ecc.).

- la voce **teorie e modelli** è nel vertice in alto; ma la figura rappresentata è un pentagono regolare, con relazioni di ciascun vertice con tutti gli altri. Quindi, detto con un vecchio adagio: cambiando l'ordine dei fattori, il prodotto non cambia;
- la voce **prodotto** è al centro in quanto, allo stesso tempo, rappresenta l'interfaccia tra i cinque vertici e costituisce il risultato delle varie interazioni. L'aggiunta di **resoconto** sta ad indicare il tipo di prodotto che, normalmente, è associato alle pratiche di ricerca: dai report scientifici, pubblicati e no, alle presentazioni per i clienti aziendali.

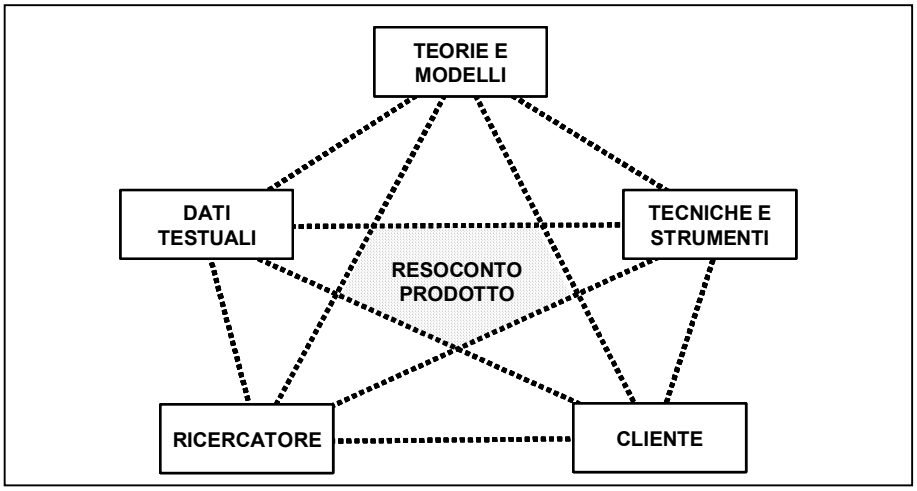



*Fig. 1 – Un contesto di riferimento*

## 1.2. Le pratiche di ricerca e i loro resoconti

Per districare la rete delle relazioni in Fig. 1, cioè per dipanare ciò che è tessuto insieme (il contesto), il miglior modo che io conosco è cominciare con il riflettere sul **prodotto** cioè, nel caso specifico, sui resoconti delle pratiche di ricerca che hanno come oggetto materiali testuali. Non avendo fatto indagini empiriche su questo tema, mi limiterò a proporre qualche ipotesi.

In genere, la struttura di questi resoconti è organizzata da due processi argomentativi:

- raccontare e spiegare **cosa si è fatto**;
- raccontare e spiegare **cosa si è capito**.

Il primo tipo di processo include la definizione degli obiettivi, la descri-

zione dei criteri e dei metodi per la raccolta dei dati, le annotazioni sulle tecniche e sugli strumenti usati per analizzarli.

Il secondo tipo di processo include la dichiarazione dei propri modelli di riferimento e la ricostruzione delle inferenze che, in base ai risultati, hanno portato a capire "***x***".

La "***x***", detta anche incognita, è in questo caso una label che sta per *valore aggiunto*: quella cosa che i clienti si aspettano e che, in genere, sono anche capaci di verificare. Ma quali tipi di clienti?

In generale, la cultura dei ricercatori è che i loro clienti primari sono le comunità scientifiche, in gran parte costituite dai loro stessi colleghi. D'altro canto, è all'interno delle comunità scientifiche che viene codificato il modo di far ricerca e il modo di resocontarla. In questo contesto, per ragioni facili da capire, il valore aggiunto viene spesso identificato con metodi e tecniche di analisi; con il rischio che qualcuno possa ironicamente affermare «L'operazione è perfettamente riuscita; ma il paziente è morto».

Fuor di metafora, il paziente morto è in questo caso la possibilità di avere un mercato. Eppure, la domanda di capire qualcosa a partire dall'analisi dei testi è in continua espansione. Basti pensare alle implicazioni dell'uso di Internet. Molte cose che prima venivano solo dette o scritte a qualcuno, oggi sono testi che circolano in rete: e-mail, chat, forum, newsgroup ecc. Articoli di quotidiani e riviste, così come i "classici" libri, sono sempre più accessibili in formato digitale. Interviste e questionari possono essere realizzati ed analizzati via web.

Ciò premesso, una buona domanda sembra essere la seguente: a quali condizioni un resoconto di ricerca può diventare un buon prodotto? Per tentare di rispondere, propongo una scaletta argomentativa costituita da cinque affermazioni:

**1-** ogni resoconto è un testo;
**2-** i contenuti dei testi veicolano rappresentazioni;
**3-** i resoconti che includono l'analisi dei testi propongono interpretazioni (cioè rappresentazioni di rappresentazioni);
**4-** i resoconti, se ben costruiti, attivano processi di pensiero sulle rappresentazioni;
**5-** i processi di pensiero sulle rappresentazioni facilitano la comprensione dei cambiamenti culturali e possono attivare nuove strategie dell'azione sociale.

Glossario: in questo contesto, i **cambiamenti culturali** pertinenti sono quelli in assenza dei quali qualche soggetto sociale si trova a fronteggiare *costi* rilevanti. Pensiamo ai costi della non riuscita integrazione degli immigrati, del gap tra formazione scolastica e mondo del lavoro, dell'insoddisfazione dei

cittadini verso le pubbliche amministrazioni, di quella dei clienti per i servizi di una qualunque azienda, delle abitudini di comportamento di guida nel traffico, della diffusione della droga nelle nuove generazioni, del mancato sviluppo del mezzogiorno, del decadimento dei media, della mancata assistenza agli anziani, del degrado ambientale, degli abusi nei consumi dell'acqua, della disaffezione dalla politica, dei pregiudizi sugli organismi geneticamente modificati e di tanti altri ancora. Per ridurre questi tipi di costi c'è sempre qualcuno interessato a investire in qualche buon progetto di ricerca.

Proviamo ora a sviluppare le implicazioni delle cinque affermazioni di cui sopra. La prima («ogni resoconto è un testo») è in qualche modo scontata; le altre no.

Dire «i contenuti dei testi veicolano rappresentazioni» significa già fare un'ipotesi sulla natura dei *contenuti*. In questo caso, i modelli a cui mi riferisco sono mutuati dalla linguistica e dalla semiotica, ambiti nei quali la parola "contenuto", almeno in prima istanza, è sinonimo di "significato": una delle due facce del segno. L'altra faccia, come si ricorderà, è costituita dall'"espressione", ovvero dal "significante".

Secondo F. de Saussure (1922, tr. it., p. 85), il legame "arbitrario" che unisce il significante al significato è determinato (codificato) dalla lingua intesa come sistema sociale. Diversamente, riferendoci alla *semiotica cognitiva*[6] di C. S. Peirce, possiamo affermare che i contenuti, ad esempio quelli costruiti (o ri-costruiti) attraverso un processo di analisi dei testi, sono il risultato di qualche *inferenza*.

> Un segno, o *representamen*, è qualcosa che sta a qualcuno per qualcosa sotto qualche rispetto o capacità... Definisco un *Segno* come qualcosa che da un lato è determinato da un *oggetto* e dall'altro determina un'idea nella mente di una persona, in modo tale che quest'ultima determinazione, che io chiamo l'*interpretante* del segno, è con ciò stesso mediatamente determinata da quell'*oggetto* (C. S. Peirce, 1980, p. 132 e 194).

Detto in altri termini: il segno, anche considerato come realtà a due facce (espressione e contenuto), "diventa" rappresentazione solo in virtù della mediazione di un *interpretante* che lo mette in relazione con un "qualcosa d'altro".

La Fig. 2 illustra una ricostruzione del modello triadico proposto da C.S. Peirce, valido sia per i processi dell'inferenza che per quelli della significazione; ciò in quanto, a giudizio dell'autore, pensiero (inferenza) e segno (significazione) sono strutturalmente la stessa cosa.

[6] La dizione "semiotica cognitiva" è stata proposta da M. Bonfantini nell'introduzione all'edizione italiana di alcuni scritti di C. S. Peirce (1980).

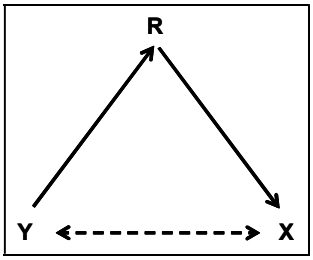



*Fig. 2 – Il modello triadico di C. S. Peirce*

Nel caso della significazione, le chiavi di lettura sono le seguenti:

X = *oggetto*, inteso come ciò a cui il segno si riferisce[7];
Y = *segno*, inteso come unità di significante e significato;
R = *interpretante*, inteso come "terza" istanza che, mettendo in relazione segno e oggetto, "trasforma" il segno in *rappresentazione*[8].

Più avanti ci soffermeremo a considerare altre implicazioni di questo modello. Per ora solo due precisazioni:

- l'*interpretante* non è un soggetto (persona o gruppo), bensì una funzione mentale;
- ogni rappresentazione o estrazione di contenuti, compresa la mia ricostruzione del pensiero di Peirce, è anche una interpretazione.

In effetti, quanto l'antropologo C. Geertz (1973, tr. it., p. 45) ha scritto a proposito della "densità" delle descrizioni etnografiche, può essere applicato a molteplici altri casi:

ciò che chiamiamo i nostri dati sono in realtà le nostre interpretazioni delle interpretazioni di altri.

Veniamo quindi alla terza affermazione della nostra scaletta: «i resoconti che includono l'analisi dei testi propongono interpretazioni (cioè rappresentazioni di rappresentazioni)».

La Fig. 3 illustra il modello a cui mi riferisco.

[7] Cioè un qualunque aspetto della "realtà" (fisica, mentale, culturale, ecc.), altri segni compresi.

[8] Il fatto che un segno, in quanto entità a due facce, sia "di per sé" una rappresentazione è materia di dibattito.

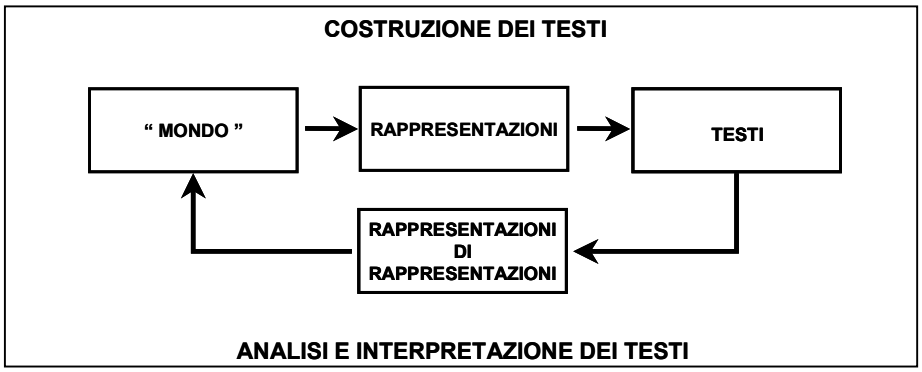



*Fig. 3 – Costruzione e interpretazione dei testi*

Come si può osservare, sia nella costruzione dei testi che nella loro analisi-interpretazione, le rappresentazioni mediano il rapporto tra testi e "mondo"[9]. Quanto alla "natura" di quest'ultimo, in questa sede possiamo pensare alle prime cose che ci vengono alla mente: fatti, emozioni, sistemi sociali, culture, ecc.

Con una sottolineatura: chi costruisce testi non si limita a riferirsi al mondo o a rappresentarlo. Di fatto costruisce *mondi possibili*. La teoria di questi mondi[10], elaborata nell'ambito della logica e della semantica modale[11], è stata ripresa e utilizzata da Umberto Eco (1979) in un suo noto testo sulla cooperazione interpretativa.

Cito da questo:

> Definiamo come mondo possibile uno stato di cose espresso da un insieme di proposizioni dove per ogni proposizione o p o ~ p ( cioé, per il principio di non contraddizione, o è vera la sua affermazione, o è vera la sua negazione). Come tale un mondo consiste di un insieme di *individui* forniti di *proprietà*. Siccome alcune di queste proprietà o predicati sono *azioni*, un mondo possibile può essere visto anche come un *corso di eventi*. Siccome questo corso di eventi non è attuale, ma appunto possibile, esso deve dipendere *dagli atteggiamenti proposizionali di qualcuno*, che lo afferma,

[9] Le relazioni rappresentate in Fig. 3 sono diversamente "interpretate" dalle varie scienze sociali. Si pensi, ad esempio, alle implicazioni della nozione di "mondo interno" nelle teorie psicoanalitiche e alle diverse teorie del simbolismo nell'ambito dell'antropologia culturale.

[10] Secondo questa teoria, la differenza tra mondo "reale" e mondi possibili è fondata sul fatto che nel primo l'esistenza dei suoi "abitanti" è indipendente dalle credenze (o rappresentazioni) di chicchessia; ad es. gli individui in carne ed ossa possono sempre sconfermare le ipotesi sulle "caratteristiche" che gli vengono attribuite. Per dirla nei termini di S. Kripke (1980, tr. it. 46), i mondi possibili sono "stipulati", mentre il mondo reale può essere "scoperto".

[11] Cfr. G.E. Hughes e M.J. Cresswell, 1968; S. Kripke, 1980.

lo crede, lo sogna, lo desidera, lo prevede, eccetera (ibid. p. 128; la nota esplicativa è mia).

Notare bene: un mondo possibile è non soltanto quello de *Il Signore degli Anelli*, di *Dune* o di *Matrix*. Un "tema" di un liceale, la trascrizione di una seduta psicoterapeutica, l'intervista di un manager, l'editoriale di un giornalista e il resoconto di un "evento" ne sono altrettanti validi esempi.

Per inciso: la logica dei mondi possibili consente di trasformare i testi in tabelle (righe e colonne), piene di "1" e di "0" (1 = vero, 0 = falso), che possono essere oggetto di trattamenti statistici. Come vedremo più avanti, T-LAB costruisce ed analizza anche tabelle di questo genere.

Tornando alla Fig. 3: così come le rappresentazioni possono essere trasformate in testi, questi a loro volta possono diventare delle rappresentazioni; ciò a causa del fatto che entrambi sono "realtà" costituite da segni.

Sulla distinzione tra *analisi* e *interpretazione* mi soffermerò più avanti. Quanto alla distinzione tra rappresentazioni e "rappresentazioni di rappresentazioni", può esser letta come la differenza tra comunicazione e metacomunicazione[12], tra linguaggio e metalinguaggio, o analoghi.

Chi ama la vertigine della ricorsività avrà già pensato che le interpretazioni di testi producono altri testi, che questi possono essere nuovamente interpretati, e così via *ad infinitum*; un po' secondo la logica della semiosi illimitata proposta da C. S. Peirce.

Certo, anche questa è una forma di pensiero; ma per quale mercato?

Nella quarta affermazione della scaletta («i resoconti, se ben costruiti, attivano processi di pensiero sulle rappresentazioni») ho inserito una condizione («se ben costruiti») che, detta così, vuol dire tutto e niente allo stesso tempo.

Riconosco che non riesco a spiegarmi senza chiamare in causa la quinta e ultima affermazione: «i processi di pensiero sulle rappresentazioni facilitano la comprensione dei cambiamenti culturali e possono attivare nuove strategie dell'azione sociale».

Evidentemente la condizione «se ben costruiti» non si riferisce alle proprietà stilistiche o retoriche di un paper o di una presentazione, né alla mera "correttezza metodologica". Tutte queste caratteristiche, anche se auspicabili, si riferiscono infatti alla logica del resoconto come *prodotto tecnico*; ma, non dimentichiamolo, il cliente si aspetta un *servizio*.

La letteratura su questi temi (prodotto-servizio) è ampia e variegata. Io stesso, più di qualche anno fa, ho scritto qualcosa al riguardo[13]. Per dirla nel modo più semplice: c'è servizio quando chi fornisce un prodotto si occupa dell'*uso* che gli altri (i clienti) ne fanno e ne possono fare.

[12] Cfr. P. Watzlawick, J.H. Beavin e D.D. Jackson, 1967.

[13] Cfr. F. Lancia, 1996.

Quanto ai resoconti concernenti analisi dei testi, gli usi possibili sono fondamentalmente di due tipi:

- usarli come strumenti per **incrementare la conoscenza**;
- usarli come strumenti per **elaborare strategie di azione**.

Questi usi, complementari tra loro, variano in funzione dei tipi di contesti e dei tipi di clienti. Ad esempio, i clienti interni alle comunità scientifiche possono essere più o meno interessati all'uso dei modelli interpretativi proposti, ai metodi e alle tecniche utilizzate; non ultimo per verificare ipotesi di lavoro e progettare ulteriori attività di ricerca. Quanto agli *altri* tipi di clienti, si pongono due problemi di *traduzione* che, nell'ottica del servizio, dovrebbero essere anche a carico dei ricercatori.

In primo luogo, la classica costruzione argomentativa dei resoconti di ricerca (premesse-obiettivi-metodi-risultati-conclusioni) va tradotta – punto per punto – in un linguaggio che tenga conto della *cultura* di cui il cliente è portatore. Ad esempio, una cosa è interloquire con chi si occupa di strutture sanitarie altra è interloquire con chi si occupa di marketing, una cosa è interloquire con i sindacati altra è interloquire con gli insegnanti. In questi casi, anche le tabelle e i grafici dovrebbero essere costruiti con una diversa "grammatica".

In secondo luogo, la struttura del resoconto deve includere delle *indicazioni operative*. La parola "deve", ovviamente, non sta per una prescrizione normativa, bensì serve a marcare l'orientamento al *servizio*. In effetti, chi commissiona ricerche è in primo luogo interessato a rispondere a questa semplice domanda: "cosa me ne faccio?", ovvero "come posso tradurre i risultati in strategie di intervento?".

La chiave di volta, se così si può dire, è in due espressioni contenute nella quinta affermazione della nostra scaletta: "cambiamenti culturali" e "nuove strategie" in interazione tra loro.

Come è noto, teorie dei cambiamenti culturali sono state elaborate in vari ambiti disciplinari. Dalla sociologia alla psicologia, dalla storia della scienza alla storia del quotidiano: teorie dei sistemi, teorie della complessità, teorie della mente, teorie delle rappresentazioni sociali, cambiamenti di paradigmi, cambiamenti di mentalità, cambiamenti lineari, cambiamenti discontinui, ecc.

In questo contesto, i riferimenti concettuali che utilizzo sono sintetizzati in due citazioni: la prima di un antropologo, la seconda di un semiologo.

> La **cultura** è l'intelaiatura di significato nei cui termini gli esseri umani interpretano la loro esperienza e guidano la loro azione (C. Geertz, 1973, tr. it., p. 188).

> La cultura, come sistema complesso, è formata di strati sviluppatisi a diverse velocità, così che qualunque suo taglio sincronico mostra la simultanea presenza di vari

stadi. Le esplosioni in alcuni strati possono unirsi al graduale sviluppo in altri. Questo tuttavia non esclude la loro interazione… Sia i processi esplosivi che quelli graduali, in una struttura funzionante sincronicamente, adempiono a importanti funzioni: gli uni assicurano l'innovazione, gli altri, la continuità (J.M. Lotman, 1992, tr. it., pp. 24-25).

Notare: le interazioni tra i due processi individuati da Lotman portano a concepire l'"intelaiatura" dei significati che costituiscono le culture come una realtà dinamica e in cambiamento, anche quando questa viene rappresentata come "statica".

Quanto poi alla nozione di **strategia**, la intendo come riferita a una sequenza di decisioni, siano esse anticipate che realizzate, ciascuna delle quali implica l'adozione di criteri di categorizzazione concernenti gli *stati del mondo* (cioè dati e informazioni sul contesto) e gli *obiettivi* da perseguire: i primi categorizzati in base a criteri di rilevanza, i secondi in base a criteri di adeguatezza.

A partire da questi riferimenti, la mia ipotesi è che tanto la comprensione dei cambiamenti culturali quanto l'elaborazione di nuove strategie sono attività inferenziali che implicano riformulazioni dei *criteri di categorizzazione*, cioè dei criteri che usiamo per «rendere equivalenti cose distinguibilmente differenti, aggruppare gli oggetti, gli eventi e la gente intorno a noi in classi, e rispondere ad essi nei termini della loro appartenenza ad una classe, anziché della loro singolarità» (J. Bruner e R.M. Brown,1956, tr. it., p. 15).

Come corollario di questa ipotesi, ritengo che un resoconto "ben costruito" sia caratterizzato dal fatto che può funzionare come strumento per fare due tipi di inferenze, più o meno "creative": conoscitive e pratiche.

Torniamo quindi al modello di Peirce (Fig. 2).

Nel caso dell'*inferenza conoscitiva*, le chiavi di lettura sono le seguenti:

X = *regola* (*type)*, intesa come "legge" generale o come "modello" interpretativo;
Y = *premessa* (*token*), intesa come constatazione di un fatto o "risultato" di una ricerca;
R = *conclusione*[14], intesa come interpretazione del risultato e come "valore aggiunto" della conoscenza.

La diversa combinazione dei tre elementi (X, Y, R) porta Peirce a distinguere tre tipi di inferenze: abduttiva, induttiva e deduttiva[15]. Delle tre, prende-

[14] Secondo C.S. Peirce (1980, p. 142), la conclusione "rappresenta" l'Interpretante.
[15] Un'analoga tipologia è stata proposta da Aristotele.

rò in considerazione solo la prima, in quanto è direttamente collegata alla "logica" della categorizzazione.

L'"argomento" abduttivo, nella sua forma semplificata ha la seguente struttura:

*premessa*: l'evento "a" ha caratteristiche così-e-così;
*regola*: gli eventi della categoria "x" hanno caratteristiche così-e-così;
*conclusione*: l'evento "a" appartiene alla categoria "x".

In altri termini, l'attribuire un evento a una classe (categorizzazione), il riconoscere un caso come appartenente a un "tipo" o il codificare una unità di analisi come appartenente a una categoria, sono altrettanti esempi di inferenza abduttiva.

Per quanto concerne l'*inferenza pratica*[16], vari studiosi[17] l'hanno analizzata in relazione con la cosiddetta spiegazione teleologica (da *télos* = fine o scopo), intesa quest'ultima come «una ricostruzione del calcolo fatto dall'agente dei mezzi da adottare per il fine proposto, alla luce delle circostanze in cui si trova» (W. Dray, 1957, tr. it., p. 170).

La ricostruzione dell'inferenza pratica, nel suo schema semplificato, è la seguente:

*premessa*: il soggetto "a" ha l'intenzione di realizzare "p";
*regola*: "a" ritiene che per realizzare "p" deve fare "q";
*conclusione*: "a" si dispone a fare "q"[18].

In termini di processo psico-logico, la *regola* dell'inferenza pratica è fondata sulla *conclusione* (il criterio di categorizzazione) di una precedente inferenza cognitiva. Quest'ultima, nel caso delle inferenze pratiche che sostengono le "pratiche forti" (ad es. ingegneria e medicina) può avere la forma dell'induzione[19]; diversamente, nel caso delle "pratiche deboli" (vedi scienze umane) ha la forma dell'abduzione[20].

[16] Assimilabile alla "credenza pratica" in base alla quale – secondo C. S. Peirce (1984, p. 246) – l'uomo si dispone ad agire.

[17] Cfr. G.H. Wright (von), 1971; G. Di Bernardo, 1983.

[18] "p" designa un obiettivo, "q" un'azione.

[19] La forma tipica dell'argomento induttivo è la seguente:
*premessa*: il caso "a" del fenomeno "x" è accompagnato dalle circostanze "C";
*regola* (o *legge*): in generale i casi "a", "b", "c", "...", "n" sono accompagnati da circostanze "C";
*conclusione*: "tutti" i casi del fenomeno "x" sono accompagnati dalle circostanze "C".

[20] Nella seconda edizione del suo manuale sull'analisi di contenuto, K. Krippendorff (2004, p. 38) afferma che «the whole enterprise of content analysis may well be regarded as an argument in support of an analyst's abductive claims»; tuttavia, la sua impostazione metodologica (vedi la sua "typology for validating evidence") continua ad essere improntata a una logica di tipo induttivo.

Ne deriva che, nel caso delle "pratiche forti", le regole delle inferenze sono fondate su leggi generali o su previsioni altamente probabili (vedi, ad esempio, le prescrizioni farmacologiche e le verifiche di stabilità degli edifici); quindi le relazioni tra conoscenza e azione sono, per così dire, ipercodificate. Al contrario, quando si tratta di ancorare le "indicazioni operative" a un resoconto di ricerca che includa l'analisi di testi, le relazioni tra i due tipi di inferenze (cognitive e pratiche) sono ipocodificate[21].

Ciò porta ad interrogarsi sui criteri di *validità* della conoscenza prodotta, cioè sui criteri in base ai quali i risultati dei processi inferenziali possono – per così dire – essere presi sul serio, sia dai committenti-clienti che all'interno delle comunità scientifiche. Poiché interrogativi di questo tipo chiamano in causa complesse questioni epistemologiche, in questa sede mi limito a proporre una "semplice" distinzione tra due tipi di validità: la prima, concernente le inferenze sul *significato* dei testi basate sulle loro relazioni con le caratteristiche del contesto inscritte nel corpus analizzato (ad es. le caratteristiche socioculturali dei locutori); la seconda riferita ad inferenze generalizzanti concernenti le relazioni tra i testi analizzati ed altri testi o contesti. Mentre il primo tipo di validità è direttamente connesso connesso con la logica degli strumenti di indagine (vedi T-LAB), la seconda discende dai "disegni di ricerca" e dai modelli teorici a cui i ricercatori fanno riferimento.

In ogni caso, l'esplicitazione dei criteri (le *regole*) fondanti i vari tipi di inferenze è un compito ineludibile. Solo in questo modo, infatti, i resoconti saranno "ben costruiti", cioè potranno essere utilizzati dai clienti per *rendere conto* ad altri clienti, sia essi interni alla loro organizzazione o utilizzatori di qualche loro prodotto-servizio.

## 1.3. Testo, contesto e corpus

Nel disegno della ricerca, prima dell'analisi dei testi si procede alla loro raccolta e alla predisposizione di un corpus. Con quali criteri?

Tornando alla Fig. 1 (paragrafo 1.1) ne possiamo individuare cinque:

- la domanda del cliente;
- gli interessi e gli obiettivi del ricercatore;
- il tipo di prodotto che si intende ottenere;
- le teorie e i modelli a cui il ricercatore fa riferimento;
- le tecniche e gli strumenti che si intendono utilizzare.

[21] La distinzione tra abduzioni ipercodificate e ipocodificate è stata proposta da U. Eco (1983).

Ciascuno di essi contribuisce a definire il contesto della ricerca, cioè un insieme di vincoli e di opportunità per le attività di analisi e di interpretazione[22]. Prima di entrare nel merito di quest'ultime, e per delimitare il campo della loro applicazione, proporrò alcune definizioni e alcune argomentazioni concernenti le parole chiave e i nodi concettuali che ritengo più rilevanti. In particolare, dopo aver chiarito le nozioni di testo, contesto e corpus, mi soffermerò a considerare alcune implicazioni metodologiche di un dibattito entro il quale i ricercatori costruiscono il significato delle loro prassi di analisi e di interpretazione: quello articolato dall'opposizione *qualitativo vs quantitativo*.

In primo luogo, qualche breve (e provvisoria) definizione[23]:

- il **testo** è una parte di un evento comunicativo, espressa nel linguaggio verbale e fissata nella scrittura;
- il **contesto** è costituito dall'insieme delle dimensioni linguistiche ed extralinguistiche a cui, di volta in volta, l'interpretazione dei testi fa riferimento;
- il **corpus** è un testo (o una collezione di testi) assunto come oggetto di analisi, selezionato e "preparato" per essere trattato con opportuni metodi e tecniche allo scopo di fare inferenze sui suoi contenuti.

Nella definizione di **testo**, i riferimenti all'evento comunicativo e al linguaggio verbale chiamano in causa le relazioni tra aspetti semantici, sintattici e pragmatici, cioè le relazioni tra il testo e ciò che esso significa (semantica), le relazioni tra gli enunciati che lo costituiscono e che organizzano la sua forma (sintattica), le relazioni tra il testo e chi lo produce e/o lo interpreta (pragmatica).

Uno dei più noti modelli dell'**evento comunicativo**, quello proposto da R. Jackobson (1963), prevede sei fattori in relazione tra loro: mittente, messaggio, destinatario, canale (o contatto), codice e contesto. Di primo acchito il testo viene identificato con messaggio, ma, a ben vedere, in esso sono presenti anche tracce degli altri fattori: del mittente (o locutore), in quanto non esiste testo che, in qualche modo, non sia *firmato*; del destinatario, sia come interlocutore presupposto dagli atti linguistici (J.L. Austin, 1962), sia come attore sociale che media la rappresentazione dei contenuti veicolati dal testo; del canale, ad esempio attraverso le varie trasformazioni dell'editing; del codice,

[22] In particolare, i criteri adottati nella *costruzione* del corpus (selezione dei testi, loro codifica, ecc.) determinano l'ambito di *validità* dei processi inferenziali.

[23] Evidentemente, le tre definizioni non pretendono di avere validità generale. Ad esempio, la definizione di testo fornita dai semiologi può includere il riferimento a linguaggi non verbali. Oppure, quando un lessicografo costruisce un corpus può avere obiettivi diversi da quelli di "fare inferenze sui suoi contenuti".

attraverso la lingua in cui è il testo è scritto; del contesto, attraverso vari tipi di mediazioni culturali.

Ne consegue che ogni testo può essere analizzato stabilendo relazioni tra esso e i vari fattori dell'evento comunicativo, oppure, secondo altre teorie, tra esso e i fattori dell'agire comunicativo (S.J. Schmidt, 1973). Ovviamente, le pratiche di ricerca che sono orientate a fare inferenze sui contenuti (o rappresentazioni) tengono conto solo di alcuni tipi di relazioni. In ogni caso, le relazioni tra testo, codice e contesto sono ineludibili.

Quanto alla nozione di **codice**, faccio riferimento a due definizioni proposte da semiologi:

> Un codice… consiste di due universi del discorso, il campo semantico e il campo noetico, le cui divisioni in classi complementari sono corrispondenti (L. J. Prieto, 1966, tr. it., p. 56).

> Un codice stabilisce equivalenze semantiche tra elementi di un sistema di significanti e elementi di un sistema di significati (U. Eco, 1973, p. 73).

La logica di queste definizioni è illustrata dal un modello (vedi Fig. 4) mutuato dalla teoria degli insiemi. Entro questa teoria, dato un insieme *A*, il suo *insieme complementare* – in genere denominato *A'* – include tutti gli "oggetti" che non sono elementi di *A* e che sono il risultato di un'operazione di differenza. Nel caso specifico, dato un codice, ogni elemento dell'insieme (o classe) dei significanti *A* "ammette" un insieme *B* di significati e ne esclude altri (*B'*); reciprocamente, ogni elemento dell'insieme (o classe) dei significati *B* ammette un insieme di significanti *A* e non altri (*A'*). Ad esempio, nella nostra lingua, la parola "uomo" (*A*) articola la differenza tra l'insieme dei "maschi umani adulti" (*B*) e tutto ciò che non corrisponde a questa definizione (*B'*).

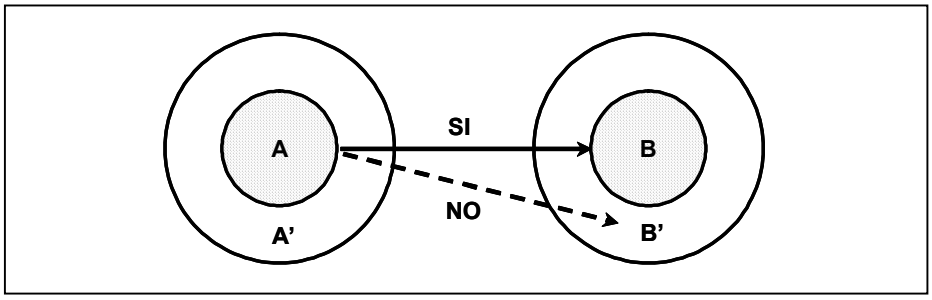



*Fig. 4 – I codici come regole di corrispondenza*

In altri termini, i codici linguistici funzionano come *regole* per fare inferenze; più precisamente, come propone U. Eco (1983), per fare delle *abduzioni ipercodificate*[24].

Consideriamo ora la nozione di **contesto**. Se questo è un tipo di ambiente, metaforicamente possiamo utilizzare il riferimento ai sistemi operativi dei computer (ad es. Microsoft Windows e Macintosh OS) cioè gli ambienti software che stabiliscono le regole per usare altri software.

Lo studio dell'epistemologia insegna che molti modelli della scienza sono in qualche modo delle metafore[25], cioè dispositivi semiotici in grado di trattare l'"ignoto" come se fosse "noto", dei modi di «parler d'une chose dans les termes d'une autre qui lui ressemble»[26] (P. Ricoeur, 1975, p. 250). Basti pensare ai modelli dell'atomo in forma di sistemi planetari e ai modelli della mente in forma di reti.

Il punto è che mentre alla nozione di codice corrispondono vari modelli, la nozione di contesto – soprattutto quando non è riferita a entità linguistiche – evoca solo metafore[27]. Ma se, come ho proposto (vedi paragrafo 1.1), la nozione di contesto corrisponde a *un modo di stabilire relazioni* tra qualcosa e qualcosa d'altro (dove, in genere, il qualcosa è un sottoinsieme del qualcosa d'altro), le differenze tra codice e contesto sembrano assottigliarsi.

Più precisamente, la mia ipotesi è che tanto i codici quanto i contesti (culturali e/o sociali) – così come sono "rappresentati" all'interno di un processo di analisi – sono sistemi di regole per fare delle inferenze. Diversamente non saprei come connettere le cosiddette rappresentazioni sociali al cosiddetto comportamento individuale. La differenza tra abduzione ipercodificata e abduzione ipocodificata è a questo proposito un utile modello di riferimento; in effetti, quando codifichiamo-decodifichiamo il significato delle parole siamo piuttosto certi di quello che stiamo facendo, al contrario quando *contestualizziamo* sappiamo che le regole a cui facciamo riferimento sono solo "ipotesi plausibili".

Domanda: le teorie e i modelli che usiamo per interpretare i testi sono codici o contesti?

La domanda diventa meno imbarazzante se la traduciamo nei termini se-

[24] Secondo Eco, nel caso dell'abduzione ipercodificata la *regola* è data in modo automatico o semi-automatico; diversamente, l'abduzione ipocodificata richiede una selezione tra «regole equiprobabili messe a nostra disposizione» e «Poiché la regola è selezionata in quanto è la più plausibile tra molte, ma non è certo la più "corretta", la spiegazione (ovvero la *conclusione* dell'inferenza) è solo presa in considerazione, in attesa di successive verifiche» (U. Eco, 1983, p. 245, il testo tra parentesi è mio).

[25] Cfr. M. Black, 1962; A. Ortony, 1979; M. B. Hesse, 1970.

[26] «Parlare di una cosa nei termini di un'altra che le rassomiglia».

[27] Vedi le molteplici variazioni sulla dicotomia parte/tutto.

guenti: a partire dai dati che sto analizzando (*premessa*) le teorie e i modelli che sto utilizzando (*regola*) che tipo di inferenze mi consentono di fare (*conclusione*)?

Infine qualche considerazione sulla nozione di **corpus**.

Questo, anche quando è costituito da una collezione di testi, per quanto piccoli e frammentati essi siano, è *un* testo e delimita *un* contesto. Sia il testo che il contesto, ovviamente, sono in questo caso quelli definiti dal ricercatore: *un* testo perché si presume che i vari sottoinsiemi del corpus abbiano a che fare con tipi di contenuti analoghi, vuoi perché veicolano rappresentazioni riferite agli stessi tipi di oggetti, vuoi perché propongono un «insieme di enunciazioni tematicamente coerente con funzione comunicativa riconoscibile» (S.J. Schmidt, 1973, tr. it., p. 257); *un* contesto in quanto, parafrasando una nota affermazione di L. Wittgenstein (1961), potremmo affermare che i limiti del nostro corpus significano i limiti della nostra analisi[28].

In effetti, a partire dal momento in cui il ricercatore ha selezionato e "preparato" un corpus alcune "premesse" delle inferenze cognitive e pratiche (ad esempio le decisioni circa i metodi di analisi) sono già diventate dei paletti per montare una tenda, per ancorare una barca o per costruire qualche recinzione.

In particolare, quando si usano software, si pongono problemi di formato e di "taglia". Problemi questi che, essendo specifici di ogni software, tratterò più avanti. Segnalo tuttavia che il problema della "taglia" non è di poco conto. Anni fa F. Fornari (1977) ha scritto un libro di circa 200 pagine analizzando la trascrizione di un consiglio di classe che ne "durava" poco più di 5, cioè, in termini informatici, circa 10 Kb di file testo. Ebbene, un corpus di questo tipo è "troppo piccolo" per essere utilmente analizzato con l'ausilio di software, soprattutto se si tratta di software che usano algoritmi di tipo statistico.

Qualcuno dirà: nel caso dei metodi qualitativi, questo tipo di problema non si pone.

## 1.4. Sull'opposizione "quantitativo" *vs* "qualitativo"

Nell'ambito delle scienze sociali l'opposizione quantitativo *vs* qualitativo è un *tòpos* particolarmente interessante: una sorta di concentrato di luoghi comuni, di questioni epistemologiche e metodologiche che meriterebbe uno studio a parte[29]. In questa sede, mi limiterò ad esplorare alcune delle sue implica-

[28] La forma originale della proposizione 5.6 del *Tractatus logico-philosophicus* è la seguente: «*I limiti del mio linguaggio* significano i limiti del mio mondo».

[29] In Italia, i motivi dell'"incerta distinguibilità" fra analisi quantitativa e analisi qualitativa sono stati lucidamente esposti da E. Campelli (1996). La mia ipotesi è che il dibattito su questi

zioni concernenti l'analisi dei testi e, per essere più chiaro, farò riferimento a due volumi di autori italiani[30]: *La ricerca qualitativa* (a cura di L. Ricolfi, 1997) e *Metodi qualitativi in psicologia sociale* (a cura di B. Mazzara, 2002).

Prima osservazione: nel primo volume le differenze tra metodi quantitativi e metodi qualitativi vengono tematizzate e articolate, nel secondo – pubblicato cinque anni più tardi – sono sfumate; già nelle parole usate dai rispettivi curatori.

L. Ricolfi (1997), proponendo una tassonomia dei metodi, scrive:

> La ricerca "qualitativa" è stata caratterizzata attraverso tre tratti:
> a) l'assenza della matrice dati;
> b) la non ispezionalità della base empirica;
> c) il carattere (relativamente) informale delle procedure di analisi dei dati (ibid., p. 32).

Simmetricamente, lo stesso autore afferma che la *ricerca quantitativa* è stata caratterizzata attraverso tre tratti: l'impiego della matrice dati; la presenza di definizioni operative dei "modi" della matrice dati (perlopiù casi e variabili); l'impiego della statistica o dell'analisi dei dati (ibid., p. 30).

B. Mazzara (2002), dopo aver sottolineato il fatto che «si tratta di opposizioni non dicotomiche bensì graduabili», afferma che è «non solo legittimo, ma anche molto produttivo usare in maniera combinata i diversi approcci». Inoltre rileva che «è sul piano metodologico che la valorizzazione di soluzioni intermedie ha avuto più fortuna, favorita anche, negli ultimi anni, dal notevole sviluppo di strumenti informatici» (ibid., p. 28 e 39).

Seconda osservazione: in entrambe i volumi la *Grounded Theory*[31] è presentata come un paradigma della ricerca qualitativa, cioè, per usare una delle accezioni proposte da T.S. Kuhn (1962, tr. it., p. 10), come «un modello di problemi e soluzioni accettabili a coloro che praticano un certo campo di ricerca».

A. Strati (1997), scrivendo sulla *Grounded Theory* (G.T.), ne sottolinea le

problemi è tanto più "acceso" quanto più le comunità scientifiche sono centrate sui loro problemi "interni" e perdono di vista il fatto che possono avere un cliente "esterno". In altri termini, ritengo che i vari modi di articolare le differenze (quantitativo *vs* qualitativo, oggettivo *vs* soggettivo, ecc.) siano prevalentemente usati come modi per dichiarare la propria identità scientifica e la propria appartenenza ideologica e/o culturale.

[30] Chi, più in generale, sia interessato ad approfondire gli aspetti epistemologici di questo dibattito, può utilmente consultare il volume *Il sociologo e le sirene. La sfida dei metodi qualitativi* (a cura di C. Cipolla e A. De Lillo, 1966).

[31] Metodologia di ricerca, inizialmente proposta da B.G. Glaser e A.L. Strauss (1967), caratterizzata dall'intento di fondare (*to ground*) la "generazione" di concetti e teorie sull'attenta esplorazione dei dati e dei contesti in cui essi vengono prodotti.

seguenti caratteristiche: l'orientamento alla *scoperta*, il riferimento alla *sociologia interpretativa*, l'assunto che il fenomeno sociale non può essere tradotto nel linguaggio delle variabili. Aggiunge poi che le caratteristiche dello studio di tipo qualitativo sono: l'*immersione nel contesto*, l'acquisizione di familiarità ad esso, lo svolgere un'analisi in *profondità*, «la ricerca di comprendere quel contesto in termini che sono vicini a quelle persone che in quel determinato contesto vivono ed operano e sono a loro intelligibili» (ibid., p. 131).

Simmetricamente, E. Cicognani (2002) afferma che la G.T. è un esempio di ricerca qualitativa di tipo interpretativo, caratterizzata dall'orientamento alla scoperta e da una concezione "circolare" o "iterativa" (a spirale) del processo di ricerca. In questo ambito, quindi «le fasi della raccolta e dell'analisi dei dati non sono separate ma strettamente interconnesse; in pratica, ciò significa che l'analisi (e la costruzione della teoria) inizia non appena si rendono disponibili i primi dati, e i risultati delle analisi preliminari orienteranno a loro volta la raccolta dei dati seguente» (ibid., pp. 43-45).

Al di là di specifici aspetti di tipo tecnico-metodologico, la concezione dell'analisi dei testi come un processo ciclico[32] fondato sull'attività interpretativa riconducono la G.T. entro il modello metodologico del *circolo ermeneutico*[33], che H. G. Gadamer (1960, tr. it., p. 314) ha descritto nel modo seguente:

> Chi si mette a interpretare un testo, attua sempre un progetto. Sulla base del più immediato senso che il testo gli esibisce, egli abbozza preliminarmente un significato del tutto. E anche il senso più immediato il testo lo esibisce solo in quanto lo si legge con certe attese determinate. La comprensione di ciò che si dà da comprendere consiste tutta nella elaborazione di questo progetto preliminare, che ovviamente viene riveduto in base a ciò che risulta dall'ulteriore penetrazione del testo… l'interpretazione comincia con dei pre-concetti i quali vengono via via sostituiti da concetti più adeguati[34].

In effetti, il riferimento alla Grounded Theory ci consente di meglio capire che l'opposizione quantitativo *vs* qualitativo ha più di qualche radice nell'antica querelle spiegazione *vs* comprensione[35], la quale – a partire dalla fine dell'ottocento – è stata alimentata dal bisogno di fondare l'autonomia del-

[32] La natura ciclica del processo riguarda due tipi di "movimenti": dall'osservazione all'interpretazione, e viceversa; dalle parti al tutto, e viceversa.

[33] In realtà, la G.T. è una tradizione di ricerca molto variegata e, tramite uno dei suoi fondatori (A.L. Strauss), viene fatta risalire all'interazionismo simbolico, alla scuola di Chicago e al pragmatismo di J. Dewey.

[34] Secondo P. Ricoeur (1969, tr. it., p. 30), il circolo ermeneutico implica anche una correlazione tra la comprensione del testo e l'autocomprensione dell'interprete.

[35] O spiegazione *vs* interpretazione.

le *scienze dell'uomo*[36] rispetto alle *scienze della natura*. Evidentemente non è questo il luogo per ricostruire un secolo di dibattiti; basti ricordare quanto, in anni più recenti e sul nostro tema specifico, ha scritto P. Ricoeur:

> La nozione di spiegazione si è modificata: non viene più derivata dalle scienze della natura, ma da modelli propriamente linguistici. Quanto alla nozione di interpretazione, essa ha subito nell'ermeneutica moderna profonde trasformazioni che l'hanno allontanata dalla nozione psicologica di comprensione, nel senso indicato da Dilthey… Ne risulta che sarà sullo stesso terreno, all'interno della sfera del linguaggio, che spiegazione e interpretazione si confronteranno (1970, tr. it., p. 133 e 147).

Secondo Ricoeur, i modelli scientifici che consentono di spiegare le "relazioni interne" ai testi scritti sono quelli elaborati dallo strutturalismo linguistico e dalla semiotica; diversamente l'ermeneutica, trattando i testi come discorsi parlati e restituendoli alla "viva comunicazione", si occupa di interpretarli. Ma i due "atteggiamenti" sono tra loro complementari; infatti, se consideriamo l'analisi strutturale una fase, necessaria, tra un'interpretazione ingenua ed un'interpretazione critica, allora – afferma l'autore – sarà possibile considerare la spiegazione e l'interpretazione come due fasi diverse all'interno del circolo ermeneutico (ibid., p. 151).

Personalmente, quando penso alla pratiche di ricerca realizzate con il supporto di software, trovo più utile distinguere tra momenti di **analisi** e momenti di **interpretazione**, ma non come fasi che si succedono linearmente. Inoltre trovo più utile riservare il termine **spiegazione** per indicare tutti i processi argomentativi tramite i quali costruiamo resoconti delle varie attività di analisi e di interpretazione.

Dal mio punto di vista, il motivo per cui l'opposizione tra i due "atteggiamenti" diventa in molti casi un luogo comune e non pensato è nel fatto che, in genere, le teorie dell'interpretazione non includono una teoria dell'analisi. Di conseguenza, poiché l'analisi tende ad essere identificata con il puro "fare", lo statuto dell'interpretazione rischia di essere impropriamente legittimato come una forma autocompiacente di "pensiero debole"[37].

Quanto alla fondazione teorica dell'analisi, ritengo che i modelli della linguistica e della semiotica offrano più di qualche opportunità, soprattutto per il fatto che, come argomenterò nei capitoli successivi, consentono di legittimare e di "pensare" il passaggio dai testi alle unità di analisi e dalle parole ai numeri.

[36] All'epoca di Dilthey e del suo primo scritto su questo tema (1896) erano definite "scienze dello spirito".

[37] Che poi, in quanto fondato sulla logica dell'abduzione, lo statuto dell'interpretazione venga fatto corrispondere alla forma epistemologica del "pensiero debole" è altra cosa.

Anche delle tecniche statistiche, di quelle usate in T-LAB, dirò più estesamente altrove; tuttavia, poiché il loro impiego nell'ambito dell'analisi dei testi ha alimentato più di qualche pregiudizio, ritengo opportuno premettere alcune considerazioni.

Secondo H. G. Gadamer (1967, tr. it., p. 81), «I pregiudizi non sono necessariamente ingiustificati ed erronei per il fatto che mascherano la verità». Essi, nel senso etimologico del termine (pre-giudizi), costituiscono «le linee orientative provvisorie che rendono possibile ogni nostra esperienza»[38]. Probabilmente, molti ricercatori che si riconoscono nel paradigma ermeneutico condividono qualche pregiudizio nei confronti della tecniche statistiche, vuoi perché implicano una "presa di distanza" rispetto ai testi (provvisoria e reversibile), vuoi perché sembrano dar poco spazio all'"immaginazione" del ricercatore[39]. A questi, come anche a coloro che condividono pre-giudizi opposti (ad es. coloro secondo i quali «non c'è scienza senza statistica») può essere utile ricordare quanto lo stesso Gadamer, riflettendo sulle relazioni tra ermeneutica e statistica, ha affermato:

> Un esempio utile che mostra come la dimensione ermeneutica interessi anche l'insieme dei procedimenti scientifici, ci è offerto dalla statistica… Ciò che viene constatato (tramite l'applicazione della statistica) appare come linguaggio dei fatti; ma proprio su tali interrogativi: a quali problemi questi fatti adducono una risposta? e: quali sono i fatti che comincerebbero a parlare se altri problemi fossero posti? , interviene la riflessione ermeneutica. Sarebbe forse più opportuno porre l'interrogativo ermeneutico per giustificare il significato di quei fatti e delle conseguenze che derivano dalla loro esistenza… Chi intende apprendere una scienza deve apprendere a possederne alla perfezione il metodo. Ma noi sappiamo anche che il metodo, in quanto tale, non basta da solo a garantire la fecondità delle sue applicazioni. Esiste piuttosto una sterilità metodica, nel momento in cui si applica a qualcosa che non è stato assolutamente proposto come oggetto di ricerca, a partire da una posizione autentica del problema (H.G. Gadamer, 1967, tr. it., pp. 83-84; la nota tra parentesi è mia).

Detto ciò, per i non addetti ai lavori, valgono le seguenti precisazioni.

La costruzione di buone tabelle (o matrici) dati è il presupposto fondamentale per il corretto uso delle tecniche statistiche. Tutto sta a vedere cosa si mette in riga, cosa in colonna, e quali tipi di valori (numeri) sono nelle celle. Nel nostro caso, i valori sono costituiti da occorrenze o da co-occorrenze di parole, sia come forme grafiche che come lemmi o come classi semantiche,

[38] Secondo P. Ricoeur (1987, pp. 70-71) «il pregiudizio è semplicemente la proiezione, sul piano del giudizio, di una categoria ermeneutica fondamentale: la tradizione».

[39] Personalmente condivido l'affermazione secondo la quale «l'immaginazione è il dono più importante del ricercatore» (H. G. Gadamer, 1967, tr. it., p. 85).

mentre le label in riga e/o in colonna possono essere parole, contesti elementari (frasi o quasi-frasi) e sottoinsiemi del corpus[40].

Ogni tecnica statistica consente di analizzare alcuni tipi di tabelle (e non altre), e di rispondere ad alcuni tipi di domande (e non altre) circa le relazioni tra i dati. In particolare, quando si tratta di analizzare tabelle di grandi dimensioni (vedi analisi dei testi), vengono utilizzate tecniche statistiche di tipo multivariato, dette anche multidimensionali[41]. Quest'ultime si distinguono in due tipi:

- quelle orientate ad evidenziare **relazioni di dipendenza** tra le variabili, quindi atte a **verificare ipotesi** e a supportare argomentazioni ispirate al paradigma del *Covering Law Model*[42] (casi tipici: regressione ed analisi della varianza);
- quelle orientate ad evidenziare **relazioni di interdipendenza** tra le variabili, quidi atte a **costruire ipotesi** interpretative e a supportare argomentazioni ispirate al *paradigma indiziario*[43] (casi tipici: analisi fattoriale delle corrispondenze, multidimensional scaling e cluster analysis).

Nei software che offrono strumenti di statistica testuale (vedi T-LAB) le tecniche elettive sono quelle del secondo tipo e, come vedremo più avanti, i risultati che esse producono sono in gran parte *icone* da interpretare, cioè delle *rappresentazioni* dei dati iniziali costruite secondo le regole della matematica e della geometria. Più precisamente, nel processo di analisi i testi subiscono due tipi di *trasformazioni*[44]: il primo, mediato da modelli linguistici, esita nella costruzione di tabelle dati (dalle parole ai numeri); il secondo, mediato da algoritmi di tipo statistico, esita nella costruzione di altri tipi di tabelle che possono essere rappresentate in forma di grafici (dai numeri alle icone).

Da un punto di vista epistemologico, la "vera" questione concerne la fondatezza delle trasformazioni messe in atto. Tra queste, quelle più opinabili derivano dall'uso dei modelli linguistici, cioè dal modo in cui questi consentono di "ritagliare", classificare e conteggiare le *unità di analisi*. Diversamente, le trasformazioni del secondo tipo sono più attendibili in quanto gli algoritmi statisti-

[40] Per la spiegazione di questi termini si veda il paragrafo 2.1.

[41] Cfr. L. Lebart, A. Morineau, M. Piron, 1995; J.F. Hair, R.E. Anderson, R.L. Tatham, W.C. Black, 1995; S. Bolasco, 1999.

[42] Letteralmente: "modello della legge di copertura". Secondo questo modello, presupposta la validità di una o più leggi generali, gli eventi (o "fatti") possono essere spiegati o previsti in funzione del verificarsi di specifiche condizioni. Cfr. C.G. Hempel, 1952.

[43] Questa espressione, coniata C. Ginzburg (1979), indica un'insieme di pratiche conoscitive caratterizzate dall'attenzione ai particolari ("tracce" o indizi) e dall'uso di inferenze di tipo abduttivo (sulla logica dell'abduzione vedi paragrafo 1.2 di questo libro).

[44] Vedi capitolo *La logica analitica di T-LAB*.

ci sono delle regole inferenziali che, a partire dagli *stessi* dati e dagli *stessi* parametri (premesse) producono *sempre* gli *stessi* risultati (conclusioni). Quanto poi alla "correttezza" del loro uso, questa dipende dalle competenze dei ricercatori, cioè dalla loro capacità di capire la *logica* dei processi messi in atto.

Ad esempio, per iniziare a ragionare sul nesso tra algoritmi e *rappresentazioni* dei dati, può essere utile riferirsi alla nozione di **trasformazione**, così come è usata nell'ambito della geometria.

Uno dei casi più semplici è costituito dall'omotetia (Fig. 5): una trasformazione lineare che non comporta rotazioni, composta da una traslazione e da una dilatazione centrale.

Nella Fig. 5 ciascuno dei due triangoli (ABC e A'B'C') costituisce una *buona rappresentazione* dell'altro; ciò in base a una regola di trasformazione che è condensata in una semplice formula. Ovviamente, nel caso delle analisi statistiche le trasformazioni messe in atto sono meno intuitive e, nei vari passaggi, non riproducono la completezza dell'informazione; tuttavia, la logica è dello stesso tipo. Ad esempio, usando l'analisi delle corrispondenze, tramite alcune formule (algoritmi) si possono ottenere alcune rappresentazioni dei dati e reciprocamente, usando le "formule inverse", dalle rappresentazioni si può tornare ai dati originari.

In ogni modo, sia che usino o che non usino tecniche statistiche, i ricercatori costruiscono delle rappresentazioni dei dati testuali attraverso *trasformazioni progressive* che fanno capo a due attività distinte e complementari: l'analisi e l'interpretazione.

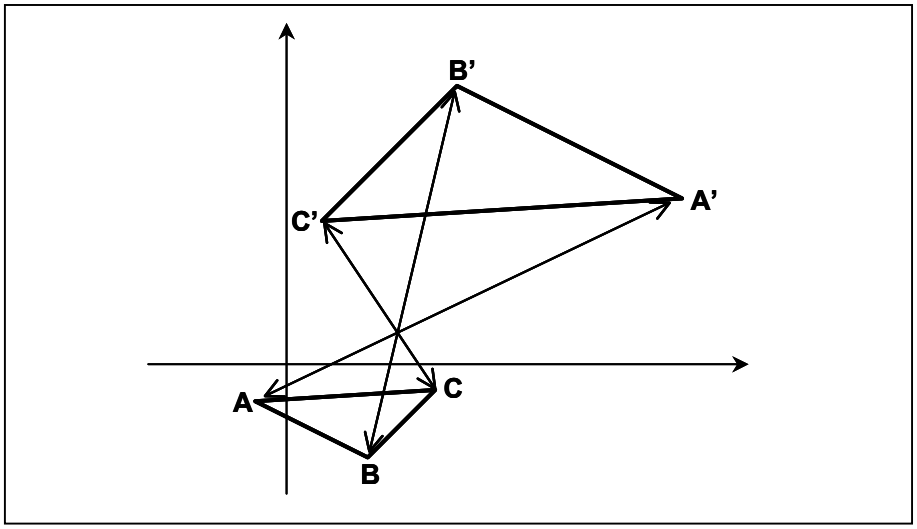



*Fig. 5 Esempio di trasformazione*

## 1.5. Analisi, interpretazione e software

I campi semantici evocati dai termini *analisi* e *interpretazione* sono fluidi e in gran parte sovrapposti. Per articolare le differenze tra i due, parto da un'affermazione di P. Ricoeur (1969, tr. it., p. 96): «L'insieme dei segni, per poterlo sottomettere ad analisi, deve essere considerato come un sistema chiuso». Quindi, assumendo (provvisoriamente) che il corpus sia un sistema chiuso di parole che rinviano solo ad altre parole, considero come operazioni di analisi tutte le trasformazioni che consentono di rappresentare il corpus su un foglio quadrettato o in un database. Queste, dal mio punto di vista, possono essere distinte in tre tipi:

a- **scomporre** il corpus in **unità di analisi**[45] che, a seconda dei casi, possono essere testi, paragrafi, enunciati o singole parole;
b- **classificare** le unità di analisi, attribuendo a ciascuna di esse una *label* che corrisponde a una **categoria**[46];
c- **stabilire relazioni** tra le unità di analisi, applicando qualche **regola** che corrisponda a operazioni logiche o a calcoli di tipo statistico[47].

Fin qui l'analisi. Con una sottolineatura: tutti e tre i processi (*a*, *b*, *c*) sono attività inferenziali[48], più o meno codificate. Per intenderci: anche quando si decide di scomporre un testo in paragrafi, la *premessa* è costituita dall'osservazione del testo come "catena", la *regola* è costituita dal punto e a capo, la *conclusione* è l'operazione di scomposizione in unità di contesto.

Quanto poi al fatto che i tre tipi di operazioni possono essere effettuate con l'ausilio di software o tramite algoritmi in essi implementati, nulla toglie alla struttura logica del processo analitico. Prendiamo ad esempio l'operazione del

[45] «Il primo compito dell'analisi – scrive L. Hjelmslev (1961, tr. it., p. 33) – è di affrontare una partizione del processo testuale. Il testo è una catena, e tutte le parti… sono similmente catene, tranne le eventuali parti ultime che non si possano sottoporre ad analisi».

[46] Ogni categoria è una classe di equivalenza, cioè un insieme di oggetti (in questo caso, unità di analisi) con qualche caratteristica in comune. La label è costituita da una o più parole che designano la classe. Il fatto che, a seconda dei casi, le categorie possono essere fondate su criteri lessicali o possono derivare da modelli teorici non cambia la struttura logica delle operazioni messe in atto.

[47] Nel caso dell'applicazione di calcoli, un'ulteriore operazione analitica è costituita dalla definizione delle regole per il "conteggio" (presenza/assenza, frequenze, co-occorrenze, etc; vedi L.Bardin, 1977, pp. 108-114).

[48] K. Krippendorff (2004, p. 36) sostiene una posizione analoga: «Content analytic inferences may be hidden in the uman process of coding. They may be built into analytical procedures, such as the dictionaries in computer-aided text analyses or well-established indices. Sometimes, especially after complex statistical procedures have been applied, inferences appear in the analyst's interpretations of statistical finding».

classificare: che sia il ricercatore ad affermare «questa parola (o questa frase) è simile a quest'altra» o che sia il software a produrre qualcosa di analogo, la vera questione è «in base a quali criteri?», oppure «in base a quali algoritmi?». Notare bene: quando si parla di analisi siamo nella sfera *immanente* del linguaggio, cioè stiamo considerando il corpus come un sistema di segni che rinviano solo ad altri segni. Secondo P. Ricoeur, la *chiusura dell'universo dei segni* è il postulato fondamentale della linguistica strutturale[49]: una conquista della scientificità e una «digressione necessaria» preliminare a ogni attività ermeneutica; anche per chi è convinto che «l'essenziale del linguaggio comincia al di là della chiusura dei segni» (1969, tr. it., p. 110).

In effetti, la distinzione tra linguistica strutturale ed ermeneutica, sia letta come opposizione che come possibilità di integrazione, è un buon modello per riflettere sulle relazioni tra analisi ed interpretazione.

La Tab. 1 ne riassume le componenti principali.

*Tab. 1 – Linguistica ed ermeneutica*

| *linguistica*[50] | *ermeneutica* |
|---|---|
| referenza interna | referenza esterna |
| fuori del tempo | nel tempo |
| senza soggetto | relazione tra locutori |
| struttura | senso |
| codice | discorso |
| enunciato | enunciazione |
| spiegazione | comprensione |
| *analisi* | *interpretazione* |

Un adeguato commento di questa tabella richiederebbe molte pagine[51]. La-

[49] Postulato che rinvia ad una affermazione di L. Hjelmslev (1959, p. 21): «Intendiano per linguistica strutturale un insieme di ricerche fondate sull'ipotesi che è scientificamente legittimo descrivere il linguaggio come se fosse essenzialmente una entità autonoma fatta di dipendenze interne, o, in una parola, una struttura».

[50] Uso il termine linguistica come sinonimo di linguistica strutturale.

[51] Notare: questo schema classificatorio, in gran parte tratto da P. Ricoeur, è condiviso da alcuni rappresentanti dello strutturalismo. Ad es., nel dizionario di semiotica curato da A.J. Greimas e J. Courtés (1979, p. 171), alla voce *ermeneutica* si legge: «essa mette in gioco il rapporto tra testo e referente, chiamando in causa le dimensioni extra-linguistiche dei discorsi, le condizioni della loro produzione e quelle della loro lettura. A differenza dell'approccio semiotico, per il quale – ad esempio – l'enunciazione può essere ricostruita secondo un simulacro logico-semantico elaborato a partire dal solo testo, l'ermeneutica fa intervenire il contesto socio-storico, compreso quello della comprensione attuale».

scio quindi al lettore il compito di fare le sue inferenze, ovviamente includendo – voce per voce – il termine *testo*: ad es. "referenza interna del testo", "testo senza soggetto", ecc. Quanto alla voce *contesto*, coerentemente con quanto affermato in precedenza, non è inclusa nella Tab. 1 perché – colonna per colonna – la considero come risultato delle interazioni tra tutte le altre voci in elenco. Tuttavia, volendo distinguere, direi che nel caso dello strutturalismo il contesto è esclusivamente linguistico (co-testo), mentre, nel caso dell'ermeneutica, è costituito anche da eventi extra-linguistici[52].

Nel paragrafo precedente (1.4) ho affermato che i *momenti* dell'analisi e quelli dell'interpretazione non si succedono linearmente. Aggiungo: non solo in base a un criterio temporale, cioè per il fatto che gli uni possono essere "prima" degli altri o viceversa. Usando una metafora, direi che le operazioni dell'analisi consentono di rappresentare i dati testuali su superficie piana (in ipotesi, quella di un foglio quadrettato), mentre l'interpretazione interviene su dimensioni che intersecano la superficie piana e che danno "profondità" al lavoro di analisi, vuoi perché connettono punti e linee con eventi extralinguistici, vuoi perché ne costruiscono il senso e tentano di darne una spiegazione.

In ogni caso, quello che avviene nelle concrete pratiche di ricerca, cioè – per così dire – la semiodiversità delle operazioni di analisi e di interpretazione, è ben più vario delle tassonomie che ne vengono proposte. Basti, ad esempio, esplorare la galassia dei software per l'analisi dei testi. Molto spesso, la struttura delle cosiddette *directory* in cui questi sono rubricati e/o presentati non è molto dissimile dalla curiosa enciclopedia cinese citata da Borges (vedi paragrafo 1.1).

Quindi, volendo parlare di software, mi limiterò a proporre alcune considerazioni sulle *logiche di funzionamento* di due tipi di strumenti abbastanza noti al pubblico dei ricercatori italiani: due "famiglie" che, in differenti modi, supportano le attività di analisi dei testi, siano questi trascrizioni di interviste, di focus group, di sedute psicoterapeutiche, di risposte a domande aperte, oppure articoli di giornale, libri, storie di vita ecc.

Le logiche di funzionamento di questi due tipi di "famiglie", che indico con due lettere dell'alfabeto ("A" e "B") si differenziano per il fatto che ciascuna di esse è in sintonia con un modo di costruire il percorso della ricerca:

- l'uno ("A") caratterizzato dal fatto che ogni passo dell'attività analitica è mediato da un lavoro di interpretazione e di elaborazione teorica;

[52] Il che, dal punto di vista dello strutturalismo, significa mettere in gioco relazioni tra vari sistemi semiotici (linguaggio scritto, linguaggio dei gesti, modelli culturali ecc.).

- l'altro ("B") caratterizzato dal fatto che l'attività di interpretazione è prevalentemente concentrata sugli output dell'analisi, in gran parte automatica e mediata dall'applicazione di tecniche statistiche.

Nel primo caso ("A"), il testo va letto e riletto, dall'inizio alla fine; nel secondo ("B"), la lettura del testo non è un prerequisito necessario.

Nella famiglia "A" includo programmi *theory-driven* quali *Atlas.ti*, *Nud.Ist*, *Aquad*, *Kwalitan* e *HyperResearch*; nella famiglia "B" includo programmi *word-driven* quali *Spad-T*, *Sphinx*, *Taltac*, *Alceste* e *T-LAB*. Evidentemente l'elenco non è esaustivo. Né mi interessa approfondire somiglianze e differenze interne a ciascuna famiglia. Aggiungo che, per semplificare il discorso, farò riferimento a due "prototipi"[53]: *Atlas.ti* e *T-LAB*.

Per molti versi, i "principi" che ispirano la metodologia di questi due software sono quelli indicati all'inizio del manuale[54] *Atlas.ti*: Visualization, Integration, Serendipity[55], Exploration. Tuttavia, nei due casi e in funzioni di specifiche operazioni, ognuno dei quattro termini assume significati molto diversi tra loro[56].

Le notevoli differenze tra i due sono ben evidenti fin dall'inizio, cioè a partire dal momento in cui si intraprende un nuovo progetto di lavoro:

- nel caso di *Atlas.ti*, il progetto è un container aperto – denominato *Hermeneutic Unit*[57] – che, nella fase iniziale, include solo dei link ai dati testuali[58]. Come a dire: «sono pronto, ma attendo istruzioni per organizzare i mio database»;
- diversamente, nel caso di *T-LAB*, l'importazione di un nuovo corpus produce un database già organizzato, in cui le varie unità di analisi sono già

[53] Uso questo termine nel senso "figurato" riportato nei dizionari della lingua italiana: "chi riassume in sé, chi manifesta in più alto grado i caratteri tipici di una categoria". Sul giudizio di valore ("chi manifesta in più alto grado"), propongo di sospendere il giudizio.

[54] Il link per il manuale Atlas.ti è: http://www.atlasti.de/demo.shtml. Quello T-LAB è: http://www.tlab.it/it/download.asp.

[55] Definizione tratta dal manuale Atlas.ti: «To find something without having searched for it. The term *serendipity* stands for an intuitive approach to data. A common operation making use of the serendipity is *browsing*».

[56] Ad esempio, la nozione di integrazione, che in Atlas.ti è riferita alla possibilità di «Not to loose the feeling for the whole when working on details», in T-LAB concerne sia l'architettura del database che le relazioni tra i vari strumenti di analisi.

[57] Il fatto che si usi questa dizione non significa che l'uso di Atlas.ti sia sempre associato a un lavoro di tipo ermeneutico. Ad esempio, quando le categorie di analisi sono precostituite in base a una teoria, il processo di codifica è del tipo *top-down* (dall'alto verso il basso).

[58] I *primary documents* di Atlas.ti possono essere costituiti da tre tipi di dati: testuali, video e audio.

definite e in gran parte classificate[59]. Quindi, fin dall'inizio, l'utilizzatore può "vedere" alcuni risultati.

In effetti, nel caso di *Atlas.ti*, le operazioni di analisi (vedi i punti "a", "b" e "c" all'inizio di questo paragrafo) sono tutte a carico dell'utilizzatore o, quanto meno, richiedono una sua puntuale regia. Da questo punto di vista, la sua logica di funzionamento è una trasposizione del "vecchio" e nobile metodo carta e matita: fin quando non si procede alla quadrettatura del foglio e al riempimento delle varie celle, il database è vuoto. In altre parole: senza *code*, niente *retrieve*.

Ancora:

- nel caso di *Atlas.ti*, i segmenti di testo (*quotation*) che corrispondono alle unità di analisi, oltre ad avere lunghezza variabile (dalla singola parola a un intero documento), possono essere continuamente ri-definiti e ri-classificati;
- diversamente, nel caso di *T-LAB*, le unità di analisi sono solo di tre tipi: sottoinsiemi del corpus, contesti elementari (CE) e unità lessicali (UL)[60]. Nessuna di queste, dopo l'importazione, può essere ri-definita (vedi i criteri della *scomposizione*); mentre due di esse (CE e UL) possono essere ri-classificate.

In entrambi i software, l'operazione analitica del classificare è di particolare importanza, ma i rispettivi "oggetti" (cioè le unità di analisi) e i rispettivi criteri sono affatto diversi.

- nella logica di *Atlas.ti*, ad ogni unità di analisi può essere attribuito uno o più *code*, ciascuno dei quali può essere anche collegato a un *super-code*; inoltre, il ricercatore dispone di *memo* in cui può annotare il significato dei codici utilizzati, informazioni sugli autori dei testi e/o sui contesti in cui questi sono stati prodotti, riflessioni teoriche, decisioni metodologiche ecc.;
- nella logica di *T-LAB*, l'opzione della classificazione manuale è riservata alla codifica delle unità di contesto iniziali nella fase di preparazione del corpus[61], quindi, in corso d'opera, alle unità lessicali[62]. Quanto poi alla classificazione dei contesti elementari, è il risultato di alcune operazioni che passano attraverso l'applicazione di tecniche statistiche.

[59] Le modalità operative per la "preparazione" del corpus sono illustrate nel manuale T-LAB. Quanto all'organizzazione del database e ai possibili interventi dell'utilizzatore, rinvio alla trattazione specifica contenuta nel paragrafo 2.2.2.

[60] I particolari di questi aspetti saranno chiariti nel prossimo capitolo.

[61] Eventuale uso di variabili (fino a 10) e rispettive modalità (max. 150 per ogni variabile).

[62] Ciascuna unità lessicale può essere codificata e ri-codificata "n" volte, ma sempre con una sola label; cioè, ogni label sostituisce la precedente.

Ma veniamo al *valore aggiunto* che i due software, attraverso i loro processi di elaborazione e i rispettivi output, offrono al ricercatore e alla sua capacità interpretativa.

- in *Atlas.ti* i risultati (quelli del "suo" lavoro) sono di tre tipi: a) estrazioni di unità di analisi secondo criteri logici, semantici e di contiguità spaziale; b) *network view* con relazioni tra i vari componenti selezionati (codici, memo, ecc.); c) tabelle di contingenza, con le occorrenze dei vari codici, che possono essere esportate e trattate mediante vari software di tipo statistico;
- in *T-LAB*, fatta eccezione per alcune funzioni di consultazione, i risultati sono sempre tabelle e/o grafici che, in modo interattivo, mostrano output di analisi statistiche, vuoi orientate a ricostruire i significati contestuali delle "parole", vuoi orientate ad esplorare somiglianze e differenze tra le varie unità di analisi, sia all'interno del corpus che all'interno dei suoi sottoinsiemi.

Per molti versi, si può affermare che *Atlas.ti* immagazzina le informazioni prodotte dall'utilizzatore e gliele restituisce in modo ordinato; mentre *T-LAB* genera "nuovi dati" che vanno interpretati.

Nel caso di T-LAB, quindi, il "vero" lavoro interpretativo comincia a valle delle elaborazioni statistiche, cioè pochi minuti o – a seconda dei casi[63] – poche ore dopo l'importazione del corpus. I suoi output sono delle **icone** (i grafici) o possono diventarlo (le tabelle); il che significa che l'*interpretazione iconologica* deve essere sempre preceduta da una attenta *analisi iconografica*.[64] In altri termini, prima di interpretare il senso dei diagrammi, dei cluster o dei fattori, il ricercatore è chiamato a decifrare il significato statistico di cose quali "rami del dendrogramma", "polarità fattoriali", "indici di associazione" e quant'altro. Altrimenti le icone diventano *immaginette* più o meno sacre, atte a evocare suggestioni o reverenze, più nella logica delle emozioni che in quella del pensiero.

Notare: l'analisi e l'interpretazione degli output T-LAB consente di elaborare delle categorie che, volendo, possono essere re-immesse nel circolo ermeneutico o essere utilizzate in ulteriori analisi di contenuto, anche del tipo carta e matita.

In un paper pubblicato su Internet[65] ho dichiarato che le teorie "implementate" in T-LAB sono mutuate dalla linguistica e dalla semiotica. Non poteva

[63] Pochi minuti se si usano le "impostazioni automatiche", qualche ora se si decide di intervenire sul dizionario del corpus e/o sulla selezione delle parole-chiave ("impostazioni personalizzate").

[64] Mi riferisco alla distinzione proposta da E. Panofsky (1955, tr. it., pp. 29-57) in un suo noto testo sul significato delle arti visive.

[65] *La logica di un testoscopio* (F. Lancia, 2001).

essere diversamente: un software analizza dati e produce icone, ma non interpreta. I limiti del *suo* mondo, per riprendere la nota affermazione di L. Wittgenstein, sono definiti dalla logica binaria dei linguaggi informatici, i quali sono un perfetto esempio di chiusura dell'universo dei segni. Il fatto poi che un ricercatore possa decidere di utilizzare un software (o più di uno) per analizzare e per interpretare alcuni materiali testuali, dipende dai suoi interessi, dai suoi obiettivi e dalle sue competenze. Uno strumento è solo uno strumento: se serve lo si usa, altrimenti se ne fa a meno.

Quanto all'uso degli strumenti che chiamiamo teorie, la questione è più complessa; non fosse altro perché tanto i criteri di analisi quanto i criteri di interpretazione vanno fondati su qualche teoria, sia essa "abbozzata" o molto articolata. Ad esempio, nella ricerca citata nel paragrafo 1.3, F. Fornari (1977) ha utilizzato le teorie semiotiche per analizzare e la teoria coinemica per interpretare. Lui, come si sa, era uno psicoanalista; e la teoria coinemica propone una specifica articolazione tra codici affettivi e funzionamento del sistema inconscio. Altri ricercatori hanno lavorato e lavorano con altri criteri: grounded theory, analisi emozionale del testo, analisi delle rappresentazioni sociali, analisi delle strutture narrative ecc. In ogni caso, vale la pena di ricordare che quando il ricercatore non dichiara esplicitamente i suoi criteri di analisi e di interpretazione, i testi che lui costruisce sono strutturalmente orientati ad alimentare un tipo di cultura fondata sulla *riproduzione*[66].

Per restare nell'ambito della semiotica, ricordo la distinzione proposta da J.M. Lotman e B.A. Uspenskij (1973) tra due tipi di culture, la cui coesistenza si articola in vario modo nei nostri sistemi sociali, compresi quelli delle comunità scientifiche. La prima, denominata *cultura testualizzata*, «non tende a distinguere un particolare metalivello (le regole del suo costruirsi) né propende per le autodescrizioni» (tr. it., p. 70); proponendosi come *intraducibile* e trasformando i propri testi in stereotipi, essa si caratterizza per essere mitica, mono-linguistica e orientata all'autocomunicazione. La seconda, denominata *cultura grammaticalizzata*, introduce la possibilità di usare un metalinguaggio, anche per parlare di sé stessa; quindi è poli-linguistica, de-mitologizzante e aperta all'*altro*.

Evidentemente, la costruzione e l'uso dei testi scritti nella *grammatica della ricerca* non appartengono per definizione alla cultura del secondo tipo, neanche quando propongono sofisticate argomentazioni epistemologiche. Ad esempio, il fatto che in alcuni ambienti si tenda ad identificare la prassi ermeneutica con l'uso di alcuni tipi di software, oppure che si tenda ad escludere la possibilità di *comprendere* i testi anche con l'ausilio di trattamenti statistici, che tipo di cultura comunica?

[66] Cfr. B. Bourdieu e J.C. Passeron, 1970.

Da qualche parte ho sentito affermare che l'uso di software tipo T-LAB può trasformare l'analisi dei testi in una "scorciatoia", una sorta di *via brevis* contrapposta alla *via longa* che caratterizza ogni prassi ermeneutica. Stando che ogni ricercatore può scegliere la velocità con la quale percorrere le sue strade, la *via brevis* è solo quella che termina con gli output del programma. Quanto poi al ri-tornare al testo (o al contesto), sia per verificare che per incrementare le proprie interpretazioni, è una questione che, in generale, riguarda il modo in cui ciascuno articola le relazioni tra le teorie e le "credenze condivise" che, nell'ambito della sua comunità scientifica, regolano la costruzione e la legittimazione degli itinerari di ricerca[67].

[67] Secondo Y. Elkana (1981), le pratiche di ricerca comportano sempre un'interazione tra i "pensieri sul mondo" e i "pensieri sulla conoscenza" (*second-order thinking*). I primi fondano il *corpo del sapere* e si articolano in teorie e discipline scientifiche; i secondi fondano l'*immagine del sapere* e si articolano in credenze concernenti le fonti della conoscenza (tradizione, esperimenti, ecc.), gli obiettivi della scienza (comprensione, previsione, ecc.) e altri aspetti del contesto culturale.

# 2. La logica analitica di T-LAB

## 2.1. Introduzione

Quando siamo alla guida di un'autovettura le conoscenze per noi più rilevanti sono quelle concernenti il codice della strada, la topografia dei percorsi e le abitudini di comportamento nel traffico, cioè i tipi di conoscenze che in questo contesto funzionano come regole per le nostre azioni. Diversamente, quando usiamo un software le regole sono in gran parte dettate dalla logica interna del programma, cioè da una *architettura* che prevede ed orienta specifiche azioni atte a conseguire specifici scopi.

Evidentemente, come tutti i software, T-LAB si presta a molteplici usi, molti dei quali sono al di là dell'immaginazione di chi lo ha progettato; tuttavia la sua architettura prevede dei *percorsi logici* che, per quanto aperti e reversibili, sono anche regolati da semafori. In altri termini, alcune cose si possono fare e altre no, alcune cose vanno fatte prima e altre dopo.

Un semplice schema (vedi Fig. 1) può esserci di aiuto per iniziare a capire "come funziona".

Le *componenti* principali sono tre: l'**interfaccia utente**, costituita dal cosiddetto menu e da tutte le funzioni che restituiscono dati e/o informazioni (help contestuale, "videate" con grafici o tabelle, lettura e scrittura di file, ecc.); il **database**, costituito da tabelle che includono dizionari e archivi delle varie unità di analisi (parole, frasi, sottoinsiemi del corpus); gli **algoritmi**, siano essi orientati a specifici trattamenti linguistici (vedi lemmatizzazione), a predisporre tabelle di input per le analisi statistiche, a fare i calcoli o a determinare la forma degli output.

Tutte e tre queste componenti sono suscettibili di evoluzioni: l'interfaccia utente, ad esempio, ha già subito qualche cambiamento rilevante[1]; il database

[1] La prima versione di T-LAB, rilasciata nell'autunno 2001 aveva un menu completamente diverso da quello attuale (aprile 2004).

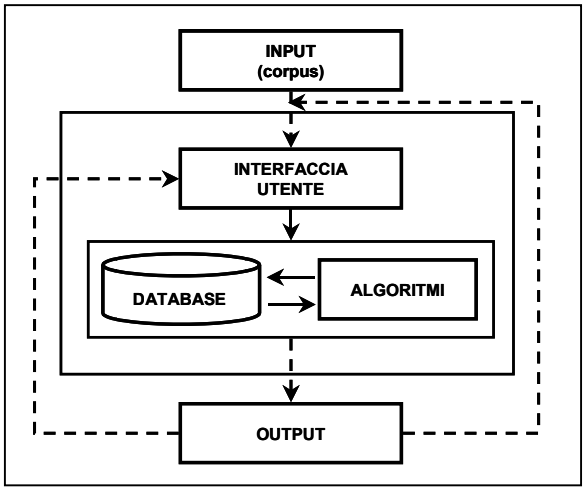



*Fig. 1 – T-LAB: diagramma di flusso*

può essere ulteriormente arricchito o trasformato; gli algoritmi possono essere modificati e incrementati[2]. In altri termini, l'architettura di T-LAB è aperta e flessibile, con la possibilità di realizzare vari tipi di prodotti; tuttavia il modello generale è quello illustrato in Fig. 1 che, a ben vedere, può essere applicato a molti altri tipi di software.

Come si può osservare, due feedback consentono di tornare sui propri passi: per usare altri strumenti, per sperimentare ulteriori strategie di analisi, o per creare nuovi file input (corpus o sub-corpus); inoltre, ogni processo è costituito dalle interazioni tra database e algoritmi, vuoi per costruire tabelle, vuoi per analizzarle, vuoi per produrre output.

Nelle pagine seguenti di questo capitolo saranno descritti i vari processi che presiedono al funzionamento di T-LAB come "macchina analitica". Le modalità operative del suo uso, corredate da puntuali riferimenti alle varie opzioni del *menu*, saranno trattate nella seconda parte del libro.

## 2.2. Analisi linguistica: scomporre e classificare

Durante la fase di importazione del corpus T-LAB realizza il suo primo compito analitico: la **scomposizione** del corpus in **unità di analisi**[3].

[2] Ad esempio, in una prossima versione di T-LAB sarà implementato un metodo di multidimensional scaling per rappresentare le relazioni di associazione sia tra parole che tra nuclei tematici.

[3] Sulla nozione di "unità di analisi" vedi il paragrafo 1.5.

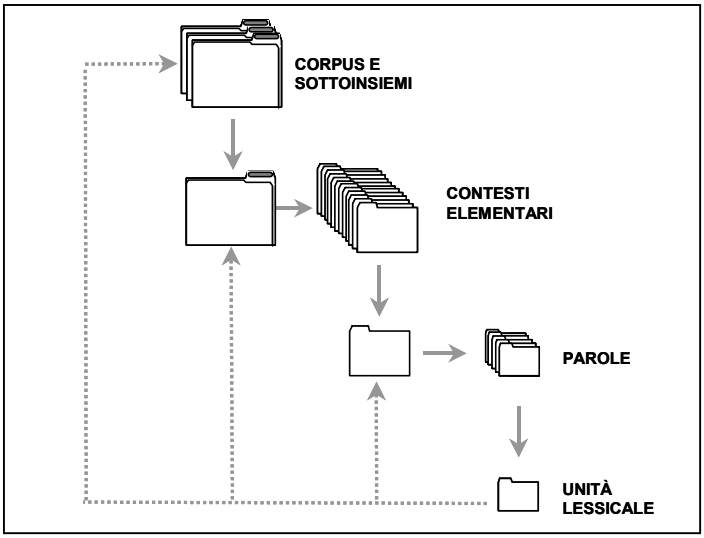



*Fig. 2 – T-LAB: scomposizione in unità di analisi*

Questo processo, anche rispettando eventuali criteri fissati dall'utilizzatore, prevede vari step (Fig. 2). Il primo, opzionale, esita nel riconoscimento e nella classificazione delle **unità di contesto** (UC) di primo livello: i **sottoinsiemi** del corpus, cioè le partizioni determinate dall'eventuale uso di variabili[4]. Ad esempio, se il corpus è costituito da alcune interviste, ciascuna di esse potrà essere classificata mediante il riferimento alle caratteristiche dell'intervistato o dell'intervistatore (sesso, professione, ecc.), così come anche a quelle del contesto di provenienza (luoghi, tempi, ecc.). Ovviamente ciascuna di queste caratteristiche si declinerà in due o più modalità (ad es. variabile "sesso"; modalità "maschio" o "femmina")[5], cioè in label definite dall'utilizzatore interessato ad ancorare le analisi a una o a più dimensioni del *contesto extralinguistico*.

Lo step successivo comporta la segmentazione del corpus in UC di secondo livello: i **contesti elementari**, cioè frammenti di testo di lunghezza comparabile[6], intervallati da punteggiatura e corrispondenti a una o più frasi. Nel database ciascuno di essi è catalogato con un numero di serie e con una label che

[4] Quando il corpus importato è costituito da un *unico* testo (ad es un libro o una intervista; oppure da più interviste "assemblate" senza l'uso di variabili classificatorie) non si danno sottoinsiemi.

[5] Al momento T-LAB consente che ogni sottoinsieme del corpus sia classificato mediante max 10 variabili; ciascuna delle quali può articolarsi fino a max 150 modalità.

[6] Attualmente, in T-LAB, la lunghezza massima di ogni contesto elementare è fissata a 400 caratteri.

indica la sua eventuale appartenenza a uno o più sottoinsiemi. Ad esempio, se una intervista è classificata per regione (es. Marche) e per fascia di reddito (es. alta), tutti i contesti elementari che la compongono apparterranno a due sottoinsiemi (es. regione_Marche; fascia_alta).

Il terzo e ultimo step è costituito dal riconoscimento delle *parole* e dalla loro classificazione come **unità lessicali** (UL). Più precisamente, entro la logica di T-LAB, le parole sono sequenze di caratteri[7] separate da spazi o da segni di punteggiatura; e, per ciascuna di esse, nel database sono registrate le seguenti caratteristiche:

a) una label corrispondente alla sua forma grafica, al lemma[8] di appartenenza o ad altra categoria definita dall'utilizzatore (classe semantica, categoria di contenuto ecc.);
b) un numero di serie corrispondente alla sua posizione nella sequenza del discorso;
c) il contesto elementare di appartenenza, con tutte le specificazioni di quest'ultimo (eventuali sottoinsiemi di riferimento e relative variabili);
d) una categoria binaria che indica la sua presenza/assenza nella lista della parole-chiave selezionate per le analisi.

Una volta effettuata l'importazione, tutte le caratteristiche concernenti le *unità di contesto* (UC: sottoinsiemi e contesti elementari) restano invariate[9]. Al contrario, le caratteristiche "a" e "d" delle *unità lessicali* (UL: parole e loro label) possono subire interventi ripetuti da parte dell'utilizzatore.

Prima di entrare nella logica di questi ultimi è opportuno soffermarsi a considerare le implicazioni dei due processi finora menzionati: la *scomposizione* del corpus in unità di analisi (UC e UL) e loro *classificazione*. Da un punto di vista formale, questi processi determinano il modo in cui il corpus è rappresentato nel database, cioè come un insieme di oggetti (le unità di analisi) a ciascuno dei quali corrisponde una o più caratteristiche (le label); diversamente, da un punto di vista sostanziale, la loro funzione è quella di *ancorare le analisi statistiche a una teoria del significato*.

[7] La lunghezza massima accettata da T-LAB è pari a 50 caratteri.

[8] Per *forma grafica* si intende la singola parola, così come è scritta (ad es. *"correva"*). Per *lemma* si intende la sua "forma canonica" (es. *"correre"*), così come è riportata nei dizionari linguistici.

[9] Nel momento in cui l'architettura del programma dovesse prevedere la possibilità di classificare manualmente le singole frasi (i contesti elementari) questa affermazione non sarebbe più valida. Inoltre: dopo alcuni tipi di analisi, T-LAB archivia i cluster di contesti elementari come ulteriori sottoinsiemi del corpus.

### 2.2.1. Unità di analisi e teoria del significato

Per chiarire le implicazioni della precedente affermazione, ricordo che nell'ambito della linguistica strutturale[10], l'attività analitica della scomposizione si applica ai **rapporti sintagmatici**, cioè alle *combinazioni* delle entità linguistiche entro i contesti: l'una accanto all'altra nella successione del discorso. Mentre, l'attività analitica della classificazione si applica ai **rapporti paradigmatici**, cioè alle *associazioni* delle entità linguistiche entro famiglie[11] morfologiche e semantiche.

Le due attività analitiche, in interazione tra loro, fondano la possibilità di fare inferenze sul significato delle entità linguistiche, non prese "in sé", ma nelle loro reciproche relazioni; cioè inferenze sul "valore" che le varie entità assumono in determinati contesti. «Il valore di un qualunque termine – scrive F. de Saussure (1916, tr. it., p. 141) – è determinato da ciò che lo circonda», cioè dai significanti e dai significati degli altri termini con i quali esso è in relazione; quindi, continua l'autore,

> Quando io affermo semplicemente che una parola significa qualche cosa, quando io mi attengo all'associazione dell'immagine acustica (significante) col concetto (significato), faccio un'operazione che può in una certa misura essere esatta e dare un'idea della realtà; ma in nessun caso io esprimo il fatto linguistico nella sua essenza e nella sua ampiezza» (ibid., p. 142, le aggiunte tra parentesi sono mie).

Da un punto di vista metodologico, il dispositivo analitico dei due assi (sintagma e paradigma), oltre a consentire di interpretare il significato degli enunciati, consente di spiegare come essi vengono prodotti e costruiti. Infatti, secondo R. Jakobson (1966, tr. it., p. 24), ogni atto linguistico implica la *selezione* di certe entità linguistiche (entro paradigmi) e la loro *combinazione* in unità linguistiche maggiormente complesse (entro sintagmi). Ad esempio, la produzione di una semplice frase come «Stasera vado al cinema», oltre a "combinare" quattro parole in un certo ordine (sintagma), implica che ciascuna di esse sia scelta all'interno di un repertorio di possibili alternative (paradigma)[12].

[10] Cfr. R. Barthes, 1964; F. de Saussure, 1916; L. Hjelmslev, 1943; R. Jakobson, 1963.

[11] "Associazioni" e "famiglia" sono termini usati da F. de Saussure (1916). Secondo questo autore, «Mentre un sintagma richiama immediatamente l'idea di un ordine di successione e di un numero determinato di elementi, i termini di una famiglia associativa non si presentano né in numero definito né in ordine determinato» (tr. it., p. 152). La famiglia (paradigma), quindi, può essere costituita dall'insieme delle flessioni di un lemma, ma anche da un insieme di lemmi con significato simile.

[12] Ad esempio, alternative di "stasera" sono "questa sera", "in serata" ecc. Evidentemente, l'analisi dell'*evento* comunicativo – chi, in quale contesto, a chi e perché sta dicendo «Stasera vado al cinema» – non rientra nei compiti dell'analisi linguistica (vedi paragrafo 1.5).

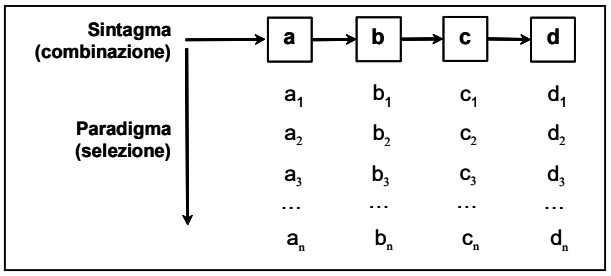



*Fig. 3 – Sintagma e Paradigma*

Di fatto, tanto la selezione quanto la combinazione sono vincolate al sistema linguistico e alla **cultura** dei locutori. Non a caso, quando F. de Saussure ha provato a spiegare la nozione di "valore" ha chiamato in causa le differenze culturali e ha sottolineato l'impossibilità di tradurre termine con termine. Un suo classico esempio è il seguente:

> Il francese *mouton* può avere la stessa significazione dell'inglese *sheep*, ma non lo stesso valore, e ciò per più ragioni, in particolare perché parlando di un pezzo di carne cucinato e servito a tavola, l'inglese dice *mutton* e non *sheep*. La differenza di valore tra *sheep* e *mouton* dipende dal fatto che il primo ha accanto a sé un secondo termine, ciò che non è il caso della parola francese (1916, tr. it. pp. 140-141).

Un esempio analogo è quello, proposto da B. Whorf (1956), degli eschimesi che hanno molti termini per indicare la realtà che noi, con un unico termine, chiamiamo "neve". Come è noto, a questo autore e al suo maestro E. Sapir (1949), si fa risalire l'ipotesi del cosiddetto relativismo linguistico, la quale è fondata sull'affermazione che le differenze tra le culture sono largamente determinate dalle differenze tra i rispettivi linguaggi[13].

> L'essenza della questione – scrive E. Sapir (1949, tr. it., p. 59) – è che il "mondo reale" viene costruito, in gran parte inconsciamente, sulle abitudini linguistiche del gruppo. Non esistono due lingue che siano sufficientemente simili da essere considerate come rappresentanti della stessa realtà sociale. I mondi in cui vivono differenti società, sono mondi distinti, non sono semplicemente lo stesso mondo con etichette differenti… Noi vediamo e udiamo e facciamo altre esperienze in un dato modo, in gran parte perché le abitudini linguistiche della nostra comunità ci predispongono a certe scelte di interpretazione.

Da un punto di vista semiotico (R. Barthes, 1964), il modello paradigma/sintagma (selezione/combinazione) può essere usato anche per analizzare

[13] Ovviamente vale anche l'ipotesi inversa: le differenze tra culture determinano i differenti modi in cui usiamo le parole e il linguaggio. Inoltre, il linguaggio *è* cultura.

le differenze tra modelli culturali. Basti pensare alle diverse regole che presiedono alla scelta e alla combinazione delle pietanze nella cucina italiana da un lato e in quella cinese dall'altra; oppure alle diverse regole che presiedono alla scelta e alla combinazione dei capi di abbigliamento nella cultura dei manager e in quella dei "ragazzi di strada". In effetti, se le culture corrispondono ai diversi modi in cui i gruppi sociali costruiscono le loro rappresentazioni del mondo, costruiscono le loro esperienze, organizzano le loro relazioni e attribuiscono valore alle cose, le differenze tra culture non riguardano solo le "tribù" o le nazioni. Differenti culture sono anche quelle elaborate e proposte dai gruppi di adolescenti, dalle aziende, dalle professioni, dai gruppi politici e da altri gruppi sociali.

In ogni caso, tutte le culture costruiscono diverse rappresentazioni dei contesti, cioè diversi "mondi" abitati da oggetti e individui, con specifiche proprietà e in specifiche relazioni reciproche. Mondi che vengono parlati e che, quindi, si traducono in testi che possiamo analizzare[14].

A rigor di termini, nell'analisi dei testi, ogni volta abbiamo a che fare con un contesto diverso: quello definito dal corpus che abbiamo deciso di analizzare, più esattamente – per tornare alla logica di T-LAB – quello definito dalle relazioni tra unità di contesto (UC) e unità lessicali (UL).

Per capire come la struttura di queste relazioni consente di ricostruire il significato contestuale delle parole o i "mondi" dei testi, dobbiamo considerare tutte le fasi del processo analitico: la scomposizione del corpus in unità di analisi (UC e UL), la loro classificazione, la costruzione delle tabelle, l'uso delle tecniche statistiche e l'interpretazione degli output.

La prima fase, come si è già detto, riguarda l'asse del sintagma, la seconda quella del paradigma, le altre fasi riguardano le relazioni tra entrambe le dimensioni (sintagma e paradigma).

In linea generale, l'ancoraggio alle relazioni sintagmatiche è garantito dalla partizione del corpus in unità di contesto (UC); tuttavia va considerato che queste ultime sono di due tipi: contesti elementari e sottoinsiemi. Ora, mentre i primi sono in ogni caso sintagmi, cioè frasi o enunciati di lunghezza comparabile, i secondi possono corrispondere a realtà diverse tra loro. Infatti, quando i sottoinsiemi sono determinati dall'uso di un solo criterio di partizione, e questo rispetta la suddivisione "naturale" del corpus[15], ogni sottoinsieme conserva il suo statuto di "vero" sintagma, cioè continua ad essere una "catena" di parole costruita dai locutori; diversamente, quando i criteri di partizione sono due o più di due, la realtà sintagmatica di ogni sottoinsieme è determinata dai

[14] Secondo questa ipotesi, anche i testi prodotti da una singola persona propongono uno o più "mondi".

[15] Cioè da una sola variabile le cui modalità corrispondono a realtà linguistiche quali "un" capitolo di un libro, "una" intervista o "un" articolo di quotidiano.

criteri di classificazione usati dall'analista. In quest'ultimo caso, cioè, la costruzione della "catena" discorsiva che caratterizza ogni sottoinsieme è il risultato di un processo interpretativo tramite il quale l'analista ha deciso di considerare come sintagmi delle sommatorie di contesti elementari supposti simili tra loro, vuoi perché prodotti dagli stessi tipi di locutori, vuoi perché prodotti negli stessi tipi di contesti, vuoi perché accomunati dal riferimento agli stessi temi.

Per chiarire questo punto, supponiamo che un corpus sia costituito da cinquanta interviste ad altrettanti soggetti, e che ciascuna di esse sia codificata tramite l'uso di tre variabili: una che prevede il riferimento al nome degli intervistati (50 modalità), una al loro sesso (2 modalità), l'altra alla loro organizzazione di appartenenza (ad es. 5 modalità). In questo modo otteniamo tre tipi di sottoinsiemi, di cui solo il primo, quello che prevede cinquanta unità di contesto, rispecchia la successione "naturale" degli enunciati entro le varie interviste. Diversamente il sottoinsieme dei "maschi" o quello degli appartenenti all'organizzazione "X" risultano costituiti da un certo numero di interviste messe insieme dall'analista allo scopo di esplorare somiglianze e differenze riconducibili a fattori del contesto extralinguistico. In ogni caso, tutti e tre i tipi di sottoinsiemi sono "combinazioni" di contesti elementari e di unità lessicali; quindi, da un punto di vista metodologico, le loro diversità non inficia la logica del processo analitico.

### *2.2.2. Parole, label e dizionari*

Consideriamo ora il particolare processo di classificazione che esita nell'attribuzione di label alle unità lessicali, cioè il processo che – entro T-LAB – chiama in causa l'asse del paradigma. La Fig. 4 ne illustra le vari fasi. Al primo step viene generato il *vocabolario* del corpus, cioè una tabella con tutti i tipi di "parole" in esso presenti e le rispettive occorrenze. Successivamente, se viene utilizzata l'opzione della lemmatizzazione automatica[16], entrano in opera i *dizionari* T-LAB, cioè tabelle in cui ogni parola è messa in corrispondenza con uno o più lemmi della lingua selezionata[17].

[16] Come è noto, il processo della lemmatizzazione automatica, anche nei software specificamente orientati a questo scopo, non è esente da errori. Quindi questo aspetto dell'architettura T-LAB è, in modo particolare, suscettibile di essere discusso e perfezionato.

[17] Ad esempio, la versione attuale del dizionario italiano include 170.944 forme, corrispondenti a poco meno di 21.000 lemmi, selezionati tramite l'uso di lessici di frequenza.

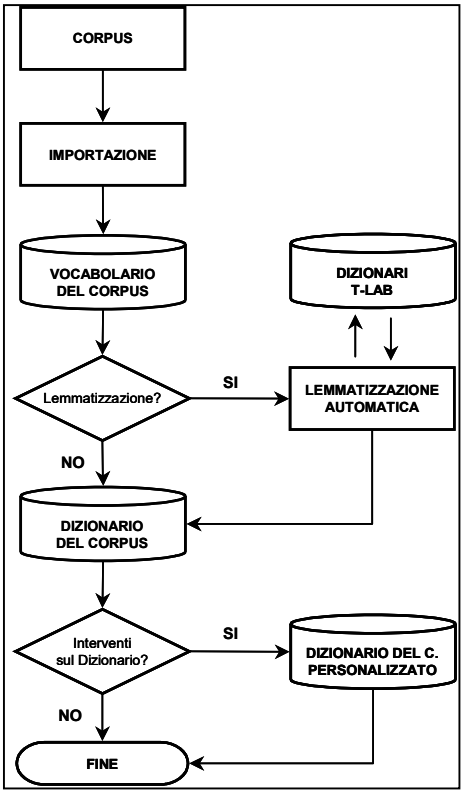



*Fig. 4 – Costruzione del dizionario del corpus*

Come risultato, viene prodotto il *dizionario del corpus*, cioè una tabella in cui ad ogni parola del corpus viene attribuita una label che, a seconda dei casi, corrisponde alla forma grafica della parola originaria o a un lemma[18].

Più precisamente, i casi in cui nel dizionario del corpus la label attribuita non è un lemma sono i seguenti:

a) la parola risulta "sconosciuta", cioè non inclusa nei dizionari T-LAB;
b) la parola risulta essere un caso di omografia rilevante (es. "costa");
c) l'uso del lemma (es. "accidente") comporta una perdita significativa della specificità semantica della parola (es. "accidenti")[19].

[18] Nei casi in cui non viene utilizzata la lemmatizzazione automatica, nel dizionario del corpus tutte le label replicano le forme originarie delle parole. Per ulteriori approfondimenti, si veda l'Appendice 1.

[19] L'applicazione delle categorie (b) e (c) può risultare discutibile e, in qualche caso, inesatta; tuttavia l'utilizzatore può fare le sue verifiche e decidere di usare le label che ritiene più adeguate.

In ogni caso, e in qualunque fase del processo analitico, il dizionario del corpus può essere modificato e personalizzato, anche tramite l'importazione di dizionari esterni.

Tutte queste operazioni hanno a che fare con l'attribuzione di label alle unità lessicali che, normalmente, chiamiamo **parole**. Queste, intese come *parole nella scrittura*[20], sono le unità minime per i conteggi e per le analisi statistiche; ma il loro statuto scientifico non è privo di ambiguità. In effetti, nella linguistica strutturale la nozione di parola – intesa come unità di analisi – è spesso evitata a causa del fatto che non può essere considerata un segno, cioè un'entità a due facce (significante/significato).

Secondo L. Hjelmslev (1943, tr. it., p. 48), a causa delle asimmetrie tra il piano dell'espressione (o significante) e quello del contenuto (o significato), le "parole" possono essere analizzate in parti che, a loro volta, sono portatrici di significato: radici, elementi di derivazione, elementi inflessionali. Ad esempio, la parola *"scriviamo"* è costituita da due elementi: il primo ("*scriv"*-) designa un certo tipo di azione, il secondo (-"*iamo"*) designa il soggetto dell'azione stessa. Per indicare queste unità minime dotate di significato, i linguisti[21] usano il termine *morfema*[22] che, in quanto segno, è un'unità a due facce: una faccia significata, costituita dal suo senso o valore, e una faccia significante, costituita dai fonemi che la compongono.

Evidentemente, il morfema è una unità di analisi difficilmente trattabile tramite software come T-LAB; esso, infatti, può essere una semplice parola ("*oggi"*), un radicale ("*parl"*-iamo), un prefisso ("*auto"*-mobile), un suffisso (amministra-"*tore"*) o una desinenza (compr-"*erà"*). Diversamente, la nozione di **lessia**, così come è stata proposta da B. Pottier (1974, pp. 265-269), permette di ricomporre l'unitarietà della parola[23] e, per molti versi, ne ampia i confini. Secondo questo autore, la lessia è una unità lessicale memorizzata determinata dalla forma del contenuto e che, sul piano dell'espressione (significante), corrisponde a unità di dimensione variabile. I suoi tipi fondamentali sono tre: *semplice*, corrispondente alla "parola" nel senso tradizionale del termine (es. "*cavallo"*, "*mangiava"*); *composta*, costituita da due o più parole integrate in un'unica forma (es. "*biotecnologie"*, "*mangianastri"*); *complessa*,

[20] Cioè come stringhe di uno o più caratteri, separate da spazi o da punteggiatura.

[21] Cfr. J. Dubois et al. 1973, B. Pottier, 1974.

[22] In una accezione ormai consolidata, i morfemi vengono distinti in due tipi: "morfemi lessicali" (o lessemi) e "morfemi grammaticali" (o grammemi). Per indicare la stesse entità A. Martinet (1960) ha usato il termini "monemi lessicali" e "monemi grammaticali".

[23] Secondo B. Pottier (1992, p. 38) la parola è «une unité construite, intermédiaire entre le morphème, unité de construction, e la lexie, unité memorisée de fonctionnement» («una unità costruita, intermediaria tra il morfema, unità di contruzione, e la lessia, unità memorizzata di funzionamento»). Per evitare inutili tecnicismi, da adesso in poi userò spesso "parola" come sinonimo di "lessia".

costituita da una sequenza in via di lessicalizzazione[24] (es. "*a mio giudizio"*, "*complesso industriale"*, "*O.N.U."*).

Così definita, la lessia può trovare agevole corrispondenza nella nozione di "parola nella scrittura" propria dell'analisi statistica[25]; solo che i tipi complessi vanno preventivamente trasformati in stringhe unitarie (es. "*sangue_freddo"*, "*a_mio_giudizio"*). Questa operazione, denominata lessicalizzazione, in T-LAB può essere effettuata sia in modo automatico che "pilotato" dall'utente[26].

Nel dizionario del corpus, così come è mostrato in Fig. 5, le parole intese come lessie e risultanti dal processo di scomposizione sono riportate nella colonna *forma*; diversamente, le unità lessicali classificate mediante l'attribuzione di una label sono riportate nella colonna *lemma*, anche se – in molti casi – le label non sono propriamente lemmi.

Concretamente, usando la funzione *Personalizzazione del dizionario* (Fig. 5), l'attribuzione di label alle unità lessicali (UL) può declinarsi secondo due modalità[27]: *distinguere* e *assimilare*. Nel primo caso è orientata a risolvere casi di ambiguità o a salvaguardare la specificità semantica di alcune UL; nel secondo è orientata a creare classi di equivalenza, cioè a trattare due o più UL (lessie) come espressioni che si riferiscono a contenuti simili tra loro.

In particolare, nel caso delle operazioni del tipo *assimilare*, tramite l'opzione della lemmatizzazione[28] automatica, T-LAB propone l'uso di una *semantica a dizionario*[29]; quindi l'utilizzatore può restare entro questi confini, può decidere di conservare l'individualità di specifiche forme, può usare label che seguono la logica dell'iponimia-iperonimia (Fig. 6); ma può anche decidere di usare una *semantica a enciclopedia* [30]. In altri termini, il lavoro sui pa-

[24] «Processo linguistico per cui un sintagma o un raggruppamento di parole diventa un solo elemento lessicale o come tale si comporta» (G. L. Beccaria, 1994, p. 426).

[25] Fatta eccezione per le cosiddette parole "vuote" (articoli, preposizioni, ecc.). Queste, essendo prive di contenuto, non possono essere considerate lessie.

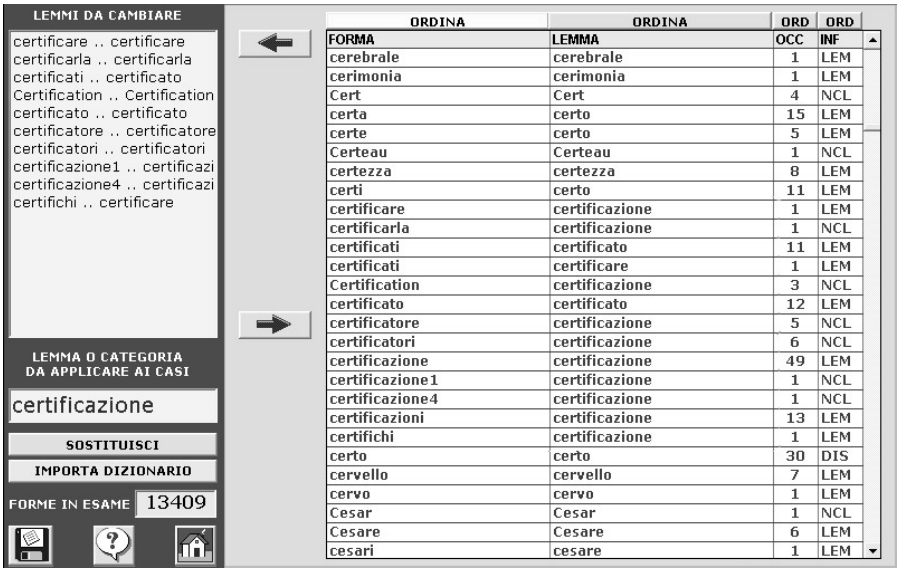
[26] Evidentemente, non si tratta di risolvere "tutti" i casi, ma solo quelli che – nel corpus in analisi – hanno particolare rilievo, vuoi per la loro tipicità, vuoi per il loro numero di occorrenze.

[27] Entrambe le operazioni possono essere effettuate manualmente o tramite l'importazione di dizionari esterni.

[28] L'uso di un lemma come label di varie forme (es. "parlare" per "parlava", "parliamo", "parleranno", ecc.) è un esempio di classe di equivalenza.

[29] Di fatto nelle diverse versioni di T-LAB sono implementati dizionari specifici per ogni lingua (attualmente: italiano, inglese, francese, spagnolo, latino).

[30] La distinzione tra semantica a dizionario e semantica a enciclopedia è stata proposta da U. Eco (1981, pp. 831-875). Secondo questo autore, la prima propone la logica tassonomica del cosiddetto albero di Porfirio, la seconda si caratterizza per una struttura a reticolo («È proprio la capacità che noi abbiamo di riorganizzare continuamente e contestualmente le unità di contenuto che fonda la possibilità del reticolo enciclopedico» ibid. p. 864). Nel primo caso (semantica a dizionario) il significato delle parole può essere rappresentato come un insieme finito di componenti; di-




| ORDINA FORMA | ORDINA LEMMA | ORD OCC | ORD INF |
|---|---|---|---|
| cerebrale | cerebrale | 1 | LEM |
| cerimonia | cerimonia | 1 | LEM |
| Cert | Cert | 4 | NCL |
| certa | certo | 15 | LEM |
| certe | certo | 5 | LEM |
| Certeau | Certeau | 1 | NCL |
| certezza | certezza | 8 | LEM |
| certi | certo | 11 | LEM |
| certificare | certificazione | 1 | LEM |
| certificarla | certificazione | 1 | NCL |
| certificati | certificato | 11 | LEM |
| certificati | certificare | 1 | LEM |
| Certification | certificazione | 3 | NCL |
| certificato | certificato | 12 | LEM |
| certificatore | certificazione | 5 | NCL |
| certificatori | certificazione | 6 | NCL |
| certificazione | certificazione | 49 | LEM |
| certificazione1 | certificazione | 1 | NCL |
| certificazione4 | certificazione | 1 | NCL |
| certificazioni | certificazione | 13 | LEM |
| certifichi | certificazione | 1 | LEM |
| certo | certo | 30 | DIS |
| cervello | cervello | 7 | LEM |
| cervo | cervo | 1 | LEM |
| Cesar | Cesar | 1 | NCL |
| Cesare | Cesare | 6 | LEM |
| cesari | cesare | 1 | LEM |

*Fig. 5 – Personalizzazione del dizionario*

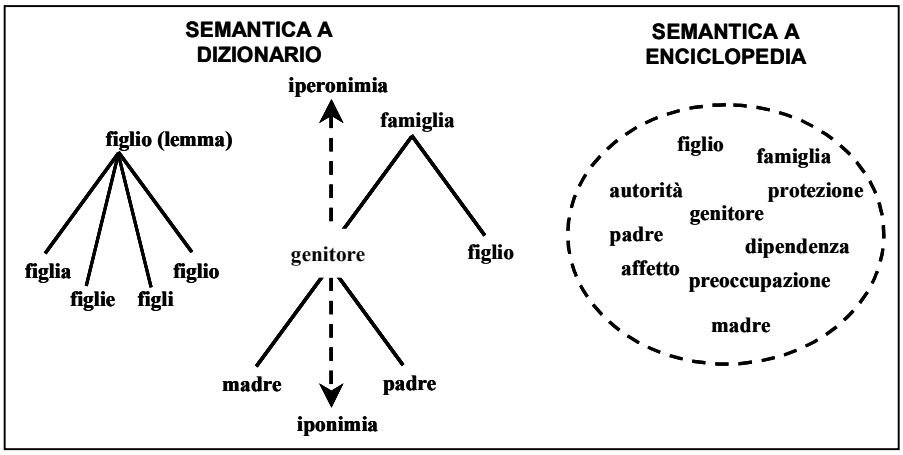



*Fig. 6 – Semantica a dizionario e semantica a enciclopedia*

radigmi e sulle "famiglie" di parole può essere effettuato in modo automatico, semi-automatico e "personalizzato".

versamente, nel secondo caso (semantica a encilopedia), le componenti del significato sono potenzialmente illimitate e dipendono dalle inferenze suggerite dai differenti contesti d'uso.

Proviamo a spiegarci con un esempio. Anche quando si sceglie l'opzione della lemmatizzazione automatica, T-LAB riconosce la parola "*figlia*" come caso di omografia e quindi la lascia invariata. Il ricercatore può accettare questa scelta oppure può decidere di attribuirgli la label "*figlio*" (lemma) o "*famiglia*" (iperonimo); ma può anche attribuirgli altre label (ad es. "*dipendenza*") derivanti da inferenze basate sui suoi modelli teorici o sui suoi contesti di riferimento.

### 2.2.3. Occorrenze e co-occorrenze

Personalmente sono dell'avviso che, nella fase iniziale del lavoro, convenga restare nella logica della *semantica a dizionario*[31]; infatti gli strumenti analitici implementati in T-LAB, quelli orientati a *stabilire relazioni* tra le unità di analisi, offrono più di un'opportunità per fare inferenze sul significato contestuale delle parole. Ciò a partire dal fatto che le reciproche relazioni tra le unità di analisi (UC e UL) possono essere rappresentate come matrici (o tabelle) i cui valori numerici indicano fenomeni di occorrenza e co-occorrenza.

*Tab. 1 – Occorrenze*

| | $uc_1$ | $uc_2$ | $uc_3$ | … | $uc_m$ |
|---|---|---|---|---|---|
| $ul_1$ | 21 | 3 | 5 | … | 12 |
| $ul_2$ | 11 | 1 | 34 | … | 50 |
| $ul_3$ | 1 | 32 | 4 | … | 8 |
| …. | …. | …. | …. | …. | …. |
| $ul_n$ | 61 | 9 | 16 | … | 24 |

*Tab. 2 – Co-occorrenze*

| | $ul_1$ | $ul_2$ | $ul_3$ | …. | $ul_n$ |
|---|---|---|---|---|---|
| $ul_1$ | | 5 | 13 | …. | 8 |
| $ul_2$ | 5 | | 2 | …. | 11 |
| $ul_3$ | 13 | 2 | | …. | 6 |
| …. | …. | …. | …. | …. | …. |
| $ul_n$ | 8 | 11 | 6 | …. | |

[31] Segnalo il fatto che nell'ambito dell'analisi di contenuto il termine *dictionary* è spesso usato per indicare "griglie" di categorie per la codifica delle unità di analisi. In particolare, nella tradizione inaugurata dal General Inquirer (http://www.wjh.harvard.edu/~inquirer/) un *dictionary* è costituito da una lista di label (o tag) indicanti categorie tematiche, a ciascuna delle quali viene fatto corrispondere un insieme di parole (ad es. "ANOMIA" = "anarchia", "caos", "caotico", "scoraggiato", "disillusione", ecc.). Questi tipi di dizionari, con T-LAB, possono essere importati, costruiti e utilizzati; tuttavia, poiché implicano una "riduzione" della varietà linguistica (ad es. uso della label "ANOMIA" al posto di decine di parole che ricadono nella sua "categoria"), comportano l'uso di strategie di analisi poco orientate all'esplorazione di "inattese" connessioni di senso.

Ad esempio, dato un insieme di unità lessicali ($ul_1$, $ul_2$, $ul_3$, …$ul_n$) e un insieme di unità di contesto ($uc_1$, $uc_2$, $uc_3$, … $uc_m$), rispettivamente con "n" e "m" elementi, è possibile costruire una matrice rettangolare "n x m", con "n" righe e "m" colonne (vedi Tab. 1 oppure, nella seconda parte del libro, la Tab. 2 a paragrafo 2.3.), ai cui singoli incroci ($ul_i$ $uc_j$) corrispondono valori di *occorrenza* che indicano quante volte ciascuna unità lessicale (ad esempio una parola) è presente in ciascuna unità di contesto (ad esempio un articolo di quotidiano). Ugualmente, se le UC sono costituite da contesti elementari, possiamo costruire una matrice quadrata "n x n" (vedi Tab. 2[32]), ai cui singoli incroci ($ul_i$ $ul_j$) corrispondono valori di *co-occorrenza*, ovvero numeri che indicano la quantità di contesti elementari in cui ciascuna unità lessicale (ad esempio una parola) è presente insieme a ciascuna delle altre[33].

In entrambe i casi, cioè sia nelle matrici rettangolari che in quelle quadrate, ciascuna riga e ciascuna colonna riassume il *profilo* di una unità di analisi (UC o UL). Ad esempio, nelle matrici rettangolari "n x m" ogni riga può essere usata per rappresentare il profilo di una parola inteso come *distribuzione*[34] delle sue occorrenze entro i contesti considerati. Diversamente, nel caso delle matrici quadrate, ogni riga e ogni colonna rappresenta un profilo di co-occorrenze.

Le nozioni di occorrenza e di co-occorrenza, di profilo e di distribuzione, sono quindi le chiavi per capire le teorie del significato che sono implementate negli strumenti T-LAB.

In effetti, nella logica del software ogni unità di analisi (UC o UL) è "conosciuta" solo in base al suo profilo. Più esattamente:

- ogni UC è conosciuta attraverso il profilo delle occorrenze di ogni UL in essa presente;
- ogni UL è conosciuta attraverso il profilo delle sue occorrenze all'interno delle varie UC, come dalle sue co-occorrenze con ciascuna delle altre UL.

Evidentemente, *il software non conosce propriamente i significati* (o contenuti), bensì solo i significanti, cioè le stringhe e le label che individuano le

[32] Le stesse informazioni contenute nella Tab. 2 possono essere rappresentate mediante una tabella rettangolare ("m X n") con valori di presenza/assenza ("1" e "0"): contesti elementari (UC) in riga e unità lessicali (UL) in colonna.

[33] Il calcolo delle co-occorrenze può risultare molto diverso a seconda della *lunghezza* delle unità di contesto che vengono prese in considerazione. Queste, infatti, possono essere costituite da sintagmi di due sole parole, da frasi, da paragrafi e da interi testi (o documenti). In T-LAB, fatti salvi eventuali interventi dell'utilizzatore, la lunghezza standard è quella dei "contesti elementari": circa 400 caratteri, pari a circa 30 parole "piene" (nomi, verbi, aggettivi, avverbi, pronomi).

[34] A questo proposito è pertinente il riferimento alla linguistica distribuzionale (L. Bloomfield, 1933), ambito in cui la *distribuzione* di un'unità è costituita dall'insieme dei contesti in cui la si può incontrare.

varie UC e UL; tuttavia, le relazioni tra significanti (ovvero i rapporti sintagmatici) assumono delle forme significative che, in qualche modo, propongono una *rappresentazione contestuale dei significati*.

Consideriamo, ad esempio, l'unità lessicale per eccellenza: la singola "parola". Se prescindiamo dalle categorie morfologiche e dalle funzioni sintattiche [35], ogni singola parola che utilizziamo si differenzia da tutte le altre in base alle "unità minimali" o "tratti distintivi" in cui può essere scomposta, sia sul versante del significante (o espressione) che su quello del significato (o contenuto): nel primo caso (significante/espressione) il risultato della scomposizione è costituito dai fonemi, nel secondo caso (significato/contenuto) dai *sémi* o tratti semantici. Questi ultimi, nel modello proposto da A. J. Greimas (1966) risultano dalla combinazione di due tipi di elementi:

- il **nucleo semico**, supposto invariante, condiviso da tutte le occorrenze della parola considerata[36];
- i **semi contestuali**, variabili e condivisi con altre parole co-occorrenti negli stessi contesti.

Ebbene, nella logica di T-LAB, il significato di ogni singola parola (UL) è conosciuto solo attraverso le sue relazioni con i contesti, cioè attraverso le distribuzioni delle sue co-occorrenze e delle sue occorrenze all'interno delle unità di contesto (UC)[37]. Più precisamente, nel caso delle co-occorrenze il significato contestuale di ogni parola è conosciuto attraverso le relazioni di associazione e di prossimità con altre parole entro gli stessi contesti elementari; mentre, nel caso delle occorrenze, è conosciuto attraverso il diverso modo in cui, rispetto alle altre parole, la sua numerosità si distribuisce all'interno dei sottoinsiemi del corpus. Usando la metafora della moda, potremmo dire che nel primo caso (co-occorrenze), l'informazione utilizzata dal software concerne gli abbinamenti tra capi di vestiario nelle *mise* dei vari soggetti (ad es. quali calze con quale cravatta), mentre nel secondo caso (occorrenze) concerne l'uso dei singoli capi da parte di specifiche popolazioni (ad es. l'uso dei jeans all'interno delle varie "fasce" di età).

Poiché ogni contesto elementare è prodotto da "un" locutore, l'analisi delle co-occorrenze consente di fare inferenze sui modelli mentali e culturali che presiedono alle associazioni delle parole[38]; diversamente, nel caso dell'analisi

[35] Le categorie morfologiche vengono utilizzate per classificare le parole in sostantivi, aggettivi, verbi ecc. Le funzioni sintattiche sono quelle del tipo "soggetto" e "predicato".

[36] Nella terminologia di Greimas le parole, considerate sotto l'aspetto del significato, sono "sememi".

[37] Il riferimento ai semi contestuali di Greimas è dunque più che pertinente.

[38] Per questa ragione, agli albori delle prime applicazioni informatiche, C. E. Osgood

delle occorrenze le inferenze sono fondate sull'osservazione di somiglianze e differenze tra i *profili* dei sottoinsiemi, cioè tra le partizioni del corpus predefinite dal ricercatore.

Prima di passare a considerare gli algoritmi attraverso i quali T-LAB assolve il compito di *stabilire relazioni* tra le unità di analisi (UC e UL), ancora una nota di metodo. Quando si usano tecniche statistiche, in primo luogo si tratta di decidere quali tabelle sottoporre ad analisi; il che, nel nostro caso, significa *selezionare le unità di analisi* da mettere in riga e quelle da mettere in colonna.

In qualche modo, la selezione delle unità di analisi è allo stesso tempo un'*operazione analitica* e un'*operazione interpretativa* (vedi paragrafo 1.5). In effetti, dal momento che T-LAB consente di selezionare le unità di contesto (sottoinsiemi e contesti elementari), l'operazione analitica si esplica come una forma automatica di scomposizione; diversamente, la selezione delle unità lessicali, ad esempio la *costruzione di liste* di parole-chiave, è in qualche modo un'attività interpretativa.

Delle due operazioni, la seconda (la costruzione di liste) è certo la più *impegnativa*: non tanto perché è difficile da mettere in atto[39] (Fig. 7), quanto perché sollecita l'analista a definire dei criteri di rilevanza e, per tornare alla metafora del ricercatore-pescatore (vedi paragrafo 1.1), a decidere quale rete usare per quali tipi di pesci. Fuor di metafora: un corpus o un suo sottoinsieme possono includere diverse centinaia di unità lessicali (UL) che, a seconda dei casi, possono essere usate per definire le righe o le colonne delle tabelle dati[40]. Ora, da un lato i vincoli del software[41] dall'altro quelli derivanti dagli obiettivi dei ricercatori portano a ridurre la quantità di UL da includere nell'analisi. Inoltre: non è detto che l'uso di una quantità elevata di UL sia garanzia di una migliore qualità dei risultati. Gli "sbilanciamenti" delle occorrenze o delle co-occorrenze possono infatti generare più di un'anomalia nei risultati delle analisi.

Evidentemente un criterio ottimale per la selezione delle unità di analisi non può essere definito a priori. Anche per questa ragione, T-LAB consente sempre di tornare sui propri passi: di provare delle strade, di costruire un percorso lineare o circolare, e di ricominciare daccapo, anche con la possibilità di "salvare" e riutilizzare il lavoro già fatto (vedi personalizzazione dei dizionari).

(1959) propose uno specifico metodo di analisi delle co-occorrenze, da lui denominato *contingency analysis*.

[39] In T-LAB è possibile scegliere le "impostazioni automatiche" oppure gestire agevolmente le "impostazioni personalizzate".

[40] Ad esempio: in alcuni casi le unità lessicali sono "oggetti" in riga; in altri casi, quando funzionano come "caratteristiche" dei contesti elementari, sono in colonna.

[41] In T-LAB alcune tecniche statistiche consentono l'uso di max 10.000 tipi di unità lessicali, altre di max 500. In un caso (analisi delle sequenze) di max 250.

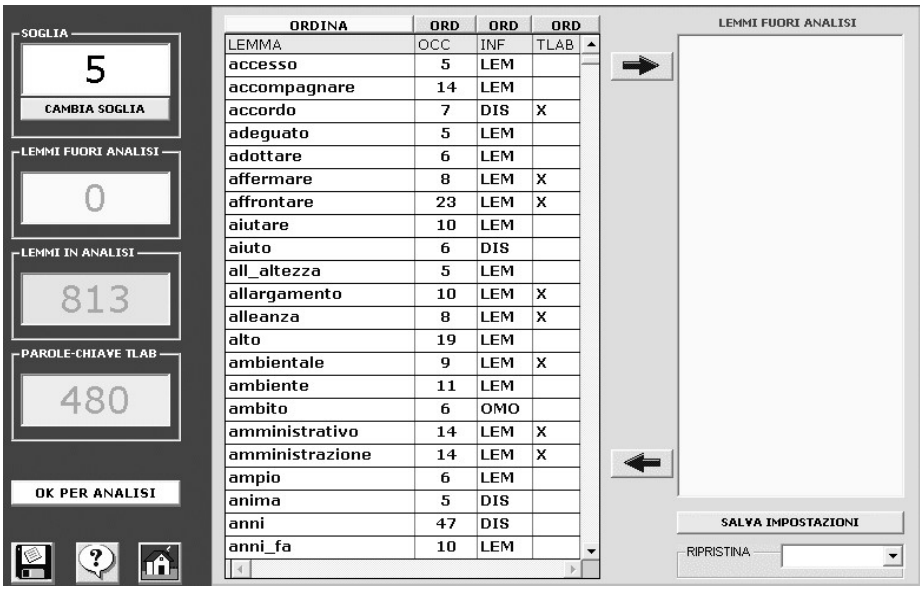



*Fig. 7 – Impostazioni di analisi*

## 2.3. Analisi statistica: stabilire relazioni tra i dati

Ogni ricercatore, quale che sia il percorso che segue o che intende seguire, ha sempre delle specifiche *domande* circa le **relazioni tra i dati** che sta analizzando. Queste domande, per diventare operative, devono tradursi in "comandi" dati al software, cioè nella scelta di qualche opzione di analisi.

Nel momento in cui una di queste viene attivata, in T-LAB entrano in opera le relazioni tra database e algoritmi; più precisamente, nel caso delle analisi statistiche ogni click del mouse (il comando) si trasforma in una o più query che – in prima istanza – genera una tabella dati (Fig. 8).

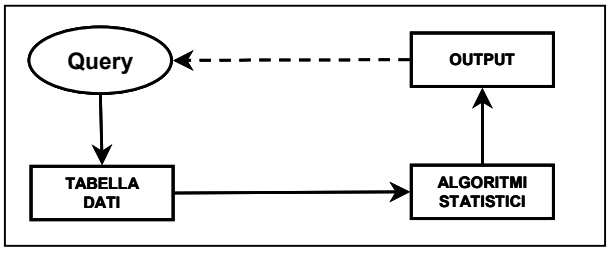



*Fig. 8 – Query e tabella dati*

Ogni tabella costituisce una *rappresentazione* delle relazioni tra le unità di analisi (UC e UL) poste in riga e/o in colonna: relazioni che, essendo espresse in forma di numeri, possono essere opportunamente trattate dagli algoritmi di calcolo.

Nelle pagine che seguono, tanto le *logiche* che presiedono alla costruzione delle tabelle quanto quelle implementate negli algoritmi di calcolo verranno descritte e commentate con l'esclusivo obiettivo di rendere comprensibili i processi messi in atto.

Quindi, per non appesantire il discorso, da un lato ho deciso di non soffermarmi a considerare puntualmente le relazioni tra l'insieme degli strumenti[42] T-LAB e quello degli algoritmi utilizzati (vedi Tab. 3), dall'altro ho deciso di far ricorso a formalismi di tipo matematico solo nei casi in cui la comprensione delle "operazioni" non richiede competenze di tipo specialistico[43].

*Tab. 3 – Strumenti T-LAB (menu Analisi e Mappe)*

| *Strumenti di analisi* | *Algoritmi statistici* |
|---|---|
| Associazioni di parole | Coefficiente del coseno (Cos.) |
| Confronti tra coppie di parole chiave | Test del chi$^2$ |
| Analisi delle specificità | Test del chi$^2$ |
| Analisi delle sequenze | Catene markoviane |
| Mappe dei nuclei tematici[44] | Cos, AFC[45] Cluster analysis (CLU) |
| Analisi delle corrispondenze | AFC |
| Cluster analysis | CLU, Test del chi$^2$ |
| Tipologie dei contesti elementari | Cos., CLU, AFC, Test del chi$^2$ |

## 2.3.1. Coefficiente del coseno

Il coefficiente del coseno è una *misura* proposta da G. Salton (1984), nell'ambito degli studi sull'*information retrieval*, per valutare il grado di *as-*

[42] Indicazioni sull'uso dei vari strumenti sono contenute nel manuale disponibile via internet (http://www.tlab.it/it/download.asp) e negli esempi riportati nella seconda parte di questo libro.

[43] In particolare, per quanto concerne l'analisi delle corrispondenze e la cluster analysis, il lettore interessato ad approfondire gli aspetti matematici e statistici potrà utilmente consultare la letteratura citata nelle note dei rispettivi paragrafi.

[44] In una versione di T-LAB attualmente (aprile 2004) in fase di test, l'uso di questo strumento include una ulteriore tecnica statistica: il *Multidimensional Scaling* (MDS).

[45] AFC sta per Analisi Fattoriale delle Corrispondenze.

*sociazione*[46] tra documenti e parole, e successivamente per realizzare una classificazione automatica degli uni e/o delle altre (document clustering e keyword clustering).

In T-LAB il coefficiente del coseno è usato per misurare l'associazione tra coppie di unità lessicali (parole, lemmi ecc.) co-occorrenti all'interno dei contesti elementari del corpus o di suoi sottoinsiemi; ciò anche allo scopo di individuare "nuclei tematici".

Data una qualunque coppia di unità lessicali (X e Y), l'algoritmo calcola il rapporto tra la quantità delle loro co-occorrenze ($X \cap Y$) e quella ottenuta moltiplicando le radici quadrate delle rispettive occorrenze[47].

La sua formula è la seguente:

$$(1) \quad C(X, Y) = \frac{X \cap Y}{\sqrt{X} * \sqrt{Y}}$$

Il valore massimo è 1, il minimo è 0[48].

Consideriamo ad esempio il testo (corpus) della Costituzione Italiana: al suo interno il lemma *"regione"* (X) e quello *"provincia"* (Y) hanno rispettamente 58 e 11 occorrenze; mentre la quantità delle loro co-occorrenze ($X \cap Y$) è 9. Ne deriva che, in questo caso, il coefficiente del coseno (C(X,Y)) è pari a 0,356 (vedi Tab. 4).

*Tab. 4 – Coefficiente del coseno: esempio di calcolo*

$$C(X,Y) = \frac{9}{\sqrt{58} * \sqrt{11}} = \frac{9}{7{,}616 * 3{,}17} = \frac{9}{25{,}262} = 0{,}356$$

Evidentemente, l'associazione tra due sole parole non da informazioni sufficienti per fare inferenze sul loro significato contestuale. Per questa ragione gli strumenti T-LAB applicano questi calcoli a tabelle con molte unità lessicali (UL). Ad esempio, di seguito vengono riportati i coefficienti del coseno, ordinati in modo decrescente, corrispondenti alle prime 20 UL che nella Costituzione Italiana risultano associate con *"regione"*.

[46] Altre note misure di associazione sono i coefficienti di Dice, di Jaccard e il coefficiente di equivalenza proposto da B. Michelet (1988).

[47] La quantità delle co-occorrenze è data dal numero di contesti elementari (frasi o segmenti ad essi equivalenti) in cui X e Y sono contemporaneamente presenti. La quantità delle occorrenze è data dal numero di volte in cui X e Y sono presenti nel corpus o in un suo sottoinsieme.

[48] In genere, quando l'indice è molto alto (prossimo a 1), la coppia considerata è un sintagma con significato unitario (lessia).

provincia (0,356); comune (0,351); amministrativo (0,295); repubblica (0,271); ente_locale (0,268); richiesta (0,232); popolazione (0,232); sentire (0,232); legge (0,230); funzioni (0,223); attribuire (0,208); consiglio_regionale (0,203); interessato (0,197); Valle_d_Aosta (0,197); Stato (0,192); organo (0,181); locale (0,176); giunta (0,176); esercitare (0,172); referendum (0,166).

Utilizzando questi dati, T-LAB produce in modo automatico grafici come quello in Fig. 9[49].

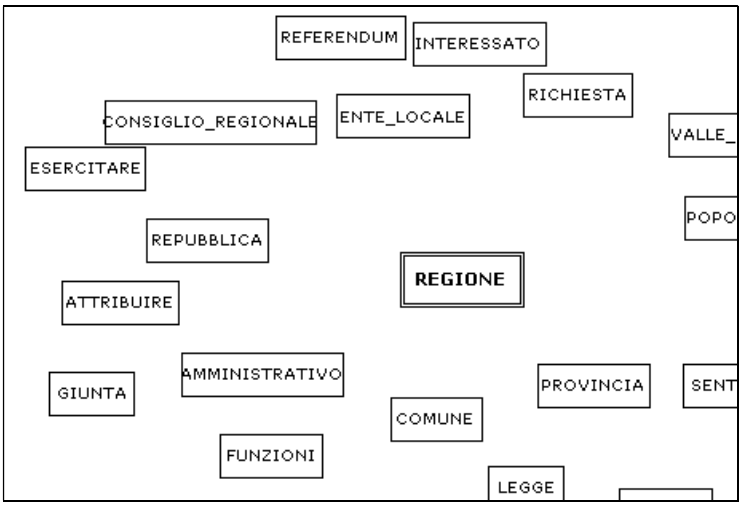



*Fig. 9 – Associazioni relative a "regione"*

Inoltre, da alcuni strumenti T-LAB la formula del coseno è usata per costruire matrici di prossimità da trattare mediante cluster analysis. Poiché queste matrici sono in genere molto grandi (anche 500 x 500), propongo un esempio "fittizio" con 5 elementi (a, b, c, d, e) corrispondenti ad altrettante unità lessicali (UL) presenti nel sottoinsieme "X".

La matrice quadrata sulla sinistra (Tab. 5) riporta i valori di co-occorrenza per ogni coppia: ad esempio, nel caso della coppia "a-b" la quantità delle loro co-occorrenze è pari a 15. La tabella sulla destra (Tab. 6) riporta i valori di occorrenza delle stesse 5 UL all'interno del sottoinsieme considerato ("X")[50].

[49] Si tratta di un grafico a "raggiera" (o "radar"), in cui le distanze di ogni punto (label) dal centro sono inversamente proporzionali al loro grado di associazione (coefficiente del coseno) con la parola chiave selezionata.

[50] Nell'esempio, per ogni oggetto-riga (ad es "a"), il numero delle occorrenze non corrisponde al totale delle sue co-occorrenze con gli altri quattro. Ciò in quanto il sottoinsieme "X" include molti contesti elementari in cui "a" è presente insieme ad unità lessicali diverse (f, g, h, ecc.).

*Tab. 5 e Tab. 6 – Esempi di matrici (co-occorrenze e occorrenze)*

| | *a* | *b* | *c* | *d* | *e* |
|---|---|---|---|---|---|
| *a* | 0 | 15 | 1 | 8 | 15 |
| *b* | 15 | 0 | 2 | 14 | 14 |
| *c* | 1 | 2 | 0 | 11 | 14 |
| *d* | 8 | 14 | 11 | 0 | 25 |
| *e* | 15 | 14 | 14 | 25 | 0 |

| | *X* |
|---|---|
| *a* | 505 |
| *b* | 256 |
| *c* | 66 |
| *d* | 119 |
| *e* | 190 |

In base a questi dati, utilizzando la formula (1) viene generata una matrice di prossimità (Tab. 7) i cui valori sono altrettanti coefficienti del coseno.

*Tab. 7 – Matrice quadrata con coefficienti del coseno*

| | *a* | *b* | *c* | *d* | *e* |
|---|---|---|---|---|---|
| *a* | 0,000 | 0,042 | 0,005 | 0,033 | 0,048 |
| *b* | 0,042 | 0,000 | 0,015 | 0,080 | 0,063 |
| *c* | 0,005 | 0,015 | 0,000 | 0,124 | 0,125 |
| *d* | 0,033 | 0,080 | 0,124 | 0,000 | 0,166 |
| *e* | 0,048 | 0,063 | 0,125 | 0,166 | 0,000 |

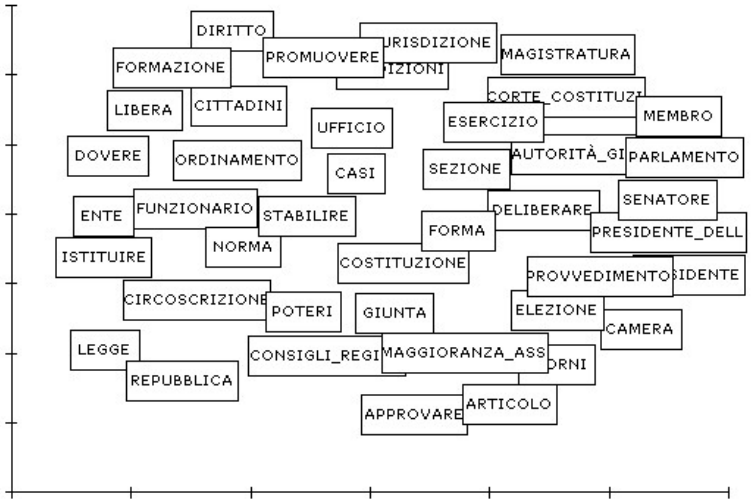



*Fig. 10 – Costituzione Italiana: mappa dei nuclei tematici*

Poiché una proprietà di queste tabelle (o matrici) è quella di essere simmetriche, ai fini dei calcoli successivi vengono utilizzati solo i valori corrispondenti a una delle matrici triangolari (bassa o alta) separate dalla diagonale.

Ad esempio, applicando un algoritmo di multidimensional scaling[51] (MDS) alla matrice di prossimità della Costituzione Italiana, è possibile ottenere un grafico come quello in Fig. 10[52].

### 2.3.2. Test del chi quadro

Data una qualunque tabella a doppia entrata, con incroci tra due o più categorie e con valori di frequenza, il *test* del chi quadro consente di stabilire se esiste una *differenza significativa* tra i valori osservati (O) e quelli teorici (E) sui quali si basa l'ipotesi nulla[53].

Negli strumenti T-LAB il test del chi quadro[54] è usato per verificare la significatività delle *differenze tra coppie* di valori che indicano occorrenze di unità lessicali (UL). Quindi, anche quando le tabelle in input sono composte da centinaia di righe e di colonne, i calcoli vengono effettuati "sempre e soltanto" sugli incroci 2 x 2, con un solo grado di libertà e un valore di soglia corrispondente a 3.84 (p. 0.05).[55]

Proviamo a spiegarci con un esempio. Supponiamo che il corpus sia costituito da quattro articoli di altrettanti quotidiani (A, B, C, D) e che ci interessi testare le differenze tra due di essi (A e B). In questo caso le differenze sono date dalle distribuzioni delle occorrenze delle UL al loro interno, e la lista di quest'ultime può includere centinaia di elementi. Per ciascuno di essi, considerando la coppia data dalle occorrenze in "A" e in "B", il chi quadro ($\chi^2$) si calcola nel modo seguente:

$$(2) \quad \chi^2 = \Sigma \frac{(O-E)^2}{E}$$

[51] Si tratta di una tecnica statistica che consente di analizzare matrici di similarità e di rappresentare le relazioni tra i dati entro uno spazio di dimensioni ridotte.

[52] In questo tipo di grafico ogni label (es. "costituzione") rappresenta un piccolo cluster ("nucleo tematico") di parole associate.

[53] Secondo l'ipotesi nulla i valori osservati delle categorie in riga e in colonna non hanno alcuna relazione di dipendenza reciproca e, in qualche modo, la loro distribuzione è casuale.

[54] Più esattamente, si tratta del chi quadro di associazione (K. Pearson).

[55] Quindi, valori di soglia maggiori o uguali a 3.84 consentono di respingere l'ipotesi nulla.

Nella Tab. 8 è riportato un esempio in cui una unità lessicale ("a") ha 19 occorrenze nell'unità di contesto "A" e 5 occorrenze nell'unità di contesto "B". Per effettuare il calcolo vengono anche considerate le somme delle occorrenze delle altre UL presenti in "A"(1360) e in "B" (1848). I quattro valori finora elencati (19; 5; 1360; 1848) sono le cosiddette frequenze "osservate" (vedi simbolo "O" nella formula 2).

*Tab. 8 – Esempio: frequenze osservate*

| | *in "A"* | *in "B"* | | |
|---|---|---|---|---|
| *UL "a"* | 19 | 5 | 24 | *($N_j$)* |
| *altre UL* | 1360 | 1848 | 3208 | |
| | 1379 | 1853 | 3232 | *($N_{ij}$)* |
| | *($N_i$)* | | | |

Per ogni cella, le frequenze "attese" o "teoriche" (simbolo "E" della formula) sono ottenute dividendo il prodotto dei totali marginali per il totale generale ($(N_i \times N_j) / N_{ij}$). Ad esempio (vedi Tab. 9), nel caso della cella con frequenza osservata pari a "19", la frequenza attesa è calcolata nel modo seguente:

$$(24 \times 1379) / 3232 = 10{,}24$$

*Tab. 9 – Esempio: frequenze attese*

| | *in "A"* | *in "B"* |
|---|---|---|
| *UL "a"* | 10,24 | 13,76 |
| *altre UL* | 1368,76 | 1839,24 |

Usando la formula 2, si ottengono i valori del chi-quadro per ogni cella. Ad esempio (vedi Tab. 10), il valore 7,49 si ottiene nel modo seguente:

$$(19 - 10{,}24)^2 / 10{,}24 = (8{,}76)^2 / 10{,}24 = 76{,}74/10{,}24 = 7{,}49$$

*Tab. 10 – Esempio: valori del chi quadro*

| | *in "A"* | *in "B"* |
|---|---|---|
| *UL "a"* | 7,49 | 5,58 |
| *altre UL* | 0,06 | 0,04 |

Quindi si calcola la loro somma (7,49 + 5,58 + 0,06 + 0,04).

Ne risulta che il valore del chi-quadro è 13.17. Stando che in questo caso (df = 1; p. 0.05) il valore critico è 3.84, l'ipotesi nulla viene respinta e la UL in esame può essere considerata come un elemento che determina una significativa differenza tra "A" e "B".

D'altro canto, come si può rilevare dalla Tab. 10, gli scarti corrispondenti alla UL considerata (7,49 e 5,58) sono significativi – cioè più elevati del valore critico (3.84) – anche presi uno alla volta.

### *2.3.3. Catene markoviane*

Una catena markoviana[56] è costituita da una *successione* di eventi, generalmente indicati come *stati*, caratterizzata da due proprietà:

- l'insieme degli eventi e dei loro possibili esiti è finito;
- l'esito di ogni evento dipende solo (o al massimo) dall'evento immediatamente precedente.

Con la conseguenza che ad ogni transizione da un evento all'altro corrisponde un valore di probabilità[57].

Un semplice esempio aiuterà meglio a capire la sua logica di funzionamento.

Supponiamo che sei ragazzi giochino a passarsi una palla e che qualcuno registri tutti i passaggi (transizioni) da un ragazzo all'altro fino alla fine del gioco. Al termine, tutte le transizioni possono essere riportate in una tabella quadrata dove ogni valore ($s_{ij}$) corrisponde al numero dei passaggi effettuati da ciascun ragazzo verso ciascuno degli altri cinque. Ad esempio (vedi Tab. 11) i passaggi dal ragazzo "$s_1$" al ragazzo "$s_4$" risultano essere 11.

[56] Dal nome del matematico russo Andrei Andreiëvich Markov (1856-1922).

[57] Detto "probabilità di transizione" e che, in T-LAB, concerne solo le catene markoviane del primo ordine, cioè quelle in cui il calcolo prende in considerazione – una alla volta – le transizioni tra due "stati" contigui. Ulteriori approfondimenti su questo tema sono contenuti nei manuali di calcolo delle probabilità (cfr. S. Lipschutz, 1965).

*Tab. 11 – Transizioni*

| | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | TOT |
|---|---|---|---|---|---|---|---|
| $s_1$ | 0 | 8 | 7 | 11 | 2 | 1 | 29 |
| $s_2$ | 6 | 0 | 24 | 5 | 10 | 8 | 53 |
| $s_3$ | 9 | 24 | 0 | 3 | 28 | 16 | 80 |
| $s_4$ | 3 | 7 | 5 | 0 | 6 | 14 | 35 |
| $s_5$ | 4 | 5 | 26 | 11 | 0 | 7 | 53 |
| $s_6$ | 7 | 9 | 18 | 5 | 7 | 0 | 46 |

*Tab. 12 – Probabilità di transizione*

| | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | TOT |
|---|---|---|---|---|---|---|---|
| $s_1$ | 0,00 | 0,28 | 0,24 | 0,38 | 0,07 | 0,03 | 1 |
| $s_2$ | 0,11 | 0,00 | 0,45 | 0,09 | 0,19 | 0,15 | 1 |
| $s_3$ | 0,11 | 0,30 | 0,00 | 0,04 | 0,35 | 0,20 | 1 |
| $s_4$ | 0,09 | 0,20 | 0,14 | 0,00 | 0,17 | 0,40 | 1 |
| $s_5$ | 0,08 | 0,09 | 0,49 | 0,21 | 0,00 | 0,13 | 1 |
| $s_6$ | 0,15 | 0,20 | 0,39 | 0,11 | 0,15 | 0,00 | 1 |

Successivamente, mediante un semplice calcolo (divisione di ciascun valore per il corrispondente totale di riga), la Tab. 11 può essere trasformata in una tabella con valori di probabilità, detta anche matrice stocastica. In quest'ultima, ad esempio (vedi Tab. 12), la probabilità dei passaggi del ragazzo "$s_1$" al ragazzo "$s_4$" risulta essere 0,38.

A partire da questi dati, è possibile effettuare ulteriori calcoli, costruire ulteriori tabelle e, tramite opportuni grafici, ricostruire l'intera *rete* (network) dei passaggi da un ragazzo all'altro.

In ambito scientifico, il modello delle catene markoviane è utilizzato per analizzare le successioni di eventi economici (ad es. gli andamenti dei mercati), biologici (ad es. le mutazioni delle cellule), fisici (ad es. i processi implicati nella fusione nucleare), ecc. Nell'ambito degli studi linguistici le sue applicazioni hanno come oggetto le possibili combinazioni delle varie unità di analisi sull'asse delle relazioni sintagmatiche; su questo, infatti, «Il testo è una *catena*, e tutte le parti (per esempio proposizioni, parole, sillabe, e così via) sono similmente delle *catene*, tranne le eventuali parti ultime che non si possono sottoporre ad analisi» (L. Hjelmslev, 1943, tr. it., p. 33).

In T-LAB, l'uso delle catene markoviane è attualmente un work in progress: nelle versioni in commercio è circoscritto all'analisi delle successioni tra parole chiave all'interno del corpus, in alcuni prototipi è esteso all'analisi delle successioni delle frasi (i contesti elementari) e dei loro contenuti[58].

Quanto alla successione (o catena) delle parole chiave, per ciascuna di esse gli output includono l'elenco dei predecessori e dei successori[59] con i rispetti-

[58] Sull'uso delle catene markoviane nell'ambito dell'analisi di contenuto, si veda P. Maranda (1995).

[59] Poiché l'elenco è costituito da "tipi" (word type), secondo la teoria dei grafi, per ogni nodo (o parola chiave) la sommatoria dei predecessori corrisponde al "semigrado interno" (inde-

vi valori di probabilità[60]. Per molti versi, questo tipo di applicazione può essere definito come uno strumento per analizzare, nel dettaglio, il fenomeno delle associazioni (co-occorrenze).

Consideriamo ad esempio le associazioni relative al lemma *"sicurezza"* all'interno del corpus dell'esercizio N. 2 (parte seconda, paragrafo 3.3,)[61]. Usando l'indice del coseno (vedi paragrafo 2.3.1) otteniamo la lista in Tab. 13; mentre, usando il modello delle catene markoviane otteniamo due liste: quella dei predecessori (Tab. 14) e quella dei successori (Tab. 15).

In un caso (indice del coseno) le relazioni in esame sono quelle tra la parola chiave "sicurezza" e tutte altre parole presenti all'interno dei contesti elementari in cui essa occorre, nell'altro (catene markoviane) le relazioni in esame sono tra "coppie", cioè tra *"sicurezza"* e ogni parola che la precede o la segue. In ogni caso, i valori degli indici sono misure delle relazioni tra la parola selezionata (*"sicurezza"*) e ciascuna delle altre, prese una ad una.

Al momento, nella funzione *Analisi delle sequenze*, i grafici che T-LAB rende disponibili (vedi Fig. 11) sono del tipo a raggiera (o radar), analoghi a quelli della funzione *Associazioni*; il che significa che le relazioni mostrate sono del tipo una-ad-una, tra la parola selezionata (ad es. "sicurezza") e ciascuna delle altre.

Tuttavia, col metodo carta e matita oppure servendosi di un semplice software grafico, l'utilizzatore può realizzare diagrammi che mostrano reti di relazioni.

Ad esempio il grafico in Fig. 12[62] è stato ottenuto nel modo seguente:

I primi tre predecessori e successori di *sicurezza* sono stati disposti rispettivamente sulla sua sinistra e sulla sua destra. Successivamente, cliccando sulla tabella a sinistra di Fig. 11, sono stati individuati i predecessori e i successori delle label in Fig. 12. Ad esempio, poiché il primo predecessore di *"problema"* è risultato essere *"rischi",* la sua label è stata aggiunta a sinistra, con una freccia "orientata" (o arco) che funziona da connessione.

gree) e quella dei successori al "semigrado esterno" (outdegree).

[60] In questo caso, la somma dei valori di probabilità è pari a 1 solo quando la lista delle parole in analisi include tutti i tipi (type) di parole presenti nella catena del corpus. Negli altri casi, la somma è pari a 1-p, dove "p" indica il numero di volte in cui le transizioni riguardano parole non incluse nella lista di quelle analizzate.

[61] Le lista utilizzata include 236 unità lessicali: quelle con valori di occorrenza pari o superiori a 5. Le tabelle (13, 14, 15) includono i 15 lemmi con gli indici più elevati. Quanto all'interpretazione dei vari indici (di associazione e di probabilità), ricordo che ciascuno dei suoi valori misura le relazioni all'interno di un "sistema chiuso", quello definito dalla lista utilizzata.

[62] Secondo la teoria dei grafi (O. Ore, 1976), poiché le connessioni tra i vari elementi ("spigoli" o "vertici") sono "frecce" direzionate (o "archi"), la Fig. 12 rappresenta un "grafo orientato" o "digrafo".

*Tab. 13 – Associazioni*

| *Cos.* | *Unità Lessicali* |
|---|---|
| 0,344 | informazione |
| 0,314 | sistema |
| 0,308 | informatica |
| 0,280 | nuovo |
| 0,279 | rete |
| 0,278 | problema |
| 0,275 | azienda |
| 0,260 | sistemi |
| 0,260 | protezione |
| 0,257 | tecnologico |
| 0,251 | internet |
| 0,245 | garantire |
| 0,240 | comunicazione |
| 0,239 | livello |
| 0,237 | certificare |

*Tab. 14 – Predecessori*

| *Prob.* | *Unità Lessicali* |
|---|---|
| 0,039 | problema |
| 0,033 | certificare |
| 0,030 | governo |
| 0,024 | politica |
| 0,018 | aspetti |
| 0,018 | termine |
| 0,018 | gestione |
| 0,015 | attacchi |
| 0,012 | livello |
| 0,012 | garanzia |
| 0,012 | caratteristica |
| 0,012 | garantire |
| 0,012 | privacy |
| 0,012 | valore |
| 0,012 | minimo |

*Tab. 15 – Successori*

| *Prob.* | *Unità Lessicali* |
|---|---|
| 0,075 | informatica |
| 0,051 | informazione |
| 0,033 | rete |
| 0,030 | sistema |
| 0,021 | nazionale |
| 0,021 | riservatezza |
| 0,015 | comunicazione |
| 0,012 | problema |
| 0,012 | azienda |
| 0,009 | telematico |
| 0,009 | organizzazione |
| 0,006 | consentire |
| 0,006 | considerare |
| 0,006 | controlli |
| 0,006 | diventare |

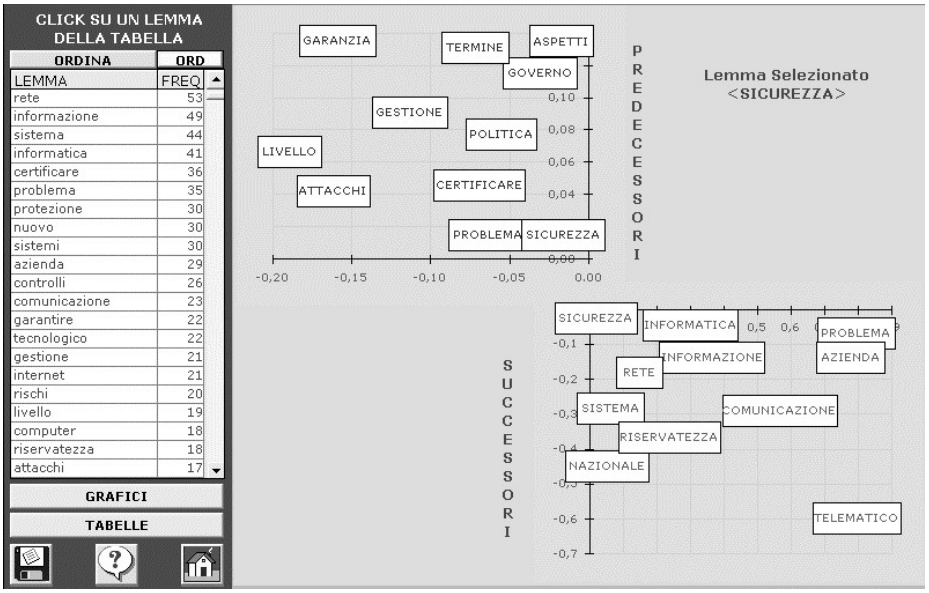



*Fig. 11 – Sequenze: esempio di grafico*

Evidentemente, la ricostruzione delle rete può essere estesa. In ogni caso, stando che – secondo il modello utilizzato – l'esito di ogni evento dipende solo dall'evento immediatamente precedente, la sequenza va scandita a "coppie". Ad es. rischi → problema, problema →sicurezza; sicurezza→ informatica; informatica → azienda. Inoltre, ogni transizione può essere corredata dal suo valore di probabilità. Ad esempio, in Fig. 12 la transizione da "problema" a "sicurezza" ha un valore di probabilità pari a 0.039.

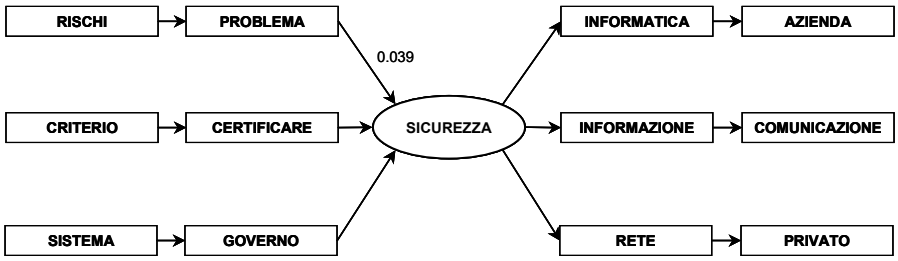



*Fig. 12 – Sequenze: esempio di grafico*

### *2.3.4. Analisi delle corrispondenze*

L'analisi delle corrispondenze è una tecnica per rappresentare graficamente le relazioni tra i profili riga e i profili colonna delle matrici (tabelle) con valori di frequenza o con valori di presenza/assenza.

In T-LAB questa tecnica è utilizzata per analizzare tre tipi di tabelle:

- unità lessicali in riga e sottoinsiemi del corpus in colonna, i cui profili sono distribuzioni di occorrenze;
- contesti elementari in riga e unità lessicali in colonna, i cui profili sono co-occorrenze espresse con valori di presenza/assenza ("1" e "0");
- unità lessicali in riga e in colonna, i cui profili sono co-occorrenze espresse con valori di frequenza.

Per capire cosa significa rappresentare graficamente le relazioni tra profili riga e profili colonna, possiamo partire da un semplice esempio: una tabella costituita da 5 righe e 2 colonne (vedi Tab. 16), cioè da 5 casi (o oggetti) e 2 variabili (o caratteristiche).

Mediante l'uso di un semplice algoritmo possiamo costruire un grafico (Fig. 13) che – *in modo esaustivo* – rappresenta le relazioni tra questi dati. Ad

esempio, il profilo riga "A" (1, 3) è puntualmente identificato dalle sue coordinate "X" (1) e "Y" (3).

*Tab.16 – Esempio di tabella dati e sua rappresentazione grafica*

| | X | Y |
|---|---|---|
| A | 1 | 3 |
| B | 2 | 2 |
| C | 3 | 2 |
| D | 4 | 1 |
| E | 1 | 1 |

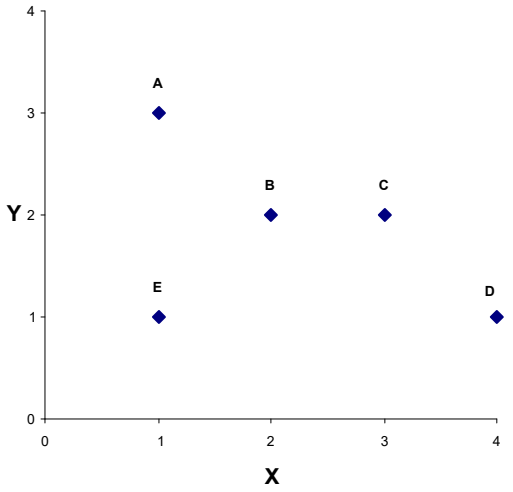



*Fig. 13*

Nel caso di tabelle con molti casi ("n") e molte variabili ("m"), la costruzione di *una* mappa che – *in modo esaustivo* – rappresenti le posizioni di tutti i dati (i "punti") è di fatto impraticabile. Possiamo solo immaginare che – se fosse possibile – sarebbe costituita da nuvole "n" o "m" dimensionali, al cui interno i dati sarebbero dispersi; quindi, ogni intento di costruire modelli interpretativi delle loro relazioni sarebbe di fatto velleitario.

L'analisi delle corrispondenze è appunto una tecnica statistica che consente di ridurre le dimensioni entro le quali i dati possono essere rappresentati ed esplorati. Per descrivere la logica del suo funzionamento, almeno in prima istanza, si può prescindere da nozioni matematiche e statistiche[63]. Basti pensare che le operazioni implementate nei suoi algoritmi consentono di ottenere due tipi di risultati, complementari tra loro:

[63] Il lettore interessato ad approfondire questi aspetti può consultare i seguenti volumi: J. P. Benzecri (1984), M. J. Greenacre (1984), L. Lebart, A. Morineau, M. Piron (1995), S. Bolasco (1999).

- **rintracciare regolarità** nelle tabelle dati, ciò attraverso un confronto incrociato di tutti i profili (righe e colonne) nelle reciproche relazioni di somiglianza-differenza, con il risultato che – attraverso una serie di permutazioni – le tabelle possono essere ri-ordinate e le informazioni in esse contenute diventano "leggibili";
- **ridurre le dimensioni** entro le quali i dati possono essere rappresentati, ciò attraverso la creazione di nuove variabili (i fattori) i cui valori corrispondono alle coordinate spaziali dei profili (righe e colonne). In questo modo i dati, inizialmente dispersi in uno spazio multi-dimensionale, risultano "agglomerati" entro uno spazio a dimensioni ridotte, quelle definite dai pochi fattori che, in modo statisticamente significativo, spiegano la loro variabilità.

Proviamo a spiegarci con un esempio. La Tab. 17 riporta gli addetti per settore di attività economica nelle venti regioni italiane.

*Tab. 17 – Dati ISTAT – Censimento 2001*
*Numero di addetti per settore di attività nelle venti regioni italiane*

| | *INDUSTRIA* | *COMMERCIO* | *ALTRI SERV.* | *ISTITUZIONI* |
|---|---|---|---|---|
| *ABRUZZO* | 136.641 | 65.264 | 112.172 | 85.167 |
| *BASILICATA* | 45.614 | 23.980 | 42.754 | 40.971 |
| *CALABRIA* | 81.233 | 82.869 | 112.867 | 157.436 |
| *CAMPANIA* | 288.763 | 225.549 | 367.164 | 316.111 |
| *EMILIA-ROMAGNA* | 645.648 | 303.469 | 524.867 | 254.359 |
| *FRIULI-VENEZIA GIULIA* | 165.892 | 76.337 | 131.067 | 86.671 |
| *LAZIO* | 316.367 | 284.008 | 595.622 | 395.050 |
| *LIGURIA* | 107.072 | 107.187 | 198.664 | 107.409 |
| *LOMBARDIA* | 1.488.019 | 642.074 | 1.043.835 | 507.691 |
| *MARCHE* | 232.396 | 96.543 | 158.375 | 98.231 |
| *MOLISE* | 28.694 | 14.392 | 23.420 | 18.826 |
| *PIEMONTE* | 612.539 | 272.653 | 467.142 | 273.460 |
| *PUGLIA* | 274.293 | 183.075 | 260.570 | 247.148 |
| *SARDEGNA* | 106.111 | 86.476 | 126.127 | 116.950 |
| *SICILIA* | 195.202 | 202.319 | 265.484 | 340.354 |
| *TOSCANA* | 470.603 | 234.657 | 397.411 | 228.027 |
| *TRENTINO-ALTO ADIGE* | 117.706 | 65.925 | 126.714 | 90.128 |
| *UMBRIA* | 94.619 | 49.804 | 78.885 | 60.017 |
| *VALLE D'AOSTA* | 14.787 | 7.381 | 17.796 | 14.132 |
| *VENETO* | 774.803 | 310.064 | 484.431 | 269.291 |

Una prima rappresentazione grafica di questi dati può essere ottenuta tramite semplici istogrammi (vedi Figg. 14-17).

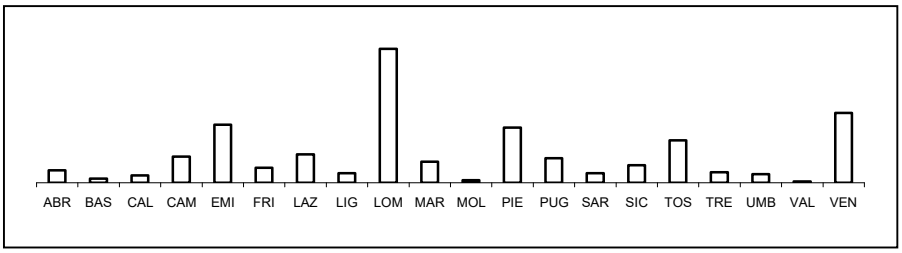



*Fig. 14 – Dati ISTAT 2001 – Addetti Industria*

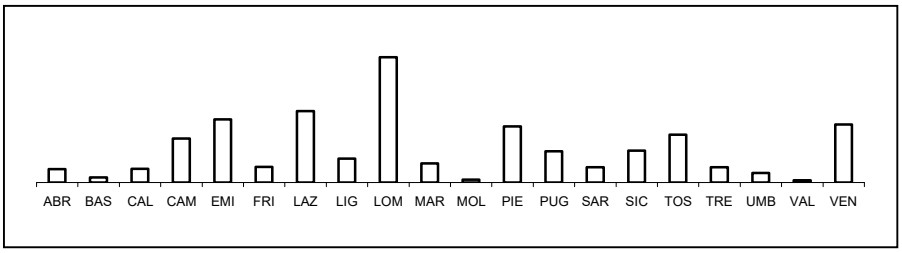



*Fig. 15 – Dati ISTAT 2001 – Addetti Commercio*

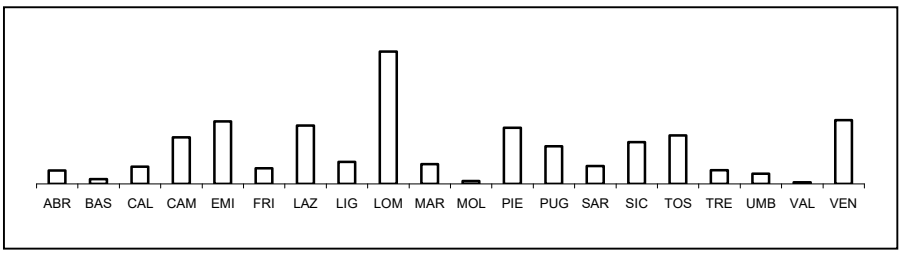



*Fig. 16 – Dati ISTAT 2001 – Addetti Altri Servizi*

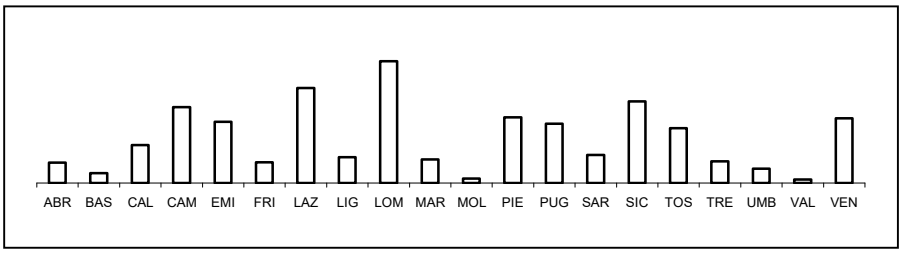



*Fig. 17 – Dati ISTAT 2001 – Addetti Istituzioni*

Come si può verificare, la rappresentazione tramite istogrammi consente di apprezzare le differenze tra singole righe e/o tra singole colonne, ma non le eventuali regolarità nella struttura complessiva dei dati. Ad esempio, è possibile rilevare che i valori della Lombardia (LOM) sono in genere molto elevati e che quelli del Molise (MOL) sono in genere molto bassi, ma ogni tentativo di costruire dei "raggruppamenti" è in qualche modo destinato al fallimento.

Diversamente, se applichiamo agli stessi dati di Tab. 17 un'analisi delle corrispondenze, e poi proviamo a riordinare gli istogrammi secondo i valori (le coordinate) del primo fattore estratto, otteniamo una *rappresentazione* completamente diversa. Come si può rilevare (vedi Figg. 18-21), i raggruppamenti delle regioni sui due estremi dell'asse orizzontale cominciano ad acquistare "senso": da una parte Veneto-Lombardia-Marche-Emilia, dall'altra Sardegna-Lazio-Sicilia-Calabria. E la "bilancia" di ciascun asse, man mano che dalla situazione degli addetti nell'industria (Fig. 18) passiamo a quella degli addetti nelle istituzioni (Fig. 21), "pesa" sempre più verso destra.

In effetti, la tecnica statistica applicata è orientata a valutare somiglianze e differenze tra ciascun profilo (riga o colonna) e tutti gli altri in base a come si distribuisce la proporzione dei "pesi", cioè delle frequenze (numero degli addetti) ponderate, all'interno di ciascuno di essi. In altri termini, il fatto che in Fig. 18 il Veneto (VEN) è all'estremo di sinistra non sta ad indicare che in questa regione è presente il più alto numero di addetti nell'industria, bensì che in questa regione – più che nelle altre 19 – la "bilancia" degli addetti (ad esempio, il rapporto tra industria e istituzioni) "pesa" di più sull'industria. Il ragionamento inverso vale per la Calabria (CAL): pur non avendo il punteggio minimo nel numero di addetti nell'industria, è collocata sull'estremità destra.

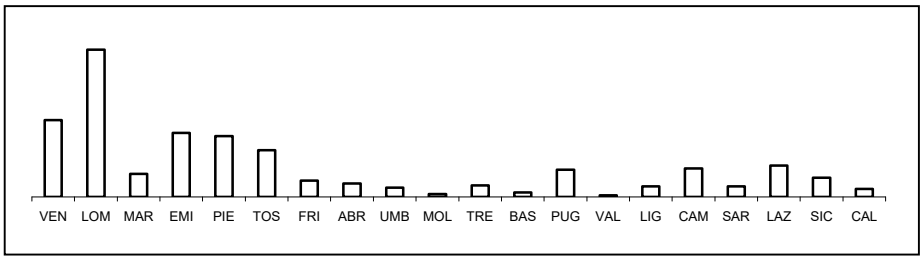



*Fig. 18 – Dati ISTAT 2001 – Addetti Industria*

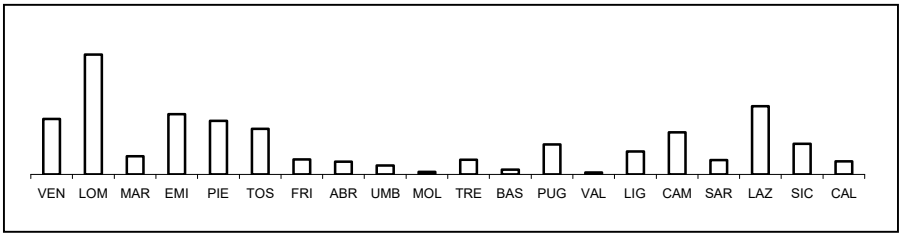



*Fig. 19 – Dati ISTAT 2001 – Addetti Commercio*

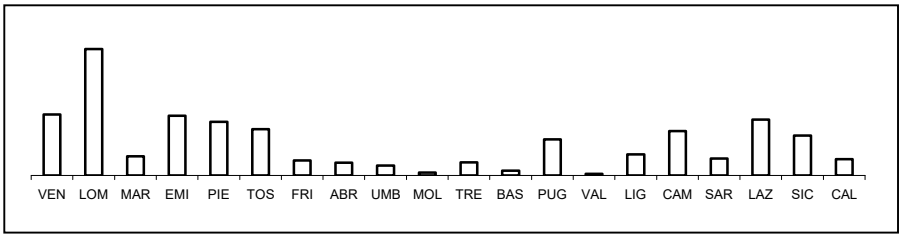



*Fig. 20 – Dati ISTAT 2001 – Addetti Altri servizi*

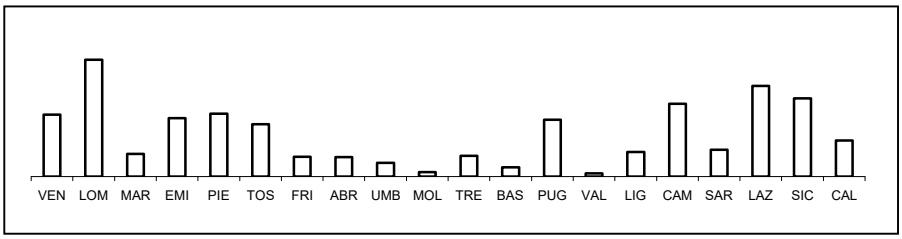



*Fig. 21 – Dati ISTAT 2001 – Addetti Istituzioni*

Sempre utilizzando i risultati dell'analisi delle corrispondenze (primi due fattori estratti e relative coordinate) è possibile generare un grafico come quello di Fig. 22.

A questo punto la *struttura delle relazioni* tra profili riga (le 20 regioni) e profili colonna (i 4 settori di attività) è "visibile", pronta per essere analizzata e interpretata.

Prima di procedere, una precisazione: le tabelle che T-LAB costruisce (ad es. lemmi x variabile) e che tratta mediante analisi delle corrispondenze hanno la stessa struttura di quella ISTAT, e i numeri sono dello stesso tipo: frequenze, ovvero occorrenze. Ad esempio, la Tab. 2 del paragrafo 2.3 (seconda parte del libro) riporta le occorrenze[64] di alcune unità lessicali in un corpus costitui-

[64] Per motivi di impaginazione, la tabella include solo le prime e le ultime parole chiave della lista selezionata.

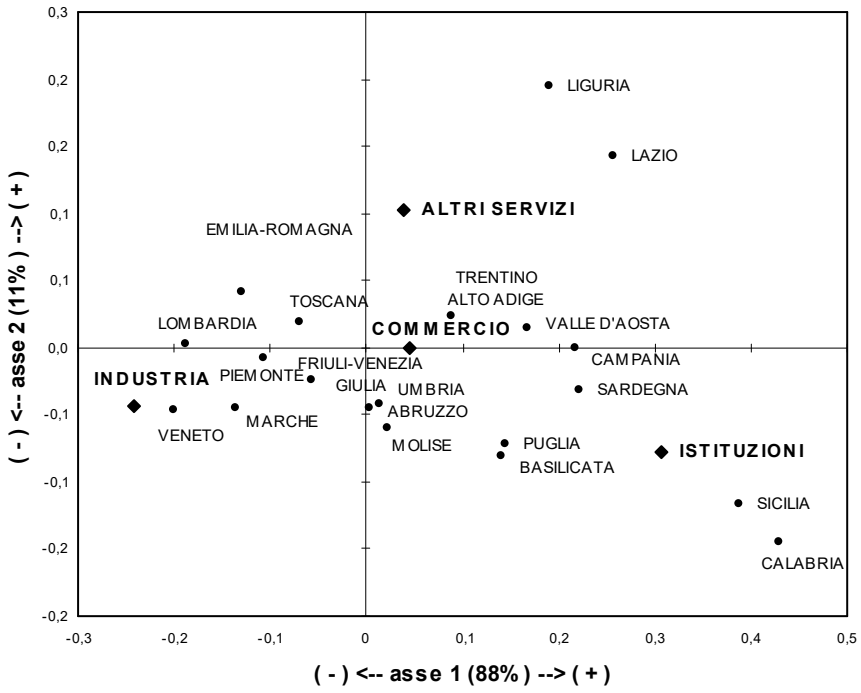



*Fig. 22 – Dati ISTAT 2001 – Analisi delle corrispondenze*

to dai discorsi programmatici degli ultimi cinque presidenti del consiglio della Repubblica Italiana: Amato, Berlusconi (primo e secondo), D'Alema, Prodi.

Nel capitolo 1.5 ho affermato che i grafici prodotti da T-LAB sono delle *icone* e che per interpretarli bisogna prima procedere a un'attenta *analisi iconografica.* Torniamo quindi al grafico (Fig. 22) con i dati ISTAT.

In primo luogo un rilievo: il primo asse, quello orizzontale, spiega da solo l'88% dell'inerzia (o varianza). Affermazione quest'ultima che può essere parafrasata in almeno due modi: a) le somiglianze e le differenze tra i dati (20 righe x 4 colonne) sono all'88% determinate dal primo fattore estratto; b) la dispersione dei punti rispetto al baricentro della "nuvola"[65] è, per l'88%, determinata dal primo fattore estratto.

Soffermiamoci a considerare questo primo dato e facciamo un ulteriore passo. Per definizione, ogni fattore ha un baricentro (lo "0") e due polarità ("-" e

[65] In questo caso, la nuvola ha solo 3 dimensioni. Infatti, per definizione, il numero dei fattori estratti è "*n*-1"; dove "*n"*, in questo caso, sta per il totale (4) dei settori di attività.

"+"); usando la metafora dell'altalena, possiamo verificare in che misura ogni "punto" (riga o colonna, cioè regione o settore di attività) contribuisce, con il suo "peso", a bilanciare la situazione mostrata dal grafico.

La Tab. 18 riporta i valori di cui abbiamo bisogno: le *coordinate* e i *contributi assoluti*[66].

I primi (coordinate) consentono di verificare le posizioni di ogni punto su-

*Tab. 18 – Dati ISTAT 2001 – Coordinate e contributi assoluti*

| | *COORDINATE* | | | *CONTRIBUTI ASSOLUTI* | | |
|---|---|---|---|---|---|---|
| | *1* | *2* | *3* | *1* | *2* | *3* |
| *ABRUZZO* | 0,004 | -0,045 | -0,038 | 0,000 | 0,009 | 0,069 |
| *BASILICATA* | 0,141 | -0,080 | -0,071 | 0,004 | 0,011 | 0,093 |
| *CALABRIA* | **0,429** | -0,144 | -0,011 | **0,110** | 0,098 | 0,007 |
| *CAMPANIA* | 0,218 | -0,001 | 0,005 | 0,078 | 0,000 | 0,004 |
| *EMILIA-ROMAGNA* | -0,131 | 0,041 | 0,009 | 0,040 | 0,032 | 0,016 |
| *FRIULI-VENEZIA GIULIA* | -0,056 | -0,024 | -0,025 | 0,002 | 0,003 | 0,034 |
| *LAZIO* | **0,257** | 0,142 | -0,024 | **0,144** | 0,349 | 0,112 |
| *LIGURIA* | 0,191 | 0,195 | 0,055 | 0,026 | 0,215 | 0,193 |
| *LOMBARDIA* | **-0,187** | 0,003 | 0,011 | **0,176** | 0,000 | 0,057 |
| *MARCHE* | -0,136 | -0,044 | -0,019 | 0,015 | 0,013 | 0,026 |
| *MOLISE* | 0,022 | -0,059 | -0,026 | 0,000 | 0,003 | 0,007 |
| *PIEMONTE* | -0,107 | -0,008 | -0,015 | 0,025 | 0,001 | 0,043 |
| *PUGLIA* | 0,145 | -0,073 | 0,016 | 0,028 | 0,055 | 0,031 |
| *SARDEGNA* | 0,221 | -0,032 | 0,032 | 0,029 | 0,005 | 0,052 |
| *SICILIA* | **0,387** | -0,116 | 0,022 | **0,206** | 0,148 | 0,058 |
| *TOSCANA* | -0,069 | 0,020 | 0,004 | 0,009 | 0,006 | 0,003 |
| *TRENTINO-ALTO ADIGE* | 0,089 | 0,024 | -0,043 | 0,004 | 0,002 | 0,091 |
| *UMBRIA* | 0,014 | -0,042 | -0,006 | 0,000 | 0,006 | 0,001 |
| *VALLE D'AOSTA* | 0,167 | 0,015 | -0,126 | 0,002 | 0,000 | 0,103 |
| *VENETO* | **-0,200** | -0,046 | -0,003 | **0,101** | 0,043 | 0,002 |
| *ALTR_SERV* | 0,039 | 0,102 | -0,010 | 0,012 | 0,625 | 0,069 |
| *COMMERCIO* | 0,045 | -0,001 | 0,045 | 0,009 | 0,000 | 0,813 |
| *INDUSTRIA* | **-0,243** | -0,044 | -0,007 | **0,501** | 0,131 | 0,038 |
| *ISTITUZIONI* | **0,307** | -0,078 | -0,013 | **0,479** | 0,244 | 0,080 |

[66] Gli output T-LAB includono ulteriori valori, quali i contributi relativi (o coseni quadrati) e i valori test. I primi indicano in che misura la posizione di ogni punto (ad es. regione o settore di attività) è determinata e "spiegata" dal contributo dei vari assi fattoriali. I secondi sono misure (vedi L. Lebart, A. Morineau, M. Piron, 1995) che hanno due proprietà rilevanti: un valore di soglia (2) per il rifiuto dell'ipotesi nulla (p = 0.05) e un segno (-/+). Ciò significa che ordinando i loro valori in modo crescente o decrescente, a seconda che si considerino i valori sul polo "negativo" (-) o su quello "positivo" (+), è possibile verificare in che misura ciascun oggetto (ad es. regione o settore di attività) "caratterizza" le opposizioni che organizzano i vari spazi bi-dimensionali.

gli assi della Fig. 22[67], i secondi sono misure del "peso" che ogni punto ha nell'organizzazione delle bi-polarità fattoriali. Ad esempio, se consideriamo la polarità "-" (o semiasse negativo) del primo fattore, la tabella ci informa che i contributi maggiori sono dati dal settore Industria e dalle regioni Veneto e Lombardia; reciprocamente, se consideriamo la polarità "+" (o semiasse positivo), i contributi maggiori sono dati dal settore Istituzioni e dalle regioni Calabria, Sicilia e Lazio.

Ancora una nota: il fatto che il Lazio, pur avendo un contributo assoluto maggiore della Calabria, ha una posizione (coordinata) meno "estrema" sull'asse in questione, è determinato dal modo in cui vengono calcolati i contributi assoluti. Infatti, la formula che viene applicata include la "massa", cioè il peso relativo di ogni punto. In altri termini, facendo ricorso al modello (metafora) del sistema planetario, il "pianeta" Lazio ha una massa più grande del "pianeta" Calabria; per questa ragione il suo peso sull'altalena del fattore è più rilevante.

Analoghi rilievi possono essere fatti, ad esempio, per le regioni Lombardia e Veneto.

Ancora qualche nota iconografica: mentre il settore Industria risulta prevalentemente aggregato con regioni del "Nord", il settore Istituzioni risulta prevalentemente aggregato con regioni del "Sud".

A partire da queste descrizioni, anche completandole con quelle relative al secondo fattore (quello verticale), è possibile passare all'interpretazione dei dati, cioè – in qualche modo – all'iconologia. Evidentemente quest'ultima porterà a formulare ipotesi relativamente diverse a seconda che l'interprete sia uno storico, un economista, un politico o un sociologo. In altri termini, l'interpretazione dell'analisi iconografica è funzione dei modelli che ciascuno, nel suo settore di competenza, utilizza per fare inferenze.

Come sa chi ha pratica di queste cose, quando si interpretano i risultati di un'analisi fattoriale delle corrispondenze spesso si rischia di "perdersi" negli anfratti delle varie tabelle. Tenendo conto di questo fatto, e per facilitare il lavoro di analisi, T-LAB mostra in automatico delle tabelle ordinate con i valori test. Ad esempio la Tab. 19 riporta i valori test concernenti il primo fattore estratto mediante analisi delle corrispondenze (contesti elementari x unità lessicali) applicata al testo della Costituzione Italiana[68].

Come si può rilevare, su un lato "pesano" i temi della cosiddetta "società civile", sull'altro quelli delle "regole istituzionali". Evidentemente, l'espressione "come si può rilevare" è quanto meno impropria. Ancora una volta, si tratta di inferenze: osservo un gruppo di parole, stabilisco relazioni di somi-

[67] A questo scopo è sufficiente proiettare ogni punto sull'asse, cioè tirare una perpendicolare che parte dal punto considerato e incide sull'asse orizzontale.

[68] Per motivi di impaginazione, si riportano solo 25 unità lessicali per ogni polarità fattoriale.

*Tab. 19 – Costituzione Italiana: valori test del primo fattore*

| *Semiasse (–)* | | *Semiasse (+)* | |
|---|---|---|---|
| *Unità Lessicale* | *V.TEST* | *Unità Lessicale* | *V.TEST* |
| repubblica | -6,235 | camera | 7,515 |
| lavoro | -5,880 | componente | 6,585 |
| lavoratore | -5,742 | eleggere | 5,408 |
| diritto | -5,637 | presidente | 5,391 |
| economico | -5,606 | approvare | 4,976 |
| sociale | -5,425 | Presidente_della_Repubblica | 4,858 |
| libera | -5,410 | giorni | 4,449 |
| privato | -5,313 | maggioranza | 4,428 |
| riconoscere | -5,259 | elezione | 4,293 |
| diritti | -4,904 | parlamento | 4,158 |
| famiglia | -4,510 | mese | 4,026 |
| mezzi | -4,357 | maggioranza_assoluta | 3,826 |
| assicurare | -4,220 | prima | 3,678 |
| organizzazione | -4,074 | Governo | 3,585 |
| libertà | -4,068 | Camera_dei_Deputati | 3,526 |
| tutela | -3,991 | elettore | 3,345 |
| cittadino | -3,952 | votazione | 3,279 |
| favorire | -3,809 | nuovo | 3,252 |
| adempimento | -3,734 | senatore | 3,247 |
| internazionale | -3,588 | giorno | 3,245 |
| formazione | -3,566 | riunire | 3,190 |
| dovere | -3,459 | dimissione | 3,162 |
| rapporto | -3,416 | deputato | 3,153 |
| scuola | -3,393 | luogo | 3,069 |
| grado | -3,265 | Consiglio_Regionale | 3,053 |

glianza tra i loro significati, costruisco e denomino una categoria. Il tutto in base a modelli culturali più o meno condivisi.

Per molti versi, l'analisi delle corrispondenze *costruisce* dei risultati secondo una logica affine con un modo di funzionare della nostra mente: i fattori, infatti, possono essere considerati dei **principi di classificazione** (C.L. Burt, 1940), ovvero degli organizzatori delle relazioni tra dati che, secondo la logica della categorizzazione, mettono insieme cose simili, le distinguono dalle diverse, e mostrano "parentele" tra categorie di esse.

Scrive J.P. Benzecri (1984, p. 302), un matematico che ha molto contribuito a definire il modello di questa tecnica:

> Interpretare un asse fattoriale significa trovare ciò che vi è di analogo, da una parte

tra tutto ciò che è situato a destra dell'origine (o baricentro), dall'altra tra tutto ciò che è alla sinistra di questo, ed esprimere poi con concisione ed esattezza l'opposizione tra i due estremi.

Con questa affermazione, mentre descrive un metodo di interpretazione, di fatto l'autore ci comunica una specifica concezione dei fattori intesi come organizzatori di relazioni oppositive tra insiemi o classi ("tutto ciò" che sta a destra e "tutto ciò" che sta a sinistra dell'origine).

In effetti, nonostante la parola fattore evochi un significato di causa, le analisi di tipo fattoriale consentono "soltanto" di individuare un ordine nella complessità dei dati oggetto di trattamento, ovvero consentono di ridurre le dimensioni spaziali in cui questi possono essere rappresentati. Ma, evidentemente, una cosa è il senso statistico (o geometrico) dei fattori, altra cosa sono i modelli che, all'interno di ogni disciplina, fondano la possibilità di interpretarli. D'altro canto, se le scienze non cercassero di spiegare i fattori che generano un qualche tipo di ordine nei fenomeni oggetto di studio, esse stesse non avrebbero alcuna ragione di esistere.

Anche a partire da questo rilievo, può essere utile distinguere tra due tipi di fattori:

- quelli evidenziati dalle elaborazioni statistiche, corrispondenti alle strutture di tipo matematico e geometrico (le dimensioni) che organizzano le relazioni tra dati, e che sono stati denominati **principi di classificazione**;
- quelli che, attraverso l'uso di processi inferenziali e facendo riferimento a modelli teorici, vengono evocati per fondare interpretazioni e/o spiegazioni dei fattori intesi come principi di classificazione. Questi fattori – di secondo tipo – possono essere denominati **principi di spiegazione** e sono specifici di ogni disciplina scientifica.

### *2.3.5. Cluster analysis*

La denominazione *cluster analysis* indica una famiglia piuttosto numerosa di tecniche statistiche che analizzano le relazioni di somiglianza/differenza tra oggetti[69] di vario tipo allo scopo di classificarli secondo dei criteri non noti *a priori*. Più esattamente, l'obiettivo di queste tecniche è quello di costruire e distinguere gruppi che abbiano due caratteristiche complementari: al loro interno, la massima omogeneità (somiglianza) tra gli elementi che li costituiscono; tra di loro, la massima eterogeneità.

[69] Molte tecniche consentono di clusterizzare anche le variabili, cioè i profili che in genere corrispondono alle colonne delle tabelle dati.

Le differenze tra le tecniche di clusterizzazione (o classificazione) sono molteplici: alcune riguardano i tipi di variabili (categoriali, metriche, ecc.) utilizzate nelle tabelle dati, altre le misure di associazione tra i singoli elementi[70], altre le regole per costruire i gruppi, altre ancora i criteri per determinare una loro partizione ottimale. In ogni caso, le procedure di classificazione funzionano per approssimazioni successive e, come ricorda S. Bolasco (1999, p. 275), «non producono un risultato "certo" (in quanto molte scelte sono dovute all'analista), e non soddisfano condizioni di ottimalità assoluta, ma forniscono soluzioni con carattere di "ottimi" *relativi* e "locali"». Aspetto quest'ultimo che qualifica le tecniche di clustering come strumenti particolarmente utili nell'ambito di ricerche di tipo esplorativo e indiziario.

Prima di procedere nell'esposizione, un po' per "mostrare" quanto la logica della cluster analysis sia simile a un modo di funzionare della nostra mente, un po' per rassicurare il lettore circa i tipi di risultati che da essa ci si può attendere, propongo un esperimento che ha a che fare con la teoria dei *mondi possibili*[71].

Supponiamo di avere a disposizione solo quattro categorie di tipo binario (A, B, C, D) e che con queste dobbiamo organizzare la nostra conoscenza del mondo. Nel mondo possibile così definito, di ogni oggetto in esso presente o rappresentato potremmo solo dire se è o non è "A", se è o non è "B", se è o non è "C", se è o non è "D". Ne deriva che in questo mondo potremmo conoscere solo sedici tipi di oggetti e non altri[72]. In altri termini, gli oggetti "conoscibili" (a, b, c, d, e …p) corrisponderebbero ai vettori-riga della matrice in Tab. 20.

Ebbene, applicando a questa tabella un algoritmo di cluster analysis, otteniamo come risultato il dendrogramma in Fig. 23.

Come si può rilevare, a partire da sinistra verso destra, le aggregazioni dei 16 elementi è scandita regolarmente dalla binarietà delle loro caratteristiche. Inoltre, lo stesso algoritmo segnala che la migliore partizione degli elementi in questione è a quattro gruppi: tanti quante, nel nostro mondo possibile, sono le caratteristiche disponibili (A, B, C, D).

Nell'esempio appena fatto è stato applicato un algoritmo di tipo gerarchico. Come è noto, in letteratura[73], quando si parla di cluster analysis, in genere si usa distinguere tra due tipi di metodi, spesso usati in combinazione tra loro[74]:

[70] Una di queste è il coefficiente del coseno (vedi paragrafo 2.3.1.).

[71] Per la definizione di *mondo possibile* vedi paragrafo 1.2.

[72] Il numero 16 corrisponde a 2 elevato a potenza 4, dove 2 sta per la coppia degli eventi possibili (presenza /+/ o assenza /-/) e 4 per il numero delle categorie stipulate (A,B,C,D). Su questo argomento, vedi J. Hintikka (1973).

[73] Il lettore italiano può utilmente consultare i seguenti volumi: V. Cinanni (1990), A. Rizzi (1990), R. Biorcio (1993).

[74] L'uso combinato dei due metodi è caratteristico delle cosiddette "strategie miste".

*Tab.20 – Oggetti possibili*

| | A | B | C | D |
|---|---|---|---|---|
| *a* | 0 | 0 | 0 | 0 |
| *b* | 1 | 0 | 0 | 0 |
| *c* | 0 | 1 | 0 | 0 |
| *d* | 1 | 1 | 0 | 0 |
| *e* | 0 | 0 | 1 | 0 |
| *f* | 1 | 0 | 1 | 0 |
| *g* | 0 | 1 | 1 | 0 |
| *h* | 1 | 1 | 1 | 0 |
| *i* | 0 | 0 | 0 | 1 |
| *j* | 1 | 0 | 0 | 1 |
| *k* | 0 | 1 | 0 | 1 |
| *l* | 1 | 1 | 0 | 1 |
| *m* | 0 | 0 | 1 | 1 |
| *n* | 1 | 0 | 1 | 1 |
| *o* | 0 | 1 | 1 | 1 |
| *p* | 1 | 1 | 1 | 1 |

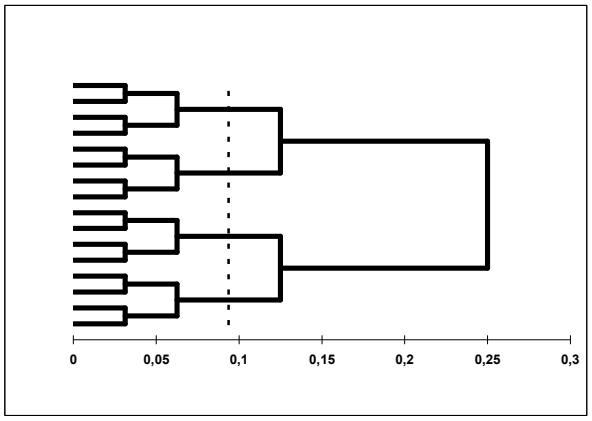



*Fig. 23 – Dendrogramma della Tab. 20*

- *hierarchical methods*, i cui algoritmi ricostruiscono l'intera gerarchia degli oggetti in analisi, vuoi in senso ascendente (associazioni progressive dei singoli elementi), vuoi in senso discendente (divisioni progressive dell'insieme considerato). In questo modo, ad ogni livello del cosiddetto "albero", ogni cluster risulta dalla fusione o dalla divisione di precedenti agglomerazioni;
- *partitioning methods*, i cui algoritmi prevedono che l'analista abbia preventivamente definito il numero di cluster in cui l'insieme degli oggetti in analisi va diviso. In questo modo, ogni singolo oggetto viene inizialmente attribuito a uno dei cluster, quindi subisce una serie di "traslochi" che termina quando lo spostamento dei singoli elementi da un cluster all'altro non migliora i valori dei parametri fissati.

Gli algoritmi di cluster analysis implementati negli strumenti T-LAB sono tre, due dei quali sono di tipo gerarchico[75]. Dei tre solo uno produce output visibili e interrogabili da parte dell'utilizzatore; quindi, in questa sede, mi soffermerò a considerare la logica di quest'ultimo.

Si tratta di un metodo di tipo gerarchico, più esattamente di tipo gerarchico-ascendente: cioè, parte dai singoli oggetti e – via via – utilizzando delle misure

[75] Più esattamente: nella funzione *Mappe dei nuclei tematici* è implementato un algoritmo gerarchico che, per trattare matrici di similarità (coefficiente del coseno) usa la tecnica del legame completo (*complete linkage*); nella funzione *Cluster analysis*, che analizza tabelle con le coordinate fattoriali generate da una precedente analisi delle corrispondenze, un diverso algoritmo gerarchico usa la tecnica di Ward; nella funzione *Tipologie dei contesti elementari* (e nella speculare *Estrazione dei contesti significativi*), oltre alle precedenti tecniche, è implementato un algoritmo del tipo K-Means.

di prossimità, li aggrega fino a ricomporre l'intero insieme. Detto in altri termini, parte dalle singole foglie per arrivare al tronco (o *radice*) dell'albero.

Gli "oggetti", in questo caso, sono costituiti da unità lessicali o da unità di contesto (i contesti elementari), e ciascuno di essi è caratterizzato da un profilo ottenuto tramite analisi delle corrispondenze: quello costituito dalle sue coordinate sui primi tre primi assi fattoriali. Il criterio di aggregazione utilizzato è il metodo di Ward[76], uno dei più adatti quando i profili riga sono costituiti da coordinate fattoriali.

Quanto al criterio di partizione che determina il numero di cluster ottenuti, in T-LAB è implementato un algoritmo che utilizza il rapporto tra varianza inter-cluster e varianza totale[77] e che – in modo automatico – assume come "partizione ottimale" quella in cui questo rapporto supera la soglia del 50%.

Per illustrare il funzionamento di questo metodo, possiamo tornare a considerare l'esempio dei dati ISTAT che, nel capitolo precedente, sono stati trattati mediante analisi delle corrispondenze. "A occhio", osservando la mappa in Fig. 22 (paragrafo 2.3.4), possiamo individuare dei raggruppamenti tra regioni: quelli più evidenti sono costituiti da coppie (Liguria-Lazio, Sicilia-Calabria), quelli meno evidenti sono costituiti da tre o più regioni (ad es. Lombardia-Piemonte-Veneto-ecc.). Per farci un'idea più precisa della situazione, possiamo tornare alla coordinate fattoriali (vedi Tab. 18): tre variabili che riassumono tutte le relazioni tra i dati in questione. In effetti, la tabella regioni x coordinate è un ottimo input per la cluster analysis[78].

La Fig. 24 riporta il dendrogramma ottenuto[79], cioè l'albero con *tutte* le sue diramazioni. I valori in basso sono indici che, a partire dalla sinistra (valore "0") consentono di verificare a quale livello del processo due o più elementi sono stati raggruppati. Il segmento tratteggiato indica la partizione ottimale suggerita dall'algoritmo utilizzato: tre cluster.

A questo punto, anche facendo ricorso ad ulteriori dati, potremmo commentare le relazioni tra il dendrogramma e la mappa ottenuta mediante analisi delle corrispondenze. Ad esempio, potremmo chiederci perché mai Sicilia e Calabria fanno gruppo a sé, mentre Liguria e Lazio risultano aggregate con altre 6 regioni. Il che ci porta a considerare almeno due aspetti:

[76] Questo metodo, ad ogni stadio, produce aggregazioni che producono il minimo aumento della varianza totale entro i gruppi.

[77] La varianza inter-cluster risulta dalla dispersione dei cluster nello spazio n-dimensionale; mentre la varianza totale risulta dalla sommatoria tra varianza inter-cluster e quella intra-cluster, la quale ultima risulta dalla dispersione degli elementi entro ciascun cluster.

[78] Ciò non significa che ogni applicazione dell'analisi delle corrispondenze deve essere seguita dall'applicazione di una cluster analysis. Ad esempio, in questo caso (dati ISTAT) il suo uso è giustificato solo da criteri didattici. Infatti, se l'obiettivo è quello di interpretare le relazioni tra regioni e settori di attività, la mappa in Fig. 22 e i dati della Tab. 18 sono più che sufficienti.

[79] La tecnica usata è quella di Ward.

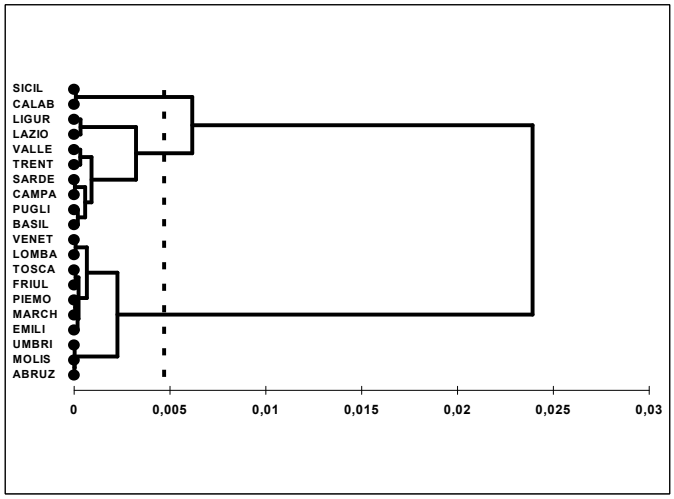



*Fig. 24 – Dati ISTAT 2001 – Dendrogramma*

a) la mappa ottenuta mediante analisi delle corrispondenze (Fig. 22) mostra un piano, cioè uno spazio a due dimensioni, con in ascissa e in ordinata rispettivamente le coordinate del primo e del secondo fattore; mentre nella cluster analysis, come *caratteristiche* di ogni regione, abbiamo preso in considerazione le coordinate di *tutti* e tre i fattori estratti;
b) il primo fattore (asse orizzontale) è quello che in modo più significativo (88%) determina le somiglianze/differenze tra le regioni. Ora, mentre Sicilia e Calabria hanno punteggi elevati (contributi assoluti) su questo fattore, Liguria e Lazio hanno punteggi elevati solo sul secondo fattore che, nell'organizzazione delle somiglianze/differenze, "pesa" solo l'11%.

Torniamo ora a T-LAB e al modello di cluster analysis che stiamo considerando.

In questo caso, le tabelle analizzate sono costituite da ben più di 20 oggetti: normalmente si tratta di centinaia di unità lessicali; quindi, per default, viene mostrato un dendrogramma che include solo le ultime nove diramazioni dell'albero (Fig. 25)[80].

Al termine di ogni elaborazione, per ogni cluster è possibile verificare quali sono gli oggetti che ne fanno parte e quali sono le loro caratteristiche; inoltre l'utilizzatore può agevolmente esplorare otto partizioni diverse, da 2 a 9 cluster, sia mediante grafici che mediante tabelle. Poiché la cluster analysis usa le coordinate fattoriali ottenute mediante una precedente analisi delle corrispondenze, i grafici sono analoghi alle mappe prodotte da quest'ultima (Fig. 26).

[80] Gli indici con i livelli di associazione sono mostrati in altri grafici T-LAB.

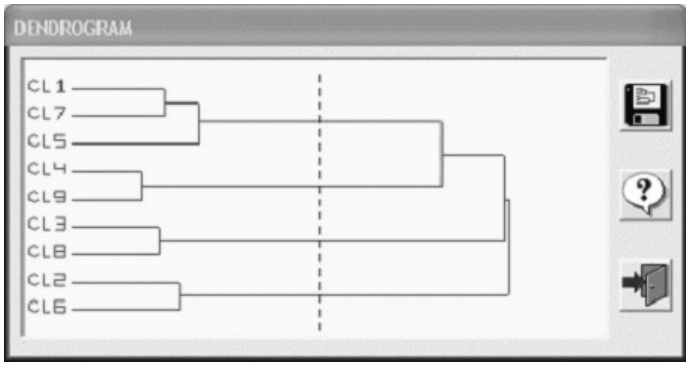



*Fig. 25 – T-LAB: esempio di dendrogramma*

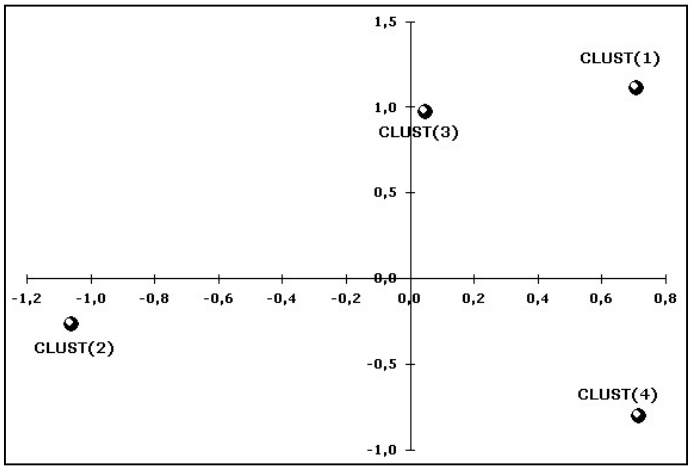



*Fig. 26 – T-LAB: cluster sul piano fattoriale*

Quanto alle tabelle, la loro struttura è di due tipi, a seconda del tipo di oggetti classificati. In alcuni casi, infatti, i profili riga sono costituiti da unità lessicali (ad es. modello lemmi x variabile), in altri sono costituiti da contesti elementari (ad es. modello segmenti x lemmi). Nel primo caso, per ogni cluster, le tabelle riportano le unità lessicali che ne fanno parte, ordinate per occorrenze decrescenti; nel secondo caso riportano le unità lessicali ordinate secondo valori decrescenti del chi quadro, che indicano la tipicità di ciascuna di esse entro il cluster corrispondente[81].

[81] In questo secondo caso, poiché gli "oggetti" clusterizzati sono i contesti elementari, le singole parole possono appartenere a due o più cluster.

*Tab. 21 – Articoli su giovani e nuove tecnologie (cluster)*

| *CLUSTER 1* | | *CLUSTER 2* | | *CLUSTER 3* | | *CLUSTER 4* | |
|---|---|---|---|---|---|---|---|
| *Unità Less.* | *Occ.* | *Unità Less.* | *Occ.* | *Unità Less.* | *Occ.* | *Unità Less.* | *Occ.* |
| computer | 105 | scuola | 139 | bambino | 155 | tecnologia | 60 |
| apprendimento | 98 | società | 54 | adulto | 75 | sapere | 52 |
| nuovo | 91 | storia | 41 | rapporto | 34 | comunicazione | 46 |
| mondo | 73 | simulazione | 39 | infanzia | 30 | corpo | 41 |
| ragazzo | 69 | sociale | 34 | cultura | 28 | capacità | 37 |
| insegnante | 59 | ambiente | 32 | media | 28 | educativo | 32 |
| diverso | 57 | linguaggio | 30 | violenza | 24 | costruire | 28 |
| esperienza | 56 | capire | 28 | multimedialità | 21 | utilizzare | 27 |
| vita | 55 | vedere | 27 | multimediale | 21 | dati | 23 |
| studente | 51 | domanda | 27 | famiglia | 20 | sviluppo | 22 |
| realtà | 48 | produrre | 25 | tv | 20 | cognitivo | 21 |
| strumento | 43 | avvenire | 25 | infantile | 18 | uomo | 21 |
| diventare | 42 | risposta | 23 | uomo | 21 | percorso | 20 |
| internet | 41 | vivere | 23 | leggere | 13 | costruzione | 20 |
| usare | 41 | piccolo | 17 | dinamico | 13 | gruppo | 19 |
| rete | 40 | intelligenza | 17 | comportamento | 13 | laboratorio | 19 |
| imparare | 39 | familiare | 16 | proporre | 12 | attività | 19 |
| giovane | 37 | fenomeno | 21 | scrittura | 12 | progetto | 18 |
| informazione | 35 | futuro | 21 | produzione | 12 | sviluppare | 17 |
| apprendere | 34 | insegnare | 21 | sentimento | 11 | collaborazione | 17 |

Ad esempio, la Tab. 21 riporta le caratteristiche (le prime 20) di quattro cluster ottenuti analizzando un corpus costituito da 12 articoli concernenti il rapporto tra giovani e nuove tecnologie. In questo caso, poiché la tabella analizzata è del tipo lemmi x variabili, per ogni cluster sono riportate le occorrenze delle prime unità lessicali in elenco[82].

La Tab. 22 riporta le caratteristiche (i primi 15 lemmi in elenco) di quattro cluster ottenuti analizzando un corpus costituito da 48 articoli concernenti l'Islam, pubblicati tra il gennaio 2001 e il febbraio 2002. In questo caso, poiché la tabella analizzata è del tipo contesti elementari x lemmi, per ogni cluster sono riportati tre valori: il chi quadro ottenuto con il metodo illustrato nel paragrafo 2.3.2 e il numero di contesti elementari – sia del cluster ("CLUS")

[82] Negli output T-LAB l'elenco delle unità lessicali è completo.

che totali (“TOT”) – in cui ogni unità lessicale risulta presente.

*Tab. 22 – Articoli sull'Islam (cluster)*

| *CLUSTER 1* | | | | *CLUSTER 2* | | | |
|---|---|---|---|---|---|---|---|
| *Unità Lessicale* | *CHI²* | *CLUS* | *TOT* | *Unità Lessicale* | *CHI²* | *CLUS* | *TOT* |
| Stati_Uniti | 89,39 | 53 | 64 | moschea | 83,09 | 47 | 61 |
| militare | 76,66 | 51 | 65 | Islam | 73,35 | 76 | 129 |
| attentato | 68,17 | 46 | 59 | cultura | 70,96 | 28 | 30 |
| aereo | 56,19 | 28 | 31 | musulmano | 67,06 | 56 | 87 |
| americano | 52,84 | 58 | 91 | Roma | 58,51 | 25 | 28 |
| terroristico | 50,41 | 27 | 31 | arabo | 51,22 | 63 | 114 |
| Israele | 47,87 | 31 | 39 | cristiano | 51,16 | 21 | 23 |
| intelligence | 41,31 | 17 | 17 | fede | 49,35 | 23 | 27 |
| attacchi | 35,85 | 18 | 20 | papa | 46,81 | 18 | 19 |
| Bush | 34,98 | 24 | 31 | scrittore | 45,68 | 16 | 16 |
| alleato | 34,73 | 16 | 17 | pregare | 41,81 | 19 | 22 |
| terrorismo | 33,06 | 40 | 65 | immigrato | 35,46 | 14 | 15 |
| sicurezza | 28,73 | 15 | 17 | chiesa | 32,62 | 13 | 14 |
| Washington | 26,85 | 18 | 23 | italiano | 32,61 | 21 | 29 |
| Bin_Laden | 24,03 | 51 | 99 | studi | 30,07 | 18 | 24 |
| *CLUSTER 3* | | | | *CLUSTER 4* | | | |
| *Unità Lessicale* | *CHI²* | *CLUS* | *TOT* | *Unità Lessicale* | *CHI²* | *CLUS* | *TOT* |
| talebani | 90,59 | 35 | 49 | donna | 69,41 | 57 | 87 |
| mullah | 84,77 | 22 | 24 | raccontare | 58,57 | 24 | 26 |
| afghano | 76,34 | 28 | 38 | bambino | 49,62 | 27 | 34 |
| scuola | 74,85 | 24 | 30 | morire | 43,04 | 21 | 25 |
| studente | 66,50 | 21 | 26 | ferito | 41,90 | 24 | 31 |
| Afghanistan | 55,61 | 29 | 48 | ragazzo | 39,37 | 23 | 30 |
| Pakistan | 55,58 | 21 | 29 | famiglia | 37,28 | 20 | 25 |
| Madrasse | 45,96 | 13 | 15 | morto | 35,43 | 28 | 42 |
| Kabul | 44,11 | 18 | 26 | esplodere | 33,71 | 15 | 17 |
| insegnamento | 42,36 | 11 | 12 | piccolo | 31,87 | 20 | 27 |
| Al_Jazeera | 38,14 | 14 | 19 | povero | 28,06 | 10 | 10 |
| insegnare | 38,03 | 10 | 11 | sangue | 22,85 | 11 | 13 |
| lingua | 35,17 | 17 | 27 | fermare | 22,45 | 8 | 8 |
| Corano | 33,83 | 21 | 38 | suicida | 22,36 | 12 | 15 |
| Sharia | 26,30 | 15 | 26 | corpo | 20,97 | 9 | 10 |

Per molti versi, in entrambi i casi esemplificati, i cluster possono essere considerati delle **isotopie** (*iso* = uguale; *topos* = luogo)[83], sia nell'accezione della semantica strutturale che da un punto di vista strettamente geometrico.

Da un punto di vista semantico, la nozione di isotopia rinvia infatti a una concezione del significato come "effetto del contesto", cioè come qualcosa che non appartiene alle parole prese singolarmente, bensì che risulta dai loro rapporti all'interno delle unità sintagmatiche. D'altro canto, come si può rilevare (Tab. 22), ciascun cluster individua un contesto di riferimento "condiviso" da più parole, ma che non deriva dai loro specifici significati. Di conseguenza, ciascun cluster consente di ricostruire "un filo" del discorso all'interno della trama complessiva costituita dal corpus in analisi o da un suo sottoinsieme.

Da un punto di vista geometrico, i cluster in questione sono ottenuti mediante analisi delle relazioni tra profili di coordinate spaziali: quelle risultanti dall'analisi delle corrispondenze. Ne deriva che – letteralmente – gli elementi di ciascun cluster (unità lessicali e unità di contesto) "abitano" in luoghi prossimi tra loro.

Tuttavia va ricordato che il riconoscimento di un'isotopia non è la mera constatazione di un "dato", bensì il risultato di un processo interpretativo (F. Rastier 1987, pp. 11-12). Ciò porta a riflettere sulla particolare natura del rapporto tra dati e strumenti di osservazione nell'ambito dell'analisi dei testi. In effetti, a partire dal momento in cui diventano oggetti di osservazione, *i dati testuali sono costruiti dagli strumenti che utilizziamo per analizzarli*[84]; cioè diventano rappresentazioni del rapporto tra due tipi di modelli culturali, quelli "contenuti" nei testi e quelli utilizzati da chi li analizza.

[83] La nozione di isotopia è stata inizialmente proposta da A. J. Greimas (1966) per definire la ricorrenza, all'interno delle unità sintagmatiche (frasi e/o testi), di più parole con tratti semantici in comune tra loro. Questi ultimi, denominati "semi contestuali" o "classemi", consentono di ricondurre varie parole entro una categoria (o classe) di contenuto.

[84] Ad esempio, le analisi di tabelle con valori di occorrenza o di co-occorrenza portano a riconoscere e a costruire isotopie di tipo diverso.

# Parte seconda

# 1. Strategie e percorsi di analisi

## 1.1. Testimonianze di utilizzatori

Questo paragrafo è stato pensato con un preciso intento: comunicare l'idea che questo libro – come T-LAB – ha un'architettura aperta.

In effetti, lo sviluppo del software deve molto alle osservazioni e ai suggerimenti degli utilizzatori. Con molti di questi intrattengo rapporti di collaborazione e, in più di un caso, sono diventato partner delle loro attività di ricerca.

Stando queste premesse, nel marzo 2004, quando il mio lavoro di scrittura era ormai a compimento, ho pensato di rivolgere un invito ad alcuni ricercatori che usano T-LAB in diversi contesti e con differenti obiettivi, in Italia e all'estero.

Via e-mail, ho chiesto loro di compilare una scheda in cui fossero evidenziati quattro punti. Testualmente:

- una presentazione di sé, del proprio contesto di appartenenza e dei progetti a cui sta lavorando;
- una descrizione dei propri interessi per l'analisi dei testi, anche facendo riferimento ai tipi di materiali che in genere si trova ad analizzare (es. articoli, interviste, risposte a domande aperte, ecc.);
- una descrizione degli obiettivi per i quali usa, o pensa di usare, T-LAB;
- una descrizione di alcuni risultati conseguiti (o attesi), eventualmente con riferimenti a pubblicazioni.

Il risultato ottenuto, oltre a superare le mie aspettative, costituisce un documento di particolare interesse, sia per la varietà degli "oggetti di indagine" che per la varietà degli approcci metodologici. Inoltre, le diciotto testimonianze[1] consentono di fare più di qualche ipotesi sulle modalità in cui i ricercatori propongono *prodotti* che rispondono a una domanda sociale.

A ciascuno degli autori, come anche a tutti gli utilizzatori di T-LAB, va un mio particolare ringraziamento.

[1] Quattro di queste, redatte da ricercatori stranieri, sono state tradotte in lingua italiana.

**Rosetta Zan**
**Università degli Studi di Pisa**

Sono docente universitario nel settore delle Matematiche Complementari, e coordino un gruppo di ricerca nel campo della didattica della matematica.

I nostri interessi riguardano le difficoltà in matematica, ed il ruolo che hanno nella nascita e nel superamento di tali difficoltà fattori quali convinzioni, emozioni, atteggiamenti. Si tratta di costrutti "importati" da altre discipline, anche se rielaborati per gestire le problematiche tipiche dell'apprendimento-insegnamento della matematica.

L'approccio metodologico a questi aspetti è stato per tanto tempo limitato all'uso di questionari con domande a scelta multipla, che consentivano l'osservazione di un gran numero di soggetti, e che rendevano particolarmente semplice l'analisi dei dati. In realtà l'uso esclusivo di tali strumenti è stato messo in discussione da molti punti di vista, ed è stata sottolineata la necessità di utilizzare (anche) strumenti di osservazione meno rigidi e chiusi, quali questionari con domande aperte, interviste, temi, diari. Naturalmente l'analisi di questo tipo di dati, soprattutto se il campione è comunque numeroso, è più problematica.

In questo momento stiamo affrontando questo tipo di problemi all'interno di due progetti.

Il primo è un progetto nazionale triennale che coordino, finanziato dal Miur nell'ambito del FIRB, cui partecipano ricercatori delle Università di Cagliari, Genova, Alessandria, Napoli, Modena, e al quale collaborano insegnanti di vari livelli di scuola. Il progetto, dal titolo *L'atteggiamento negativo verso la matematica: analisi di un fenomeno allarmante per la cultura del terzo millennio*, si pone l'obiettivo di indagare sulle origini dell'atteggiamento negativo nei confronti della matematica, e sui fattori che ne influenzano la formazione. Ciò attraverso: a) l'osservazione nell'arco dei tre anni del progetto di un campione di allievi; b) indagini su altri soggetti, quali insegnanti, adulti in generale, matematici di professione.

La metodologia prevede, oltre a osservazioni dirette, l'utilizzazione di questionari, diari, interviste, temi.

Il materiale su cui abbiamo finora utilizzato T-LAB riguarda la prima attività prevista nel progetto: un'indagine mirata ad evidenziare le conseguenze che un cattivo rapporto con la matematica, alla fine della scuola superiore, può avere sulla scelta degli studi universitari. La modalità con cui abbiamo indagato questo aspetto è un questionario. La domanda centrale (con risposte a scelta multipla) è: «Il fatto che in un corso universitario ci sia un esame di matematica come influisce sulla tua scelta dell'università?» (a). Altre domande riguardano:

b) il voto in matematica;
c) se la matematica piace (per niente / poco / abbastanza / molto);

d) tre aggettivi associati alla matematica.

Abbiamo raccolto 2.000 questionari, da studenti appartenenti agli ultimi due anni di scuole superiori del territorio nazionale.

Per quanto riguarda l'analisi realizzata con T-LAB, ogni questionario è stato rappresentato da un testo (lungo tre parole, cioè i tre aggettivi) e identificato dalle specifiche modalità relative alle variabili individuate dalle domande "a", "b", "c".

In questo modo siamo riusciti a:

- identificare, a seconda degli aggettivi inseriti, 5 cluster di studenti;
- avere informazioni sul rapporto con la matematica dei ragazzi sui quali l'influsso della matematica è stato determinante o negativo;
- ricavare ulteriori informazioni confrontando le diverse modalità delle variabili (ad esempio è emerso che il lemma più ricorrente sull'intero corpus è "interessante", ma all'interno del gruppo degli studenti corrispondenti ai voti più bassi è invece "noiosa").

Un'ulteriore analisi sarà orientata a definire una lemmatizzazione diversa degli aggettivi, ad esempio distinguendo fra aggettivi che corrispondono ad emozioni suscitate dalla materia ("noiosa", "divertente", ecc.) e convinzioni sulle caratteristiche della stessa ("utile", "infondata", ecc.).

Il secondo progetto per cui stiamo esplorando le possibilità di T-LAB è invece legato all'analisi di un campione di circa 200 temi sulla matematica (dal titolo *Io e la matematica: descrivi il tuo rapporto con la matematica*), finora condotta senza aiuto di software. In una prima utilizzazione, T-LAB ci ha permesso di individuare i lemmi più significativi, e di selezionare per ognuno di essi i temi che meglio li identificano.

---

**Priscilla Murphy**
**Temple University – U.S.A.**

---

Sono una professoressa del Department of Strategic and Organizational Communication, all'interno della School of Communications and Theater, presso la Temple University.

Precedentemente, ho lavorato come Executive Director of Research and Granting, Associate Dean for Research and Graduate Programs e Director of the doctoral program in Mass Media and Communication.

Prima di intraprendere la carriera accademica, sono stata Vice Presidente delle Pubbliche Relazioni presso PaineWebber Group Inc. (New York City), dove ero responsabile di tutte le comunicazioni esterne, incluse le relazioni con i media, le comunicazioni finanziarie e il supporto al marketing.

Mi sono laureata presso lo Smith College e ho conseguito un dottorato in letteratura inglese presso la Brown University.

Ho pubblicato molti contributi sulle strategie dei media, sugli activist groups e sull'ambientalismo. Collaboro con le redazioni del Journal of Public Relations Research e della Public Relations Review. Insegno in molti corsi sui temi della comunicazione e sono consulente di varie aziende per la gestione delle comunicazioni in momenti di crisi

Avendo lavorato in azienda sui temi della comunicazione e delle relazioni con i media, mantengo un interesse per i temi "costruiti" dai media e per la costruzione sociale delle controversie in cui sono in gioco interessi di gruppi di pressione. In generale, ho usato la computer-assisted text analysis per estrarre strutture di senso dai pubblici dibattiti su problemi politici, sia per mostrare come i gruppi di interesse li usano per manipolare le decisioni politiche, sia per evidenziare aree di convergenza tra le parti in conflitto. Inoltre, il mio interesse per l'analisi dei testi ha a che fare con una concettualizzazione di ampio respiro concernente le strutture di senso implicate nelle reti di influenza, in base alla quale la loro diffusione attraverso i media può essere ricostruita per mezzo dell'analisi dei testi.

Le ragioni per le quali prediligo il tipo di analisi consentita da T-LAB sono due. La prima: il testo viene fatto parlare "da sé", senza pre-concetti (consapevoli o inconsapevoli) imposti dal ricercatore. In particolare, questo approccio consente che pattern inerenti la massa indifferenziata dei testi emergano senza interventi esterni imposti dalle codifiche del ricercatore. La seconda: da un punto di vista teorico, un vantaggio di T-LAB è che la nozione di contesto è inscritta nello stesso metodo di ricerca, cioè è radicata nelle relazioni tra tutte le parole all'interno del testo. Per queste ragioni, gli approcci tipo T-LAB diminuiscono gli effetti delle aspettative del ricercatore e, allo stesso tempo, consentono l'emersione di connessioni inattese.

Qualche tempo fa, ho usato software per l'analisi dei testi per esaminare testimonianze congressuali di parti politiche concernenti la regolazione della pubblicità del tabacco, la dipendenza da nicotina e i test genetici. Inoltre, con i miei studenti, ho analizzato i reportage (statunitensi e europei) sui cibi geneticamente modificati e le rappresentazioni del rischio ambientale proposte da popolari quotidiani statunitensi.

Attualmente sto usando T-LAB per due progetti di ricerca:

- per la Robert Wood Johnson Foundation sto lavorando a un progetto di Web-based text-mining teso ad esaminare, nel corso degli ultimi 20 anni, i cambiamenti delle strategie manageriali determinati dall'intensificazione della legislazione anti-tabacco. In questo caso, ci aspettiamo che le analisi realizzate con T-LAB ci aiutino a comprendere meglio le strategie fondanti gli sforzi delle aziende produttrici di tabacco tesi ad influenzare i pubblici discorsi su loro stesse e sui loro prodotti.
- in un progetto finanziato dal National Cancer Institute sto usando T-LAB per estrarre temi da una serie dei focus group in cui donne di minoranza etnica, residenti in aree urbane degradate, discutono le loro opinioni sul cancro alla mammella e sui mammogrammi, ciò allo scopo di progettare una campagna di prevenzione.

**Daniele Nigris**
**Università degli Studi di Trieste**

**Daniele Nigris,** ricercatore e docente di *Metodologia e tecnica della ricerca sociale*, Università di Trieste (www.cspn.units.it/nigris, e-mail: nigris@tin.it).

*Attività scientifiche.* La mia attività di ricerca si realizza in tre ambiti:
1) gli aspetti logico-epistemologici della ricerca empirica nelle scienze sociali;
2) le problematiche metodologiche e tecniche delle ricerche non-standard;
3) la ricerca empirica nell'ambito della sociologia della salute, e sulle connesse problematiche dell'identità.

L'analisi dei testi ricade soprattutto nei punti 2 e 3.

*I miei interessi rispetto all'analisi dei testi.* I miei interessi per l'analisi testuale sono di carattere sia metodologico, sia sostantivo.

Versante metodologico: ho di recente proposto in alcuni passaggi di *Standard e non-standard nella ricerca sociale* (Angeli 2003) delle considerazioni sulla necessità di un'analisi integrata (sia ermeneutica, sia automatizzata) dei testi. Considero questo approccio la maniera migliore di completare reciprocamente le informazioni forniteci da due strategie così diverse sia per logiche che per fini.

Versante sostantivo: come Coordinatore scientifico della XII Sezione tematica della SISS (Società Italiana di Sociologia della Salute), dedicata allo studio de "*I vissuti della malattia. Quotidianità, soggettività, narrazione*", sto lavorando con i colleghi alla crescita di una piccola comunità di specialisti che si occupino di ricerca narrativa sulla salute e sulla malattia. Per tutto ciò che è *illness narrative*, e quindi interrogazione non strutturata e non direttiva, racconto di vita di due-quattro ore di registrazione, ritengo indispensabile avvalersi di entrambe le strategie di analisi del testo, a prescindere dall'ottica accademica oppure operativa di chi faccia ricerca.

*Ricerche in corso e obiettivi per i quali uso T-LAB.* Sto terminando una ricerca realizzata per conto della filiale italiana dell'*European Dialysis and Transplant Nurses Association & European Renal Care Association* sui vissuti dei pazienti dializzati, e degli infermieri di dialisi. I risultati verranno presentati nel volume *Vivere con la dialisi* (titolo provvisorio), che uscirà per FrancoAngeli. Sia per la ricerca in corso, sia per un'altra che vedrà l'avvio in autunno (anch'essa sulle persone dializzate) uso in maniera integrata T-LAB e ATLAS.ti, nell'ottica descritta sopra.

Un secondo obiettivo, di carattere metodologico, per il quale uso T-LAB, è – dato il grande numero di possibilità di analisi offerte dal programma – quello di approfondire il mio ragionamento su quali siano, per le scienze umane, le tecniche di analisi più adeguate a specifiche basi empiriche.

Non sono convinto cioè che ogni tecnica, e ogni logica, vadano egualmente bene per ogni fine cognitivo, ma credo anzi che dovrebbe essere parte essenziale del bagaglio di un ricercatore che lavori su testi, e che ne faccia un trattamento anche automatizzato, la comprensione di quali siano gli strumenti più adeguati alle sue necessità conoscitive e ai vincoli della base empirica da cui parte. Perché i casi, a mio parere, sono due: o si mantiene una visione ingenua e sacralizzata del "testo", oppure bisogna accettare serenamente che testo scritto / testo *tra*scritto; testo monolingue / testo multilingue; testo di 80 / 500 / 13.800 / 75.000 parole; testo *mono*logico / testo *dia*logico; *30 testi / 1300 testi*; sono oggetti diversi – e cioè basi empiriche testuali che pongono al ricercatore problemi differenti, e che richiedono certe – o sopportano solo certe – tecniche di analisi.

E penso soprattutto che *uno specifico* corpus che risulti dalla combinatoria di alcuni degli elementi ora accennati – insieme alle problematiche specifiche del campo di indagine dato – sia *un tipo* di base empirica a sé, per la quale potranno essere maggiormente adeguate certe tecniche e certe strategie, anziché altre. È in questa direzione, mi sembra, che la riflessione metodologica può costituire – al di là di banali guerre di religione scientifica – un terreno di incontro fecondo tra ermeneuti del testo e statistici del testo.

---

**Christophe Lejeune**
**Université de Liège (Belgio)**

---

**Christophe Lejeune:** sociologo presso il Service de Méthodologie et d'Épistémologie des Sciences Sociales dell'Università di Liegi (Belgio), membro del gruppo Sociologie Pragmatique et Réflexive della École des Hautes Etudes de Paris (Francia).

Il suo sito http://www.smess.egss.ulg.ac.be/lejeune/ presenta recensioni di molti software utili per la ricerca qualitativa nell'ambito delle scienze sociali.

Le sue ricerche hanno per oggetto le nuove tecnologie e, più precisamente, la comunicazione tra gli autori di software non commerciali. Lo studio del modo in cui questi congiungono il loro sforzi passa attraverso l'osservazione delle loro interazioni sui forum di discussione (Bulletin Boards). Il metodo di analisi prevede l'uso combinato di tre approcci: l'analisi della conversazione, il metodo qualitativo Prospéro e il metodo quali-quantitativo T-LAB. Tre strumenti che, rispettivamente, portano a focalizzare il livello (micro) del messaggio, quello (meso) della discussione e quello (macro) del corpus.

Al livello del messaggio, l' analisi conversazionale sviluppata dagli etnometodologi permette di esplicitare le istanze della comunicazione. In particolare, lo studio sequenziale dei "turni" evidenzia le attese degli attori e le conoscenze informali date per scontate. I risultati di questa indagine consentono di costruire un repertorio sia delle inferenze che gli attori giudicano *normali* sia dei loro rispettivi universi di riferimento. Inoltre, all'interno di questa analisi, il ri-

corso al dispositivo della categorizzazione dei membri (MCD) porta a definire il campo di variazione delle denominazioni delle entità evocate dagli attori (persone, avvenimenti, concetti o oggetti).

La seconda fase dell'indagine si colloca al livello delle discussioni. In questo caso, ciò che viene studiato non più è il passaggio da un messaggio all'altro, bensì la consistenza di più interventi concatenati che costituiscono il filo della discussione. Il metodo qualitativo Prospéro (software di sociologia pragmatica, sperimentale e riflessiva), utilizzando i risultati della fase precedente, permette di individuare le concatenazioni argomentative dei vari attori e di integrarle in un nuovo tipo di analisi. I risultati del primo tipo (i campi di variazione referenziale) sono utilizzati per elaborare delle costruzioni (qualificate come convenzionali) che aggregano le differenti espressioni designanti lo stesso tipo di entità. A un secondo livello, gli universi di riferimento servono ad alimentare le categorie di analisi implementate nello strumento. Il terzo tipo di risultati della prima fase (concernenti le attese e le inferenze della comunità analizzata) orienta l'esplorazione delle diverse discussioni. In questo modo, il suo interesse empirico si trasforma in funzione euristica, cioè aiuta il ricercatore a diagnosticare la consistenza o le divergenze di una discussione (il fatto che uno scambio di messaggi riguardi lo stesso soggetto o scarti dal tema introdotto nel contributo iniziale).

La terza fase del processo, che si colloca al livello del corpus, fa ricorso al metodo T-LAB. I diversi grafici del software permettono di evidenziare le tendenze comuni all'interno dell'insieme del corpus. La loro lettura e la loro interpretazione sono integrate dalle conoscenze estratte dalle fasi precedenti. Inoltre T-LAB consente un ritorno riflessivo sull'esplorazione etnometodologica e sull'analisi qualitativa. In questo modo, le tendenze estratte orientano il ricercatore nel suo percorso.

Anche se, per ragioni didattiche e analitiche, le tre fasi sono state distinte, di fatto esse non si succedono linearmente nel corso del processo di ricerca. Proprio per questa ragione sono interessanti i ritorni che T-LAB consente sui livelli empirici precedenti. In quanto tale, l'uso di questo strumento costituisce un notevole supporto euristico: la visione sinottica che offre del corpus permette di formulare ulteriori ipotesi e queste possono essere esaminate sia al livello delle discussioni che al livello degli interventi. Allo stesso modo, i livelli precedenti orientano l'interpretazione e portano il ricercatore a costruire ulteriori tabelle, le quali – a seconda dei casi – possono confermare o invalidare le ipotesi precedentemente formulate.

In conclusione, l'articolazione dei tre metodi permette una integrazione dei livelli del messaggio (micro), della discussione (meso) et del corpus (macro). Piuttosto che rappresentare una architettura lineare e per sommatorie, il percorso così definito rende conto della mutua fecondazione dei differenti livelli di analisi.

*Riferimenti bibliografici*

Lejeune Christophe, 2001, Du mode de définition de deux programmes de recherche en sociologie et en ethnométhodologie, *Carnets de bord*, 2, p. 56-66.

Lejeune Christophe, 2002, Indexation et organisation de la connaissance. La régulation des décisions sur un forum de discussion, *Les cahiers du numérique*, 3, n° 2, p. 129-144.
Marie-Christine Bureau, Christophe Lejeune, Didier Torny, Patrick Trabal, Francis Chateauraynaud, 2003, *Internet à l'épreuve de la critique. Rumeurs, alertes et controverses au cœur des nouvelles technologies*, Paris, EHESS-CNRS.
Lejeune Christophe, 2004, Représentations des réseaux de mots associés, *Le pouvoir des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, p. 726-736.

---

**Silvio Ripamonti**
**Università Cattolica del Sacro Cuore di Milano**

---

**Silvio Ripamonti**, Università Cattolica del Sacro Cuore di Milano, Dipartimento di Psicologia, Area di Psicologia Applicata, Laboratorio "Culture Organizzative e di Consumo".

All'interno dell'Area di Psicologia Applicata del Dipartimento di Psicologia Università Cattolica del Sacro Cuore di Milano ho partecipato allo sviluppo di progetti di ricerca nell'ambito della psicologia del lavoro e dell'organizzazione. I tre principali temi di ricerca che sono oggi attivi e per i quali è stato utilizzato, o si prevede di poter utilizzare, T-LAB sono i seguenti:

*1) I processi di ristrutturazione aziendale nelle narrazioni dei manager.* L'obiettivo di questa ricerca, all'interno della quale ho svolto il mio progetto di dottorato, è l'analisi e lo studio dei processi di cambiamento organizzativo di natura straordinaria (reengineering, acquisizioni, fusioni, downsizing). Attraverso interviste qualitative di tipo narrativo rivolte ai manager, si è cercato di indagare il sistema di rappresentazioni sul processo di azione e di comprendere quali condizioni stanno alla base di una efficace integrazione tra i progetti individuali ed un contesto organizzativo in transizione, che non offre elementi di sicurezza.

La prima fase della ricerca ha prodotto un'ingente mole di dati, che sono stati analizzati con T-LAB, insieme ad altri strumenti di analisi di contenuto.

*2) Apprendimento organizzativo e condivisione delle conoscenze all'interno delle imprese sociali non profit.* Il progetto si propone di comprendere e descrivere le culture lavorative, professionali ed organizzative legate al fare ed essere impresa sociale. Le nuove imprese sociali infatti richiedono una massimo di attivazione nella ridefinizione delle modalità di lavoro, dei rapporti interni, nell'attivazione di nuove modalità di scambio con l'ambiente e nuovi codici culturali, nonché affettivi e simbolici.

*3) Costruzione dell'identità professionale degli operai nelle aziende di eccellenza*. In molte realtà industriali di alto profilo la figura dell'operaio può assumere connotazione di elevata professionalità e specializzazione contribuendo in maniera sostanziale al successo dell'organizzazione. Tuttavia la percezione del ruolo operaio nell'immaginario collettivo non è "positiva" e questa professione viene scelta con molta resistenza, soprattutto da parte di chi ha una scolarità medio-alta. Alcune organizzazioni ci hanno chiesto di provare ad analizzare i sistemi simbolici e culturali diffusi nel mondo operaio, le modalità con cui si costruisce l'identità professionale.

Nell'ambito di questi progetti di ricerca-intervento, spesso ci ritroviamo ad utilizzare colloqui ed interviste non strutturate, volti ad indagare come le persone intervistate "costruiscono il senso" del processo sociale che li ha coinvolti e "negoziano una costruzione condivisa della realtà" con gli altri attori, come premessa per orientare l'azione.

Le interviste, quando possibile, sono condotte secondo il modello narrativo: si chiede agli intervistati di rievocare in forma di storie, episodi ed eventi paradigmatici, la loro esperienza.

I testi così raccolti rappresentano di per sé un veicolo di comprensione e conoscenza delle realtà organizzative e sociali con cui entriamo in contatto; tuttavia si avverte con forza l'esigenza di utilizzare strumenti euristici in grado di dare un fondamento sistematico e di rendere più chiare e comunicabili le ipotesi e le teorie emergenti. In particolare, l'analisi del testo attraverso T-LAB consente di fare esplicito riferimento a modelli psicologici di interpretazione del materiale testuale raccolto, evitando di arrestarsi ad una pre-comprensione, o una conoscenza preriflessiva della realtà osservata con questo tramite in base ad una considerazione solo parziale dei dati a disposizione, o di elementi dotati di particolare salienza o che evocano assonanze definite solo per il ricercatore individuale.

Grazie alla grande versatilità d'uso di T-LAB per le analisi di corpus testuali di ingenti dimensioni, come quelli da noi raccolti, abbiamo potuto analizzare le culture ed i sistemi di significati condivisi presenti in diversi contesti. È stata particolarmente utile, ai fini delle analisi, la possibilità di codificare secondo criteri di volta in volta diversi le parti dei corpus raccolti. In questo modo, è possibile individuare relazioni significative di somiglianza/differenza tra "gruppi strategici" di soggetti che costituiscono altrettanti distinti raggruppamenti di tipo culturale.

---

**Carmençita Serino, Annarita Celeste Pugliese**
**Università degli Studi di Bari**

---

**Carmençita Serino** (c.serino@psico.uniba.it) è professore straordinario di Psicologia Sociale all'Università degli Studi di Bari. Principali temi di ricerca: categorizzazione e categorizzazione sociale; asimmetria cognitiva nel confronto fra sé e altri; identità personale-sociale; rappresentazioni sociali.

**Annarita Celeste Pugliese** (a.pugliese@psico.uniba.it) è dottoranda di ricerca in Psicologia della Comunicazione: processi cognitivi, affettivi e linguistici, presso il Dipartimento di Psicologia dell'Università degli Studi di Bari, sotto la supervisione di Carmençita Serino. Interessi di ricerca: identità sociale; relazioni tra gruppi; psicologia politica; computer mediated communication.

*La ricerca.* Il nostro progetto di ricerca, dal titolo "Strategie di inclusione categoriale nelle relazioni tra gruppi politici", è centrato sui livelli di inclusività nella costruzione delle categorie sociali nell'ambito delle relazioni tra gruppi politici. In particolare, la ricerca esplora la dimensione strategica, oltre che quella identitaria, delle categorizzazioni sovraordinate, anche in relazione allo status di maggioranza e opposizione. L'orizzonte teorico nel quale tale ricerca si inscrive è quello della Social Identity Theory e dei suoi sviluppi (Tajfel, 1978; Tajfel & Turner 1979; Turner, Hogg, Oakes, Reicher & Wetherell, 1987). Nell'indagare i livelli di inclusività e le modalità di identificazioni sovraordinate è risultato, sin dal primo momento, molto produttivo focalizzare la nostra attenzione su dati testuali. In questa direzione abbiamo selezionato, come corpus di analisi, 16 discorsi sulle crisi internazionali in Afghanistan e in Iraq, tenuti in Parlamento da esponenti di Forza Italia e dei Democratici di Sinistra. La nostra analisi si è concentrata sulla identificazione dei livelli di inclusione delle categorie del sé, utilizzati dagli oratori dei due schieramenti, e sulla ricostruzione e il confronto tra le differenti rappresentazioni di crisi internazionali veicolate nei discorsi.

In questa nostra analisi T-LAB ci ha permesso in primo luogo di esplorare il corpus e di raffrontare tra loro i discorsi della Destra e della Sinistra attraverso una serie di strumenti di text mining (parole-chiave, contesti elementari), text mapping (associazioni, somiglianze e differenze) e analisi del contenuto (analisi delle corrispondenze). In particolare, le mappe di associazione delle categorie di volta in volta considerate ci hanno permesso di illustrare anche graficamente le relazioni tra le categorie e di interpretarne il significato secondo i differenti usi nei due schieramenti politici.

Successivamente T-LAB ci ha consentito di estrarre dal corpus principale dei corpus d'analisi ristretti, secondo i nostri obiettivi di ricerca. Attraverso un'analisi delle corrispondenze, infatti, è stato possibile, per esempio, individuare i discorsi più omogenei all'interno di ogni gruppo politico e le relazioni di vicinanza/lontananza dei diversi parlanti tra loro, nonché ricostruire lo spazio di rappresentazione entro cui si collocano temi e problemi legati all'oggetto da noi considerato (pace/guerra, maggioranza/minoranza; Occidente/Islam, ecc.).

I risultati preliminari della nostra ricerca sono inclusi nei seguenti paper:

Pugliese A. C. & Serino C. *Using superordinate categories to enlarge consensus: an analysis of Italian politicians'speeches.* 8th European Congress of Psychology, luglio 2003, Vienna.

Pugliese A. C. & Serino C. *Self-categories e bias intergruppi nei discorsi politici sulle*

*crisi internazionali: destra e sinistra a confronto.* V Congresso Nazionale AIP – Sezione di Psicologia Sociale, settembre 2003, Bari.
Pugliese A. C. & Serino C. *Re-categorization strategies and political groups*. Twenty-Seventh Annual Scientific Meeting of the International Society of Political Psychology, "The Political Psychology of Hegemony and Resistance", luglio 2004, Lund.
C. Serino & A. C. Pugliese. *Stratégies d'inclusion catégorielle face à un événement déchirant: discours des politiques italiens sur la guerre en Iraq.* Cinquième Congrès International de Psychologie Sociale de Langue Française, settembre 2004, Lausanne.

---

**Francesca Ferretti**
**Università degli Studi di Parma**

---

**Francesca Ferretti (**francescaferretti@tiscalinet.it), laureata in Psicologia all'Università di Parma discutendo la tesi di laurea dal titolo "L'analisi dei resoconti scritti come strumento di valutazione in psicoterapia: studio pilota in soggetti con disturbi alimentari psicogeni" (relatore prof. Carlo Pruneti, correlatore Francesco Rovetto; con la collaborazione della dottoressa Elisa Fiori). Tirocinante presso il Dipartimento di Scienze Neurologiche e del Comportamento a Siena. Coadiuva alla realizzazione di una tesi di laurea su soggetti donatori nei trapianti di fegato. L'obiettivo sarà quello di individuare, con l'uso di T-LAB, aspetti che definiscono le rappresentazioni mentali dei soggetti in questione.

Ho mostrato vivo interesse verso l'analisi dei testi come possibile strumento da impiegare in ambito clinico. Nello specifico la finalità con cui mi sono avvicinata a tale ambito era quella di individuare degli indici di cambiamento dello stile linguistico nei diari di soggetti nel corso della terapia come parametri attendibili di valutazione per una diagnosi multidimensionale e per monitorare gli effetti del trattamento.

T-LAB è stato impiegato, in ambito clinico, con l'obiettivo di conciliare istanze diverse, quali la valorizzazione della soggettività e parallelamente, il rispetto per l'oggettività. Questo software è stato di grande aiuto poiché ci ha permesso di svolgere un'analisi più attendibile dei dati e di individuare le differenze nei vari stili linguistici dei soggetti nel corso della terapia.

L'esplorazione delle potenzialità applicative, in ambito clinico e di ricerca, dello studio computerizzato dei testi e la scelta di un percorso standardizzato da seguire, potrebbe, infatti, incentivare un impiego di tali analisi, dando l'input ad un fenomeno in fervente espansione come quello degli aspetti quantitativi della lingua.

Questo software è stato utilizzato per analizzare diari scritti da soggetti con disturbo alimentare psicogeno, diagnosticati secondo il DSM-IV (APA, 1994) nel corso di un trattamento psicoterapico cognitivo-comportamentale. Era richiesto di registrare in modo libero idee, emozioni, comportamenti, eventi ambientali, tra un incontro e l'altro con il terapeuta.

Lo scopo della ricerca è stato quello di rintracciare "parole-chiave" associabili alle caratteristiche del disturbo alimentare e di individuare possibili elementi che permettano di verificare i progressivi cambiamenti degli schemi cognitivi, indicativi del percorso terapeutico.

Il campione, generato dai testi in analisi, risulta composto da 171.789 occorrenze da cui è stato possibile estrapolare, per ogni soggetto, le parole tipiche caratterizzanti l'inizio e la fine del trattamento e conseguentemente effettuare un confronto tra di esse. Nello specifico, è stato suddiviso il testo di ogni soggetto secondo la variabile "anno" e successivamente sono state impiegate varie procedure: l'analisi delle specificità, l'analisi delle corrispondenze, la cluster analysis e l'analisi delle associazioni. Sulla base dei risultati ottenuti con l'ausilio delle suddette procedure impiegate è stato possibile estrapolare delle categorie che hanno mostrato chiaramente il mutamento del linguaggio dei soggetti confrontando l'anno di inizio e di fine terapia.

Più specificatamente per i soggetti con anoressia nervosa è stato riscontrato all'inizio della terapia la presenza di parole che fanno riferimento alle caratteristiche tipiche del disturbo quali "ossessione per il cibo, preoccupazione per il peso e le forme corporee, emozioni negative e compulsioni", mentre alla fine le parole risultate significative sono state raggruppate nelle seguenti categorie: "emozioni positive, sintomi fisici, interessi e rapporti interpersonali e autoconsapevolezza".

Per quanto riguarda i soggetti con bulimia nervosa sono state individuate all'inizio "ossessione per il cibo, emozioni negative, preoccupazioni per il peso e le forme corporee e compulsioni", mentre alla fine "emozioni positive, sintomi fisici, interessi e rapporti interpersonali e autoconsapevolezza".

L'aspetto interessante riscontrato è la concordanza tra i risultati ottenuti con questo software, i test standardizzati e il colloquio eseguito dal terapeuta.

---

**EURISKO**
**Ricerche su consumi, comunicazione e mutamento sociale**

---

L'istituto di ricerca **Eurisko** dal 1972 realizza ricerche che coprono l'intero panorama della ricerca sociale e di mercato, sia per la molteplicità delle aree di indagine esplorate, sia per la varietà delle metodologie utilizzate. In questa prospettiva è costante l'attenzione dedicata alla sperimentazione di nuovi metodi e strumenti di ricerca e tecniche di analisi dei dati.

Il nostro gruppo di ricerca[2] è diretto da **Albino Claudio Bosio** ed è formato da **Edoardo Lozza**, **Melania Gabrieli**, **Guendalina Graffigna**, **Isa Cecchini**. Uno dei principali temi di lavoro del gruppo, in particolare, è di esplora-

[2] Per ulteriori informazioni contattare edoardo.lozza@eurisko.it.

re e verificare le condizioni di utilizzo e le specificità metodologiche della ricerca qualitativa on-line.

Il nostro interesse per T-LAB nasce dalla necessità di analizzare i testi prodotti durante forum (gruppo di discussione "time-extended", interazione asincrona) e chat (gruppi di discussione sincroni, in tempo reale). Le ricerche on-line permettono di avere con relativa facilità la trascrizione memorizzata delle verbalizzazioni dei gruppi: questo, da un lato, rende più agevole il lavoro di analisi del ricercatore; da un altro lato, richiede strumenti più efficienti ed efficaci dei tradizionali approcci a supporto dell'analisi testuale. Si è pensato, quindi, a T-LAB come ad uno strumento che permettesse una mappatura rapida e flessibile dei contenuti di un testo, in modo da supportare il lavoro di analisi e di sintesi dei dati da parte del ricercatore.

*Esempio: una ricerca sul discorso sociale sul rischio AIDS.* La ricerca[3], di tipo esplorativo, aveva lo scopo di indagare come viene costruito il discorso sull'AIDS nella comunicazione interpersonale tra pari e l'influenza del contesto comunicativo su tale lavoro di costruzione. A tal fine abbiamo deciso di utilizzare 4 condizioni sperimentali differenti di cui tre on-line (forum, chat, forum + chat) e una face-to-face. Il campione era omogeneo per tutti i gruppi (giovani studenti, 18-24 anni, non sposati, di entrambe i sessi, di Milano).

Il corpus importato in T-LAB era composto dalle trascrizioni delle verbalizzazioni di tutti i gruppi ed è stato classificato secondo 2 variabili: il contesto comunicativo e il tema del discorso (articolato sulla base di una traccia di discussione).

*Analisi effettuate e principali risultati.* Abbiamo compiuto un'analisi tipologica dei contesti elementari per poter avere un quadro descrittivo dell'organizzazione generale del corpus, nonché per cogliere i nuclei tematici principali sottostanti al discorso e la loro articolazione. In particolare, abbiamo individuato 5 cluster tematici fondamentali che abbiamo così etichettato: il discorso politico sull'AIDS, l'uso del preservativo, il discorso sulla prevenzione, il malato di AIDS, la propria esperienza con la prevenzione.

Per verificare la specificità di impatto delle condizioni di gruppo sulle produzioni discorsive abbiamo utilizzato l'analisi delle corrispondenze e l'analisi delle specificità. I risultati ci hanno permesso di iniziare a capire quali possano essere le potenzialità e le peculiarità di determinati contesti di scambio sulle produzioni discorsive. Abbiamo così visto come il forum eliciti un discorso astratto e razionale sull'AIDS (prevale il tema politico); come la chat favorisca discorsi con ampie coloriture emotive (predomina il tema del malato di AIDS); come il forum + chat stimoli i partecipanti a rappresentarsi l'AIDS come un

[3] La ricerca faceva parte di un lavoro di tesi dal titolo "Comunicazione sociale e processazione interpersonale del messaggio: il caso del rischio AIDS" svolta da Melania Gabrieli sotto la supervisione del prof. A. C. Bosio e del dott. Lozza. Per ulteriori informazioni melania.gabrieli@eurisko.it.

problema autoriferito e ad interrogarsi su come gestire il rischio; come nel focus il discorso sull'AIDS venga attenuato e ricontestualizzato nel tema più ampio delle scelte relazionali e di coppia

---

**IMED – Istituto per il Mediterraneo**

---

**L'IMED – Istituto per il Mediterraneo** (www.imednet.it) è un ente no profit che lavora sostenendo lo sviluppo culturale, socio-economico, normativo e infrastrutturale delle regioni euro-mediterranee. Pubbliche Amministrazioni centrali e locali, la Commissione Europea con i suoi organi, associazioni dei lavoratori, degli imprenditori, del terzo settore, università ed enti di formazione, sono i committenti più frequenti dell'IMED.

Il metodo di lavoro che l'IMED privilegia è la ricerca-azione, con l'uso di strumenti e tecniche quali-quantitative, tra cui l'analisi dei testi.

**Fabrizio Paloni** (fabrizio.paloni@imednet.it), collaboratore stabile, e le consulenti esterne **Nadia Battisti, Francesca Dolcetti** e **Daniela Romei**, formano il gruppo di professionisti (RisorseObiettiviStrumenti@rmnet.it) che per l'IMED si occupa di questi aspetti.

Questo gruppo ha maturato un interesse all'utilizzo dell'analisi dei testi, sin dalla propria formazione universitaria, che prosegue nelle diverse collaborazioni con la Cattedra di Psicologia Clinica, il Laboratorio di Analisi del Testo ed il Laboratorio di Analisi della Domanda della Facoltà di Psicologia 1, Università "La Sapienza" di Roma.

In particolare usa un modello di analisi (AET: Analisi Emozionale del Testo) che permette di conoscere le *culture locali* dei contesti in cui il gruppo interviene. Più precisamente, la cultura locale è un costrutto che vede un'integrazione tra quelli di *rappresentazioni sociali* (Moscovici) e di *collusione* (Carli e Paniccia). Da ciò l'ipotesi che l'analisi della cultura locale consente di rilevare tanto le dimensioni di sviluppo, quanto le dimensioni critiche, che organizzano le relazioni individuo-contesto.

T-LAB permette di raggiungere gli obiettivi di conoscenza suddetti, in particolare attraverso le funzioni "mappe dei nuclei tematici" e "tipologie di contesti elementari". Le parole, intese come delle proposte relazionali, sono scelte sulla base di modelli di lettura psicosociologici della relazione individuo-contesto.

I testi che il gruppo analizza sono costruiti attraverso interviste scritte o orali; le associazioni discorsive degli intervistati sono sollecitate attraverso un'unica domanda tematica, individuata come centrale nel rapporto fra gli intervistati, il loro contesto di riferimento e l'evento critico che ha sollecitato la domanda di intervento.

Nel corso del tempo, i vari componenti del gruppo hanno realizzato ricerche sulle rappresentazioni sociali o delle culture locali di diversi gruppi target: insegnanti di scuole medie superiori, implicati nell'autonomia scolastica; psi-

cologi del servizio sanitario nazionale, implicati nell'aziendalizzazione del servizio; studenti in formazione universitaria, interessati alla verifica del loro percorso di professionalizzazione.

Il gruppo nei vari interventi, di natura consulenziale e formativa, che realizza per IMED dal 2002, con il supporto di T-LAB, mira a costruire informazioni *ad hoc* al fine di sviluppare la competenza dei gruppi target a leggere le potenzialità e le criticità del loro contesto di appartenenza, e di quello della domanda dei fruitori dei servizi, rivedere quindi, attraverso nuove ottiche, il rapporto con questi contesti, per individuarne nuove modalità organizzative.

Questi lavori riguardano i seguenti gruppi target e relativi obiettivi di sviluppo:

- amministratori, dirigenti e funzionari di Comuni italiani preposti a realizzare il mandato sociale della "semplificazione amministrativa", attraverso l'organizzazione del servizio di Sportello Unico per le Attività Produttive, e gruppi di imprenditori fruitori, o potenziali tali, del servizio;
- dirigenti di PA e professionisti di regioni euro-mediterranee, preposti ad individuare modelli integrati di gestione del patrimonio culturale;
- gestori di progetti finanziati dalla Commissione Europea preposti all'adozione del manuale del *Ciclo di Progetto*, per monitorare e valutare domanda e soddisfazione, in itinere, dei beneficiari intermedi e dei gruppi bersaglio, rispetto alle attività progettuali, o ancora, per individuare la committenza ai progetti espressa dagli opinion leader implicati nel sostenerli;
- amministratori pubblici, dirigenti di agenzie per l'impiego, dirigenti di servizi formativi, disoccupati, persone in cerca di occupazione e imprenditori implicati nel costruire e sperimentare modelli più efficaci di incontro tra domanda ed offerta di lavoro, in regioni italiane del Centro-Sud.

---

**Leonie A. Marks[4], Nicholas Kalaitzandonakes[5], Ludmilla Zakharova[6]**
**EMAC – University of Missouri-Columbia – U.S.A.**

---

Il dr. **Leonie A. Marks** (MarksLA@missouri.edu), il dr. **Nicholas Kalaitzandonakes** (KalaitzandonakesN@missouri.edu) e ms. **Ludmilla Zakha-**

[4] Dr. Leonie A. Marks, Assistant Research Professor of Agribusiness, Department of Agricultural Economics, University of Missouri-Columbia. *Memberships of Professional Associations*: American Agricultural Economics Association (AAEA),Western Agricultural Economics Association, Committee of Woman in Agricultural Economics (CWAE), Canadian Agricultural Economics Association (CAES),The Society for Risk Analysis (SRA).

[5] Dr. Nicholas Kalaitzandonakes, Professor of Agribusiness, Department of Agricultural Economics, University of Missouri-Columbia. *Memberships of Professional Associations*: American Agricultural Economics Association, Southern Agricultural Economics Association, Western Agricultural Economics Association, International Agribusiness Management Association, American Economic Association.

[6] Ms. Ludmilla Zakharova, Graduate Research Assistant, Economics and Management of Agrobiotechnology Center, University of Missouri-Columbia.

**rova** costituiscono un team di ricercatori che nell'Economics and Management of Agrobiotechnology Center (EMAC), presso l'Università di Missouri-Columbia (http://www.emac.missouri.edu), indaga su come le informazioni influenzano la percezione degli alimenti geneticamente modificati (biotech foods) da parte dei consumatori.

Poiché, nella maggior parte dei casi, le informazioni su questi argomenti derivano dalle notizie dei media, queste stesse informazioni giocano un ruolo predominante nella formazione delle percezioni che orientano il comportamento di consumo. In particolare, assume particolare importanza la comprensione di come i differenti media trasmettono informazioni sul "rischio" connesso al consumo dei biotech foods.

Attualmente stiamo usando T-LAB per indagare come differenti tipi di reporter (cronisti, scienziati, ambientalisti, ecc.) scrivono su questo argomento.

T-LAB ci consente di creare mappe cognitive che mostrano come i vari reporter percepiscono rischi e benefici associati con i biotech foods; quindi, ci consente di fare inferenze su come i diversi tipi di reporter "costruiscono" i messaggi di rischio sui media.

Per creare questo tipo di mappe cognitive, abbiamo raccolto una notevole quantità di articoli di quotidiani, pubblicati – nell'arco di 14 anni – da cinque testate inglesi e altrettante statunitensi.

Per noi, uno dei principali vantaggi dell'uso di T-LAB consiste nel fatto che ci consente di analizzare notevoli quantità di testi con l'aiuto di dizionari personalizzati. La personalizzazione dei dizionari, infatti, riduce al minimo la codifica manuale e – allo stesso tempo – ci consente di ottenere misure oggettive dei diversi concetti di rischio. Complementariamente, T-LAB offre la possibilità di fare ricerche automatiche che identificano le parole e le frasi che sono più frequenti all'interno dei testi, con i valori di occorrenza di ogni parola in ogni testo. La combinazione di questi metodi ci consente di usare la funzione *Associazioni di parole* (*Word Associations*), cioè una tecnica basata sulle co-occorrenze, sia per fare "co-word analysis", sia per costruire delle mappe che usiamo per scoprire nessi e relazioni tra idee e temi all'interno delle nostre collezioni di testi.

Partendo dagli output T-LAB, usando una nostra tecnica, convertiamo le relazioni di associazione in ulteriori mappe bi-dimensionali che mostrano cluster di parole associate in base alla loro prossimità all'interno dei testi. In questo modo, le mappe cognitive prodotte possono essere comparate in base alla presenza di concetti e di relazioni (cioè di affermazioni), sia esse "condivise" che "uniche".

Poiché la nostra ricerca è ancora in corso, i risultati non sono ancora stati resi pubblici; tuttavia, prevediamo di pubblicarli sia in forma di published thesis che, in forma di articoli, su rilevanti riviste specializzate sui temi della comunicazione.

**Sergio Salvatore**
**Università degli Studi di Lecce**

**Sergio Salvatore**, docente di psicologia dinamica, Università di Lecce.

I miei interessi di ricerca si rivolgono allo studio dei processi socio-simbolici che mediano il rapporto tra soggetti e sistemi collettivi di azione. Tale interesse si sviluppa sia in termini di riflessione teorica, che di analisi di specifici fenomeni micro-macrosociali.

Sul primo versante la mia attenzione si concentra sulla concettualizzazione di come il funzionamento della mente sia fortemente legato e dipendente dai contesti discorsivi e di attività in cui i soggetti sono inscritti. In particolare, la mia attività di teorizzazione e di ricerca si focalizza sul ruolo dei processi emozionali di simbolizzazione del contesto, come fondamentale mediatore – implicito – delle transazioni tra psiche individuale e gruppo/collettivo/organizzazione. In questa prospettiva, centrale risulta il concetto di cultura, riletto in chiave psicodinamica come la struttura simbolica condivisa, a fondamento inconscio, matrice dei significati e delle interpretazioni nei termini delle quali gli attori danno senso al loro mondo e dunque regolano reciprocamente le attività a cui partecipano.

A questa linea di lavoro di tipo più teorico, volta in ultima istanza alla definizione di un modello del funzionamento della mente di tipo al contempo socio-costruttivista e psicoanalitico – si collega, sul secondo versante, l'interesse per specifici fenomeni/processi socio-simbolici caratterizzanti ambiti di vita sociale ed organizzativa. In ragione di questa prospettiva, in questi ultimi anni, con il mio gruppo di ricerca ho concentrato il mio impegno tra l'altro sui seguenti ambiti:

1. analisi delle culture organizzative e del lavoro (con particolare riferimento alla rappresentazione culturale delle innovazioni in campo lavorativo; ad es. lavoro interinale);
2. studio delle immagini di contesti urbani, presenti entro la popolazione e presso i mass-media;
3. analisi dei modelli culturali attivi entro il mondo scolastico;
4. analisi del processo psicoterapeutico (con particolare attenzione a come si costruiscono gli assetti di reciprocità entro lo spazio psicoterapeutico).

In tutti e tre gli ambiti adotto una metodologia quali-quantitativa di analisi, basata sull'analisi dei testi, intesi come prodotti veicolanti i repertori di significati culturali. La metodologia è volta all'individuazione di elementi di ridondanza presenti entro l'organizzazione dei prodotti culturali, e alla loro successivamente interpretazione in chiave psicodinamica. Questa logica di analisi si focalizza in particolare su testi prodotti in vario modo (interviste, registrazione di conversazioni, trascritti di sedute psicoterapeutiche, testi scritti estratti da prodotti massmediali). Paradigmatica di questa logica di analisi quali-quantitativa la metodologia dell'AET (Analisi Emozionale del Testo), elaborata da

Renzo Carli e collaboratori, centrata sulla interpretazione delle co-occorrenze tra lemmi.

Nello studio quali-quantitativo dei testi ho utilizzato T-LAB, preferendolo ad altri strumenti per la sua flessibilità di impiego.

In particolare, per quanto riguarda il punto 3, nell'ambito di un progetto più ampio volto allo studio delle culture interne al mondo scolastico, ho svolto una indagine su un campione nazionale di scuole, finalizzata all'analisi dei modelli simbolici di rappresentazione dell'Adulto entro la popolazione degli studenti delle scuole elementari e medie. La ricerca si è basata su un corpus testuale composto dai testi elaborati dagli allievi come riempitivo di una vignetta rappresentante un dialogo tra un adulto ed un bambino.

Per quanto riguarda il punto 4, invece, sono attualmente impegnato con colleghi di altri centri di ricerca nella costruzione di un metodo di studio del processo psicoterapeutico fondato sulla micro-analisi del discorso (metodo RIFLUD, Rivelatore di FLUssi Discorsivi). In particolare, sulla base dell'utilizzazione di una versione particolare di T-LAB – RIFLUD si propone di ricostruire ed interpretare i flussi discorsivi, in quanto struttura associativa costitutiva della capacità del testo-discorso di costruire il significato intersoggettivo dello scambio comunicazionale.

---

**Franco Zambelli**
**Università degli Studi di Padova**

---

Sono professore straordinario di materie pedagogiche presso la Facoltà di Psicologia di Padova (Osservazione del comportamento in classe; Organizzazione scolastica e esperienza professionale; Programmazione dei servizi educativi; Pedagogia sperimentale) e afferisco al Dipartimento di Psicologia dello Sviluppo e della Socializzazione.

Uno dei miei interessi di ricerca riguarda lo studio dell'insegnamento e dei processi formativi rivolti agli insegnanti. Per questo motivo partecipo a vari gruppi di ricerca caratterizzati da interessi parzialmente diversi per quanto riguarda i livelli scolastici considerati e le metodologie di indagine utilizzate.

Mi sono trovato in varie occasioni a far ricorso all'analisi di testi derivanti da interviste semi strutturate o da autocaratterizzazioni (è una modalità di presentazione autobiografica all'interno della quale si possono identificare contesti di esperienza, sequenze, dimensioni di significato, ecc.). Quando la loro utilizzazione non era rivolta allo studio di singoli casi, ma agli orientamenti manifestati da differenti gruppi di partecipanti nei confronti di specifiche aree di esperienza, ho fatto ricorso all'analisi delle corrispondenze in maniera decisamente proficua (me ne sono avvalso in alcune ricerche con insegnanti e dirigenti scolastici).

Ho cominciato ad usare T-LAB dopo averne visitato il sito per le possibilità che offre nell'esplorazione, l'individuazione e il riordino del materiale linguistico raccolto, nella rappresentazione grafica dei dati, e ovviamente nei vari tipi di analisi fattoriale.

Attualmente stiamo usando il T-LAB all'interno di un Progetto di Ricerca di Interesse Nazionale – ex 40% finanziato dal Ministero ("La valutazione multidimensionale della docenza universitaria", direttore prof. Raffaella Semeraro – Università di Padova).

La valutazione dell'insegnamento è stata adottata in Italia, soprattutto ai fini del monitoraggio del Sistema Universitario Nazionale, mediante un questionario di valutazione basato su un numero molto limitato di aspetti e rivolto indifferentemente a tutte le Facoltà. Ritenendo, invece, soprattutto per il miglioramento dell'insegnamento, necessario disporre di strumenti per il rilievo e la valutazione dell'insegnamento specificamente adatti a cogliere le grandi differenze presenti negli insegnamenti realizzati nelle diverse Facoltà, per la loro costruzione, abbiamo cominciato a raccogliere circa ottanta interviste (della durata di 30-40 minuti ciascuna) a docenti e studenti di diverse Facoltà di Padova (Scienze, Psicologia, Lettere e Filosofia, Scienze della Formazione). In questa fase, l'utilizzazione del T-LAB ha consentito di individuare (tramite i Contesti elementari) contesti narrativi di termini considerati cruciali, in modo da predisporre item di questionari diversificati per Facoltà e basati sulle stesse espressioni utilizzate dagli intervistati.

In una seconda fase prevediamo anche di procedere all'analisi delle interviste semi strutturate mediante l'analisi delle corrispondenze per verificare se ed eventualmente in base a quali aspetti della didattica i docenti delle differenti Facoltà considerate siano tra loro distinguibili.

Una ulteriore applicazione del T-LAB per la quale stiamo raccogliendo il materiale narrativo (autocaratterizzazioni e interviste semi strutturate) da insegnanti, dirigenti, e altro personale, riguarda la possibilità di studiare le "culture" di differenti scuole e dei diversi gruppi professionali presenti. In particolare, ci siamo proposti di analizzare le autocaratterizzazioni individuando specifici segmenti in base ad aspetti strutturali (ad es., inizio – fine) e non solo di contenuto (ad es., medesimi contesti di esperienza).

---

**Liborio Dibattista**
**Università degli Studi di Bari**

---

**Liborio Dibattista**, laureato in medicina e filosofia, dottore di ricerca in Storia della Scienza, Seminario di Storia della Scienza dell'Università di Bari.

Da tempo il Seminario si occupa di epistemologia informatica con progetti di ricerca riguardanti l'analisi computazionale del linguaggio del *Discorso sui massimi sistemi* di Galilei e di opere di scienze della vita dell'Ottocento fran-

cese. Il settore specifico a cui sto lavorando da alcuni anni è l'uso di tecniche di linguistica computazionale come strumento di storiografia scientifica. In particolare ho pubblicato recentemente una monografia dal titolo: *J.-M. Charcot e la lingua della neurologia* per i tipi dell'editore Cacucci di Bari. In essa si esplora la possibilità di usare strumenti di analisi testuale (sono stati usati software specifici prodotti dal Seminario e il processore INTEX di Max Silberztein, Université de Franche-Comté) per verificare l'ipotesi che un autore – Charcot appunto – accreditato dalla storiografia classica come fondatore di una disciplina, la neurologia, sia stato anche un innovatore del linguaggio specifico della materia. In altri termini, nel libro intendo dimostrare che il creatore di una nuova scienza sia individuabile anche grazie al suo contributo come "costruttore" di un linguaggio specifico. Ne consegue che i testi sui quali mi trovo di solito ad operare sono corpora scientifici di robuste dimensioni (il lavoro su Charcot è stato fatto su tre tomi con un totale di oltre 350.000 parole). Inoltre, la lingua usata da Charcot nelle sue opere di carattere neurologico è stata messa a confronto con quella di altri due medici francesi, Duchenne de Boulogne e Jules Dejerine, rispettivamente anteriore e successivo a Charcot, al fine di evidenziare la novità del linguaggio specialistico rispetto al primo e l'accoglimento di tale linguaggio nel secondo. Quest'ultimo aspetto del lavoro potrebbe essere appunto rivisitato facendo uso degli strumenti messi a disposizione dal software T-LAB. Da quanto detto si evince che di solito lavoro su testi antichi, che generalmente non esistono in formato digitale e che quindi richiedono un faticosissimo lavoro preliminare di acquisizione e conversione in formato digitale (.txt) con software di OCR.

Il software T-LAB è stato da poco acquisito fra gli strumenti di lavoro del Seminario. Il prossimo progetto di ricerca riguarderà lo sviluppo della fisiologia in Francia nella seconda metà dell'Ottocento, in particolare nei medici titolari di insegnamento presso il Collège de France. Penso di utilizzare T-LAB per l'individuazione di nuclei tematici nelle opere che Etienne Jules Marey dedicò al tema del movimento, argomento a cui dedicò tutta la sua opera di scienziato, con lo scopo di sottolineare o enucleare l'evoluzione dell'approccio a questo tema nell'arco di oltre trent'anni. Inoltre, se lo permetterà la disponibilità dei testi in formato digitale, sarebbe interessante cogliere l'evoluzione diacronica dei temi principali che interessarono la scuola di fisiologia francese da Pierre Flourens a Charles François-Franck, dagli anni quaranta dell'Ottocento ai primi anni del Novecento. Le mappe dei nuclei tematici e la cluster analysis di T-LAB dovrebbero fornire strumenti particolarmente efficaci per il confronto fra opere successive dello stesso Autore o fra Autori differenti, con l'obiettivo di mostrare quali nuclei concettuali furono sviluppati per la fondazione della disciplina scientifica e quali altri invece furono abbandonati andando a costituire, così, un relitto della storiografia scientifica.

**OCP: Osservatorio sugli eventi della comunicazione politico-culturale**
**Università degli Studi di Roma "La Sapienza"**

L'**OCP** è un Osservatorio di ricerca e sperimentazione nato nella Facoltà di Scienze della Comunicazione dell'Università degli Studi di Roma "La Sapienza".

Presieduto da **Alberto Abruzzese**, diretto da **Stefano Cristante** e coordinato da **Rossella Rega** e **Marco Bigotto**, nasce con l'intento di analizzare, da una prospettiva mediologica, gli eventi della comunicazione politica e culturale.

L'obiettivo dell'osservatorio è di costituirsi quale realtà di ricerca e di intervento che possa analizzare la comunicazione da un punto di vista qualitativo, aperto alle innovazioni linguistiche e tecnologiche. Una realtà situata all'interno di una struttura pubblica universitaria che sia in grado di confrontarsi, su un piano di efficienza e soprattutto di innovazione, con il mercato.

L'OCP si occupa di:

- Comunicazione politica e culturale,
- Consumi culturali.

Per maggiori informazioni sull'OCP è possibile consultare il sito web www.ocp.too.it oppure contattarci via mail: marco.binotto@uniroma1.it o rossella.rega@uniroma.it.

Il nostro primo utilizzo di T-LAB risale alla ricerca sull'evento del G8 di Genova, condotta tra il 2001 e il 2002, realizzata con l'obiettivo di individuare le strategie comunicative sia della parte istituzionale che delle realtà che questa hanno contestato e dei risultati di questa nuova "battaglia mediale". La ricerca ha analizzato i mass media nei mesi precedenti e durante il vertice; in particolare, sono stati analizzati, attraverso schede di analisi del contenuto, quotidiani, periodici, telegiornali, speciali televisivi. Si è prestata grande attenzione anche all'uso delle nuove tecnologie da parte del movimento dei movimenti, analizzando siti web e mailing list.

La ricerca è stata condotta da una équipe di 14 persone, con l'aiuto di studenti della Facoltà di Scienze della Comunicazione.

Nell'ambito dell'analisi sui quotidiani (11 testate nazionali nel periodo 15 maggio – 29 luglio 2001) abbiamo pensato di ricorrere all'analisi dei testi tramite T-LAB sia per "ridurre la complessità" di un materiale molto vasto e avere così delle utili indicazioni per l'impostazione della fase di analisi della ricerca, sia per sottoporre a verifica delle ipotesi emerse durante il lavoro.

L'uso di T-LAB è stato limitato ai soli titoli di 3091 articoli di 11 quotidiani, riferiti ad un periodo di tempo focalizzato sull'evento G8, quindi dal 15 al 29 luglio 2001. Solo per ragioni di tempo e di economia della ricerca non abbiamo analizzato anche il testo degli articoli (che non erano disponibili in formato digitale), come invece abbiamo fatto per una ricerca successiva, condotta per conto della CGIL, relativa alle strategie comunicative nell'ambito della battaglia mediale sull'Articolo 18 dello Statuto dei Lavoratori.

I titoli dei quotidiani, sfruttando il lavoro precedentemente svolto dai ricercatori con la scheda di analisi, sono stati ricodificati con l'aggiunta di cinque variabili, delle quali si sono rivelati utili soprattutto quelle relativa alla testata e alla data di pubblicazione.

I risultati della ricerca (una sintesi dei quali è stata pubblicata nel testo a cura di Stefano Cristante, *Violenza Mediata*, Editori Riuniti, Roma, 2003) possono essere così sintetizzati:

- attraverso la mappa di nuclei tematici abbiamo individuato le tematiche che hanno maggiormente orientato le differenze nella rappresentazione mediale;
- con l'analisi delle corrispondenze abbiamo individuato una mappa con la "posizione" delle singole testate, analizzando somiglianze e divergenze attraverso l'analisi dei fattori che la componevano;
- sempre con l'analisi delle corrispondenze abbiamo individuato, creando una nuova variabile che incrociava testata e data, gli "spostamenti" delle singole testate analizzando le differenze di posizione tra i giorni "caldi" dell'evento (20 e 21 luglio) e i periodo immediatamente precedenti e successivi;
- abbiamo poi costruito una serie di mappe di associazioni per verificare, per alcuni termini chiave soprattutto riferiti agli attori in gioco, alcune ipotesi di lavoro e in particolare le associazioni con il concetto di "violenza" nelle sua varie declinazioni.

---

**Maria Luisa Farnese**
**Università degli Studi di Roma "La Sapienza"**

---

**Maria Luisa Farnese**, docente di Psicologia del Lavoro e delle Organizzazioni nel Corso di Laurea in Servizi Sociali e ricercatrice presso la cattedra di Psicologia del Lavoro (F. Avallone) della Facoltà di Psicologia2 dell'Università di Roma "La Sapienza" (marialuisa.farnese@uniroma1.it).

I principali ambiti di ricerca sviluppati con il gruppo di ricerca sono relativi allo studio dei processi di convivenza all'interno di contesti organizzativi. Nello specifico, negli ultimi anni sono stati avviati progetti sui temi della fiducia e della creazione di relazioni fiduciarie, della giustizia organizzativa e della percezione di equità in contesti lavorativi, e delle dinamiche di convivenza a livello organizzativo e sociale.

Con riferimento ad un approccio di tipo psicosociale, le ricerche realizzate generalmente si fondano su dati costituiti in parte da materiale testuale. Si tratta perlopiù di interviste realizzate in fase iniziale e svolte con diverse metodologie (focus group, interviste semistrutturate o non strutturate a testimoni privilegiati o agli attori della vita organizzativa, ecc.). Questo tipo di materiale

informativo consente infatti di far emergere con maggiore evidenza e immediatezza le dinamiche intersoggettive e i processi di interazione sociale e, dunque, di fare riferimento a chiavi interpretative afferenti da un lato alle dimensioni emozionali e di natura psicodinamica, dall'altro – se viene studiato uno specifico contesto organizzativo – ai fondamenti rappresentazionali e valoriali della cultura organizzativa di riferimento.

I corpus narrativi sono stati studiati utilizzando diverse metodologie e strumenti per l'analisi dei testi: da analisi tipo "carta e matita" attraverso un processo induttivo e progressivo di categorizzazione secondo il modello della Grounded Theory e strumenti quali Atlas.ti, alla mappatura di storie ed eventi contenuti nella struttura narrativa. L'utilizzo di T-LAB è stato privilegiato per la capacità dello strumento di far emergere la struttura narrativa dei testi, evidenziando le relazioni di prossimità e analogia tra gli elementi linguistici, nell'ipotesi che tale struttura riproduca omomorficamente i processi di produzione simbolica rappresentati nelle narrazioni. In questo senso ci consente di utilizzare il processo interpretativo in modo euristico, in quanto applicato a dati che risultano organizzati da modelli statistici che sono indipendenti dalle categorie teorico-concettuali del ricercatore.

Ad esempio gli studi sul tema della giustizia organizzativa sono stati avviati realizzando trenta interviste semistrutturate a persone che lavoravano in diversi contesti organizzativi, al fine di indagare, in modo esplorativo, la loro percezione di giustizia ed equità [Farnese M.L., Avallone F. (2003) "Perceptions of Organizational Justice: Cultural Models that Structure Everyday Work Experiences", in Quaderni di Psicologia del Lavoro, vol.11, p.144-151]. Le interviste prevedevano una domanda introduttiva ("Cosa pensa della giustizia nella sua organizzazione?") e due domande relative al racconto di episodi in cui le persone hanno sentito che il loro sentimento di giustizia è stato positivamente confermato ovvero disatteso. Dall'analisi del corpus testuale complessivo (19.819 occorrenze, 2.460 lemmi) condotta attraverso T-LAB (strumento Tipologie di Contesti Elementari) sono emersi cinque diversi modelli culturali, che rappresentano altrettante modalità di simbolizzazione affettiva del contesto lavorativo in relazione agli elementi fondanti l'equità dei processi lavorativi.

---

**Maurizio Lana**
**Università del Piemonte Orientale "A. Avogadro"**

---

**Maurizio Lana**, Dipartimento di Studi Umanistici, Università del Piemonte Orientale "A. Avogadro" (m.lana@lett.unipmn.it).

Sono laureato in lettere classiche e fin dagli anni '80 mi sono interessato di filologia computazionale (partendo dagli scritti e dall'opera di dom Froger, dom Quentin, G. P. Zarri) e di studio di testi con il computer (incominciando con J. Abercrombie, *Computer programs for the analysis of texts*). I miei inte-

ressi hanno riguardato in passato la stilometria (analisi dei testi allo scopo di individuarne caratteristiche formali e di contenuto tali da permetterne l'attribuzione, cioè il riconoscimento dell'autore), l'attribuzione di testi apocrifi o anonimi, l'analisi statistica dei testi in senso lato. Attualmente i miei interessi di ricerca, pur senza lasciare completamente da parte quelli precedenti, si orientano soprattutto verso l'analisi del contenuto come evoluzione dell'analisi dei testi e l'analisi delle corrispondenze lessicali su testi contemporanei.

I miei interessi riguardano soprattutto i testi, in senso stretto. Ho lavorato su testi antichi (la Costituzione degli Ateniesi attribuita a Senofonte, le Storie di Tucidide, l'opera di Senofonte), e lavoro ora su testi contemporanei letterari (Collodi, Le avventure di Pinocchio), e no (gli articoli pubblicati dai principali quotidiani italiani dal 12 al 29 luglio 2001 sul G8 di Genova).

In particolare, in relazione agli articoli dei quotidiani sul G8 mi interessa mettere in evidenza, attraverso lo studio del lessico utilizzato, l'evoluzione temporale del linguaggio usato, dal 12 al 29 luglio in relazione agli avvenimenti che si verificarono; mi interessa altresì caratterizzare, sempre in base al lessico utilizzato, i giornalisti e le testate, così da individuarne *distanze* e *vicinanze.* La ricerca è in corso.

Sto anche (marzo-aprile 2004) lavorando per prepararmi all'*ad-hoc attribution contest* bandito da P. Juola della Duquesne University di Pittsburgh. Il *contest* consisterà nel riconoscere quali tra una serie di campioni di testo anonimi siano di un determinato autore del quale pure viene fornito un campione di testo. Anche questa ricerca è in corso. I risultati saranno resi noti da P. Juola in occasione della ALLC-ACH Conference di giugno 2004 a Göteborg.

Nell'*ad-hoc attribution contest* sto utilizzando T-LAB per vedere se riesco ad ottenere classificazioni corrette dei testi basate sull'analisi delle corrispondenze dei segmenti ripetuti. Nel lavoro sul G8 lo utilizzerò per mettere a fuoco aree semantiche e nuclei concettuali specifici dei vari quotidiani e dei loro giornalisti.

Dal punto di vista dello studio dei testi, mi aspetto di trovare conferma alla mia convinzione che i metodi statistici (analisi delle corrispondenze) conducano a due tipi di esiti:

- da una parte dare corpo ed evidenza fattuale (quantitativa, numerica!) a intuizioni che nella lettura tradizionale, sequenziale del testo rimangono inafferrabili (ricerca di conferme ad ipotesi preesistenti);
- dall'altro, con una logica di tipo *text mining*, estrarre e porre in evidenza una serie di fenomeni linguistici e di contenuto (es.: negli articoli sul G8 del quotidiano A il nome del ministro Scajola è associato ai termini *a, b, c*; mentre nel quotidiano B è associato ai termini *r, s, t*) tra i quali il ricercatore può talora individuare quelli che alla sua analisi e interpretazione appaiono essere fatti rilevanti (esplorare materiali testuali troppo estesi per poter essere afferrati nel loro senso complessivo).

Le ricerche descritte sono ancora in corso e le relative pubblicazioni, non appena definite, potranno essere indicate a chi mi scriverà in posta elettronica.

**Felicia Pelagalli, B.S.D. Ricerche**
**Carlo Longo, Telecom Italia**

**Felicia Pelagalli** è Psicologa Clinica, Consulente in Ricerche di Marketing, Partner B.S.D. Ricerche. E-mail: pelagalli@media.unisi.it.

**Carlo Longo** è Responsabile Metodologie di Loyalty di Telecom Italia. E-mail: carlo.longo@telecomitalia.it.

L'interesse per l'analisi dei testi si colloca nell'ambito delle ricerche di mercato ed in particolare nasce dall'esigenza di trarre informazione da un consistente numero di risposte a domande aperte.

Telecom Italia ha in corso ormai da alcuni anni un'indagine di monitoraggio della "loyalty" dei suoi clienti. L'indagine prevede rilevazioni telefoniche su circa 70.000 clienti/anno, ripartiti per segmento di marketing e territorio. Le interviste, realizzate con il sistema C.A.T.I., si basano su un questionario semistrutturato di circa 200 domande, 90 delle quali sono domande a risposta aperta.

È evidente che siamo di fronte a un patrimonio informativo notevole, fatto di parole, dichiarazioni, racconti, rilasciati dai clienti nel corso dell'intervista; dichiarazioni che se trattate con le usuali codifiche, basate su sintetiche griglie di lettura, rischiano di perdere tutta la loro complessità e ricchezza.

L'obiettivo era quello di analizzare le risposte alle domande aperte del questionario "Indagine Loyalty" al fine di individuare i modelli culturali che orientano la fedeltà all'operatore TLC. In particolare, approfondire le motivazioni di fedeltà/abbandono nei confronti dell'operatore Telecom Italia monitorando:

- i motivi di insoddisfazione verso Telecom Italia;
- i motivi di soddisfazione verso altri operatori TLC;
- i motivi che orientano l'intenzione a rivolgersi ad altri operatori;
- i motivi che orientano l'intenzione a "tornare" con Telecom Italia.

Attraverso il software T-LAB sono state analizzate le risposte ad alcune domane aperte, rilasciate dagli intervistati negli anni 2002-2003. In tutto circa 30.000 risposte, classificate rispetto a più variabili di riferimento: zona geografica, segmento di clientela, periodo temporale, ecc.

Usando varie funzioni di T-LAB (Associazioni di parole, Analisi delle corrispondenze, Tipologie dei contesti elementari), sono stati ottenuti i seguenti risultati:

- sono state individuate le aree di soddisfazione/insoddisfazione maggiormente connesse a dimensioni di fedeltà/abbandono nei confronti dell'operatore di TLC;
- sono stati rappresentati graficamente i contenuti delle risposte attraverso pochi e significativi cluster e sono state evidenziate le differenze presenti tra le variabili prese in esame, come ad esempio: le differenze tra i periodi

temporali di rilevazione; le differenze nelle percezioni/esigenze dei vari segmenti di clientela, ecc.;
- sono state inoltre individuate le parole più frequentemente associate ai diversi operatori TLC da parte dei clienti intervistati.

L'analisi condotta ha consentito di esplorare il vocabolario e i contenuti delle risposte fornite dagli intervistati alle domande aperte, facendo emergere alcune importanti dimensioni di problematicità nel rapporto con l'operatore TLC, aree non previste dalla struttura "chiusa" del questionario.

La possibilità di costruire dizionari personalizzati ha consentito di applicare lo stesso tipo di lemmatizzazione ai materiali testuali (corpora) derivati dalle varie domande.

Elementi di criticità incontrati:
- nelle fasi di pre-trattamento, data la specifica origine dei testi (trascrizione di interviste telefoniche), è stato dedicato molto tempo alla revisione ortografica;
- nelle fasi di analisi, quando sono state utilizzate tecniche multivariate che prevedono l'uso di tabelle con valori di co-occorrenza (ad es. Tipologie dei contesti elementari), il limite di inclusione nell'analisi solo degli elementi in cui risultano presenti almeno due unità lessicali o nuclei tematici, è risultato problematico perché molte risposte erano costituite solo da una o due parole, spesso molto dense di significato (ad es. "la bolletta").

---

**Petri Vasara**
**Jaakko Pöyry Consulting**

---

**Petri Vasara**, dirigente presso l'agenzia di consulenza manageriale Jaakko Pöyry Consulting (Finlandia), è anche ricercatore attivo nella progettazione e nella direzione di vari progetti, in particolare nel campo del data mining. Aree di interesse: tecnologie per il monitoraggio e l'elaborazione di strategie, segnali deboli (weak signals), trend ambientali, politiche industriali.

M.Sc. in Technical Physics, Lic.Tech. in Computer Science, Dr.Tech in Neural Networks presso l' Helsinki University of Technology.

E-mail: Petri.vasara@pp.inet.fi.

*Perché l'analisi dei testi.* Anni fa, in un progetto di ricerca ho usato le reti neurali per analizzare simultaneamente immagini, dati e testi. Era l'epoca delle *.com* e l'esempio analizzato riguardava la vendita di capi di abbigliamento tramite Internet. Fu così che feci il mio primo "textual mapping". Ancora prima, sempre via Internet, avevo fatto alcune analisi semantiche di gruppi di discussione sui temi dell'ambiente e alcuni risultati mostravano il mio stesso nome con "sorprendenti" associazioni a persone e a temi.

A partire da queste esperienze, decisi di predisporre un toolkit per l'analisi dei testi, in parte acquistato, in parte custom-made.

Attualmente, per quanto riguarda l'analisi dei testi, il nostro sistema di data-mining prevede l'uso di algoritmi proprietari e T-LAB costituisce il nostro principale strumento per le analisi semantiche. La maggior parte dei nostro lavoro consiste nel trattare materiali "pescati" via Internet o resi disponibili, in formato elettronico, da varie organizzazioni. Quindi, i materiali tipici e multiformi che ci troviamo ad analizzare sono annunci, articoli di giornali, newsletter, commenti e brevetti. In misura minore, analizziamo risposte ad interviste; ciò a causa del fatto che i materiali ottenuti con questo metodo sono spesso sottodimensionati e non consentono l'uso di modelli matematici.

*Perché T-LAB.* Nel predisporre il toolkit per l'analisi dei testi, ho esplorato e testato molti tipi di strumenti. Al confronto, i vantaggi offerti da T-LAB mi sono sembrati essere i seguenti: l'ampia varietà delle analisi offerte, i continui aggiornamenti del prodotto, la possibilità di utilizzare grafici e tabelle per ulteriori elaborazioni, la stabilità del software e la possibilità di contatti personali con il produttore.

*Attività di ricerca.* Mentre scrivo, sono appena rientrato da un seminario sui media "ibridi" (area di combinazioni e convergenze tra i media digitali e quelli in formato cartaceo), in cui ho presentato alcune conclusioni di una ricerca. Poiché la definizione dei media "ibridi" è ancora scarsamente condivisa, in una parte dello studio è stata utilizzato T-LAB per analizzare i testi prodotti dai ricercatori (quelli concernenti il tema in questione), sia per trovare gli elementi comuni che per verificare le differenze tra varie "scuole di pensiero".

Molti degli studi da me realizzati sono stati effettuati per conto di clienti aziendali; quindi, per motivi di riservatezza, non posso descriverli. Uno che è stato reso pubblico riguarda le modalità di discutere (adottate da media, esperti, giornali, enti non governativi e gruppi di attivisti) sulla tecnologia RFID[7] (Radio Frequency ID), con particolare attenzione ai suoi trend e alle sue percezioni. Ciò tenendo conto del fatto che questa tecnologia "smart tagging", mentre promette notevoli progressi nel flusso delle informazioni concernenti i rapporti di fornitura, alimenta le proteste di molti fautori della privacy.

In sintesi: il mio uso di T-LAB è principalmente orientato all'analisi discussioni altamente de-strutturate, per tentare di separare le informazioni importanti dal "rumore di fondo" e per scoprire pattern dove – apparentemente – sembrano assenti.

[7] RFID è un sistema di identificazione elettronica dei prodotti che si avvale di etichette con microchip capaci di emettere segnali radio a corto raggio. In questo modo, attraverso speciali lettori, è possibile monitorare tutti i passaggi della catena di distribuzione.

## 1.2. Quasi per gioco

Nei capitoli che seguono proporrò alcuni esempi, costruiti come *itinerari guidati* all'uso degli strumenti T-LAB.

Ciascun itinerario parte dall'importazione del corpus e si arresta in un punto di confine: quello che segna il passaggio nel territorio dell'*interpretazione produttiva*, intesa come attribuzione di *senso* ai risultati e come costruzione di un resoconto per un cliente[1].

Detto in altri termini: gli esempi proposti sono resoconti di *esercizi*[2], costruiti per aiutare il lettore a meglio capire le risorse offerte dal software e i limiti dei suoi possibili usi. Quanto ai resoconti di "vere" pratiche di ricerca, ciascuno – in altri luoghi – può cercare o costruire gli esempi che più gli sono congeniali. Non ultimo, può consultare le testimonianze contenute nel paragrafo precedente (1.1) di questo libro.

La Fig. 1 illustra schematicamente un itinerario tipo dell'uso di T-LAB.

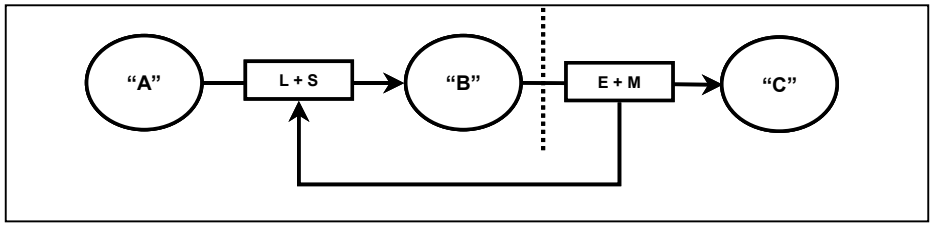



*Fig. 1- Dal corpus al resoconto*

A partire dal corpus ("A"), tramite l'uso di strumenti linguistici e statistici (L+S), vengono ottenuti dei risultati ("B"). Quindi, facendo ricorso a "enciclopedie" di conoscenze[3] e a modelli interpretativi (E+M), viene generato un prodotto-resoconto ("C"). Nella stessa figura, la linea tratteggiata indica il confine delle *simulazioni* a scopo di esercizio, mentre un feedback mette in relazione gli strumenti T-LAB con i modelli interpretativi del ricercatore.

Ma la differenza tra gli *esercizi* e le "vere" pratiche di ricerca non sta solo nel fatto che queste vanno oltre la line di confine segnata nella Fig. 1. Infatti,

[1] Sui problemi connessi al resoconto, si veda il paragrafo 1.2 (Parte prima).

[2] La loro comprensione, che richiede la lettura del capitolo "La logica analitica di T-LAB" (prima parte del libro), può risultare più interessante dopo aver sperimentato la versione demo del software (http://www.tlab.it/it/download.asp).

[3] Siano esse riconducibili a "saperi" di discipline scientifiche che derivanti da un'analisi del contesto.

soprattutto e in primo luogo, esse prendono l'avvio in contesti ben diversi da quelli definiti dal corpus ("A"); e, prima ancora di ipotizzare l'uso di qualsivoglia strumento software, esse devono fare i conti con la definizione di obiettivi, strategie e piani di lavoro.

In questi casi si suol dire "bisogna darsi un metodo". Ma, come ricorda H.G. Gadamer (1967, tr. it., p. 84), il metodo, in quanto tale, non basta da solo a garantire la fecondità delle sue applicazioni. Ad esempio, l'uso di tecniche statistiche di tipo esploratorio può anche essere inscritto in contesti di ricerca in cui la ripetitività delle operazioni si sostituisce all'interrogazione sugli obiettivi e sui possibili modi per perseguirli.

Da qui la constatazione che la possibilità di costruire una propria strategia di analisi, che – allo stesso tempo – tenga conto della specificità degli oggetti di indagine e delle particolari domande a cui si cerca di rispondere, in molti casi implica il passaggio dalla logica del rituale a quella del gioco.

Come è noto, nella **logica del rituale** una sequenza di azioni, che originariamente serviva ad altri scopi, diventa fine a sé stessa: l'importante, quindi, è che essa venga realizzata secondo un ordine precostituito e ripetitivo. Di conseguenza, entro questa logica, la discrezionalità degli attori sociali (nel nostro caso, i ricercatori) si esplica tutta secondo modalità codificate.

Se poi consideriamo che il rito implica una sorta di dichiarazione di guerra della forma contro l'indeterminatezza (S. Moore e B. Myerhoff, 1977, p. 16), possiamo ragionevolmente supporre che nel caso dell'analisi dei testi questo tipo di problema diventa di particolare importanza.

Diversamente, nella **logica del gioco** il risultato dell'azione o dell'interazione è *imprevisto* per definizione: da qui la possibilità di elaborare strategie e tattiche. Le *regole* in questo caso sono *costitutive* e non normative[4], ovvero hanno la forma di definizioni e non di imperativi.

Ad esempio, le regole degli scacchi definiscono le possibili mosse del cavallo o dell'alfiere, ma la successione delle mosse reali è determinata dalla strategia o dalla tattica di ogni giocatore. Strategia e tattica che si esplicano attraverso attente valutazioni dello stato della partita, delle proprie condizioni e di quelle dell'avversario, del tempo disponibile e di altre cose ancora.

Nell'ambito delle attività di ricerca, i motivi per i quali le regole del gioco possono a volte trasformarsi in "prescrizioni metodologiche" sono riconducibili alle *culture* delle varie comunità scientifiche, cioè ai diversi modi attraverso i quali – al loro interno – viene gestita quella che T.S. Kuhn (1977) chiamava la "tensione essenziale" tra tradizione e innovazione; ciò tenendo conto del fatto che l'acculturazione alla ricerca, cioè la formazione, general-

[4] Cfr. J. R. Searle, 1969; R. Harré e P. F. Record, 1972; N. Bobbio, 1980.

mente passa attraverso forme di dipendenza che certamente non favoriscono l'elaborazione autonoma di piani e strategie.

Ancora qualche precisazione: la nozione di **piano** rinvia a una concezione dell'azione regolata da scopi e da gerarchie di scopi; quindi "definire un piano" significa decidere quali operazioni vanno fatte e in quale ordine. Secondo alcuni studiosi[5], un piano è per l'uomo ciò che è un programma per un calcolatore, cioè, possiamo aggiungere, qualcosa che – per definizione – non tiene conto della "non prevista" variabilità del contesto.

In effetti, come ben sa chi si occupa di software, il contesto d'uso delle applicazioni è considerato solo nelle fasi di analisi che precedono la scrittura del programma e dei suoi eventuali aggiornamenti. Diversamente, la strategia[6] è fortemente ancorata al contesto e, in funzione della variabilità di quest'ultimo, definisce (o ridefinisce) le priorità delle azioni e la successione delle operazioni.

Quanto all'analisi dei testi, la definizione di una strategia richiede un'attenta valutazione dei "fattori di contesto" (vedi Fig. 1 del paragrafo 1.1, parte prima), per ogni corpus e per ogni progetto. Ad esempio: una cosa è analizzare dati testuali ottenuti tramite interviste, altra è analizzare risposte a domande aperte; una cosa è avere a che fare con un cliente "esterno", altra è pensare di presentare i risultati solo all'interno del proprio gruppo di appartenenza; una cosa è fare riferimento alle teorie e ai modelli dell'antropologia, altra a quelli della psicologia o della sociologia, con tutte le articolazioni in "scuole" e tradizioni di ricerca che queste discipline propongono.

Inoltre, la definizione di una strategia richiede un'accurata ricognizione delle risorse disponibili: persone, competenze, tempi e strumenti. Ad esempio: una cosa è lavorare da soli o con pochi intimi, altra è lavorare in un progetto che implica interazioni tra vari gruppi di lavoro; una cosa è essere impegnati a consegnare i risultati entro una settimana, altra cosa è avere a disposizione più di qualche mese.

E poi ci sono i vincoli posti dai software. In molti casi, questi propongono solo alcune opzioni: quelle che – in genere – corrispondono alle strategie di analisi previste dagli autori. Anche per questa ragione, quando si resoconta di una ricerca, invece di spiegare cosa si è fatto e perché, in molti casi ci si limita a dichiarare di aver usato il software "X" o il software "Y". Certo, questo modo di resocontare può essere attribuito alla cultura di alcuni ricercatori, ma – ad onor del vero – in alcuni software l'unica scelta strategica consentita consiste nello scegliere uno dei *piani di analisi* suggeriti dall'autore.

[5] Cfr. G. A. Miller, E. Galanter e K. H. Pribram, 1960.

[6] Per la definizione di strategia, vedi paragrafo 1.2 (Parte prima).

In alternativa, il "ricercatore esperto" può adottare una strategia che implica l'uso di due o più software; ma, dato che ciascuno di questi propone una diversa logica di funzionamento, il dispendio di risorse può essere considerevole.

Anche a partire da un'analisi di questi problemi, l'architettura di T-LAB è stata progettata per offrire in un *unico ambiente* un'ampia gamma di strumenti integrati tra loro. L'interfaccia è marcatamente user friendly e *le scelte strategiche sono tutte a carico dell'utilizzatore.*

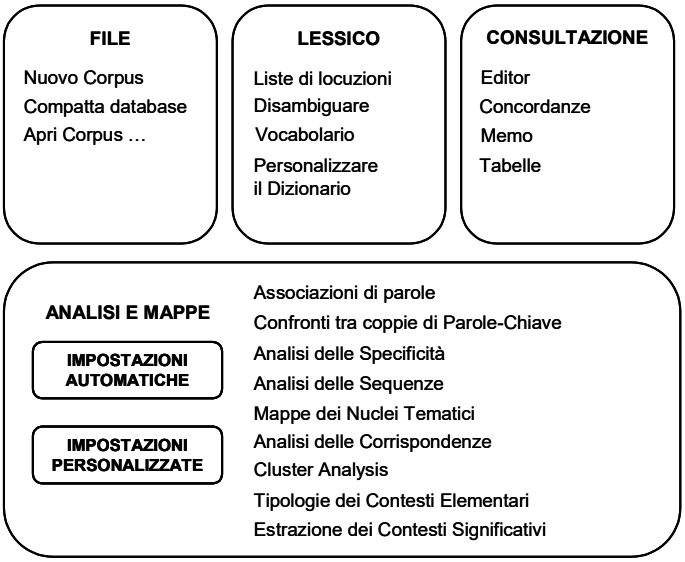



*Fig. 2 – Il menu T-LAB*

La Fig. 2 è un'immagine del menu T-LAB[7]. La Fig. 3 propone un diagramma di flusso che, a partire dall'importazione di un nuovo corpus, ricostruisce alcuni percorsi logici determinati dalle scelte dell'utilizzatore.

[7] Il menu della Fig. 1 è quello della versione T-LAB PRO 4.0, a cui si farà continuo riferimento nel corso delle prossime pagine.

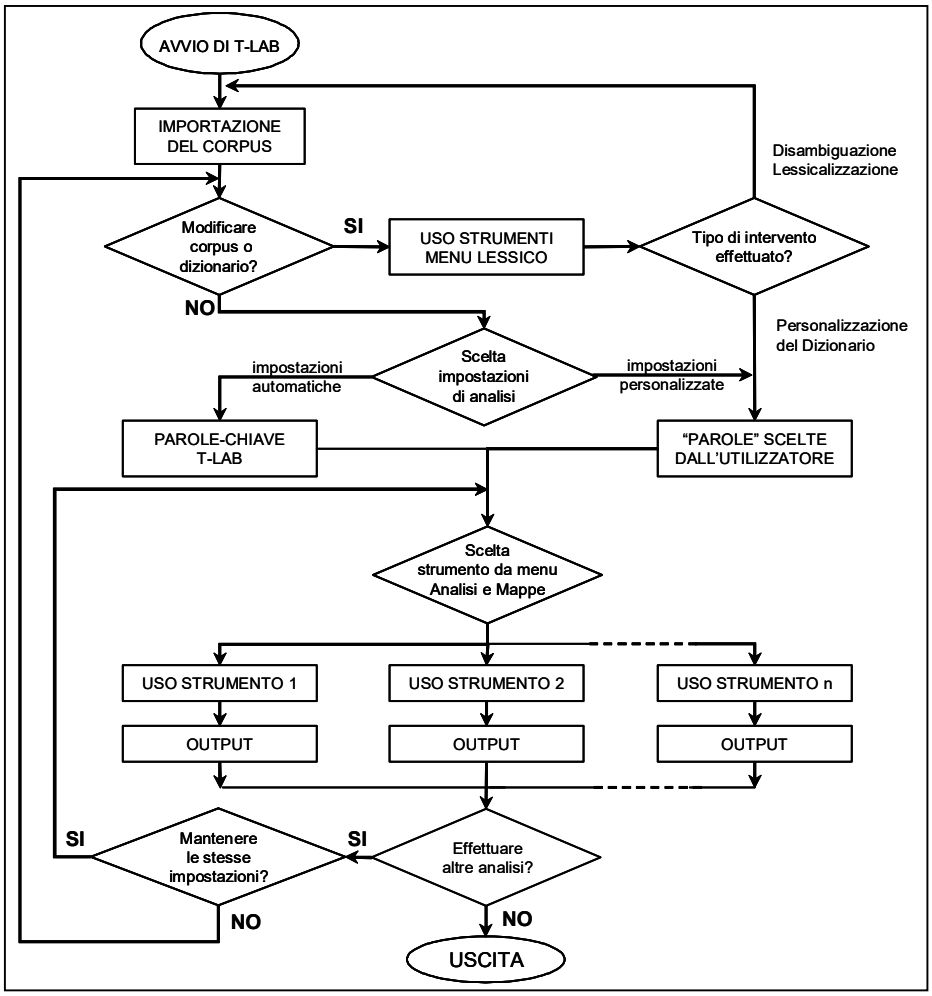



*Fig. 3 – T-LAB: diagramma di flusso delle operazioni*

All'interno di questi percorsi, i "nodi" che in modo più rilevante contribuiscono a definire le strategie di analisi sono i seguenti:

**a-** la personalizzazione del dizionario, cioè l'attribuzione di label alle unità lessicali (UL)[8];
**b-** la personalizzazione delle impostazioni, cioè la selezione delle UL da includere nell'analisi[9];

[8] T-LAB offre l'opzione della lemmatizzazione automatica; tuttavia, l'utilizzatore può proceder a interventi mirati o importare dizionari esterni (vedi Appendice n. 1).

**c-** la scelta degli strumenti di analisi;
**d-** l'interpretazione e l'uso degli output prodotti.

A ben vedere, punto per punto possono essere individuate delle analogie con l'uso di software per l'analisi dei dati numerici. Ad esempio, l'attribuzione di label (punto "a") si esplica attraverso operazioni analoghe alla codifica e alla ri-codifica[10] delle *variabili*, la selezione delle UL (punto "b") è analoga a quella dei *casi*, e così via.

In particolare, il punto "c" implica che, di volta in volta, il ricercatore valuti la congruenza tra i suoi obiettivi e le tecniche analitiche atte a perseguirli. In ipotesi, le valutazioni[11] più rilevanti sono riassunte nella Tab. 1, che propone una classificazione degli strumenti analitici di T-LAB mediante l'incrocio di due tipologie di obiettivi.

Ulteriori scelte strategiche sono messe in atto prima della fase di importazione. In particolare, l'eventuale partizione del corpus mediante l'uso di variabili e l'eventuale disattivazione del processo di lemmatizzazione automatica[12].

Quanto alla logica analitica dei vari strumenti, il fatto che – a seconda dei casi – possa essere giudicata affine a quella del text mining, dell'analisi di contenuto o dell'analisi del discorso è un aspetto assolutamente secondario. La vera differenza è data dalle domande a cui ciascuno di essi consente di rispondere.

E se, come afferma H. G. Gadamer[13] (1960, tr. it., p. 349), l'essenza della domanda sta nel porre e mantenere aperte delle possibilità, il "dialogo" con il corpus in analisi può iniziare con un qualunque strumento T-LAB. Quindi, inbase a quello che si riesce a "vedere" e a comprendere, ulteriori domande possono aprire alla ricerca di ulteriori risposte.

[9] Per facilitare le esplorazioni iniziali dei dati, T-LAB offre anche l'opzione "impostazioni automatiche".

[10] Un classico esempio di ri-codifica è quello che esita nella costruzione di *classi* di età (es. 18-24, 25-34, 35-44 anni, ecc.). Questa, da un punto di vista logico, è assimilabile alla costruzione di classi di equivalenza semantica.

[11] Tali valutazioni possono derivare da quelli che K. Krippendorff (2004, pp. 340-357) chiama "starting points" dell'analisi e che propone di distinguere in tre tipi: *text-driven*, *problem-driven* e *method-driven*. Il primo motivato dall'interesse a "familiarizzarsi" con il contenuto dei testi; il secondo prevalentemente orientato dall'esigenza di trovare risposte a specifiche domande; il terzo mosso dall'interesse ad applicare specifiche procedure.

[12] Le modalità operative sono illustrate nell'help del programma.

[13] Secondo questo autore (ibid., p. 418) «Non si fanno esperienze senza porre delle domande. Il pervenire a riconoscere che le cose stanno in modo diverso da come si credeva inizialmente presuppone ovviamente che si sia passati attraverso la fase della domanda, che ci sia chiesti se le cose stiano in questo o quel modo».

*Tab. 1 – Gli strumenti analitici di T-LAB*[14]

| | *ELABORARE UNA VISIONE DI SINTESI* | *METTERE A FUOCO DEI PARTICOLARI* |
|---|---|---|
| *ANALIZZARE RELAZIONI TRA PROFILI DI CO-OCCORRENZE* | Mappe dei nuclei tematici;<br>Analisi delle corrispondenze CE x UL;<br>Cluster analysis;<br>Tipologie dei CE | Associazioni di UL;<br>Confronti tra coppie di UL;<br>Sequenze di UL |
| *ANALIZZARE RELAZIONI TRA PROFILI DI OCCORRENZE* | Analisi delle corrispondenze UL x UC;<br>Cluster analysis | Specificità delle UC |

[14] Le abbreviazioni utilizzate vanno lette nel modo seguente: CE = Contesti Elementari; UL = Unità Lessicali; UC = Unità di Contesto. Le nozioni di profilo, occorrenza e co-occorrenza, sono spiegate nel paragrafo 2.2.2 (Parte prima).

# 2. Esercizio N. 1: cinque discorsi politici

## 2.1. Lucciole e lanterne

Il corpus "*governi*", costituito dalle dichiarazioni programmatiche di cinque Presidenti del Consiglio della Repubblica Italiana[1], è uno dei file inclusi nella versione demo di T-LAB. Al di là del suo interesse storico, il motivo per cui ho deciso di utilizzarlo come materiale esercitativo è di tipo metodologico e didattico; infatti, chi sia interessato, potrà "replicare"[2] le mie stesse analisi e ottenere gli stessi risultati. A questo scopo, è sufficiente effettuare il download della demo e quello di alcuni file[3] in cui sono archiviate tutte le operazioni descritte in questo paragrafo.

Mentre mi accingo a costruire e a descrivere il mio percorso di esplorazione, so che alcune delle mie scelte strategiche saranno determinate dal modo in cui gestirò i cosiddetti trattamenti preliminari[4]. Quindi, le prime operazioni[5] che metto in atto sono orientate da due obiettivi:

- evitare di prendere lucciole per lanterne;
- scegliere le lucciole per la mia lanterna.

Fuor di metafora: in primo luogo intendo risolvere i casi di ambiguità più

[1] Berlusconi 1° (maggio 1994), Prodi (maggio 1996), D'Alema (ottobre 1998), Amato (aprile 2000), Berlusconi 2° (giugno 2001).

[2] In linea generale, per replicare le stesse analisi e ottenere gli stessi risultati, due o più ricercatori che usano T-LAB devono condividere le seguenti risorse: lo stesso corpus (nella sua forma definitiva), lo stesso dizionario e la stessa selezione delle parole-chiave. In alternativa, possono condividere il database di sessione prodotto da T-LAB.

[3] Il link della versione demo di T-LAB è http://www.tlab.it/it/download.asp. Il link dei materiali relativi a questo esercizio (corpus "governi") può essere richiesto a info@tlab.it.

[4] Revisione ortografica, eventuale codifica di sottoinsiemi, disambiguazioni, ecc. In generale, molte di queste operazioni possono essere effettuate anche in un secondo momento, cioè dopo una prima esplorazione del corpus con l'opzione "Impostazioni automatiche".

[5] In questo caso, il corpus è già "preparato", e la revisione ortografica non è necessaria.

rilevanti[6]; in secondo luogo intendo costruire una *lista* delle unità lessicali (UL) da utilizzare come "griglia" nelle analisi successive.

Operativamente, la riduzione delle ambiguità comporta due fasi.

La prima passa attraverso la lessicalizzazione[7] di sintagmi costituiti da due o più parole e che rinviano a un significato unitario. Ad esempio, il riconoscimento di una lessia ("*finanza pubblica*") e la sua successiva trasformazione in stringa unitaria ("*finanza_pubblica*"), oltre ad individuare uno specifico concetto, consente di risolvere molti casi di ambiguità inerenti le occorrenze delle singole parole che la compongono (es. "*pubblica*", a seconda che sia verbo o aggettivo, è un'espressione che può avere significati molto diversi tra loro).

La seconda fase della disambiguazione concerne invece le singole parole; in particolare, i casi di omografia. Ad esempio, usando un apposito strumento T-LAB[8], la forma "*legge*", a seconda che corrisponda a un sostantivo o a una voce verbale, può essere trasformata in "*legge_sost*" o "*legge_verb*".

Il motivo per cui l'ordine di successione tra le due fasi (lessicalizzazione e disambiguazione) risulta sensato è abbastanza intuitivo; in questo modo, infatti, nella seconda fase i casi del tipo "*pubblica*" sono diventati meno numerosi e la loro eventuale disambiguazione richiede meno tempo.

Quanto ai casi di lessie complesse, la versione italiana di T-LAB, in modo automatico, ne riconosce poco più di 2.000; ma, come si sa, il linguaggio è mobile e ogni testo ha le sue specificità. Quindi, dopo aver effettuato l'importazione[9] del corpus "*governi*", decido di utilizzare uno strumento del menu *Lessico* (*Liste di locuzioni*) che consente di trovare lessie non incluse nel database T-LAB.

La lista ottenuta, con soglia a 4 occorrenze, include 45 elementi. La scorro rapidamente e decido di applicare[10] al corpus una lista ridotta costituita dai seguenti 28 elementi:

Banca d'Italia, Chiesa cattolica, conti pubblici, criminalità organizzata, documento di programmazione, finanza pubblica, giustizia sociale, legge elettora-

[6] Evidentemente, i criteri di rilevanza sono "soggettivi". I miei: soffermarmi sulle parole chiave del corpus e valutare il loro peso in termini di occorrenze.

[7] Per la definizione di questa nozione, vedi paragrafo 2.2.2 (Parte prima).

[8] Lo strumento in questione è *Disambiguare* (menu *Lessico*).

[9] L'importazione del corpus è un'operazione preliminare all'uso degli strumenti T-LAB. Poiché alcuni di questi sono orientati a risolvere i casi di ambiguità e a "trasformare" alcune caratteristiche del corpus, il passaggio attraverso due (o più) importazioni serve ad ottimizzare i risultati delle analisi successive. Tuttavia, l'utilizzatore può iniziare il processo di analisi anche dopo la prima importazione e – volendo – può utilizzare le *Impostazioni automatiche* del programma.

[10] Le modalità operative sono illustrate nel manuale d'uso T-LAB. Il risultato è che viene generato un nuovo corpus in cui gli elementi presenti nella lista sono trasformati in stringhe unitarie (es. "*Banca d'Italia*" → "*Banca_d_Italia*").

le, legge finanziaria, mercato del lavoro, ministro del tesoro, mondo cattolico, mondo politico, patto di stabilità, politiche sociali, Presidente del Consiglio, Presidente della Repubblica, riforma della scuola, riforme istituzionali, Romano Prodi, senso dello Stato, servizi sociali, società civile, solidarietà sociale, Stato sociale, sviluppo economico, testi unici, Unione Europea.

Ad operazione ultimata, effettuo una nuova importazione del corpus.

Dopo una manciata di secondi, T-LAB mi restituisce una serie di informazioni che trovo riassunte nella funzione *Memo* (Fig. 1). In questa fase prendo nota solo della soglia di frequenza[11] (7), cioè della quantità minima di occorrenze che caratterizza le unità lessicali rilevanti.

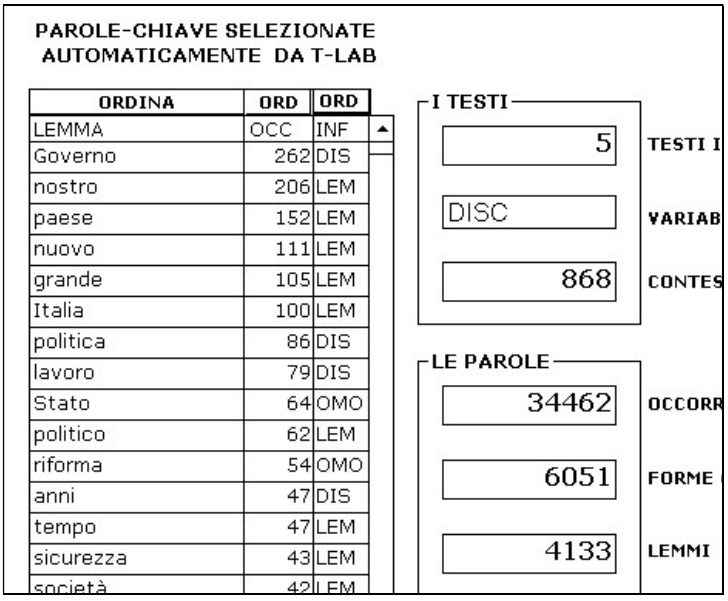



*Fig. 1 – Memo*

[11] In T-LAB il calcolo della soglia di frequenza è realizzato mediante un algoritmo documentato da S. Bolasco (1999), e che prevede i seguenti passi: a) individuazione del range delle frequenze basse, che – a partire dalla frequenza minima ("1") – è definito dal primo "salto" nei valori crescenti delle occorrenze (es. 1, 2, 3,... 18, **19**, 21); b) scelta del valore di soglia che, a seconda delle dimensioni del corpus, viene fatto corrispondere al valore minimo nel primo decile (10%) o nel secondo decile (20%) del range delle frequenze basse.

Per fare uno screening dei casi di omografia, uso la funzione *Personalizzare il dizionario*[12] e salvo una tabella con alcune informazioni concernenti tutte le unità lessicali (UL) del corpus appartenenti alla categoria delle "parole piene"[13]. Quindi, apro il file con la tabella salvata, seleziono le UL con almeno 7 occorrenze, classificate come omografe o ambigue e le ordino per valori decrescenti. In questo modo ottengo una lista di 62 elementi. Un rapido esame di questa lista, mediato dalla mia conoscenza del linguaggio dei politici italiani, mi porta a ridurre a 14 il numero delle forme da verificare ed eventualmente da disambiguare.

La lista ridotta è la seguente (i valori tra parentesi sono occorrenze):

stato (64); rispetto (40); diritto (31); diritti (25); pubblico (20); legge (19); interessi (18); pubblica (18); condizioni (17); processo (16); ricerca (15); personale (12); piano (10); processi (7).

Prima di procedere, verifico ulteriormente la tabella con il dizionario del corpus alla ricerca di eventuali forme non riconosciute da T-LAB[14], con occorrenze uguali o maggiori di 7 e omografe: nessun caso.

A questo punto, passo a usare la funzione *Concordanze* per verificare i contesti d'uso delle 14 UL selezionate (Fig. 2). Di queste, solo 4 sono effettivamente da disambiguare[15]: "*rispetto*", "*interessi*", "*ricerca*" e "*processi*".

Quindi, tramite la funzione *Disambiguare* (Fig. 3), procedo alle seguenti trasformazioni:

- "*rispetto*" ➔"*rispetto_sost*", per marcare tutti gli usi in cui "*rispetto*" è un sostantivo (es. "*il rispetto delle regole*");
- "*interessi*" ➔"*interessi_econ*", per marcare gli usi in cui gli "*interessi*" sono di tipo economico-finanziario (es. "*riduzione degli interessi*");
- "*ricerca*" ➔ "*ricerca_scien*", per marcare gli usi in cui si parla di "*ricerca*" scientifica (es. "*sforzo di ricerca e innovazione*");
- "*processi*" ➔"*processi_giud*", per marcare gli usi in cui i "processi" sono di tipo giudiziario (es. "*i processi pendenti*")[16].

[12] Questo strumento T-LAB, oltre a consentire di consultare e di modificare il dizionario del corpus, consente l'importazione di dizionari esterni (vedi Appendice n. 1).

[13] La distinzione tra parole "piene" e "vuote" dipende dalla teoria del significato a cui si fa riferimento. In genere, sono considerate "piene" le parole che appartengono alle categorie dei nomi, dei verbi, degli aggettivi e degli avverbi.

[14] Le forme non riconosciute, cioè non incluse nei dizionari T-LAB, vengono classificate come "NCL".

[15] I motivi per i quali le altre 10 non risultano ambigue dipendono da due fattori: dal funzionamento di T-LAB (ad es. la lemmatizzazione automatica riconosce i casi in cui "*stato*" è participio passato del verbo "*essere*") e dalle peculiarità del corpus importato (ad es. "*legge*" non è mai voce del verbo "*leggere*").

[16] Gli altri usi delle quattro unità lessicali (ad es. "*rispetto* al 1997", "*interessi* generali del paese", "*ricerca* di un equilibrio", "*processi* di mondializzazione") li lascio invariati.

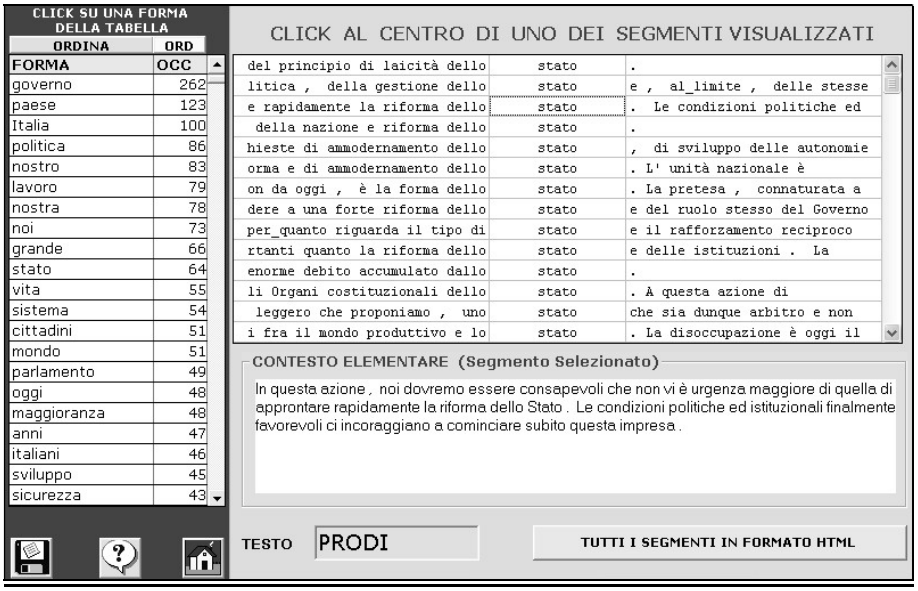



*Fig. 2 – Concordanze*

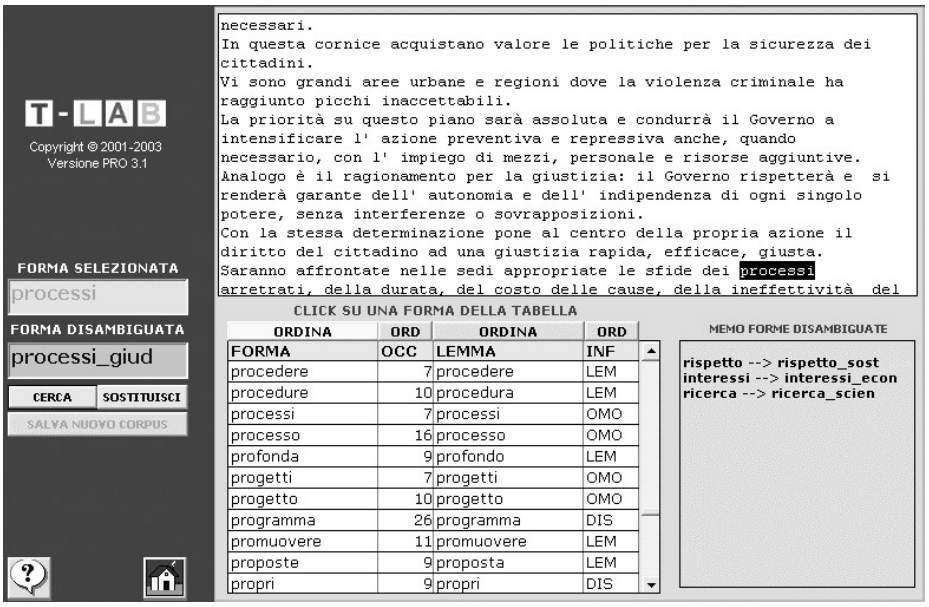



*Fig. 3 – Disambiguare*

Al termine di queste operazioni preliminari, cioè dopo circa una mezz'ora, procedo alla definitiva importazione del corpus: quella orientata a organizzare il database di sessione e ad utilizzare gli strumenti del menu *Analisi e mappe*. Scelgo l'opzione con *lemmatizzazione automatica* e mi appresto a verificare la lista delle unità lessicali.

A questo punto, per riprendere una metafora già utilizzata, il mio obiettivo è quello di *scegliere le lucciole per la mia lanterna*, cioè di costruire una *lista* delle unità lessicali (UL) da utilizzare come "griglia" nelle analisi successive.

Scelgo *Impostazioni personalizzate* (Fig. 4) e ottengo una lista con 561 lemmi, 482 dei quali risultano selezionati da T-LAB come *parole-chiave*[17]. La soglia di frequenza è 7.

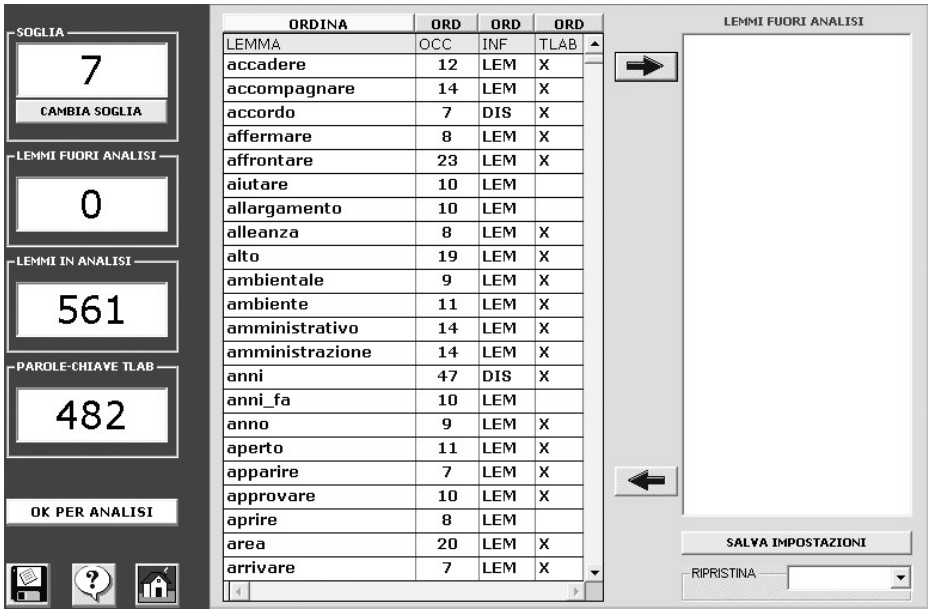



*Fig. 4 – Impostazioni personalizzate*

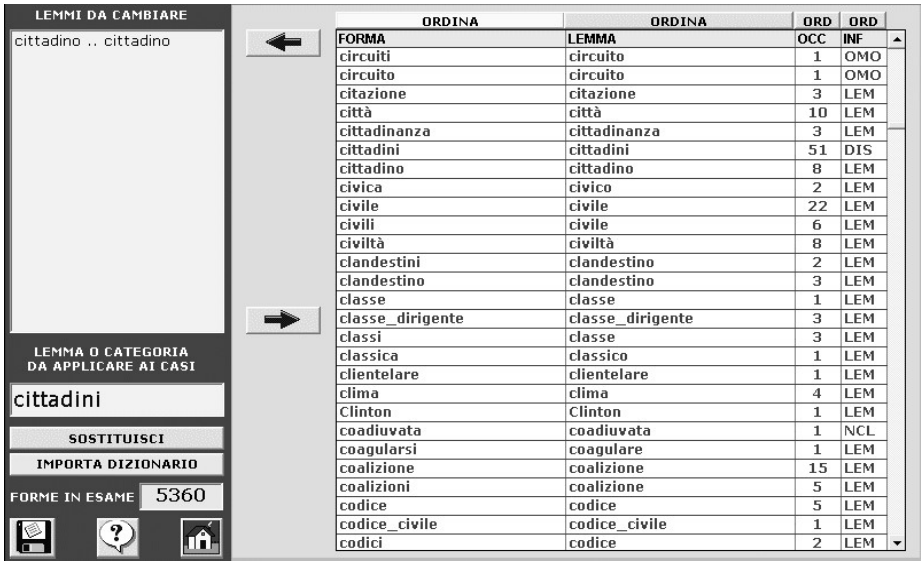
Poiché ho in mente di esemplificare il funzionamento di alcuni strumenti di analisi che consentono di usare max 500 elementi, decido di intervenire sulla lista delle parole-chiave. Per averla tutta sotto mano, ne faccio una stampa

[17] La procedura tramite la quale T-LAB seleziona una lista di parole-chiave è spiegata nell'help del programma. In questo caso (corpus con 5 sottoinsiemi), usando il test del chi quadro, sono state selezionate le unità lessicali che in modo più significativo organizzano le differenze tra i cinque sottoinsiemi, cioè tra i cinque discorsi dei Presidenti del Consiglio.

e, con il vecchio metodo carta-e-matita, prendo nota degli interventi da effettuare. Quelli che metto in atto sono i seguenti[18]:

- deseleziono alcune parole-chiave che ritengo poco pregnanti o troppo generiche (es. "*accadere*", "*certo*", "*dire*", ecc.);
- seleziono come parole-chiave nuovi elementi (UL), non segnalati da T-LAB e che, per esplorare i contenuti del corpus, mi sembrano interessanti (es. "*aiutare*", "*collaborazione*", "*difesa*", ecc.)[19].

Mentre scorro la lista, verifico che alcune parole di significato molto simile sono da "raggruppare"[20]; quindi, dopo aver concluso le operazioni di selezione (vedi sopra), salvo le *Impostazioni di analisi* ed apro la finestra *Personalizzazione del dizionario* (Fig. 5), quella che consente di cambiare le label delle unità lessicali, sia tramite interventi mirati che importando dizionari ad hoc.




| ORDINA | ORDINA | ORD | ORD |
|---|---|---|---|
| FORMA | LEMMA | OCC | INF |
| circuiti | circuito | 1 | OMO |
| circuito | circuito | 1 | OMO |
| citazione | citazione | 3 | LEM |
| città | città | 10 | LEM |
| cittadinanza | cittadinanza | 3 | LEM |
| cittadini | cittadini | 51 | DIS |
| cittadino | cittadino | 8 | LEM |
| civica | civico | 2 | LEM |
| civile | civile | 22 | LEM |
| civili | civile | 6 | LEM |
| civiltà | civiltà | 8 | LEM |
| clandestini | clandestino | 2 | LEM |
| clandestino | clandestino | 3 | LEM |
| classe | classe | 1 | LEM |
| classe_dirigente | classe_dirigente | 3 | LEM |
| classi | classe | 3 | LEM |
| classica | classico | 1 | LEM |
| clientelare | clientelare | 1 | LEM |
| clima | clima | 4 | LEM |
| Clinton | Clinton | 1 | LEM |
| coadiuvata | coadiuvata | 1 | NCL |
| coagularsi | coagulare | 1 | LEM |
| coalizione | coalizione | 15 | LEM |
| coalizioni | coalizione | 5 | LEM |
| codice | codice | 5 | LEM |
| codice_civile | codice_civile | 1 | LEM |
| codici | codice | 2 | LEM |

*Fig. 5 – Personalizzazione del dizionario*

[18] In T-LAB tutti gli interventi sul database di sessione sono "memorizzati"; quindi, ad ogni momento l'utilizzatore può verificare la composizione delle sue liste e i relativi dizionari.

[19] Il motivo per i quale queste parole non sono state selezionate automaticamente da T-LAB dipende dal fatto che il test del chi quadro non ha rilevato differenze significative nella loro distribuzione entro i cinque sottoinsiemi considerati.

[20] I criteri di questa operazione sono in gran parte empirici. Ad es.: evitare che i raggruppamenti incrementino il tasso di ambiguità, "recuperare" nel processo di analisi le forme con frequenze inferiori alla soglia stabilita, ottimizzare le "griglie" per l'analisi del contenuto ecc. Ulteriori criteri, fondati su argomentazioni di tipo tecnico, sono illustrati in un volume di S. Bolasco (1999, pp. 213-225).

In questo caso decido di intervenire il meno possibile. La lista è infatti piuttosto contenuta e le occorrenze delle parole sono distribuite in modo abbastanza equilibrato. Alcuni esempi delle fusioni[21] effettuate:

- *"cittadini" + "cittadino" = "cittadini"*
- *"competitività" + "competizione" = "competizione"*
- *"efficiente" + "efficienza" = "efficienza"*
- *"scolastico" + "scuola" = "scuola".*

Per rendere operative le mie scelte, torno in *Impostazioni personalizzate* (Fig. 4). Rilevo che, dopo i miei interventi, la lista delle parole-chiave include 478 elementi; la scorro per verificarne la composizione e do l'OK per le analisi statistiche[22].

A questo punto, mi trovo a scegliere tra due *strategie di analisi* che consentono di costruire due diverse *rappresentazioni* dei contenuti del corpus:

- la prima prescinde dalla suddivisione per sottoinsiemi (in questo caso, i cinque discorsi) e comporta l'analisi di una tabella con in riga i contesti elementari e in colonna le parole chiave;
- la seconda fa perno sulla suddivisione tra sottoinsiemi e comporta l'analisi di una tabella costituita da tante righe quante sono le parole chiave e tante colonne quanti sono i sottoinsiemi.

Mentre la prima strategia è orientata ad esplorare le *associazioni* tra parole e frasi (i contesti elementari) attraverso strumenti che analizzano le relazioni tra profili di *co-occorrenze*, la seconda è orientata ad esplorare somiglianze e differenze tra i profili delle *occorrenze*, cioè – in questo caso – tra le diverse *posizioni* dei Presidenti del Consiglio. Nel primo caso, tutti i criteri di classificazione derivano dal processo di analisi; nel secondo, almeno uno è noto e determinato a priori: quello che suddivide il corpus in cinque sottoinsiemi.[23]

Poiché mi sono assunto il compito di fare da guida all'uso degli strumenti

[21] Il termine fusione sta ad indicare raggruppamenti di due o più forme, vuoi in base a un criterio di equivalenza lessicale (lemmatizzazione), vuoi in base a un criterio di equivalenza semantica. La label di ogni raggruppamento effettuato può essere scelta tra quelle dei suoi elementi o in base ad altri criteri definiti dall'utilizzatore. Ad esempio, per marcare le differenze tra i vari tipi di interventi effettuati, alcuni ricercatori usano aggiungere dei suffissi dopo ogni label (es. "*cittadini$*").

[22] In questo caso, scelgo l'opzione "*parole-chiave*"; l'altra – prevista da T-LAB – è "*tutti i lemmi*" in elenco.

[23] Nelle due strategie, la variabile utilizzata per suddividere il corpus in cinque sottoinsiemi "gioca" un ruolo diverso: nel primo caso, in quanto caratteristica "supplementare" dei contesti elementari, è *illustrativa*; nel secondo caso, in quanto label delle colonne in tabella, è *attiva.*

T-LAB, decido di utilizzare prima una strategia e poi l'altra, evocando le loro differenze tramite due denominazioni metaforiche: *strategia del pescatore* e *strategia del fotografo*. Ciò nell'ipotesi che il pescatore, quando getta la propria rete, non sa quali tipi di pesci saranno presi e quanti; mentre il fotografo, prima di scattare le sue istantanee, ha già qualche idea sulle fisionomie dei suoi soggetti.

## 2.2. La strategia del pescatore

In T-LAB, gli strumenti per attuare questo tipo di strategia sono tre: *Mappe dei nuclei tematici* (MNT), *Analisi delle corrispondenze segmenti*[24] *x lemmi* (ACSL)[25], e *Tipologie dei contesti elementari* (TCE).

Tutti e tre implicano la costruzione e l'analisi di una tabella del tipo presenza/assenza ("0", "1"), con tante righe ("*m*") quante sono i contesti elementari e tante colonne ("*n*") quante sono le unità lessicali. Il primo strumento (MNT) consente di costruire una rappresentazione del corpus che funziona come una mappa "navigabile" (Fig. 6)[26]: i nuclei tematici sono in qualche modo tutti presenti, ma i loro contenuti e le loro relazioni non sono facili da interpretare. Il secondo (ACSL) risulta particolarmente utile quando l'interesse del ricercatore è orientato all'esplorazione dello spazio fattoriale determinato dall'analisi dell'intera tabella "*m* x *n*" senza mettere in gioco "riduzioni" ottenute mediante cluster analysis. Il terzo (TCE), usando in modo integrato vari algoritmi[27], produce una nuova variabile (cluster) che entra nell'analisi e determina una rappresentazione "sintetica" dei dati secondo una logica affine all'analisi di contenuto[28].

Decido quindi di usare lo strumento tipologie dei contesti elementari.

Dopo circa un minuto, ottengo un output con una soluzione a cinque cluster[29] (Fig. 7). Mediante grafici e tabelle verifico alcune sue caratteristiche; quindi decido di adottarla.

[24] In questo caso, *segmenti* sta per *contesti elementari*.

[25] Eventualmente seguita da cluster analysis.

[26] Ad esempio, cliccando su "*maggioranza*", è possibile verificare che questa label indica un piccolo gruppo di parole associate: *bipolarismo*, *legge_elettorale*, *maggioranza*, *maggioritario* e *opposizione.*

[27] Due modelli di cluster analysis e l'analisi delle corrispondenze.

[28] Più precisamente: in TCE lo spazio della rappresentazione è organizzato da un'analisi delle corrispondenze applicata a una tabella con le unità lessicali in riga e i cluster in colonna.

[29] Questo strumento T-LAB, per default, propone una soluzione che include da 4 a 6 cluster; tuttavia l'utilizzatore può agevolmente esplorare altre soluzioni e decidere nel modo per lui più opportuno.

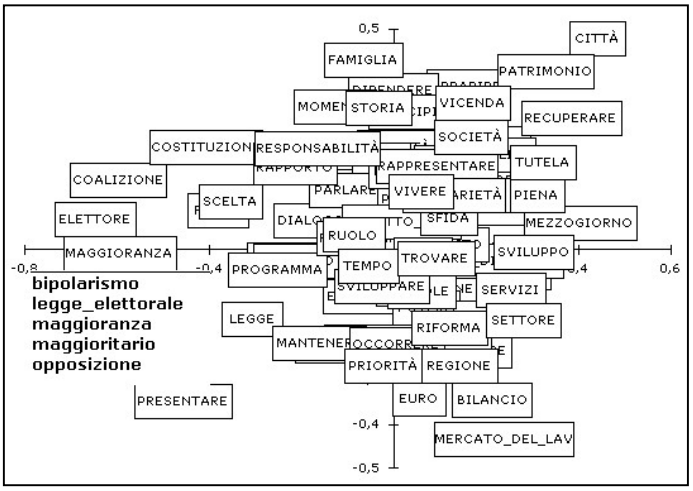



*Fig. 6 – Mappa dei nuclei tematici*

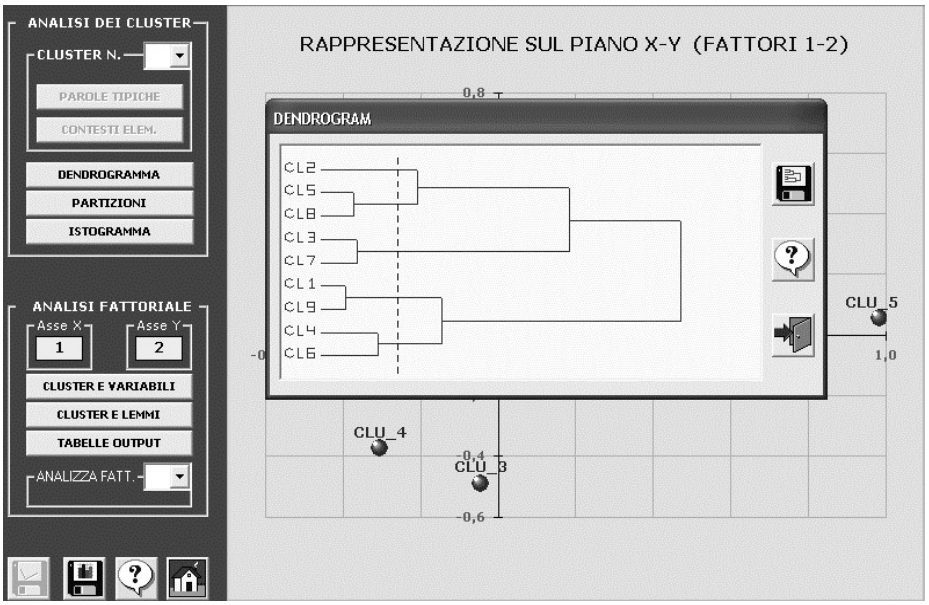



*Fig. 7 – Tipologie dei contesti elementari*

Negli output, ogni cluster è descritto mediante le frasi (contesti elementari) e le parole (unità lessicali) che più lo caratterizzano: le une e le altre selezionate e ordinate mediante il test del chi quadro[30].

In questo caso, i contesti elementari classificati sono 821 (su un totale di 868, il 95%), e la loro distribuzione (Fig. 8) risulta essere la seguente: 21,68% (clust. 1), 13,89% (clust. 2), 18,64 (clust. 3), 28,87% (clust. 4), 16,93% (clust. 5).

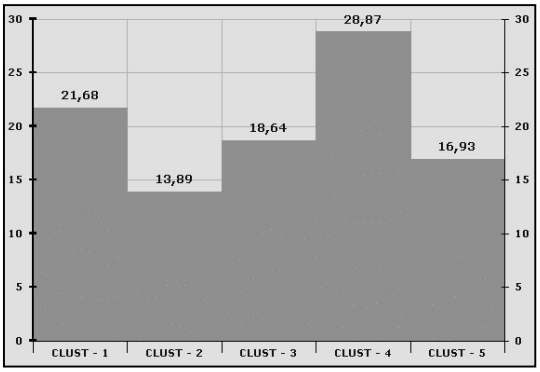



*Fig. 8 – Istogramma dei cluster*

Per non appesantire l'esposizione, di seguito riporto le prime 35 parole caratteristiche di ogni cluster[31] e, per ciascuno di essi, una delle prime frasi (c.e.) selezionate da T-LAB.

CLUSTER N. 1

**bilancio**; riforma; necessario; pubblico; **finanziario**; **fiscale**; intervento; priorità; mercato_del_lavoro; amministrazione; occupazione; **spesa**; norma; governi; amministrativo; provvedimento; risorse; **risanamento**; precedente; lavoro; ridurre; azione; attuale; euro; terreno; approvare; utilizzare; riduzione; avviare; ripetere; continuare; strutturale; flessibilità; linea; Stato.

«*La riforma dell'IRPEF e delle pensioni sociali minime non è naturalmente che uno degli elementi di una gestione innovativa della leva fiscale e della spesa pubblica, due obiettivi che perseguiremo nelle nuove condizioni istituzionali connotate dalla riunione dei Ministeri delle finanze, del bilancio e del tesoro*».

[30] Una descrizione di questo test è contenuta nel paragrafo 2.3.2.

[31] La loro sequenza è ordinata in modo automatico mediante il test del chi quadro. Le marcature in grassetto sono opera mia.

CLUSTER N. 2

**governo, parlamento**; costituzionale; parlamentare; presentare; **riforme_istituzionali**; legge; legislatura; decisione; giudicare; elettore; **regole**; **elettorale**; direzione; **legge_elettorale**; mafia; **maggioritario**; rendere; chiaro; attività; mantenere; legislativo; programma; confronto; intenzione; scegliere; aiutare; spirito; esecutivo; politico; eguaglianza; esclusione; rafforzamento; euro; **rispetto_sost**; strategia.

«*Attraverso i referendum, le proposte di legge di iniziativa popolare, l'attività parlamentare esiste un largo concorso di idee e di progetti che va verso il rafforzamento del maggioritario, magari con ipotesi diverse tra loro, ma procede in quella direzione e non mira a tornare indietro*».

CLUSTER N. 3

**internazionale**; politica; ambientale; **Unione_Europea**; nazionale; **Europa**; difesa; Mezzogiorno; **integrazione**; significare; condizione; confini; **sviluppo**; Italia; comune; sicurezza; prospettiva; storico; costruire; posti_di_lavoro; rappresentare; equilibrio; **politica_estera**; **europeo**; ultimo; fase; allargamento; stabilità; economico; crescita; rapporto; coesione; militare; migliore; permettere.

«*L'allargamento è un passaggio che abbiamo davanti, che è tanto inevitabile quanto giusto; ma l'allargamento senza una più forte anima e macchina politica dell'Europa può rappresentare un abbassamento nel livello dell'integrazione, che non possiamo permetterci*».

CLUSTER N. 4

**giovane**; **donna**; civile; **scuola**; società; uomini; **qualità**; civiltà; moderno; nostro; generazione; valorizzare; comunità; pubblica; **ricerca_scien**; giusto; grado; **famiglia**; **città**; **ragazzo**; vita; impegno; autonomia; **figli**; **lavoratore**; capire; opportunità; miliardo; professionale; livello; efficace; ragione; recuperare; libertà; sistema.

«*Se non funziona la scuola in un Paese, non c'è futuro. Ci impegniamo per costruire un complesso di sistemi formativi che funzionino in sintonia con il loro territorio e con le culture più avanzate del nostro tempo, che possano ragionevolmente esigere dai giovani il massimo impegno e che sappiano prepararli socialmente ad una competizione sana in un ambiente cooperativo*».

CLUSTER N. 5

**maggioranza**; **opposizione**; presidente; **coalizione**; forza; senatore; **bipolarismo**; centrosinistra; elezione; popolare; **democratico**; scelta; costituzione; **alleanza**; deputato; elettorale; Ulivo; sinistra; condividere; nascere; elettore; **partiti**; **dialogo**; Presi-

dente_del_Consiglio; liberale; sentire; repubblica; esecutivo; legittimo; paese; convinzione; momento; **governare**; voto; repubblicano.

«*È questo un momento di grande speranza per il nostro Paese. Abbiamo di fronte a noi prospettive ed occasioni che forse mai si sono avute in passato. Di questo noi tutti dobbiamo condividere la responsabilità: il Governo, la coalizione di maggioranza e l'opposizione*».

La Tab. 1 esemplifica l'output che descrive le caratteristiche dei cluster (in questo caso, il cluster n. 3): per ciascuna parola tipica (es. "internazionale"), insieme al corrispondente valore del chi quadro (41,44), è riportato il totale dei contesti elementari del corpus (25) e del cluster (17) in cui essa risulta presente.

A partire da questi risultati, è possibile costruire un itinerario interpretativo le cui tappe variano in funzione dei tipi di dati che vengono presi in considerazione e dei modelli teorici a cui il ricercatore fa riferimento.

In generale, un buon punto di partenza è costituito dall'osservazione dell'"insieme" dei cinque cluster e, per ciascuno di essi, dall'evidenziazione di alcuni termini chiave. Questo tipo di operazione si esplica attraverso l'individuazione di parole-indizi che possono fondare inferenze abduttive del tipo "il cluster X è caratterizzato da Y", dove "Y" sta per un insieme di caratteristiche che la "presunzione di isotopia"[32] riconduce a una classe (o categoria) di contenuto ovvero a un qualche affinità tematica[33].

In prima istanza, nell'esempio considerato i temi delle isotopie[34] individuate risultano essere i seguenti:

- "i problemi di bilancio e la riforma fiscale" (cluster n. 1);
- "le riforme istituzionali e la legge elettorale" (cluster n. 2);
- "le prospettive dell'integrazione europea" (cluster n. 3);
- "i protagonisti e i problemi della società civile" (cluster n. 4);
- "il bipolarismo e le regole della democrazia" (cluster n. 5).

Per approfondire l'analisi ci si può muovere in due direzioni: a) verificare specifiche associazioni di parole e/o frasi entro ciascun cluster; b) esplorare le relazioni tra i cluster.

[32] Questa espressione ("*présomption d'isotopie*") è stata proposta da F. Rastier (1987, p. 12) per marcare il fatto che, nel riconoscimento di un'isotopia, la visione d'insieme precede l'attribuzione di significato ai singoli elementi («où la classe determine l'élement, et où le global détermine le local»). Per la definizione di isotopia si vedano i paragrafi 2.2.1. e 2.3.5. (Parte prima).

[33] Sia la nozione di tema che quella di isotopia fanno riferimento alla costruzione di classi di equivalenza.

[34] In questo caso F. Rastier (1995) parlerebbe di "isotopie generiche dominanti".

*Tab. 1 – Cluster N. 3 (parole tipiche)*

| *Unità lessicali* | *CHI²* | *IN CLU* | *IN TOT* |
|---|---|---|---|
| internazionale | 41,439 | 17 | 25 |
| politica | 37,981 | 35 | 79 |
| ambientale | 29,620 | 8 | 9 |
| Unione_Europea | 29,620 | 8 | 9 |
| nazionale | 28,000 | 12 | 18 |
| Europa | 27,847 | 17 | 31 |
| difesa | 27,028 | 11 | 16 |
| Mezzogiorno | 23,741 | 14 | 25 |
| integrazione | 21,514 | 8 | 11 |
| significare | 21,514 | 8 | 11 |
| condizione | 20,991 | 7 | 9 |
| confini | 20,991 | 7 | 9 |
| sviluppo | 20,659 | 19 | 42 |
| Italia | 19,205 | 31 | 86 |
| comune | 17,418 | 14 | 29 |
| sicurezza | 15,793 | 17 | 40 |
| prospettiva | 14,825 | 10 | 19 |
| storico | 13,845 | 6 | 9 |
| costruire | 13,473 | 9 | 17 |
| posti_di_lavoro | 10,800 | 7 | 13 |
| rappresentare | 10,800 | 7 | 13 |
| equilibrio | 10,800 | 7 | 13 |
| politica_estera | 10,251 | 5 | 8 |
| europeo | 10,185 | 14 | 36 |
| ultimo | 9,249 | 8 | 17 |
| fase | 9,240 | 7 | 14 |
| allargamento | 8,180 | 5 | 9 |
| stabilità | 7,901 | 6 | 12 |
| economico | 7,126 | 12 | 33 |
| crescita | 7,066 | 8 | 19 |
| rapporto | 6,960 | 10 | 26 |
| coesione | 6,904 | 4 | 7 |
| militare | 6,904 | 4 | 7 |
| migliore | 6,568 | 5 | 10 |
| permettere | 6,568 | 5 | 10 |

Nel primo caso (a), focalizzando l'attenzione sul cluster n. 3, possiamo articolare le relazioni tra alcuni sottoinsiemi di parole. Ad esempio, possiamo individuare una componente di "difesa" (*"difesa"*, *"confini"*, *"sicurezza"*, *"militare"*) e una di "sviluppo" (*"sviluppo"*, *"posti_di_lavoro"*, *"economico"*), tenute insieme da un'ipotesi di "integrazione" (*integrazione*, *coesione*).

Per entrare nel dettaglio[35], utilizzando lo strumento *Associazioni di parole*, dopo aver selezionato il sottoinsieme[36] che corrisponde al cluster n. 3, possiamo verificare i significati contestuali di specifiche parole. Ad esempio, un click sulla parola-chiave "*Unione_Europea*" genera il grafico in Fig. 9.

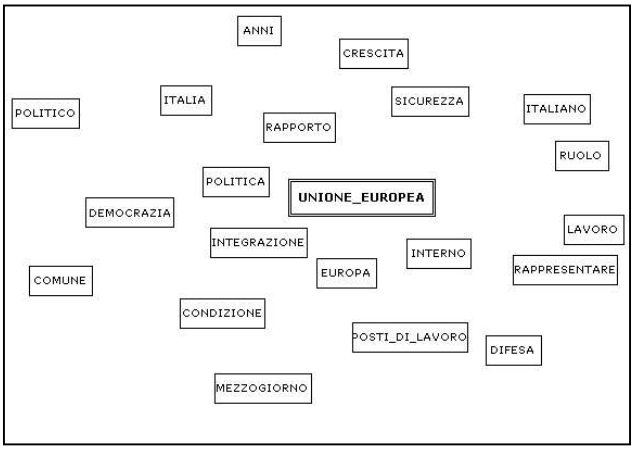



*Fig. 9 – Cluster 3: Associazioni relative a Unione_Europea*

Quanto alle relazioni tra i cinque cluster, lo strumento *Tipologie dei contesti elementari* consente di esplorarle in vario modo.

In primo luogo, osservando il dendrogramma in Fig. 7, possiamo verificare che due coppie di cluster hanno "strette" relazioni di parentela: il cluster 2 ("*le riforme istituzionali e la legge elettorale"*) con il cluster 5 ("*il bipolarismo e le regole della democrazia"*) da un parte e il cluster 1 ("*i problemi di bilancio e la riforma fiscale"*) con il cluster 4 ("*i protagonisti e i problemi della società civile"*) dall'altra. Diversamente, il cluster 3 ("*le prospettive dell' integrazione europea"*) sembra caratterizzato da una "linea" più autonoma.

In secondo luogo, attraverso le tabelle e i grafici ottenuti mediante analisi

[35] I tipi di approfondimenti che T-LAB consente sono di vario tipo, ma la "qualità" dell'interpretazione non è funzione della quantità dei dati in output.

[36] L'uso dello strumento *Tipologie dei contesti elementari* genera dei sottoinsiemi del corpus che T-LAB archivia mediante una nuova variabile ("clust_tlab"). Poiché tanto le loro "specificità" quanto l'analisi delle loro corrispondenze fanno parte dell'output, ulteriori esplorazioni di questi sottoinsiemi sono consentite solo tramite gli strumenti *Associazioni* e *Mappe dei nuclei tematici.*

delle corrispondenze[37], possiamo esplorare l'insieme delle relazioni su vari "piani" che corrispondono ad altrettante coppie di dimensioni fattoriali.

Ad esempio, osservando la Fig. 10, potremmo soffermarci a considerare che le differenze più marcate, quelle che determinano la bipolarità del primo fattore, sono tra il cluster 1 e il cluster 5[38]; potremmo inoltre osservare che il cluster 1 sembra più "parlato" dal discorso di Amato, che il cluster 3 e il cluster 4, oltre ad organizzare la bipolarità del secondo fattore, sono molto vicini tra loro; e così via[39].

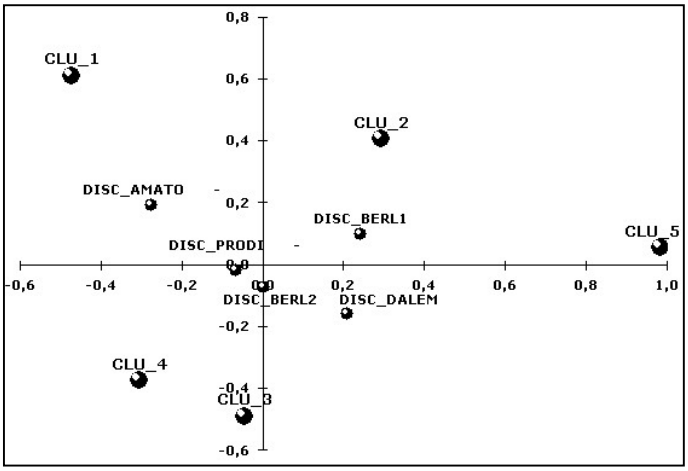



*Fig. 10 – Mappa dei cluster (assi fattoriali 1 e 2)*

In ogni caso, per dare senso alle relazioni – siano esse tra i cluster che, all'interno di ciascuno di essi, tra le differenti "parole" – bisogna andare "al di là dell'informazione data" e utilizzare un qualche modello interpretativo.

Poiché, in questa sede, mi sono assunto il compito di far da guida all'uso di T-LAB e non quello di interpretare i suoi risultati, lascio al lettore questo tipo di lavoro e passo ad esemplificare la seconda strategia di analisi che, in un modo diverso, tende a costruire una rappresentazione dei contenuti del corpus nel loro *insieme*.

[37] In questo caso, l'analisi delle corrispondenze è applicata a una tabella con in riga le 478 parole-chiave e in colonna i 5 cluster ottenuti.

[38] A questo scopo è possibile verificare i contributi assoluti e relativi di ciascun cluster.

[39] Per l'interpretazione dei grafici concernenti i cluster, si veda il paragrafo 2.3.5 (Parte prima).

## 2.3. La strategia del fotografo

Lo strumento T-LAB che, in modo elettivo, consente di perseguire questa strategia è l'*Analisi delle corrispondenze unità lessicali x sottoinsiemi*. Poiché si tratta di uno strumento che consente di analizzare tabelle con max 10.000 righe, sarei tentato di costruire una lista di unità lessicali (UL) più estesa, cioè – per tornare a una metafora già utilizzata – vorrei aumentare il numero delle lucciole nella mia lanterna. Ma, un po' per non disorientare il lettore, un po' perché trovo interessante proporre un "confronto alla pari" sullo stesso corpus, decido di tenere la stessa lista che ho utilizzato per esemplificare la precedente strategia (478 UL).

La "foto di gruppo" sarà quindi costruita analizzando la Tab. 2[40]. In questa, i *profili* dei Presidenti del Consiglio, che prendono il posto delle fisionomie e delle espressioni care ai fotografi, sono costituiti dalle distribuzioni delle varie parole all'interno dei loro rispettivi discorsi.

Gli output prodotti dall'analisi delle corrispondenze, grafici e tabelle, sono molti. In questo caso (vedi Tab. 3), poiché i primi tre fattori spiegano l'81,51% della varianza (o inerzia) determinata dalle relazioni tra i dati in analisi[41], decido di esplorare le loro varie combinazioni e di presentare tre diverse foto di gruppo, "scattate" da tre diverse prospettive.

Anche se T-LAB mi consente di generare grafici più articolati, in cui sono contemporaneamente presenti le label delle righe (unità lessicali) e quelle delle colonne (i cinque discorsi), scelgo di utilizzare quelli che mostrano solo le sigle dei Presidenti del Consiglio e le rispettive "posizioni" (Fig. 11-13).

Prima di fare altri passi, sono colpito dalla *gestalt*[42] che i tre grafici propongono: un immaginario triangolo, con ai tre vertici – alternativamente – i discorsi di Amato, Berlusconi 1°, D'Alema e Prodi; mentre al centro è sempre presente il discorso di Berlusconi 2° che – alternativamente – è prossimo, e quindi "simile", ai discorsi degli altri tre Presidenti del Consiglio.

[40] Per motivi di spazio, la tabella riporta le prime e le ultime 15 unità lessicali in elenco.

[41] Cioè le relazioni tra le 478 righe e le 5 colonne della tabella N. 2. Per una descrizione dell'analisi delle corrispondenze si veda il paragrafo 2.3.4 (Parte prima).

[42] Mentre "vedo" la forma (gestalt) dell'insieme, ricordo quanto ha scritto N. R. Hanson (1958, tr. it., p. 21 e 31): «teorie e interpretazioni sono *presenti* nella visione fin dal principio…C'è dunque un senso in cui il semplice fatto di vedere è in realtà un'impresa *carica di teoria*».

*Tab. 2 – Tabella unità lessicali per sottoinsiemi*

| | *AMATO* | *BERL. 1* | *BERL. 2* | *D'ALEMA* | *PRODI* |
|---|---|---|---|---|---|
| accompagnare | 3 | 2 | 1 | 4 | 4 |
| accordo | 0 | 0 | 0 | 1 | 6 |
| affermare | 1 | 3 | 2 | 2 | 0 |
| affrontare | 9 | 1 | 5 | 6 | 2 |
| aiutare | 3 | 1 | 1 | 3 | 2 |
| allargamento | 3 | 2 | 2 | 1 | 2 |
| alleanza | 0 | 3 | 1 | 2 | 2 |
| ambientale | 1 | 1 | 1 | 2 | 4 |
| ambiente | 3 | 1 | 1 | 2 | 4 |
| amministrativo | 4 | 3 | 1 | 1 | 5 |
| amministrazione | 0 | 2 | 7 | 1 | 4 |
| anni | 14 | 4 | 5 | 13 | 11 |
| aperto | 1 | 5 | 2 | 3 | 0 |
| apparire | 1 | 0 | 0 | 3 | 3 |
| approvare | 8 | 0 | 1 | 0 | 1 |
| … | … | … | … | … | … |
| utile | 2 | 0 | 2 | 1 | 4 |
| utilizzare | 7 | 0 | 2 | 0 | 0 |
| valore | 5 | 2 | 2 | 2 | 1 |
| valori | 4 | 3 | 5 | 5 | 0 |
| valorizzare | 1 | 0 | 3 | 2 | 8 |
| vecchio | 1 | 2 | 4 | 2 | 0 |
| vedere | 6 | 0 | 1 | 1 | 2 |
| vicenda | 3 | 0 | 0 | 4 | 0 |
| visione | 0 | 0 | 4 | 3 | 1 |
| vita | 7 | 12 | 11 | 14 | 11 |
| vita_pubblica | 0 | 5 | 2 | 0 | 0 |
| vivere | 1 | 3 | 4 | 5 | 1 |
| voce | 2 | 2 | 1 | 1 | 2 |
| volontà | 1 | 0 | 2 | 1 | 5 |
| voto | 3 | 2 | 3 | 1 | 1 |

*Tab. 3 – Varianza spiegata dai fattori*

| | Percent. | Perc. Cum. |
|---|---|---|
| **Fatt. N. 1** | 32,79% | 32,79% |
| **Fatt. N. 2** | 25,41% | 58,20% |
| **Fatt. N. 3** | 23,31% | 81,51% |
| **Fatt. N. 4** | 18,49% | 100,00% |

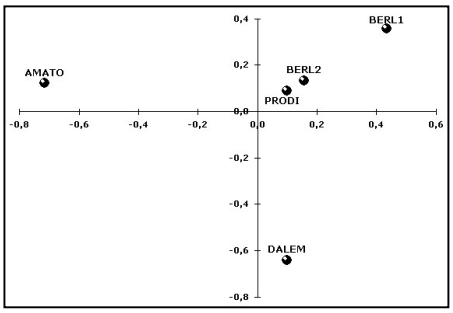



*Fig. 11 – Piano Fatt. 1 x 2*[43]

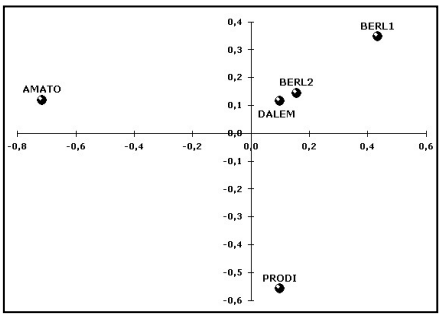



*Fig. 12 – Piano Fatt. 1 x 3*

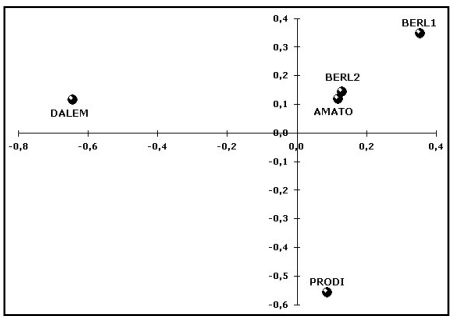



*Fig. 13 – Piano Fatt. 2 x 3*

Mi soffermo a considerare questo fatto, ipotizzo che possa dipendere dalla lista delle parole che ho utilizzato e provo a fare una verifica. Torno in *Impostazione personalizzate* e salvo la loro attuale configurazione[44]; quindi azzero le scelte precedenti e costruisco una nuova lista di parole, questa volta basata solo sul criterio della soglia che, per tenermi largo, fisso a 3 occorrenze. Ripeto l'analisi delle corrispondenze con 1365 unità lessicali e la gestalt resta invariata; quanto a dire (o a confermare) che i modelli sono già presenti nei dati[45].

Torno quindi sui risultati dell'analisi precedente, quella con 478 UL, con l'intento di decifrare il senso dei tre fattori, cioè delle dimensioni costitutive dei tre piani cartesiani nelle Figg. 11-13. In un precedente paragrafo di questo volume (2.3.4, parte prima) ho proposto una sintesi dei metodi per interpretare i risultati dell'analisi delle corrispondenze, quindi mi risparmio di tornare sugli stessi argomenti.

[43] Nelle Figg. 11-13 l'asse orizzontale corrisponde sempre al primo fattore della coppia considerata; mentre l'asse verticale corrisponde sempre al secondo.

[44] Per ogni progetto di analisi, T-LAB consente di archiviare e recuperare fino a 10 diverse impostazioni di analisi, corredate da altrettante liste di unità lessicali e relativi dizionari.

[45] J. P. Benzecri (1984) è un noto fautore di questa ipotesi.

I valori dei *contributi assoluti* (c.a.) dei cinque discorsi[46] sui primi tre fattori (Tab. 4) sono particolarmente significativi. In effetti, presi uno alla volta, il primo, il secondo e il terzo fattore risultano organizzati dalla specificità di tre discorsi: ripettivamente, quello di Amato (primo fattore: c.a. 0,726), quello di D'Alema (secondo fattore: c.a. 0,747) e quello di Prodi (terzo fattore: c.a. 0,721).

*Tab. 4 – Contributi assoluti dei cinque discorsi sui primi tre fattori*

| | *Fatt. N. 1* | *Fatt. N. 2* | *Fatt. N. 3* |
|---|---|---|---|
| *Amato* | **0,726** | 0,027 | 0,028 |
| *Berlusconi 1°* | *0,206* | 0,177 | 0,183 |
| *Berlusconi 2°* | 0,037 | 0,034 | 0,043 |
| *D'Alema* | 0,014 | **0,747** | 0,025 |
| *Prodi* | 0,017 | 0,016 | **0,721** |
| *TOTALI* | 1,000 | 1,000 | 1,000 |

Poiché uno strumento T-LAB consente di verificare le *specificità* dei sottoinsiemi, per ciascuno dei tre discorsi evidenziati (Amato, D'Alema, Prodi) provo a confrontare la lista delle sue parole tipiche (o specifiche) con quella delle parole risultate caratteristiche sul fattore in cui ha il contributo assoluto più elevato. Inoltre, poiché il discorso di Berlusconi 1° ha un contributo assoluto relativamente significativo sul primo fattore (0,206), anche nel suo caso faccio la stessa operazione.

Per motivi di spazio, riporto solo le prime 20 unità lessicali di ogni lista; un numero più che sufficiente per consentire al lettore di fare le sue verifiche.

Le chiavi di lettura sono le seguenti:

- le tabelle con le caratteristiche delle polarità fattoriali (Tabb. 5, 7, 9, 11) sono ordinate per valore test;
- le tabelle con le specificità dei vari discorsi (Tabb. 6, 8,10, 12) sono ordinate in base al valore del chi quadro. In esse, due colonne riportano le occorrenze di ciascuna unità lessicale nel discorso considerato ("SUB") e nell'insieme del corpus ("TOT");
- nelle tabelle con le caratteristiche delle polarità fattoriali, sono marcate in grassetto le parole che sono presenti anche nei corrispondenti elenchi delle specificità, quelle alla loro destra. Come si può verificare, fatta eccezione per il discorso di Berlusconi 1°, le percentuali delle corrispondenze sono sempre superiori al 75%.

[46] "Peso" dei cinque discorsi nella determinazione della varianza spiegata da ogni fattore. Per ogni fattore, il totale è pari a 1.

In genere, quando si resocontano i risultati di un'analisi delle corrispondenze, fattore per fattore ci si sofferma a considerare i vari elementi e le varie misure che consentono di dare un senso alle sue polarità (i semi-assi "-" e "+"). In questo caso, la particolare struttura dei dati ha orientato la mia esplorazione in una direzione imprevista, ma "sensata". In effetti i primi tre fattori, presi uno ad uno, sono risultati organizzati dal peso preponderante (superiore al 70%) di *un* solo discorso: prima quello di Amato (Fatt. 1), poi quelli di D'Alema (Fatt. 2) e di Prodi (Fatt. 3). Inoltre, in tutti e tre i casi, la polarità opposta è risultata bilanciata dalla "differenza specifica" del discorso di Berlusconi 1°.

Quanto al di Berlusconi 2°, che sembra caratterizzato da un profilo lessicale meno polarizzato, mi rendo conto che per me ha funzionato come un *punctum*[47] della foto di gruppo, cioè come un particolare che ha sollecitato curiosità e immaginazione.

*Tab. 5*

| *FATT. 1 – POLARITA' (–)* | |
|---|---|
| *Unità lessicali* | *V.TEST* |
| migliaio | -50,14 |
| **trasformazione** | -49,37 |
| unico | -46,31 |
| **approvare** | -45,78 |
| **tributario** | -42,43 |
| **riforma** | -42,11 |
| arrivare | -41,96 |
| **utilizzare** | -41,44 |
| **miliardo** | -38,47 |
| Mezzogiorno | -37,24 |
| **governi** | -37,21 |
| **elevato** | -36,76 |
| **ragioni** | -35,45 |
| **formazione** | -35,32 |
| euro | -34,41 |
| provvedimento | -34,29 |
| **coordinamento** | -33,78 |
| **finanziario** | -33,73 |
| **posti_di_lavoro** | -32,99 |
| **lavoro** | -32,96 |

*Tab. 6*

| *SPECIFICITA' AMATO* | | | |
|---|---|---|---|
| *Unità lessicali* | *CHI²* | *SUB* | *TOT* |
| trasformazione | 25,91 | 9 | 11 |
| approvare | 22,18 | 8 | 10 |
| unico | 20,90 | 11 | 17 |
| arrivare | 18,63 | 6 | 7 |
| tributario | 18,50 | 7 | 9 |
| utilizzare | 18,50 | 7 | 9 |
| formazione | 17,66 | 16 | 32 |
| miliardo | 15,38 | 7 | 10 |
| governi | 14,74 | 9 | 15 |
| riforma | 14,19 | 22 | 54 |
| posti_di_lavoro | 13,78 | 8 | 13 |
| ragioni | 13,78 | 8 | 13 |
| legislativo | 12,90 | 9 | 16 |
| elevato | 12,87 | 7 | 11 |
| comune | 12,00 | 14 | 31 |
| coordinamento | 11,39 | 5 | 7 |
| riguardare | 11,30 | 9 | 17 |
| essenziale | 11,02 | 10 | 20 |
| finanziario | 11,02 | 10 | 20 |
| Mezzogiorno | 10,92 | 13 | 29 |

[47] Mi riferisco all'opposizione tra *studium* e *punctum*, proposta da R. Barthes (1980, pp. 27-28) nel suo saggio sulla fotografia: il primo (*studium*) riferito all'"interesse" e al punto di vista dell'osservatore come spettatore, il secondo (*punctum*) determinato dalle specifiche caratteristiche dell'osservato (la foto) che "infrangono" le attese dell'osservatore e lo sorprendono.

*Tab. 7*

| *FATT. 1 – POLARITA' (+)* | |
|---|---|
| *Unità lessicali* | *V.TEST* |
| **opposizione** | 39,64 |
| **senatore** | 31,08 |
| introdurre | 30,10 |
| nostro | 29,95 |
| magistratura | 28,27 |
| **partiti** | 27,66 |
| esecutivo | 27,48 |
| **mafia** | 27,24 |
| liberale | 26,81 |
| **Governo** | 26,70 |
| programma | 25,47 |
| forma | 25,15 |
| **vita_pubblica** | 25,01 |
| strumento | 24,75 |
| leggi | 24,70 |
| cultura | 24,58 |
| dovere | 24,07 |
| procedura | 23,00 |
| limiti | 22,87 |
| coalizione | 22,64 |

*Tab. 8*

| *SPECIFICITA' BERLUSCONI 1°* | | | |
|---|---|---|---|
| *Unità lessicali* | *CHI²* | *SUB* | *TOT* |
| introdurre | 31,90 | 7 | 8 |
| mafia | 26,54 | 6 | 7 |
| forma | 22,24 | 13 | 27 |
| partiti | 19,63 | 7 | 11 |
| liberale | 18,14 | 6 | 9 |
| vita_pubblica | 16,85 | 5 | 7 |
| opposizione | 15,27 | 14 | 36 |
| esecutivo | 12,86 | 7 | 14 |
| produrre | 12,86 | 7 | 14 |
| limiti | 11,13 | 5 | 9 |
| Governo | 10,45 | 59 | 262 |
| ruolo | 10,33 | 13 | 38 |
| fine | 9,18 | 5 | 10 |
| repubblicano | 9,18 | 5 | 10 |
| trasformare | 9,18 | 5 | 10 |
| politico | 8,87 | 18 | 62 |
| repubblica | 8,66 | 7 | 17 |
| senatore | 8,30 | 8 | 21 |
| ministero | 8,09 | 6 | 14 |
| reddito | 8,09 | 6 | 14 |

*Tab. 9*

| *FATT. 2 – POLARITA' (–)* | |
|---|---|
| *Unità lessicali* | *V.TEST* |
| **sinistra** | -46,92 |
| **donna** | -45,00 |
| **sfida** | -43,24 |
| **colleghi** | -43,03 |
| **iniziativa** | -42,96 |
| **bipolarismo** | -39,98 |
| **scelta** | -37,55 |
| **condividere** | -36,57 |
| **società** | -34,00 |
| **serve** | -33,95 |
| **forza** | -33,85 |
| **soluzione** | -33,68 |
| **recuperare** | -33,07 |
| **Ulivo** | -31,52 |
| **elezione** | -30,33 |
| **onorevole** | -29,16 |
| **eguaglianza** | -28,60 |
| **nascere** | -27,44 |
| **crisi** | -26,76 |
| passaggio | -25,61 |

*Tab. 10*

| *SPECIFICITA' D'ALEMA* | | | |
|---|---|---|---|
| *Unità lessicali* | *CHI²* | *SUB* | *TOT* |
| sinistra | 26,71 | 10 | 13 |
| donna | 20,00 | 9 | 13 |
| sfida | 19,35 | 11 | 18 |
| bipolarismo | 19,14 | 6 | 7 |
| colleghi | 19,14 | 6 | 7 |
| iniziativa | 17,37 | 11 | 19 |
| forza | 14,09 | 18 | 42 |
| società | 14,09 | 18 | 42 |
| scelta | 13,79 | 12 | 24 |
| condividere | 12,29 | 8 | 14 |
| eguaglianza | 11,74 | 5 | 7 |
| serve | 11,74 | 5 | 7 |
| soluzione | 11,74 | 5 | 7 |
| crisi | 10,62 | 8 | 15 |
| elezione | 10,17 | 6 | 10 |
| recuperare | 10,17 | 6 | 10 |
| Ulivo | 9,18 | 5 | 8 |
| giorni | 8,03 | 7 | 14 |
| onorevole | 8,03 | 7 | 14 |
| nascere | 7,95 | 8 | 17 |

*Tab. 11*

| *FATT. 3 – POLARITA' (–)* | |
|---|---|
| *Unità lessicali* | *V.TEST* |
| **promuovere** | -51,52 |
| **risanamento** | -41,04 |
| **accordo** | -38,63 |
| **strutture** | -38,40 |
| **sacrificio** | -34,88 |
| **nostro** | -33,69 |
| **valorizzare** | -31,20 |
| **spesa** | -31,08 |
| **sentimento** | -30,56 |
| ragazzo | -30,27 |
| **grande** | -29,91 |
| **decentramento** | -29,91 |
| **ispirare** | -29,26 |
| **massimo** | -28,06 |
| **legalità** | -27,79 |
| **coalizione** | -27,61 |
| **prodotto** | -27,50 |
| **indicare** | -27,39 |
| **livello** | -26,53 |
| **perseguire** | -26,39 |

*Tab. 12*

| *SPECIFICITA' PRODI* | | | |
|---|---|---|---|
| *Unità lessicali* | *CHI²* | *SUB* | *TOT* |
| promuovere | 27,51 | 11 | 13 |
| risanamento | 17,37 | 15 | 26 |
| accordo | 15,34 | 6 | 7 |
| strutture | 15,03 | 7 | 9 |
| nostro | 13,69 | 70 | 206 |
| sacrificio | 12,04 | 6 | 8 |
| coalizione | 11,35 | 11 | 20 |
| grande | 10,02 | 38 | 105 |
| ispirare | 9,53 | 6 | 9 |
| spesa | 9,53 | 6 | 9 |
| prodotto | 9,13 | 5 | 7 |
| ragazzo | 9,13 | 5 | 7 |
| sentimento | 9,13 | 5 | 7 |
| valorizzare | 9,04 | 8 | 14 |
| decentramento | 8,30 | 7 | 12 |
| legalità | 8,30 | 7 | 12 |
| livello | 7,59 | 6 | 10 |
| massimo | 7,59 | 6 | 10 |
| perseguire | 7,59 | 6 | 10 |
| indicare | 6,93 | 5 | 8 |

# 3. Esercizio N. 2: sicurezza e privacy nelle comunicazioni

## 3.1. Preliminari

Questo esercizio inizia con una visita al sito di *Telèma*[1], prestigiosa rivista disponibile anche on-line che, alla fine del 2001, ha sospeso le sue pubblicazioni.

Il suo penultimo numero (estate 2001) è dedicato al tema *Sicurezza e privacy nelle comunicazioni*: 35 articoli di vari autori che, da differenti punti di vista, ne analizzano le implicazioni di carattere tecnologico, economico, giuridico, culturale e sociale.

La cosa mi interessa e decido di effettuare il download: un articolo alla volta, "salvato" in formato testo[2] dentro una cartella creata ad hoc. Dopo pochi minuti, ad operazione terminata, assemblo gli articoli in un unico file[3]: il corpus da trattare con T-LAB.

Una rapida verifica mi consente di capire che non ci sono particolari accorgimenti da prendere: il testo è "pulito" e la correzione ortografica non è necessaria. Provvedo all'importazione e, subito dopo, verifico che le unità lessicali con occorrenze più elevate[4] sono congruenti con la tematica in oggetto.

Le prime trenta, nell'ordine, sono le seguenti:

sicurezza (348); rete (300); informazione (288); sistemi (218); comunicazione (201); sistema (191); nuovo (183); dati (170); internet (164); informatico (136);

[1] http://www.fub.it/telema/Welcome.html.
[2] Opzione che, nei più noti browser, è inclusa nel menu "File/Salva col nome".
[3] L'operazione di assemblaggio può essere realizzata con un qualunque editore di testi.
[4] Quelle corrispondenti alle parole-chiave selezionate da T-LAB.

computer (131); utente (130); digitale (117); mondo (117); elettronico (113); nostro (113); proprio (111); vedere (110); servizi (109); hacker (107); tv (107); azienda (105); società (102); sito (100); prodotto (99); messaggio (98); grande (94); attività (89); controllo (89); tecnologia (88).

Come prima operazione, faccio un check delle lessie complesse[5] e decido di "riconoscerne" una sessantina. Le prime in elenco sono le seguenti:

pay tv (46); sicurezza informatica (23); dati personali (21); sistemi informatici (20); posta elettronica (19); protezione dei dati (19); commercio elettronico (13); società della conoscenza (12); new economy (12); tv a pagamento (12); Unione europea (12); cavalli di Troia (12); firma digitale (12); on demand (11); misure di sicurezza (11); misure di protezione (11); società dell'informazione (11); siti web (10); diritto d'autore (10).

Tramite l'opzione *Liste di locuzioni* (menu *Lessico*) le uso nella trasformazione del corpus e, di seguito, provvedo a una sua nuova importazione: quella destinata all'analisi[6].

Per ottimizzare le fasi del lavoro successivo, tramite la funzione *Personalizzare il dizionario* (Fig. 1) procedo a una serie di accorpamenti di parole (lemmi o lessie): tutti basati su un criterio di sinonimia. Tra questi:

*"calcolatore"* + *"computer"* + *"pc"* = *"computer"*
*"cellulare"* + *"telefonino"* = *"cellulare"*
*"firma_digitale"* + *"firma_elettronica"* = *"firma_digitale"*
*"pay_tv"* + *"tv_a_pagamento"* = *"pay_tv"*

Ciò fatto, la rappresentazione del corpus nel database è pronta per essere esplorata e analizzata.

[5] Per la nozione di lessia complessa e di lessicalizzazione, vedi paragrafo 2.2.2 (Parte prima).

[6] Caratteristiche del corpus: 94.256 occorrenze, 14.546 forme, 7.757 hapax (cioè, forme con una sola occorrenza).

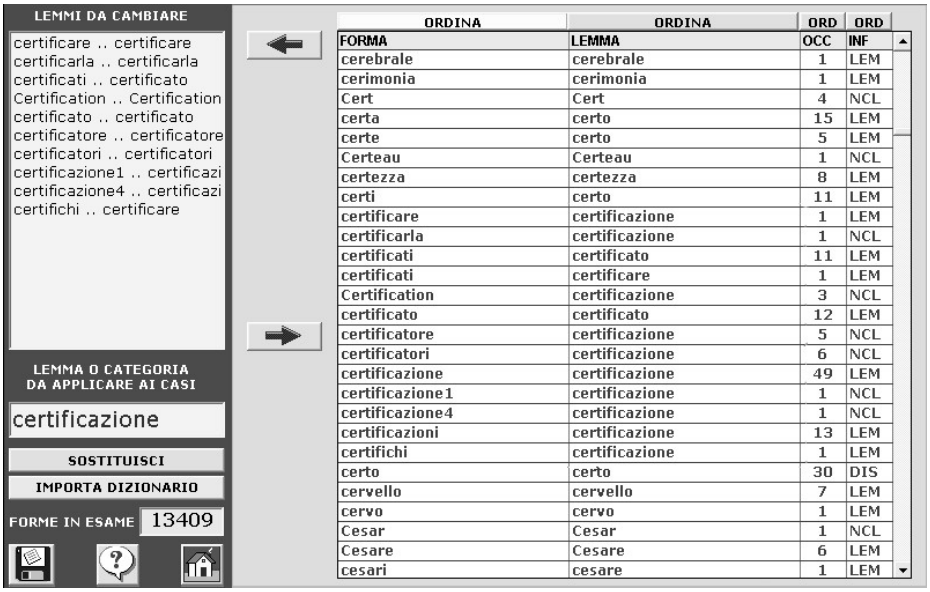



*Fig. 1 – Interventi sul dizionario del corpus*

## 3.2. Prosa e text mining

Come si ricorderà, nella pièce *Il borghese gentiluomo* (Molière) il maestro di filosofia sostiene che ci si può esprimere in due soli modi, in prosa o in versi: «tout ce qui n'est point prose est vers; et tout ce qui n'est point vers est prose». Al che Monsieur Jourdain propone la sua esilarante scoperta:

Par ma foi! il y a plus de quarante ans que je dis de la prose sans que j'en susse rien, et je vous suis le plus obligé du monde de m'avoir appris cela[7]…(Atto I, Scena IV).

Analogamente – all'insegna della *leggerezza* – chi usa gli strumenti T-LAB può scoprire che sta facendo del *text mining*. Questa espressione pressoché intraducibile, tramite una metafora mineraria indica una molteplicità di attività "estrattive" che vanno dalla ricerca di informazioni (*information retrieval*) alla scoperta di "nuove" conoscenze (*knowledge discovery*), cioè dalla "semplice" consultazione di database testuali all'applicazione di tecniche statistiche di tipo esploratorio in grado di restituire *pattern* di relazioni interpre-

[7] «Perbacco! Sono più di quarant'anni che parlo in prosa e non lo sapevo neanche. Vi sono molto obbligato per avermelo fatto scoprire».

tabili. La più elementare di queste attività (l'*information retrieval*[8]), come la prosa di Monsieur Jourdain, è praticata da tutti coloro che – in Internet – usano motori di ricerca. Ad esempio, collegandomi con Google, dopo aver digitato «What is text mining» ottengo un elenco di oltre due milioni di link ad altrettanti documenti. Il primo di questi riporta la seguente definizione:

> Text mining is about looking for patterns in natural language text, and may be defined as the process of analyzing text to extract information from it for particular purposes[9].

Il risultato, ottenuto in tempo reale, è frutto di una tecnologia che – come T-LAB – "gioca" su due componenti[10]: la struttura dei database e la logica degli algoritmi. Più precisamente, nel caso dei motori di ricerca i database sono costituiti da tabelle che incrociano "*n*" documenti con "*m*" parole; diversamente, nel caso di T-LAB, gli incroci riguardano le relazioni tra unità di contesto[11] e unità lessicali.

In T-LAB sono implementati vari strumenti di *text mining.* Quelli più elementari sono due: *Concordanze* e *Associazioni di parole*, applicabili sia all'insieme del corpus che ai suoi sottoinsiemi.

Il primo strumento fornisce liste del tipo *Key-Word-In-Context* (KWIC), sia in formato testo che in formato HTML. Ad esempio, cliccando sulla parola "*privacy*"[12], ottengo un output come quello in Fig. 2.

Il secondo strumento produce grafici e tabelle che mostrano le relazioni tra la parola chiave selezionata e tutte le altre parole presenti all'interno del contesto selezionato (il corpus o un suo sottoinsieme). Le tabelle in output riportano valori che indicano occorrenze, co-occorrenze e misure di associazione[13]. I grafici possono essere esplorati con click del mouse; ad esempio, cliccando sulla label "*tutela*" di Fig. 3, viene mostrato un file HTML (Fig. 4) con tutti i contesti elementari in cui "*privacy*" e "*tutela*" sono contemporaneamente presenti (co-occorrenze).

[8] L'affermazione che l'*information retrieval* (IR) è un tipo "elementare" di text mining (TM) può risultare discutibile. Poco importa: al di là delle classificazioni, si tratta di intenderci sulla logica di alcuni processi.

[9] URL: http://www.cs.waikato.ac.nz/~nzdl/textmining/ (ricerca effettuata il 10 marzo 2004).

[10] Vedi paragrafo 2.1.

[11] Sottoinsiemi del corpus e contesti elementari (vedi paragrafo 2.2., Parte prima).

[12] In questo caso, il contesto di riferimento è il corpus e la parola selezionata è una forma (*word type*).

[13] La misura utilizzata è il coefficiente del coseno (vedi paragrafo 2.3.1., Parte prima).

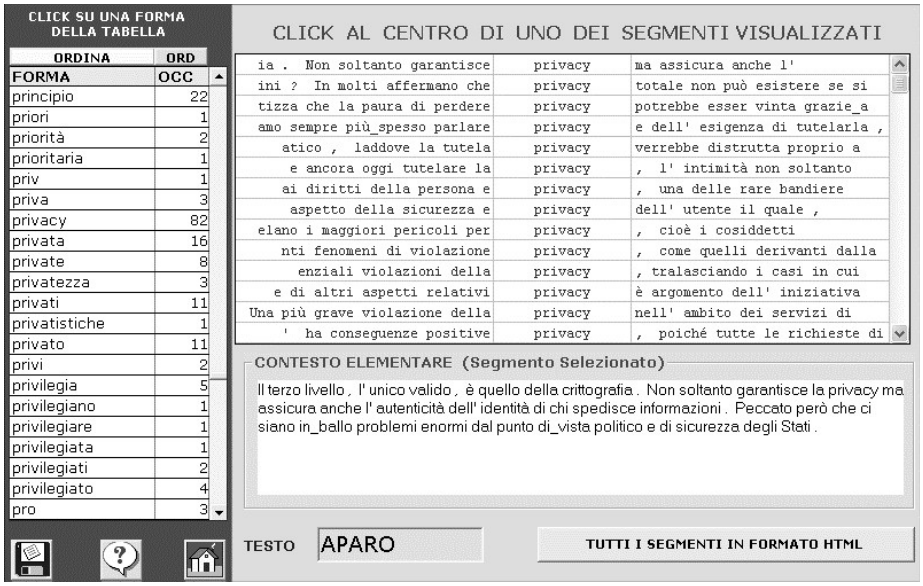



*Fig. 2 – Concordanze di "privacy"*

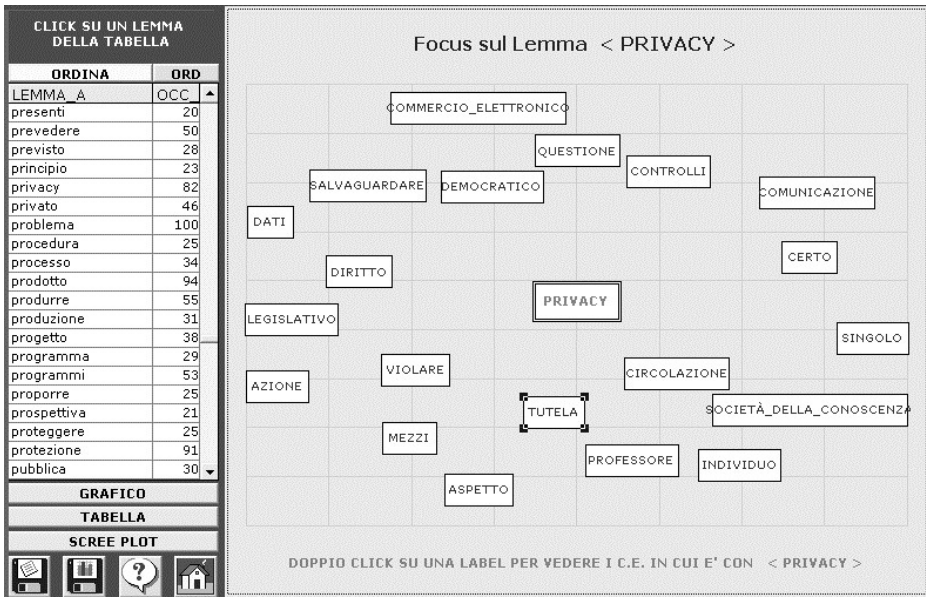
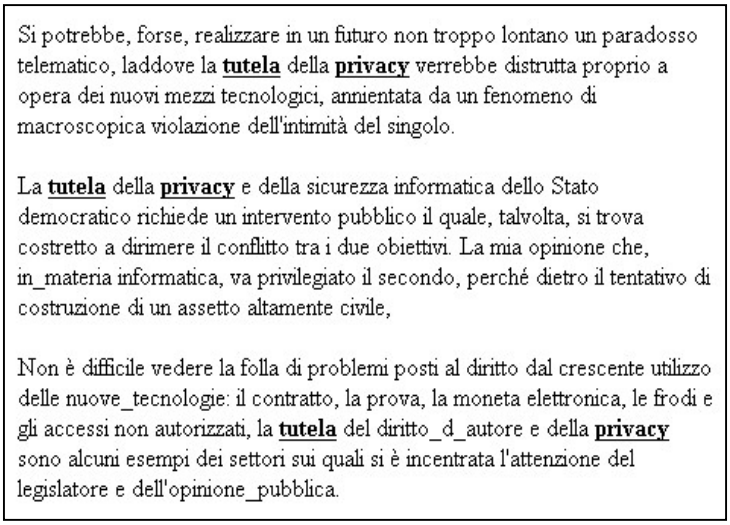



*Fig. 3 – Associazioni di "privacy"*


Si potrebbe, forse, realizzare in un futuro non troppo lontano un paradosso telematico, laddove la **tutela** della **privacy** verrebbe distrutta proprio a opera dei nuovi mezzi tecnologici, annientata da un fenomeno di macroscopica violazione dell'intimità del singolo.

La **tutela** della **privacy** e della sicurezza informatica dello Stato democratico richiede un intervento pubblico il quale, talvolta, si trova costretto a dirimere il conflitto tra i due obiettivi. La mia opinione che, in_materia informatica, va privilegiato il secondo, perché dietro il tentativo di costruzione di un assetto altamente civile,

Non è difficile vedere la folla di problemi posti al diritto dal crescente utilizzo delle nuove_tecnologie: il contratto, la prova, la moneta elettronica, le frodi e gli accessi non autorizzati, la **tutela** del diritto_d_autore e della **privacy** sono alcuni esempi dei settori sui quali si è incentrata l'attenzione del legislatore e dell'opinione_pubblica.

*Fig. 4 – Co-occorrenze di "tutela" e "privacy"*

Fin qui operazioni del tipo *retreive*: parole nei contesti e relazioni del tipo uno-a-uno. Volendo potrei approfondire l'esplorazione e fare ricerche dello stesso tipo all'interno dei singoli articoli di Telèma, ad esempio per verificare somiglianze e differenze tra i vari "punti di vista". A questo scopo il corpus dovrebbe essere stato preparato con una codifica[14] che, al momento, non è presente.

Anche per spiegare come in questi casi è possibile "rimediare" e non perdere il lavoro già fatto (vedi le operazioni sul dizionario del corpus), procedo nel modo seguente:

a) salvo il dizionario[15] su cui ho lavorato (funzione *Personalizzare il dizionario*);
b) apro il file testo che contiene la versione più aggiornata del corpus, procedo alla codifica dei vari articoli che lo compongono e lo salvo con un altro nome;
c) avvio T-LAB e importo il nuovo corpus (vedi punto precedente);
d) subito dopo, con un semplice click, attivo l'importazione automatica del dizionario su cui ho già lavorato (vedi punto "a").

Ma torniamo al *text mining*. La domanda sociale che ne sostiene lo sviluppo è fondata su due rilievi:

[14] In questo caso, all'inizio di ogni articolo dovrebbe essere stata inserita una riga del tipo **** *ARTIC_RODOTA, dove "ARTIC" indica il nome della variabile (articolo) e "RODOTA" il sottoinsieme del corpus (la "modalità" della variabile) corrispondente al nome dell'autore.

[15] Il file in output (*Dictio.diz*) contiene tutte le corrispondenze forme/label.

- l'era digitale rende disponibili quantità sterminate di testi e documenti entro i quali sono contenute molte informazioni preziose[16];
- gli strumenti software possono analizzare questi materiali per estrarne il "succo", sia organizzando criteri di classificazione che restituendo pattern che possono orientare lo sviluppo della conoscenza e la presa di decisioni.

Gli investimenti in questo settore, siano essi di aziende che di enti governativi, stanno producendo risultati notevoli; tuttavia, al di là delle dimensioni, la "logica" costruttiva dei software è fondata sugli stessi pilastri di T-LAB: trattamenti linguistici e strumenti statistici di tipo esploratorio, applicati a dati rappresentati in un database.

Una nota a proposito delle "grandi dimensioni": poiché i testi da analizzare possono essere simbolizzati come "un mare", c'è chi, vuoi per non rischiare di soccombere vuoi per fare un viaggio più rapido e confortevole, simbolizza il *text mining* come un viaggio da fare in nave. Per restare nella metafora, T-LAB è una barca che si adatta a viaggiatori che amano stare al timone; in genere ricercatori per i quali i testi non sono "un mare" che sta fuori bensì risorse selezionate "dentro" un percorso di conoscenza.

A partire da ciò, provo a definire una strategia orientata all'estrazione (mining) delle isotopie[17] che organizzano i contenuti del corpus Telèma.

Poiché ho in mente un certo uso dello strumento *Tipologie dei contesti elementari* (TCE), devo preventivamente selezionare le parole-chiave da includere nell'analisi: massimo 500. La soglia proposta da T-LAB include tutte le unità lessicali con occorrenze pari o superiori a 8: in tutto 1.275, pari a 4.010 forme (*type*) e 30.098 occorrenze (*token*)[18].

Decido di fare una rigida selezione e metto fuori analisi tutte le parole di significato generico che giudico "fuori tema". Alla fine, la mia lista include 346 parole chiave, tutte con occorrenze pari o superiori a 8. Le prime cinquanta sono le seguenti:

potere; sicurezza; rete; informazione; nuovo; comunicazione; sistema; tecnologico; computer; internet; sistemi; controlli; dati; utente; hacker; utilizzare; Europa; mondo; servizi; azienda; certificare; informatica; problema; sito_web; rischi; messaggio; segretezza; attacchi; digitale; prodotto; protezione; società; privacy; codice; americano; crimini; diffusione; garantire; Italia; consentire; tecnica; imprese; informatico; accesso; distretto; usare; richiesta; crittografia; elettronico; standard.

[16] Evidentemente, il criterio della "preziosità" varia in funzione dei tipi di testi e dei tipi di soggetti interessati alla loro analisi.

[17] Per la definizione di questa nozione vedi paragrafo 2.3.5 (Parte prima).

[18] Il corpus, incluse le parole "vuote", è costituito da 14.546 forme e 94.256 occorrenze. Le forme hapax sono 7.752.

Usando questa lista, effettuo l'analisi (TCE) e ottengo una soluzione a cinque cluster (vedi Fig. 5). Do un'occhiata alla loro distribuzione (Fig. 6) e sono colpito dal fatto che un solo cluster include oltre il 61% dei contesti elementari classificati[19].

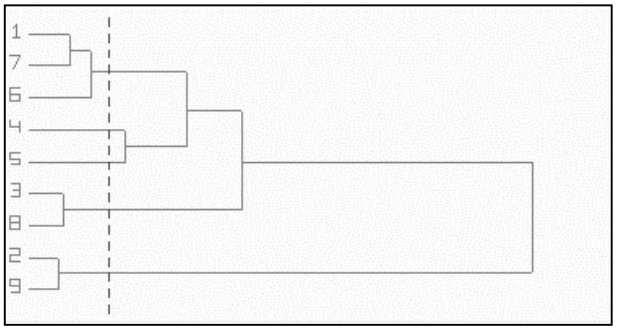



*Fig. 5 – Dendrogramma*

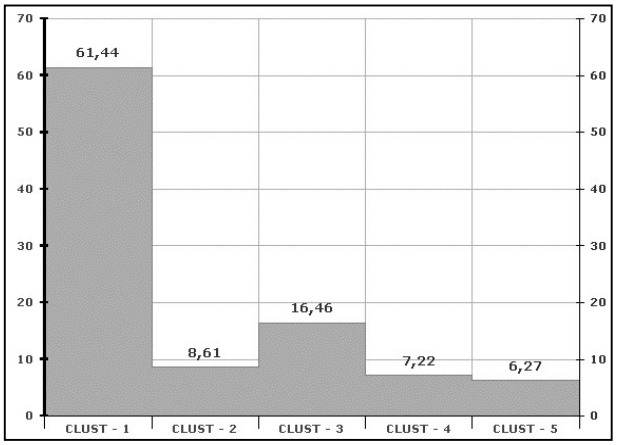



*Fig. 6 – Istogramma dei cluster*

Per illustrare meglio la situazione, realizzo un grafico a "bolle" (Fig. 7)[20].

Come si può rilevare, le relazioni tra i cluster hanno una notevole incidenza sull'organizzazione dello spazio fattoriale. In particolare, le posizioni decentrate dei cluster 2 e 5 indicano che le i loro "contenuti" sono molto diversi dagli altri.

[19] Il totale dei contesti elementari classificati è in questo caso pari a 2090, cioè l'85% del totale (2469).

[20] Per farlo, ho utilizzato Microsoft Excel. Al momento, i grafici T-LAB che mostrano i cluster negli spazi fattoriali non consentono di apprezzare le differenze di "massa" (vedi Fig. 13).

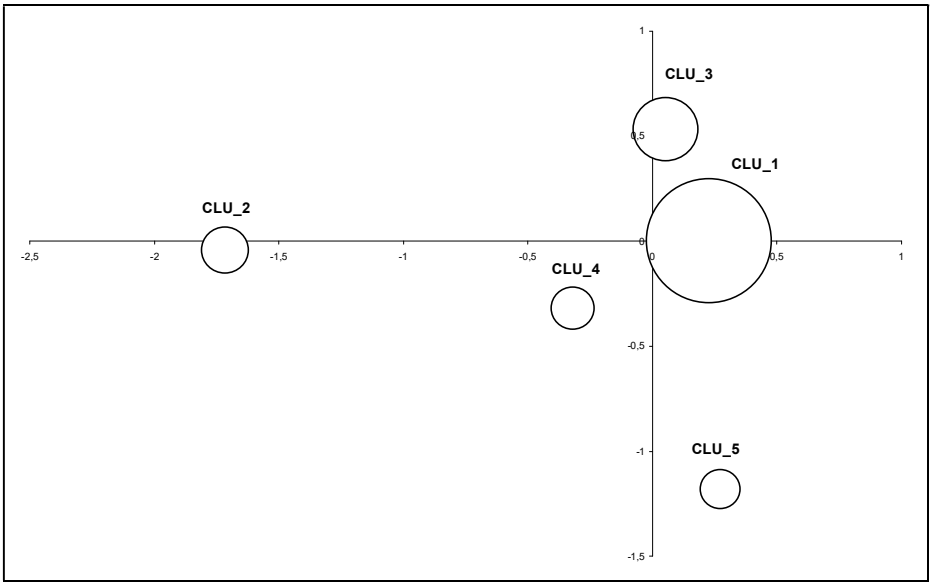



*Fig. 7 – I cluster sul piano 1x2*

Ma, al di là delle loro specificità, è l'insieme delle loro relazioni che merita di essere approfondito. Decido quindi di soffermarmi a considerare le bipolarità che organizzano i primi due fattori, i quali – complessivamente – spiegano il 77.81% dell'inerzia (o varianza).

T-LAB, in modo automatico, mi fornisce le liste ordinate con i valori test (vedi Tab. 1). A partire da queste, tramite la selezione di alcuni termini chiave, ricostruisco (vdi Tab. 2) le opposizioni fondamentali del piano fattoriale in Fig. 7.

Evidentemente, il "centro" del discorso è il cluster 1, mentre gli altri cluster – come piccoli satelliti – gli ruotano intorno. A partire da questo rilievo, volendo fare uno zoom sul cluster 1, utilizzo un'ulteriore strumento T-LAB: *Mappe dei nuclei tematici*[21].

Il risultato che ottengo è quello illustrato in Fig. 8.

[21] Questo strumento consente di costruire mappe del corpus e/o dei suoi sottoinsiemi. Poiché l'analisi precedente (TCE) ha prodotto dei nuovi sottoinsiemi (i 5 cluster), posso "mettere a fuoco" ciascuno di essi.

*Tab. 1 – Valori test delle unità lessicali sui primi due fattori*

| *Fattore N. 1* | | | | *Fattore N. 2* | | | |
|---|---|---|---|---|---|---|---|
| *Semiasse (-)* | | *Semiasse (+)* | | *Semiasse (-)* | | *Semiasse (+)* | |
| pay_tv | -22,209 | sicurezza | 5,861 | Europa | -12,231 | sito_web | 7,181 |
| televisione | -18,355 | informazione | 5,232 | nazionale | -11,394 | rete | 6,364 |
| tv | -17,866 | sistemi | 3,997 | tutela | -8,003 | utente | 5,138 |
| milioni | -15,213 | protezione | 3,645 | direttiva | -7,487 | cliente | 4,709 |
| telespettatore | -15,190 | dati | 3,406 | Italia | -6,636 | costi | 4,545 |
| abbonato | -14,515 | comunicazione | 3,283 | rispetto | -6,019 | attacchi | 4,399 |
| Stream | -13,298 | problema | 3,175 | stati_uniti | -5,640 | sistema | 4,330 |
| film | -13,186 | controlli | 3,024 | adottare | -5,624 | center | 4,143 |
| canale | -12,150 | consentire | 2,800 | giuridico | -5,482 | accesso | 4,134 |
| lira | -11,271 | diritto | 2,665 | autorità | -5,459 | server | 4,032 |
| Telepiù | -11,153 | privacy | 2,641 | certificare | -5,279 | uso | 3,870 |
| futuro | -10,883 | segretezza | 2,628 | libertà | -5,185 | prodotto | 3,857 |
| decoder | -9,397 | rete | 2,617 | normativa | -5,102 | servizi | 3,836 |
| offrire | -8,846 | giuridico | 2,582 | legislativo | -5,091 | funzionamento | 3,721 |
| unico | -8,282 | certificare | 2,505 | diritto | -4,964 | utilizzare | 3,644 |
| scegliere | -8,282 | efficace | 2,497 | legge | -4,772 | software | 3,608 |
| casa | -8,117 | riservatezza | 2,432 | criterio | -4,759 | web | 3,569 |
| digitale | -7,860 | ricerca | 2,387 | esigenza | -4,646 | elevato | 3,541 |
| programma | -7,780 | adottare | 2,365 | schema | -4,462 | e-mail | 3,483 |
| programmi | -6,737 | personale | 2,359 | americano | -4,391 | internet | 3,345 |
| scelta | -6,023 | Rispetto | 2,333 | persona | -4,332 | sistemi | 3,291 |
| offerta | -5,135 | sociale | 2,321 | internazionale | -4,247 | musica | 3,272 |
| dollari | -5,057 | attacchi | 2,298 | italiano | -4,179 | lusso | 3,165 |
| miliardo | -4,965 | tutela | 2,292 | Stati | -3,728 | gestore | 3,150 |
| spesa | -4,930 | tipo | 2,262 | politico | -3,635 | virtuale | 3,138 |
| Italia | -4,811 | richiesta | 2,194 | linguaggio | -3,556 | servizio | 3,119 |
| pirata | -4,385 | nazionale | 2,185 | termine | -3,475 | virus | 3,045 |

*Tab. 2 – Polarità dei fattori 1 e 2*

| | *Semiasse (-)* | *Semiasse (+)* |
|---|---|---|
| *Fatt. N. 1* | Pay TV, Decoder, Pirateria, *Cluster N. 2*[22] | Sicurezza, Dati, Privacy, *Cluster N. 1* |
| *Fatt. N. 2* | Tutela, Normativa, Autorità, *Cluster N. 5* | Web, e-mail, Virus, *Cluster N. 3* |

[22] Per la significatività delle relazioni tra cluster e fattori, si fa riferimento ai contributi relativi (o coseni quadrati), cioè a misure che indicano quanto ciascu cluster è "ben rappresentato" su ciascun fattore.

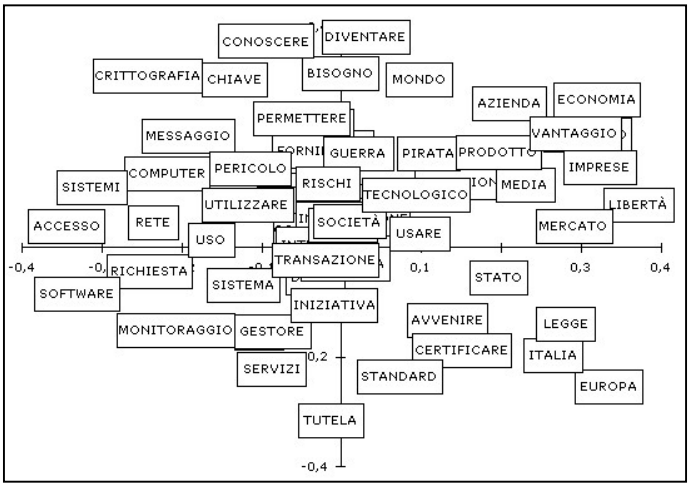



*Fig. 8 – Cluster 1: mappa dei nuclei tematici*

A questo punto, potrei soffermarmi ad esplorare questa ulteriore mappa, sia cliccando sulle singole label[23], sia decifrando le loro relazioni di prossimità e di opposizione. Ma, stando l'obiettivo esemplificativo di questo esercizio, preferisco passare ad altri tipi di *text mining*.

## 3.3. Dettagli e punti di vista

Volendo entrare nei "dettagli", penso di percorrere due strade: la prima centrata sui "punti di vista" degli autori, la seconda sulle associazioni di parole.

In base a un criterio di mera curiosità, "seleziono" tre autori: uno psicologo (Aldo Carotenuto), un esperto di informatica (Giorgio De Michelis) e l'attuale garante delle privacy (Stefano Rodotà).

La domanda che mi pongo, apparentemente banale, è la seguente: «Focalizzando gli *usi* delle parole, per quali aspetti il discorso proposto da ciascuno dei tre si differenzia da tutti gli altri?».

Lo strumento T-LAB che, in modo elettivo, consente di rispondere a questa domanda è *Analisi delle specificità*. La sua logica è fondata sulla teoria degli insiemi e l'unico strumento statistico che utilizza è il test del chi quadro[24]. In particolare, dato un sottoinsieme del corpus (in questo caso, un sin-

[23] Le opzioni di questa funzione T-LAB consentono di verificare la composizione di ogni nucleo tematico (parole e relative occorrenze) e le sue associazioni con ciascuno degli altri (coefficiente del coseno).

[24] Vedi paragrafo 2.3.2 (Parte prima).

golo articolo) è possibile verificare quali usi di parole lo rendono diverso dal "resto" del corpus o da ciascuno degli altri sottoinsiemi. Ciò analizzando in fasi diverse sia la zona di intersezione che quella della differenza, cioè sia considerando tutte le parole del corpus che quelle *esclusive* di ciascun sottoinsieme. In quest'ultimo caso (parole esclusive), non essendo possibile un confronto statistico, il test del chi quadro non viene applicato.

Poiché questo strumento non pone limiti al numero delle unità lessicali[25] da includere nell'analisi, fisso la soglia a 8 occorrenze[26] e vado avanti.

Per ciascuno dei tre articoli, analizzando le intersezioni con il resto del corpus, ottengo una lista di circa 100 "parole tipiche". Per ciascuna di esse, gli output T-LAB riportano tre valori (vedi Tab. 3): le occorrenze all'interno di tutto il corpus (TOT), le occorrenze all'interno del sottoinsieme considerato (SUB) e il valore del chi quadro.

*Tab. 3 – Specificità: esempio*

| *Articolo* | *Parola* | *CHI²* | *SUB* | *TOT* |
|---|---|---|---|---|
| Carotenuto | nemico | 525,89 | 18 | 26 |
| De Michelis | accessi | 265,57 | 13 | 17 |
| Rodotà | privacy | 67,47 | 10 | 82 |

Diversamente, nel caso delle "parole esclusive" gli elenchi sono meno lunghi e gli unici valori sono quelli delle occorrenze entro il sottoinsieme considerato. Di seguito, senza fare alcuna selezione riporto le prime 30 parole di ogni elenco. Inoltre, per non appesantire l'esposizione, decido di riportare solo alcuni valori: il chi quadro nel caso delle parole tipiche e le occorrenze nel caso di quelle esclusive (Tabb. 4, 5, 6).

A partire da questi dati, lascio al lettore il compito di rispondere a una domanda nient'affatto banale: «Stando che il contesto del discorso riguarda i temi della sicurezza e della privacy, in che misura e per quali ragioni i differenti usi delle parole sono indizi di differenze culturali tra i tre autori considerati?».

In un precedente esempio[27] ho mostrato come le "posizioni" di vari autori (o soggetti) possono essere esplorate nel loro insieme. Volendo, in questo caso potrei fare qualcosa di simile; ma, in questo momento, la mia curiosità sta prendendo altre direzioni. Sono più interessato alle associazioni tra parole.

[25] In realtà un limite c'è; ma è talmente largo (max 10.000) che risulta quasi "teorico".
[26] Soglia scelta per l'analisi delle "intersezioni".
[27] Vedi paragrafo 2.3.

*Tab. 4 – Aldo Carotenuto: specificità*

| *Parole tipiche* | *Parole esclusive* |
|---|---|
| nemico (525,89); ascoltare (437,02); paura (318,89); nostro (239,05); noi (160,14); personalità (150,26); spiare (109,66); pronto (82,17); worm (82,17); contesto (62,03); linguaggio (60,05); sentire (56,48); quotidiano (44,76); necessità (43,94); strategia (39,29); idea (36,28); minaccia (30,27); pericolo (26,86); osservare (25,12); vivere (25,12); attaccare (23,17); comunicare (23,17); pensiero (23,17); bisogno (20,17); rimanere (20,17); determinato (19,6); trasformare (19,17); rapido (18,63); riprendere (18,63); evoluzione (17,37). | Truman_show (5); delirante (3); paranoico (3); persecuzione (2); agguato (2); convivere (2); dicasi (2); fortunatamente (2); incubo (2); macroscopico (2); nolente (2); aggressività (2); patologia (2); volente (2); pronunciato (2); psichiatria (2); psicologia (2); ragionamento (2); sembianza (2); sospettoso (2); spinoso (2); trascorrere (2); udire (2); paranoidi (2). |

*Tab. 5 – Giorgio De Michelis: specificità*

| *Parole tipiche* | *Parole esclusive* |
|---|---|
| accessi (265,57); modelli (249,4); azioni (222,08); Pki (219,5); difese (179,95); autorizzare (153,83); vincoli (119,95); interazione (70,74); flessibile (69,34); identità (64,47); entità (63,39); proprietà (58,57); controlli (57,25); sistemi (54,37); compiere (39,04); sicurezza (38,71); politico (38,67); firewall (36,09); attacchi (35,11); risorse (34,55); fiducia (28,96); massimo (28,96); realizzare (28,18); sensibile (27,1); adottare (23,58); compiti (23,46); riflettere (23,46); distruttivo (21,1); management (20,57); rilevante (20,57). | Dac (9); Mac (9); Tbac (7); Rbac (6); Joy (4); policies (4); gerarchia (3); scorrettezza (3); sotto-rete (3); Ca (3); task (3); based (3); contraffatta (2); Bashir (2); incontrollabile (2); Joshi (2); authorization (2); di_ruolo (2); models (2); applications (2); negazione (2); workflow (2); robotica (2); selettivo (2); sensitivity (2); stregone (2); supportare (2); tabella (2); Thomas (2); Wired (2). |

*Tab. 6 – Stefano Rodotà: specificità*

| *Parole tipiche* | *Parole esclusive* |
|---|---|
| effettivo (117,98); stabilire (67,69); privacy (67,47); tutela (65,65); commercio_elettronico (58,83); dati_personali (58,83); dinamico (57,83); Europa (44,2); dichiarare (43,16); tendenza (39,67); regole (32,8); consenso (31,48); insaputa (31,48); negare (31,48); vicenda(31,48); Parlamento_europeo (27,56); Echelon (24,43); generare (24,43); potente (24,43); raccolto (24,43); consumatore (21,87); scorso (21,87); dimostrare (20,84); conflitto (19,74); imprese (18,49); ripetere (17,93); economico (17,75); destinatario (16,39); fonte (16,39); indispensabile (16,39). | opt (3); sgraditi (2); insidiare (2); illegittimo (2); illegittimità (2); constatazione (2); alimenti (2); vincolare (1); superficialità (1); stanco (1); spezzare (1); sorprese (1); soccombere (1); sfuggirci (1); sensibilizzazione (1); scetticismo (1); scavo (1); rivelatrici (1); riservare (1); rigetto (1); resistenze (1); rassegnare (1); puntuale (1); prudenza (1); provvisto (1); profilare (1); pro_capite (1); prevalenza (1); prematuro (1); personalizzato (1); |

Uno strumento T-LAB (*Associazioni tra coppie di parole chiave*) mi consente di fare ulteriori analisi di "dettaglio". Tramite questo, data una *coppia* di parole e i rispettivi contesti[28] in cui occorrono, è possibile verificare somiglianze e differenze sia tra i loro contesti di co-occorrenza sia tra i contesti in cui è presente solo una delle due parole selezionate. Più precisamente, dopo ogni selezione vengono mostrate (grafici e tabelle) sia le associazioni *condivise* da entrambi gli elementi della coppia, sia le associazioni *esclusive* di ciascuno dei due elementi. Anche in questo caso, la logica applicata deriva dalla teoria degli insiemi e il test utilizzato è quello del chi quadro.

In questo caso, la mia curiosità mi porta a considerare due coppie: "*sicurezza*" e "*privacy*", "*internet*" e "*tv*".

La Fig. 9 sintetizza alcune delle relazioni evidenziate nell'intersezione delle porzioni di testo (i contesti elementari) in cui "*sicurezza*" e "*privacy*" coabitano (co-occorrenze).

Come si può rilevare, nel caso di *"sicurezza"* svettano le associazioni con *"informatica"*, *"sistemi"* e *"controlli"*; diversamente, le associazioni "a favore" di *"privacy"* sono *"tutela"* e *"segretezza"*.

Quanto alle rispettive associazioni esclusive, le prime in elenco sono le seguenti: *"azienda"*, *"computer"*, *"hacker"* e *"qualità"* per *"sicurezza"*; *"lavoro", "e-mail"*, *"commercio elettronico"* e *"carta di credito"* per *"privacy"*.

Seconda coppia: *"Internet"* e *"TV"*, ovviamente entro il contesto del corpus in esame.

La Fig. 10 mostra come "*Internet"* – rispetto a "*TV"* – sia fortemente caratterizzata come nuova tecnologia associata all'"*informazione*", ai "*servizi*" e alla "*libertà*". Diversamente la "*TV*" è più associata a "*pubblico*" e a "*casa*". Quanto alle associazioni esclusive, quelle di "*Internet"* includono decine di parole che rimandano al web, alla sicurezza dei dati e agli attacchi degli hacker; mentre quelle di "*TV*" sono meno di una decina e rimandano al problema dei decoder e della tv a pagamento.

Mentre analizzavo le relazioni tra "*sicurezza"* e "*privacy"* ho verificato le rispettive occorrenze. In particolare, ho verificato che sicurezza è la parola chiave più utilizzata: oltre 329[29] occorrenze, corrispondenti a quasi altrettanti contesti elementari. Stando la peculiarità del corpus, si tratta di un dato quasi scontato; ma, ai fini del *text mining*, mi offre ulteriori possibilità di esplorazione. In effetti, facilitato dall'architettura di T-LAB, posso rapidamente costruire un sub-corpus contenente solo le frasi in cui è presente la parola chiave considerata ("*sicurezza"*).

[28] Insiemi di contesti elementari (frasi).

[29] Dopo la prima importazione erano 348 (vedi 3.1.), ma, in seguito al trattamento delle lessie complesse (es. "sicurezza dei sistemi"), la loro quantità si è ridotta.

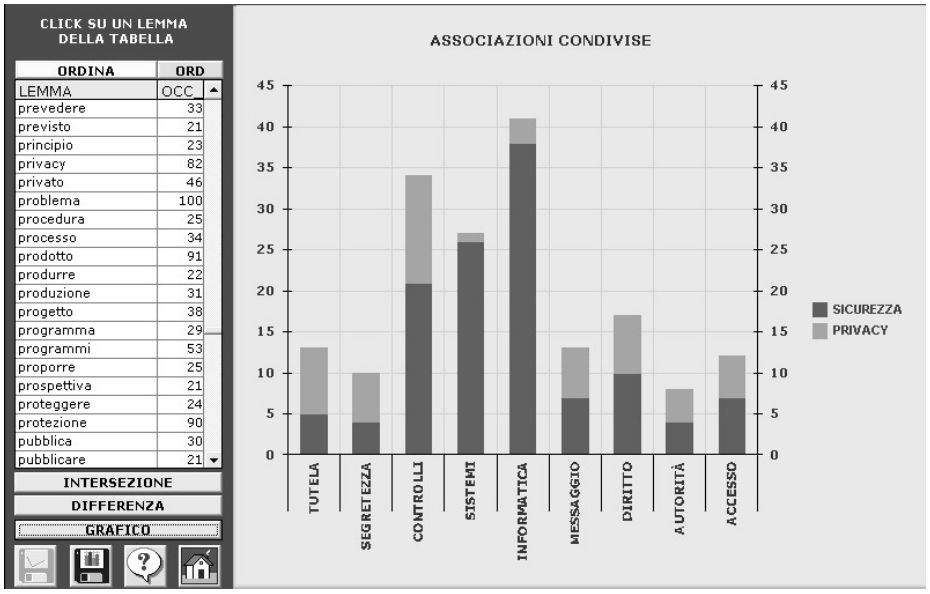



*Fig. 9 – Confronti tra coppie di parole chiave: sicurezza e privacy*

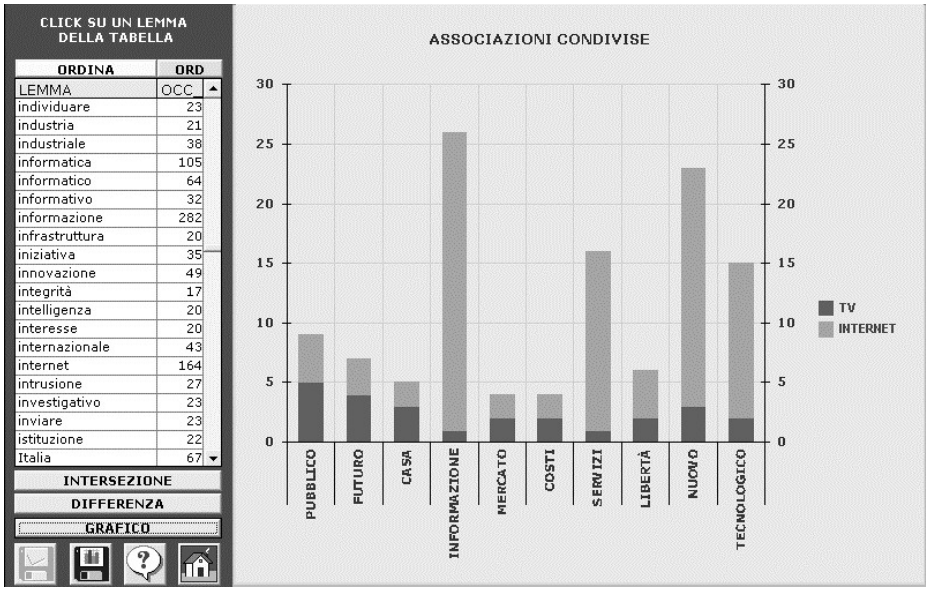



*Fig. 10 – Confronti di coppie di parole chiave: Internet e TV*

La procedura è la seguente:

- tramite un'opzione della funzione *Concordanze*, usando la parola chiave "sicurezza", salvo un file con tutti i contesti elementari in cui questa risulta presente;
- poiché il file salvato è già codificato secondo gli standard T-LAB, mi limito ad eliminare le righe iniziali che riportano data e titolo del report;
- provvedo alla sua importazione e, come già fatto in un caso analogo (vedi 3.2.), applico al nuovo corpus il dizionario su cui ho già lavorato.

La finestra *Impostazioni personalizzate* mi dà modo di verificare che nel nuovo corpus la soglia di frequenza consigliata è fissata a 4 e che l'elenco delle unità lessicali include 354 elementi: tutti pertinenti. Il che significa che posso risparmiarmi di procedere a selezioni o a "eliminazioni".

Di seguito, a titolo esemplificativo, mostro alcuni output corredati di breve note orientate a facilitare la loro analisi iconografica.

La Fig. 11 è stata ottenuta mediante la funzione *Associazioni di parole*. In questo caso le relazioni sono del tipo uno-a-uno tra la parola centrale e ciascuna delle altre. La misura delle associazioni è costituita dal coefficiente del coseno.

Le parole chiave che, nell'ordine, risultano più associate a *sicurezza* sono le seguenti[30]:

informazione (0,344); sistema (0,314); informatica (0,308); nuovo (0,280); rete (0,279); problema (0,278); azienda (0,275); protezione (0,260); sistemi (0,260); tecnologico (0,257); internet (0,251); garantire (0,245).

La Fig. 12 è stata ottenuta mediante la funzione *Mappe dei nuclei tematici*. In questo caso, gli algoritmi applicati sono di tipo multivariato[31] e le relazioni evidenziate riguardano piccoli cluster di parole chiave, ciascuno con la label della parola maggiormente rappresentativa, quella con il numero di occorrenze più elevato.

Sia in Fig. 11 che in Fig. 12 la parola "*sicurezza"* è al centro delle relazioni, ma – nei due casi – il senso di questa posizione è del tutto diverso. Infatti, in Fig. 11 la centralità è il risultato di una scelta dell'utilizzatore[32], men-

[30] I valori tra parentesi sono coefficienti del coseno.

[31] Cluster analysis di tipo gerarchico (ascendente) per individuare i "nuclei"; quindi, analisi delle corrispondenze o muldimensional scaling per rappresentare le loro relazioni entro uno spazio cartesiano.

[32] Del tipo mettere a fuoco le relazioni (una-ad-una) tra la parola selezionata e ciascuna delle altre.

tre in Fig. 12 è il risultato di un calcolo. Più specificamente, la centralità di "*sicurezza"* in Fig. 12 "significa" che il peso delle sue occorrenze, insieme alla sua funzione di relais[33], la rendono irrilevante nell'organizzazione bipolare dello spazio rappresentato.

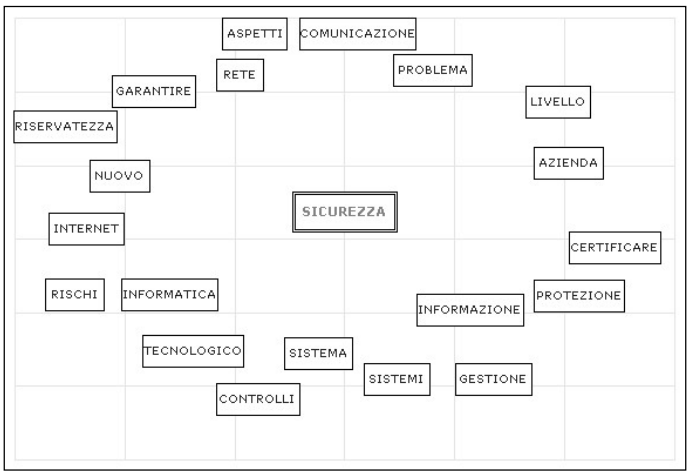



*Fig. 11 – Associazioni relative a sicurezza*

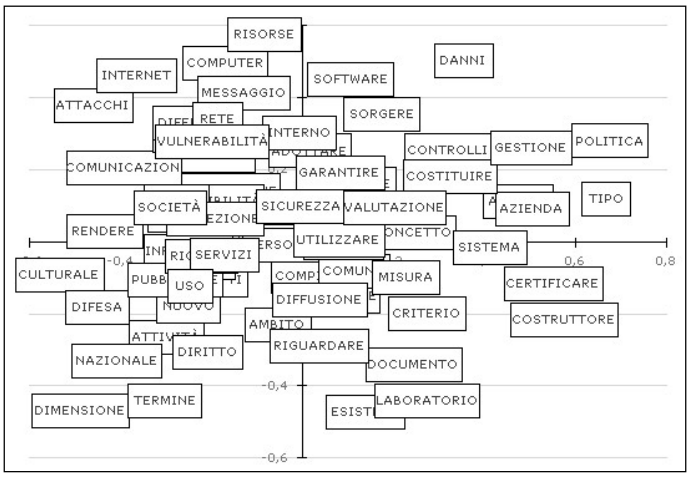



*Fig. 12 – Sub-corpus " sicurezza": mappa dei nuclei tematici*

[33] Non dimentichiamo che il sub-corpus è stato costituito a partire da una selezione "centrata" sulla parola *sicurezza*.

Inoltre: in Fig. 11 le relazioni di contiguità tra le "altre" parole (ad es. "*informazione"* e "*protezione"*) sono il risultato di un mero artificio geometrico[34] e – quindi – non possono essere interpretate; diversamente, in Fig. 12 tutti i tipi di relazione "rappresentano" risultati di calcoli fondati sul principio della classificazione (o categorizzazione). Quindi la Fig. 12 apre la possibilità di riconoscere delle isotopie e di argomentare sui loro rapporti. Ad esempio, nel primo quadrante[35] (alto a sinistra) la prossimità di "*attacchi"*, "*internet"*, "*computer"*, "*messaggio"*, "*rete"* e "*vulnerabilità"* non è certo casuale.

Infine, la Fig. 13 propone una rappresentazione ottenuta mediante uno strumento T-LAB che, come *Mappe dei nuclei tematici* (MNT), analizza le relazioni di co-occorrenza entro i contesti elementari. Lo strumento in questione è l'*Analisi delle corrispondenze segmenti[36] x unità lessicali*. A differenza di MNT, esso consente di "conservare" e di utilizzare l'informazione contenuta nelle label delle variabili. Più precisamente, tutti i punti riferiti alle modalità delle variabili vengono proiettati come "illustrativi"[37].

Ad esempio, in Fig. 13 le label a caratteri maiuscoli rinviano agli autori dei vari articoli. Quanto alle label in minuscolo, ciascuna di esse rappresenta una unità lessicale (parola, lemma o lessia).

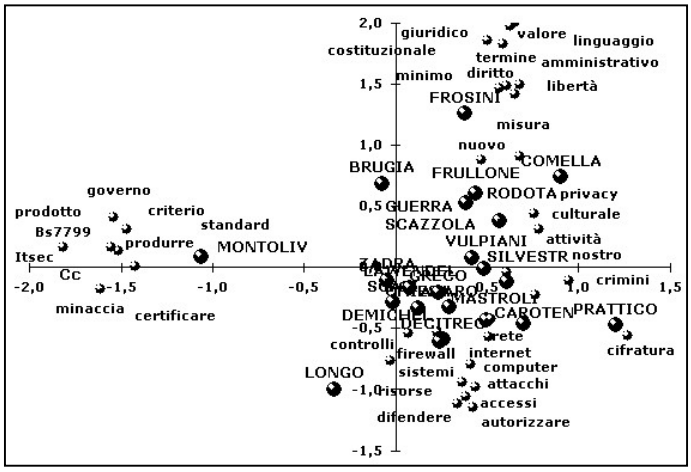



*Fig. 13 – Sub-corpus "sicurezza": Analisi delle corrispondenze*

[34] Il tipo di grafico utilizzato è a raggiera (o radar).

[35] Il riferimento ai quadranti è legittimato dal fatto che, in questo caso, la "mappa" è stata ottenuta mediante analisi delle corrispondenze.

[36] In questo caso, "segmenti" sta per "contesti elementari".

[37] Nel modello dell'analisi delle corrispondenze gli elementi "illustrativi" vengono posizionati entro uno spazio pre-definito da un'analisi in cui gli elementi "attivi" (profili riga e profili colonna) funzionano come organizzatori delle somiglianze e delle differenze. Per ulteriori approfondimenti, rinvio alla bibliografia citata nel paragrafo 2.3.4 (Parte prima).

Confrontando gli output in Fig. 12 e Fig. 13, pur rilevando che in un caso (Fig. 12) le relazioni sono tra cluster e nell'altro (Fig. 13) sono tra "individui" (singole parole), possiamo osservare che la logica delle isotopie resta la medesima. Ad esempio, le label contenute nel primo quadrante (in alto, a sinistra) di Fig. 12 sono "simili" a quelle contenute nell'ultimo quadrante (in basso, a destra) di Fig. 13. Inoltre, anche se risulta meno evidente, la struttura e il senso delle polarità fattoriali sono analoghi[38].

Mi rendo conto del fatto che, in questo esempio, l'abbondanza delle "illustrazioni" può avere un effetto disorientante; ma la pratica del *text mining* espone a questo tipo di rischio, soprattutto quando l'esplorazione delle "miniere" dei testi non è orientata da obiettivi che la radicano in un contesto che tiene conto della domanda di un qualche cliente (vedi Fig. 1 a paragrafo 1.1 della prima parte).

[38] Nei grafici ottenuti mediante analisi delle corrispondenze l'orientamento degli assi (alto-basso da una parte, sinistra-destra dall'altro) è un effetto della geometria. Ai fini interpretativi, ciò che conta sono le opposizioni bi-polari.

# 4. Esercizio N. 3: la lezione di Pinocchio

## 4.1. Le parole alla ricerca di un racconto

L'analisi del racconto di Collodi è un utile esercizio per capire le "regole del gioco" di T-LAB, cioè per capire quali sono i limiti entro i quali le possibili strategie possono produrre utili risultati. Detto in altri termini, l'analisi di *Pinocchio* dà modo di verificare quali sono le dimensioni del testo "trattabili" mediante questo tipo di software.

Il vero problema, lo dichiaro subito, non concerne gli aspetti lessicali; bensì la logica interna del racconto, cioè la *storia*, così come risulta organizzata dalla successione delle molteplici *avventure*. Problema questo che – come avremo modo di vedere – non riguarda solo *Pinocchio*, né soltanto i testi narrativi, bensì tutti i tipi di testi in cui un qualche tipo di *intreccio*[1] determina l'organizzazione sintagmatica dei contenuti.

Quanto agli aspetti lessicali, la loro trattabilità è solo una questione tattica: tipo di lemmatizzazione effettuata, riconoscimento delle lessie complesse[2], eventuali accorpamenti (fusioni) di forme, disambiguazioni, ecc. Nel caso specifico, dopo una prima importazione con opzione di lemmatizzazione automatica, mi soffermo a considerare i tipi di forme grafiche "non classificate"[3]. In effetti, il linguaggio di Collodi non corrisponde allo standard dell'italiano parlato; tuttavia, la situazione riusulta "meno grave" di quanto pensassi. L'incidenza delle forme non classificate è pari al 18% del totale (1122 su 6152). Di queste, 845 (il 75%) sono hapax, cioè forme con una sola occorrenza. Quanto alle altre, solo una quarantina hanno occorrenze pari o superiori a 5[4]. Eccole:

[1] Per il chiarimento di questo termine, rinvio al paragrafo successivo (4.2).

[2] Su questa nozione, vedi paragrafo 2.2.2 (Parte prima).

[3] Non incluse nel dizionario standard di T-LAB, ma che – in ogni caso – possono essere incluse nel processo di analisi.

[4] In questo caso, l'algoritmo implementato in T-LAB "consiglia" di usare una soglia di frequenza pari a 7; ma, in via cautelativa, decido di tenermi largo.

ciuchino (39); Lucignolo (32); Pesce-cane (27); ripetè (20); Grillo-parlante (17); turchini (13); Mangiafoco (11); maraviglia (11); Polendina (11); Balocchi (9); faine (9); babbino (8); ciuchini (8); giandarmi (8); mastr' (8); anderò (7); dovè (7); Marmottina (7); Melampo (7); par (7); Alidoro (6); allegrezza (6); Birba (6); costì (6); dè (6); donnina (6); eccolo (6); zampetto (6); bindolo (5); caldano; Can-Barbone (5); Cucù (5); fioca (5); Grillino (5); tagliola (5); veccie (5); viottola (5)[5].

Le lessie complesse che decido di trasformare in stringhe unitarie (lessicalizzazione)[6] sono le seguenti:

capelli turchini (12); mastr'Antonio (9); Campo dei miracoli (8); Paese dei Balocchi (7); mostro marino (6); Gambero Rosso (6); compagno di scuola (5); maestro Ciliegia (4); compagni di scuola (4).

Al termine di queste prime operazioni, le occorrenze totali (*word token*) risultano essere 39.591, i tipi di parole o lessie (*word type*) 6201. Selezionando quelle con frequenza uguale o superiore a 5, il tasso di copertura del testo[7] è pari all'80% (31.753/39.591).

Gli interventi sul dizionario del corpus, quelli che ritengo opportuni, non sono stati molti. In particolare: la parola *"bambina"*, generalmente usata per indicare la fata, risulta lemmatizzata (*"bambino"*), quindi sostituisco la label con la forma originaria. Quanto ai raggruppamenti effettuati, mi limito ad indicare i criteri: sinonimia in alcuni casi (ad es. *"fata"* e *"fatina"*; *"ragazzetto"* e *"ragazzino"*; tutte le varianti del *"grillo-parlante"*, cioè *"grillino"*, *"grillaccio"*, *"grillo"*, *"grillo-parlante"*), equivalenza del referente in due solo casi (ad es. *"Polendina" = "Geppetto"*, *"Mastr_Antonio" = "Maestro_Ciliegia"*)[8].

Di primo acchito provo a sondare le associazioni relative ad alcune parole-chiave. I risultati sono congruenti e, come le foto di un album di famiglia (vedi Fig. 1 e Fig. 2), evocano momenti e situazioni; ma, mi chiedo, come analizzare il racconto nel suo insieme?

[5] Per ogni parola (*type*), i valori tra parentesi indicano le rispettive occorrenze (*token*).

[6] Per le definizioni di lessia e di lessicalizzazione, vedi paragrafo 2.2.2 (Parte seconda).

[7] Il tasso di copertura del testo, così come la soglia di frequenza, è un criterio quantitativo che consente di "campionare" la popolazione delle parole del corpus. Tuttavia, quando il ricercatore ha delle sue "griglie", questi criteri passano in secondo ordine.

[8] Nel racconto di Collodi, *"Polendina"* è il soprannome di Geppetto, *"Maestro Ciliegia"* è il soprannome di Matr'Antonio.

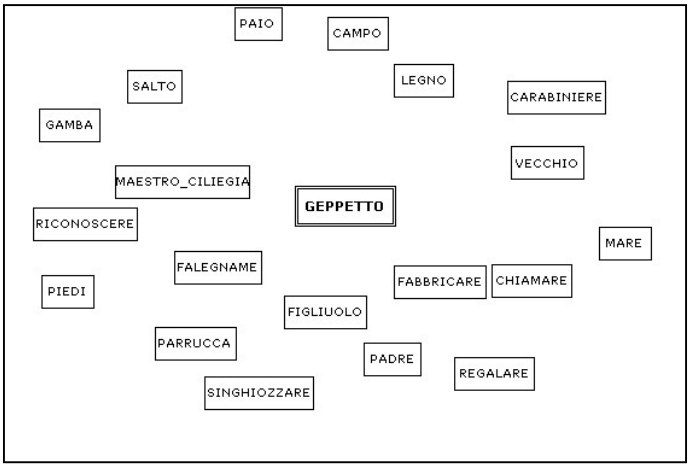



*Fig. 1 – Associazioni relative a "Geppetto"*

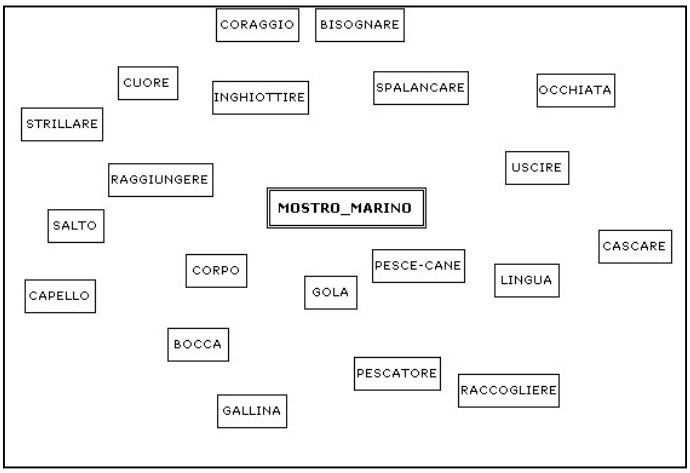



*Fig. 2 – Associazioni relative a "Mostro_Marino"*

Per avere maggiori margini di manovra, ho suddiviso il testo nei 36 episodi che lo compongono[9]; dunque, una prima possibilità di analisi è quella che passa attraverso l'analisi delle corrispondenze (unità lessicali x sottoinsiemi), tanto più che in questo caso il numero di righe (unità lessicali) non ha restrizioni particolari.

La prima mappa che ottengo è quella in Fig. 3. Risulta subito evidente che quattro episodi "sbilanciano" l'analisi sul primo e sul secondo fattore. Essi, nell'ordine – e con i titoli di Collodi – sono i seguenti:

[9] Ciò significa che all'inizio di ogni episodio ho inserito una riga di codifica (es. **** *EPIS_06).

N. 1 – Come andò che maestro Ciliegia, falegname, trovò un pezzo di legno, che piangeva e rideva come un bambino;

N. 2 – Maestro Ciliegia regala il pezzo di legno al suo amico Geppetto, il quale lo prende per fabbricarsi un burattino maraviglioso che sappia ballare, tirar di scherma e fare i salti mortali;

N. 21 – Pinocchio è preso da un contadino, il quale lo costringe a far da can da guardia a un pollaio;

N. 22 – Pinocchio scuopre i ladri e, in ricompensa di essere stato fedele, vien posto in libertà.

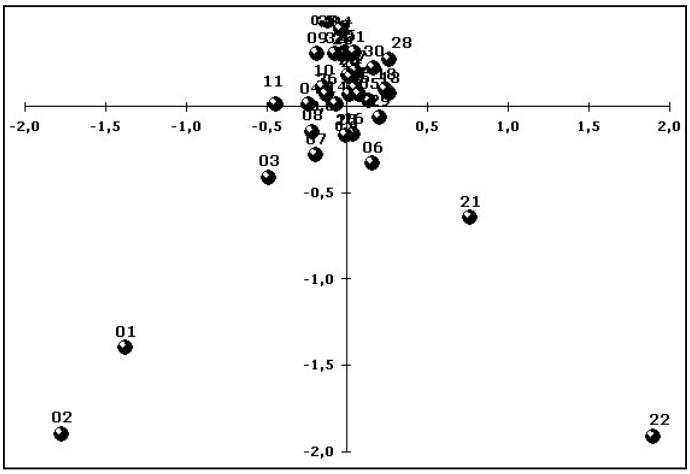



*Fig. 3 – I 36 episodi sul piano 1x2 (Analisi delle corrispondenze)*

Le tabelle con i valori test consentono una rapida verifica della situazione. Ad esempio, le "parole" che marcano l'opposizione bi-polare del *primo fattore*, risultano essere le seguenti:

Polarità "-" (semiasse negativo)

Geppetto (-12,62); falegname (-9,77); pezzo (-9,01); parrucca (-8,05); di_legno (-6,48); vocina (-5,86); fabbricare (-3,54); stanza (-3,36); per_terra (-3,31); tirare (-2,78); mano (-2,74); bambino (-2,55); ballare (-2,52).

Polarità "+" (semiasse positivo)

Melampo (13,33); faine (12,28); casotto (11,81); abbaiare (10,33); cane (8,76); contadino (8,37); pollaio (8,27); guardia (7,65); gallina (6,20); dormire (5,73); richiudere (5,46); svegliare (5,30); buca (3,81).

Confrontando i profili dei vari sottoinsiemi, le occorrenze di queste parole

dovrebbero risultare concentrate nei quattro episodi menzionati. E così è[10]. Ad esempio, *"Melampo"* è presente solo nell'episodio 22, dove anche *"faine"* esaurisce quasi il totale delle sue occorrenze (7 su 9); quanto a *"Geppetto"*, 24 delle sue occorrenze (su un totale di 83) sono nell'episodio 2; analogamente, le occorrenze di *"falegname"* sono tutte (9) all'interno dei primi due episodi.

In questo caso, quindi, la "strategia del fotografo"[11] non sembra molto adeguata, cioè le somiglianze/differenze tra i profili dei vari episodi non sembrano organizzare una visione d'insieme che possa funzionare come una rappresentazione grafica interpretabile. O meglio: l'interpretazione dei fattori sembra essere senza nessi con la possibilità di cogliere la struttura del racconto.

Per fare qualche verifica, provo ad esplorare lo spazio organizzato dal primo e dal terzo fattore (vedi Fig. 4), questa volta anche con l'aggiunta delle "parole". La situazione è un po' più articolata, ma non "vedo" utili itinerari interpretativi.

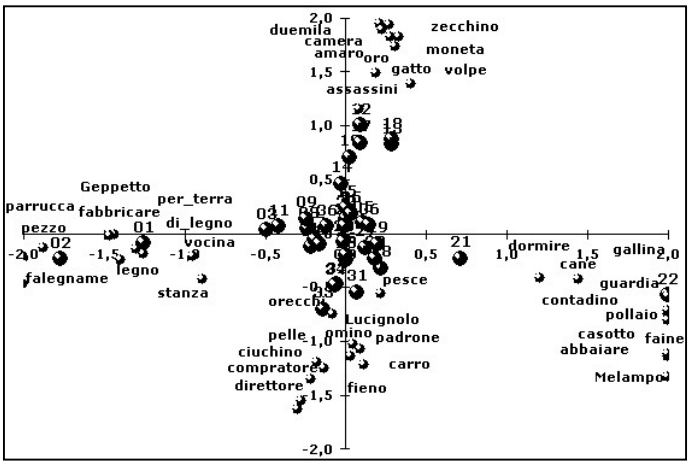



*Fig. 4 – Episodi e parole-chiave[12] sul piano 1x3 (Analisi delle corrispondenze)*

Penso quindi di cambiare strategia. In particolare, penso di passare dalla logica delle occorrenze a quella delle co-occorrenze. Non solo. Per garantirmi la possibilità di ottenere qualche risultato significativo, decido di fare un'accurata selezione delle parole da inserire nell'analisi. Il che, operativamente, significa procedere a delle "esclusioni"[13].

[10] In T-LAB questo tipo di verifica è possibile tramite la consultazione di tabelle di contingenza realizzate tramite un'apposita funzione del software (*Tabelle*).

[11] Vedi paragrafo 2.3. (Parte seconda).

[12] Le parole-chiave mostrate dal grafico sono quelle con i Valori Test più elevati. La selezione è effettuata in modo automatico da T-LAB.

[13] Vedi uso della funzione *Impostazioni personalizzate* (Fig. 4 paragrafo 2.1., Parte seconda).

I criteri che seguo sono due, entrambe orientati a incrementare il potenziale differenziante delle co-occorrenze:

- eliminare le parole con elevati valori di occorrenza che, allo stesso tempo, sono pressoché equamente distribuite nell'arco di tutto il racconto (in particolare: *"Pinocchio"* e *"burattino"*);
- eliminare le parole di significato "generico" che, in ipotesi possono essere presenti in contesti di co-occorrenza molto diversi tra loro (es: *"accadere"*, *"capitare"*).

Al termine dell'operazione, le mie parole-chiave risultano essere 317, tutte con valori di occorrenza uguali o superiori a 7.

Nell'ottica di costruire una visione d'insieme, utilizzo lo strumento *Mappe dei nuclei tematici*. Un risultato che ottengo è quello illustrato in Fig. 5.

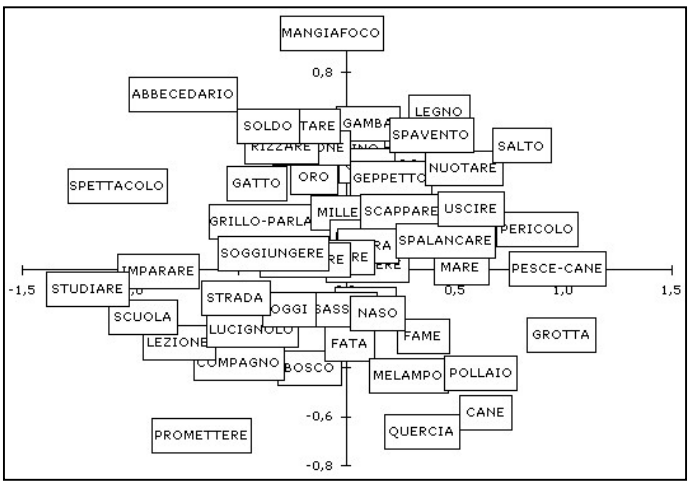



*Fig. 5 – Pinocchio: Mappa dei nuclei tematici*

Ogni label della mappa rappresenta un piccolo cluster di parole "associate". Cliccando su ciascuna di esse, posso esplorare gli elementi del cluster che portano il suo nome. Ad esempio (vedi Tab. 1), verifico che il cluster *"scuola"* include *"libro"*, *"maestro"*, *"pigliare"*, *"scuola"* e *"somaro"*. Inoltre, T-LAB mi consente di esplorare le relazioni tra le label (cluster) su vari piani cartesiani.

La rappresentazione ottenuta è suggestiva ed evocatrice: i personaggi, i luoghi, le emozioni sembrano essere quasi tutti presenti. Ma cosa ne è della loro connessione nella *storia* narrata da Collodi?

*Tab. 1 – Cluster corrispondenti ad alcune label della Fig. 33*

| *CLUSTER* | *Parole associate* |
|---|---|
| SCUOLA | libro, maestro, pigliare, scuola, somaro |
| IMPARARE | imparare, leggere, quattrino, scrivere |
| GATTO | duemila, gatto, gridare, volpe, zampa, zecchino, zoppo |
| MANGIAFOCO | Arlecchino, burattinaio, grazia, Mangiafoco |
| PERICOLO | affogare, pericolo, pesce, terribile |
| GROTTA | friggere, grotta, olio, padella |
| QUERCIA | affacciare, bambina, capelli_turchini, finestra, medico, morto, quercia |
| FATA | buona, cuore, fata, giudizio, mamma |

Il punto è che nessun "nucleo tematico" rappresenta una sequenza, né – tantomeno – la loro prossimità indica una qualche successione temporale. Ad esempio, *"Arlecchino"*, *"burattinaio"*, *"grazia"* e *"Mangiafuoco"* appartengono allo stesso cluster, ma in nessun caso sono contemporaneamente presenti all'interno della stessa frase. Il fatto che formino un piccolo cluster, e che quindi evochino una qualche isotopia, significa soltanto che nell'"albero" delle co-occorrenze le loro reciproche relazioni sono più forti delle relazioni di ciascuno di loro con una qualunque parola degli altri "nuclei tematici". Reciprocamente, le singole parole di ciascun contesto elementare (enunciati) possono appartenere a più di un cluster.

A questo punto, potrei provare ancora altre strade, ad esempio potrei tentare un'analisi con lo strumento *Tipologie dei contesti elementari*; ma, conoscendo la logica di T-LAB, so che continuerei ad avere una percezione di "assenza".

L'*esercizio* su Pinocchio, come ho anticipato, mi serviva per esplorare i limiti delle "regole del gioco", quelle che determinano le possibili strategie analitiche percorribili con T-LAB. Un limite invalicabile a questo punto mi è chiaro. Per dargli un nome e per rendere comprensibile la sua natura, mi si richiede una digressione attraverso le teorie e le tecniche per l'analisi del racconto.

## 4.2. Intrecci e figure

Secondo G. Genette (1966, tr. it., p. 273), il **racconto** è «la rappresentazione d'un evento, o d'una sequenza di eventi, reali o fittizi, per mezzo del linguaggio e, più particolarmente del linguaggio scritto». In quanto "enunciato narrativo", esso va distinto dalla **storia** (o "contenuto narrativo"), intesa come successione degli eventi reali o fittizi che formano l'oggetto del discorso, e

dalla **narrazione** intesa come l'"atto narrativo" attraverso il quale, in un determinata situazione (reale o fittizia), qualcuno racconta qualcosa a qualcun altro (G. Genette, 1972, tr. it., pp. 73-75).

Questi tre aspetti, in vario modo, sono inscritti nel tempo; tuttavia questa inscrizione assume significati molto diversi, a seconda che ci si ponga nell'ottica dello strutturalismo o in quella dell'ermeneutica.

Ad esempio, secondo R. Barthes, il racconto produce l'"illusione" del tempo e un compito della descrizione strutturale è quello di spiegare come ciò avviene:

> dal punto di vista del racconto, quello che noi chiamiamo tempo non esiste o almeno non esiste che in modo funzionale, come elemento d'un sistema semiotico; il tempo non appartiene al discorso propriamente detto, ma al referente; il racconto e la lingua conoscono soltanto un tempo semiologico; il "vero" tempo è un'illusione referenziale "realistica" (R. Barthes, 1966, tr. it., p. 24)

P. Ricoeur (1984, tr. it., p. 127), usando uno schema a tre livelli che ricorda quello proposto da Genette, distingue tra il *tempo del raccontare* (enunciazione), il *tempo raccontato* (enunciato) e l'*esperienza di finzione del tempo* (mondo del testo[14]). Secondo questa distinzione, solo il terzo tipo di tempo, quello del mondo del testo, può essere considerato "illusorio". Gli altri due, cioè quello dell'enunciazione (narrazione) e quello dell'enunciato (racconto) sono al contrario "reali". In particolare, il tempo del racconto vero e proprio, cioè quello dell'"enunciato narrativo"[15], secondo Ricoeur corrisponde ai movimenti dell'*intrigo*, cioè al dinamismo integratore che *da* una varietà di accadimenti ricava *una* storia (ibid., p. 20)[16].

Per cogliere la portata di quest'ultima affermazione, bisogna ricordare che l'analisi strutturale prevede una segmentazione del racconto in sequenze elementari[17], ciascuna delle quali articola specifiche relazioni tra atomi narrativi che, nella formulazione iniziale di W. Propp (1928), corrispondono alle *fun-*

[14] "Mondo del testo" è una categoria ermeneutica che solo in parte corrisponde alla nozione di "contenuto narrativo".

[15] L'unico che, secondo Genette, si offre direttamente all'analisi testuale.

[16] In altro luogo (1986, tr. it., p. 14), Ricoeur propone la seguente definizione: «intrigo è l'insieme delle combinazioni mediante le quali certi eventi vengono trasformati *in* storia o, correlativamente, una storia è ricavata *da* eventi. L'intrigo è il mediatore tra l'evento e la storia».

[17] I criteri di questa operazione variano in funzione del modello analitico adottato. Ad esempio, per C. Bremond (1966) la sequenza elementare è costituita dal raggruppamento di tre funzioni: quella che "apre", quella che "realizza" e quella che "chiude". Diversamente, per A. J. Greimas (1966) ad ogni sequenza deve corrispondere "un'articolazione prevedibile del contenuto" intesa come declinazione di un'isotopia.

*zioni* svolte dai personaggi[18]. Da un punto di vista modellistico, la segmentazione e l'analisi delle sequenze sono operazioni sull'asse del sintagma, mentre l'analisi delle funzioni si esplica attraverso operazioni sull'asse del paradigma[19]. Ne deriva che l'analisi si esplica attraverso due tipi di letture: una "orizzontale" (sequenze e sintagma) e l'altra "verticale" (funzioni e paradigma). Per dirla nei termini di C. Lévi-Strauss (1958), il racconto, come il mito, può essere letto come una partitura d'orchestra: con attenzione alla melodia (lettura orizzontale, da sinistra a destra) e/o all'armonia (lettura verticale, dall'alto in basso).

Ad esempio, la Tab. 2 illustra una ricostruzione del racconto *Les liaisons dangereuses* (P. C. de Laclos) proposta da T. Todorov[20].

*Tab. 2 – Les liaisons dangereuses* (Le relazioni pericolose) *secondo una ricostruzione di T. Todorov (1966, tr. it., p. 238)*

| | | | |
|---|---|---|---|
| Valmont desidera piacere | La Tourvel si lascia ammirare | La Merteuil cerca di ostacolare il primo desiderio | Valmont respinge i consigli della Merteuil |
| Valmont cerca di sedurre | La Tourvel gli accorda la propria simpatia | La Volanges cerca di ostacolare la simpatia | La Tourvel respinge i consigli di Volanges |
| Valmont dichiara il suo amore | La Tourvel resiste | Valmont insiste ostinatamente | La Tourvel respinge l'amore |
| Valmont cerca nuovamente di sedurre | La Tourvel gli accorda il proprio amore | La Tourvel fugge davanti all'amore | Valmont respinge apparentemente l'amore |
| L'amore è realizzato … | | | |

[18] Secondo Propp (1928, tr. it., p. 26), «le funzioni sono straordinariamente poche e i personaggi sono straordinariamente molti». Le funzioni, cioè, sono "categorie" di personaggi. Analogamente nella riformulazione proposta da A. J. Greimas (1966), gli *attanti* sono "classi" di attori che – nel racconto – svolgono le stesse funzioni. Ad esempio, nella sequenza elementare «Ieri sera Giulio e Roberta hanno fatto dei regali ai loro figli: un televisore a Pietro e una bicicletta a Maria» possono essere individuati tre attanti: il destinatore (Giulio e Roberta), l'oggetto (regali, televisore, bicicletta) e il destinatario (figli, Pietro, Maria).

[19] Per le nozioni di sintagma e paradigma, si veda il paragrafo 2.2.1. (Parte prima).

[20] In verità, nel testo citato, Todorov propone vari tipi di letture dello stesso racconto.

Leggendola da sinistra a destra, si dipana il "filo" dell'intreccio. Quanto al "denominatore comune" di ogni colonna, le osservazioni del semiologo bulgaro sono le seguenti:

> Tutte le proposizioni della prima colonna concernono l'atteggiamento di Valmont verso la Tourvel. Inversamente, la seconda colonna concerne esclusivamente la Tourvel e caratterizza il suo comportamento di fronte a Valmont. La terza colonna non ha per denominatore comune un soggetto ma tutte le proposizioni descrivono atti, nel senso forte del termine. La quarta, infine, possiede un predicato comune, cioè il rifiuto (nell'ultima riga si tratta di un finto rifiuto). I due membri di ogni coppia si trovano in una relazione quasi antitetica, e possiamo formulare la seguente proposizione:
> *Valmont : Tourvel : : gli atti : il rifiuto degli atti*[21] (T. Todorov, 1966, tr. it., pp. 238-239)

Da qui la conclusione che «in una narrazione, la successione delle azioni non è arbitraria, ma obbedisce ad una certa logica» (ibid., p. 239).

In realtà, i modelli per l'analisi strutturale del racconto sono molti e variegati; tuttavia tutti condividono il presupposto che la successione delle *azioni* e la costruzione dei *personaggi* obbediscono a una qualche logica. Inoltre il loro focus è sempre sulla "virtualità" delle possibili combinazioni o sulla "logica dei possibili narrativi". Per questa ragione, l'obiezione di Ricoeur non è priva di fondamento:

> La sintesi dei ruoli all'interno dell'intrigo non sta al termine di una combinazione di ruoli. L'intrigo è movimento; i ruoli sono dei posti, delle posizioni occupate nel corso dell'azione. Conoscere tutti i posti che potrebbero essere occupati – conoscere tutti i ruoli – *non vuol dire ancora conoscere l'intrigo*. Una nomenclatura, per quanto sia ramificata, non fa *una* storia raccontata (1984, tr. it., p. 77).

Evidentemente non è questo il luogo per approfondire questo problema; o meglio, in questa sede possiamo risparmiarci di chiarire ulteriormente le differenze tra le prospettive di analisi del racconto. Di fatto, tutte le analisi proposte da T-LAB sono a-croniche, ma non perché funzionano secondo la logica dell'analisi strutturale, bensì perché la logica costruttiva del database (vedi l'articolazione in unità di contesto e in unità lessicali) e quella degli algoritmi implementati nel software non sono adatte a trattare la dimensione temporale "interna" alla logica del racconto. Utilizzando la metafora proposta da C. Lévi-Strauss, potremmo dire che T-LAB non è in grado di restituire la "melodia"[22]. Aggiungo: questo limite strutturale non riguarda solo il racconto, ma

[21] Traduzione: *Valmont* "sta a" *Tourvel* "come" *gli atti* "stanno a" *il rifiuto degli atti*.
[22] Una "parvenza" di melodia può essere ricostruita attraverso un particolare uso della fun-

tutte le dimensioni che – in ogni testo – hanno a che fare con l'*intreccio*[23].

In effetti, l'intreccio non è qualcosa che riguarda solo i racconti. Un articolo scientifico, un libro di "saggistica", una relazione o una conversazione propongono sempre un ordine sequenziale dei loro contenuti: c'è sempre un prima e un dopo, così come ci sono sempre "salti" in avanti e indietro. Il motivo per cui questo aspetto dei testi, nei suoi termini generali, è stato poco analizzato, dipende forse dal fatto che *ab origine* Aristotele ha proposto la distinzione tra due diverse tecniche, la poetica e la retorica, e che a ciascuna di esse ha dedicato una trattazione specifica[24]. Come è noto, il luogo del racconto (il *muthos*) è la *Poetica* e, non a caso, tutte le tassonomie dei "generi letterari" hanno qualche radice in quest'opera[25]. Tuttavia i "generi letterari"[26] non includono tutti i tipi di testi, e non a causa del fatto che gli "altri" testi – quelli non letterari – hanno per oggetto tematiche diverse, bensì per il fatto che questi ultimi hanno diverse regole di costruzione: quelle che, secondo il modello aristotelico, riguardano l'ambito della retorica.

Per chiarire ulteriormente questo aspetto, proviamo ad esplorare le operazioni principali che fondano questo tipo di *techné*. La Tab. 3 è tratta dal testo di R. Barthes *Retorica Antica* (1970b, tr. it., p. 57).

*Tab. 3 – Le cinque parti della tecnica retorica*

| | | |
|---|---|---|
| INVENTIO | *invenire quid dicas* | trovare cosa dire |
| DISPOSITIO | *inventa disponere* | mettere ordine in quel che si è trovato |
| ELOCUTIO | *ornare verbis* | aggiungere l'ornamento delle figure |
| ACTIO | *agere et pronunciare* | recitare il discorso come attore: gesti e dizione |
| MEMORIA | *memoria mandare* | ricorrere alla memoria |

Come si può rilevare, le cinque operazioni si riferiscono alla costruzione del discorso pronunciato o da pronunciare; di queste, solo due lasciano tracce nel testo fissato dalla scrittura: la *dispositio* e l'*elocutio*.

La prima (*dispositio*), che Barthes suggerisce di tradurre con "composizione", concerne la retorica sintagmatica, cioè la collocazione delle varie "parti"

zione *Analisi delle sequenze* (vedi catene markoviane a paragrafo 2.3.5., parte prima).

[23] Uso questo temine nell'accezione proposta da U. Eco (1979, p. 102), intendendo «la storia come di fatto viene raccontata, come appare in superficie, con le sue dislocazioni temporali, salti in avanti e in indietro (ossia anticipazioni e flash-back), descrizioni, digressioni, riflessioni parentetiche».

[24] In due diversi libri: la *Poetica* e la *Retorica*.

[25] Sempre non a caso, nella trilogia dedicata a *Temps et Récit* (Tempo e racconto), Ricoeur fonda la sua ricerca su una rilettura della *Poetica* di Aristotele.

[26] Si pensi, ad esempio, ai vari "tipi" di romanzo: storico, poliziesco, autobiografico, ecc.

secondo un ordine sequenziale. La seconda concerne la retorica paradigmatica e si esplica attraverso la scelta e l'uso delle "figure"[27].

Quanto alle parti della *dispositio*, le principali sono quattro: *exordium*, *narratio* (intesa come resoconto dei fatti), *confirmatio* (intesa come resoconto degli argomenti: prove e confutazioni) e *conclusio*.

A una prima lettura, ingenuamente attenta ai "contenuti", solo la seconda parte (*narratio*) sembra imparentata con la logica del racconto; ma, se si assume l'ottica della "forma", è l'intera *dispositio* che si rivela come un tipo di *intreccio.* Inoltre, è importante sottolinearlo, come è vero che l'intreccio non è attivo nei soli testi narrativi, è altrettanto vero che esso non è attivo nella struttura di tutti i tipi di testi. Si pensi ad esempio ai dizionari e alle enciclopedie[28]: in questi casi, la *dispositio* delle varie parti o delle varie "voci", è in forma di catalogo, cioè è organizzata secondo la logica di un database. Al pari di Internet[29], questi tipi di testi sono liberamente navigabili; tuttavia, pur essendo privi di intreccio, non sono fuori del "dominio retorico".

Evidentemente, il loro essere "dentro" o "fuori", dipende dal criterio che adottiamo per segnare i confini. Quello che io ho in mente, è stato proposto da C. Perelman (1977, tr. it., p. 172):

> ogni discorso che non aspiri a una validità impersonale appartiene all'ambito della retorica. Allorché una comunicazione tende ad influenzare una o più persone, o orientare il loro pensiero, a suscitare o placare le emozioni, a dirigere un'azione, essa fa parte del campo della retorica.

Più precisamente, anche supponendo che i testi-cataloghi siano completamente privi di intreccio, resta da considerare l'impatto emozionale delle "figure", cioè di quell'insieme sterminato di dispositivi che alterano e orientano la percezione delle relazioni tra significanti e/o tra significati: iperboli, metafore, ossimori, chiasmi, ecc. Tutti dispositivi che, in genere, sono facilmente iscrivibili in piccole frasi, compresi i titoli di giornali e gli annunci pubblicitari.

A questo punto, uno dei limiti strutturali di T-LAB, quello che fondava la mia percezione di "assenza" nell'analisi di *Pinocchio*, sembra chiarito. Resta un problema: se, come ho affermato, l'*intreccio* e le *figure* sono una caratteristica sia dei testi narrativi che di quelli di "altri" generi, il fatto che T-LAB non sia in grado di trattare queste dimensioni[30] quali limiti pone ai suoi possi-

[27] In realtà, le figure retoriche propongono "giochi linguistici" che implicano l'interazione tra sintagma e paradigma.

[28] Ma anche alle risposte a questionari con domande aperte.

[29] Mi riferisco alla struttura della rete e ai modi per accedervi. Quanto ai suoi "contenuti" è un altro discorso.

[30] Solo una nota per spiegare il "limite" di T-LAB nel trattamento delle figure retoriche. Tre

bili usi? Evidentemente, questa domanda – oltre a valere per la maggior parte[31] dei software orientati al trattamento automatico dei testi – non concerne le caratteristiche dello strumento preso "in sé", ma chiama in causa il tipo di *uso* che qualcuno, in funzione dei propri obiettivi, intende farne. D'altro canto, una cosa sono i testi, più o meno organizzati dalla logica di un intreccio o di un'argomentazione, altra è l'interesse del ricercatore a ricostruire o a "seguire" la storia o la logica proposta dai loro autori.

Ogni attività di analisi comporta "guadagni" e "perdite"; in particolare, quando l'analisi è mediata da software che realizzano operazioni di scomposizione e di ricomposizione, la "perdita" più rilevante concerne l'unitarietà dell'intreccio, cioè quella "unicità del tutto" entro la quale – secondo Aristotele (*Poetica*, 1451a, 30[32]) – «le parti che la compongono devono essere coordinate per modo che, spostandone o sopprimendone una, ne resti come dislogato e rotto l'insieme».

In conclusione: se un ricercatore è soprattutto interessato alla ricostruzione dell'intreccio, oppure se ritiene che la "forma" dell'intreccio sia la caratteristica dominante del testo che si propone di analizzare (vedi *Pinocchio*), l'unica operazione sensata che deve fare consiste nel leggere il testo (o il corpus), dall'inizio alla fine. La stessa operazione, ovviamente, è "vivamente consigliata" anche nel caso dei testi poetici che, com'è noto, fanno abbondante uso di figure retoriche.

Infine, la lettura è un'esperienza piacevole e insostituibile, come quella che accompagna l'ascolto dei "movimenti" o dei "tempi" di una composizione musicale. Certo, anche queste esperienze possono essere analizzate e interpretate, ma le competenze per farlo non sono tutte traducibili nella logica del software.

## 4.3. Variazioni sul tempo

Su quale sia la natura del tempo nel racconto, o nell'esperienza della sua lettura, sospendo il giudizio. Ugualmente, per quanto le ritenga importanti, lascio aperte le questioni concernenti l'analisi delle dimensioni dell'*intreccio* tramite l'ausilio di software. Tuttavia, in margine ai discorsi fin qui fatti, trovo utile presentare tre brevi esempi[33] che in vario modo illustrano come la di-

classici esempi che, nella logica del software, sono solo "casi" di co-occorrenze: un chiasmo («Bisogna *lavorare* per *vivere* e non *vivere* per *lavorare*»); una metafora («*Beve* tutto ciò che gli *raccontano*»); un ossimoro («Questo *oscuro chiarore* che scende dalle stelle»).

[31] Forse per tutti.

[32] La numerazione rimanda alla classica edizione del testo greco curata da I. Bekker. La traduzione italiana è di Manara Valgimigli (Aristotele, 1966)

[33] Esempi che, con diversi obiettivi, sono stati pubblicati sul sito www.tlab.it.

mensione del tempo può essere trattata tramite uno strumento T-LAB: l'*Analisi delle corrispondenze unità lessicali* x *sottoinsiemi*.

Il commento che intendo fare è di tipo comparato; quindi non mi soffermerò a considerare specifici aspetti dei tre "casi". Ugualmente, non mi soffermerò sui criteri dei trattamenti linguistici, anche perché queste analisi sono state realizzate con le "impostazioni automatiche" di T-LAB, sia per la lemmatizzazione che per la selezione delle parole chiave. Particolare rilevanza assume invece la suddivisione in sottoinsiemi, nei primi due casi attraverso la codifica dei capitoli, nel terzo dei "periodi".

In particolare:

- il primo esempio è costituito da un racconto di J. Conrad (*The Shadow Line*);
- il secondo da un testo (*The Origin and Development of Psychoanalysis*) costituito da cinque conferenze che S. Freud tenne alla Clark University nel settembre 1909[34];
- il terzo da un corpus di 48 articoli, concernenti l'Islam, pubblicati su una nota rivista italiana in un arco di tempo che va dal gennaio 2001 al febbraio 2002.

L'ordine di presentazione non è casuale. Nel testo di Conrad la dimensione temporale – scandita dalla successione dei sei capitoli – è solo "interna" alla logica del racconto. Nel testo di Freud, le cinque conferenze (o lezioni), oltre a dare il nome ai vari capitoli, sono state pronunciate in tempi effettivamente diversi; inoltre, fatto non trascurabile, propongono una qualche ricostruzione della *storia* di una ricerca. Infine, nel corpus degli articoli sull'Islam, il tempo è solo una *variabile* usata per distinguere quattro periodi: il "prima" (ante) dell'11 settembre 2001, la fase intermedia tra l'attacco alle torri gemelle e l'inizio della guerra in Afghanistan (7 ottobre), il primo mese di guerra, il periodo successivo (dal novembre 2001 al febbraio 2002).

Il risultato dell'analisi applicata al testo di Conrad è in qualche modo sorprendente: la dimensione bipolare del primo fattore, percorsa da sinistra a destra, ripropone puntualmente la successione dei sei capitoli (Fig. 6).

A partire da sinistra, le parole (Fig. 7) che caratterizzano questo tipo di temporalità evocano personaggi, eventi ed emozioni che, nell'esperienza dell'autore, «si accompagnano al passaggio dalla gioventù, fervida e spensierata, al periodo più consapevole e patetico dell'età matura»[35].

[34] Sia il testo di S. Freud che quello di J. Conrad sono stati analizzati nelle versioni inglesi disponibili su Internet (ricordo che T-LAB è un software multilingue).

[35] La frase è tratta dalla "nota dell'autore" che, nella versione italiana, introduce il racconto.

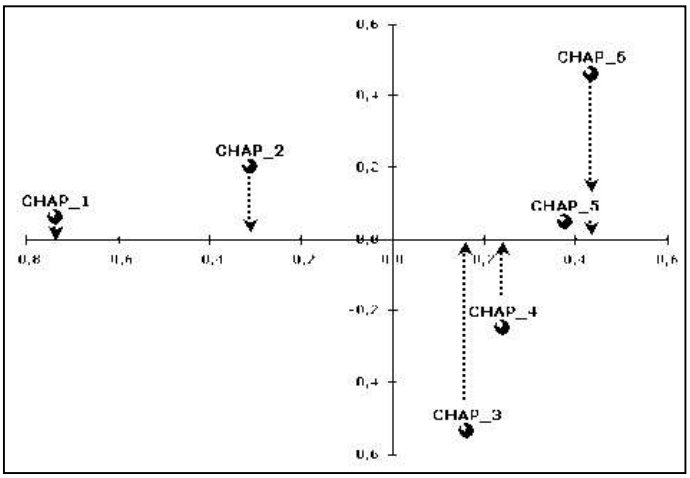



*Fig. 6 – The Shadow line (Analisi delle corrispondenze)*

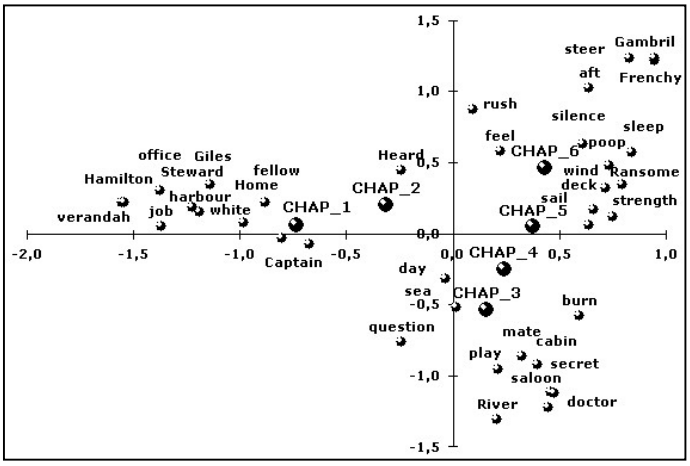



*Fig. 7 – The Shadow Line (Analisi delle corrispondenze)*

Anche nel caso del testo di Freud, la proiezione dei vari capitoli (o lezioni) sul primo asse dell'analisi delle corrispondenze[36] riproduce la loro successione temporale[37]. Inoltre, procedendo da destra verso sinistra, la lettura delle parole[38] riportate in Fig. 8 consente di ricostruire un qualche filo del discorso: dalle ricerche sull'isteria si passa all'interpretazione dei sogni, quindi alla teoria della repressione e a quella della sessualità infantile.

[36] Nella Fig. 8 sono rappresentati il primo e il secondo fattore.
[37] L'unico scarto riguarda le posizioni degli ultimi due capitoli: il quarto e il quinto.
[38] Sono quelle selezionate automaticamente da T-LAB.

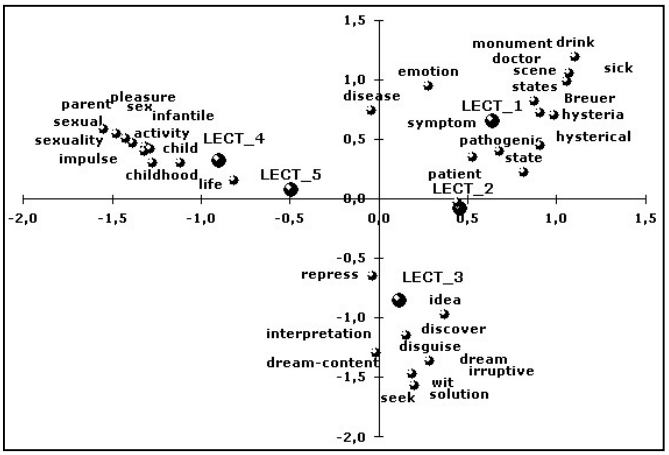



*Fig. 8 – S. Freud: The Origin and Development of Psychoanalysis (Analisi delle corrispondenze)*

Diversamente, nel caso del corpus con gli articoli sull'Islam la ricostruzione della successione temporale porta a disegnare delle linee a zig-zag che si muovono su entrambi le dimensioni del piano fattoriale (vedi Fig. 9). Inoltre, la bipolarità del primo fattore è chiaramente organizzata dall'opposizione tra un "periodo" (quello che va dall'attacco alle torri gemelle all'inizio delle operazioni militari in Afghanistan) e tutti gli altri. Quanto poi alle label del grafico (ad es. *"l'Islam come cultura religiosa"*), le ho inserite per riassumere le caratteristiche salienti dello spazio fattoriale rappresentato[39].

Il tutti e tre i casi, l'intreccio degli eventi e dei personaggi resta sullo sfondo e, per molti versi, inaccessibile. Tuttavia, il rilievo di un qualche tipo di "ordine" è possibile. Volendo lo possiamo assumere come significante di una storia; ma le tre *storie* rappresentate nei grafici hanno nature affatto diverse. Nel primo caso (Conrad) la storia raccontata è un prodotto dell'immaginazione letteraria, nel secondo (Freud) è la ricostruzione "argomentata" di una percorso di ricerca, nel terzo (Islam) è quella "cosa" a partire dalla quale i testi analizzati sono stati prodotti.

[39] Le prime 15 parole che, nell'ordine, caratterizzano (Valori Test) le quattro polarità sono le seguenti:

*L'attacco alle torri gemelle*: americano; Stati_Uniti; New_York; aereo; terrorismo; terroristico; militare; World_Trade_Center; intelligence; Washington; attentato; Bush; Bin_Laden; attacchi; Al-Qaeda;

*L'Islam come cultura religiosa*: Islam; moschea; musulmano; donna; pregare; ebreo; Corano; arabo; religione; cultura; papa; Dio; studente; famiglia; immigrato;

*La guerra in Afghanistan*: Al_Jazeera; talebani; madrasse; America; giornalista; Kabul; donna; Qatar; Haqqania; velo; Bbc; afghano; canale; foto; diritto;

*L'Islam come cultura della violenza*: Sharia, Houllebecq; palestinese; Gerusalemme; ferito; lapidazione; israeliano; Omar; Hamas; ebreo; Sharon; Jaffa; Safiya; taglio; mano.

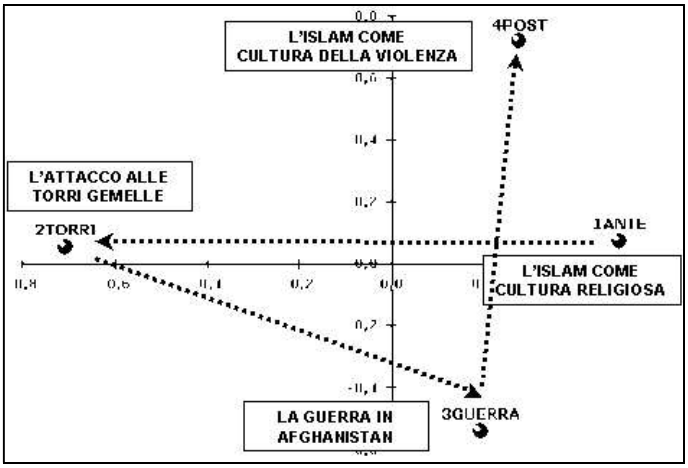



*Fig. 9 – Corpus di 48 articoli sull'Islam Analisi delle corrispondenze*

Inoltre: nei primi due casi (Conrad e Freud) siamo in presenza di "un" racconto (o resoconto) tenuto insieme dalle intenzioni di "un" autore, quindi il tempo è scandito come una composizione musicale; diversamente, nel terzo caso il tempo è scandito dalla irruzione degli eventi: attentati, guerra, ecc.

Tuttavia, in tutti e tre i casi, valgono due tipi di conclusioni:

- la rilevazione di un qualche tipo di ordine dipende dal fatto che, a monte, l'analista ha utilizzato criteri di classificazione in grado di articolare somiglianze e differenze tra le unità di contesto;
- le parole sono alla ricerca di una altro racconto: quello costruito dall'analista. Se esse, in quanto "dati", sembrano parlare è perché qualche interprete è in grado di collocarle in un altro contesto e in un'altra storia.

# Appendici

## 1. Dizionari: tra lessici e teorie

In questa appendice vengono fornite alcune informazioni utili ai ricercatori che intendano costruire, modificare o utilizzare, dizionari da importare in sessioni di lavoro T-LAB.

Nel paragrafo 2.2.2 (parte prima), è stata descritta la struttura del *Dizionario del corpus*; inoltre è stata spiegata la logica che presiede all'uso della funzione *Personalizzazione del dizionario* (menu *Lessico*).

Come si ricorderà, al termine della fase di importazione del corpus, sia che si utilizzi l'opzione della lemmatizzazione automatica, sia che si decida di escluderla, nel database di sessione (quello del corpus in analisi) ogni "parola" (forma grafica) è etichettata con una label. Più precisamente, nel primo caso (Tab. 1) la label corrisponde a un lemma, mentre nel secondo caso (Tab. 2) label e forma grafica sono identiche.

*Tab. 1 – Con lemmatizzazione*

| *Forma grafica* | *Label* |
|---|---|
| (…) | (…) |
| influenzava | influenzare |
| influenzavamo | influenzare |
| influenzavano | influenzare |
| influenzavate | influenzare |
| influenzavi | influenzare |
| (…) | (…) |

*Tab. 2 – Senza lemmatizzazione*

| *Forma grafica* | *Label* |
|---|---|
| (…) | (…) |
| influenzava | influenzava |
| influenzavamo | influenzavamo |
| influenzavano | influenzavano |
| influenzavate | influenzavate |
| influenzavi | influenzavi |
| (…) | (…) |

In entrambi i casi, intervenire sul dizionario del corpus significa modificare le label, cioè raggruppare (o distinguere) le forme grafiche secondo modalità congruenti con i propri obiettivi di analisi.

A questo scopo, fatti salvi alcuni tipi di operazioni che possono essere agevolmente realizzati tramite l'interfaccia T-LAB (vedi Fig. 5 a paragrafo 2.2.2 della prima parte), i ricercatori possono costruire, modificare e usare, diversi file dizionario; sia per applicarli – allo stesso corpus – nella stessa sessione di lavoro, sia per applicarli – in

tempi diversi – allo stesso corpus o a "n" differenti corpora.

Più precisamente, T-LAB consente di importare due tipi di file: *Dictio.diz* e *Dictionary.diz*[1].

Il primo (*Dictio.diz*) può essere importato e utilizzato sia nei casi in cui è stata scelta l'opzione della lemmatizzazione automatica, sia nei casi in cui la stessa opzione è stata disattivata. Infatti, la sua struttura prevede tanti "accoppiamenti" quante sono le parole (forme grafiche) da classificare.

Le sue regole di costruzione sono le seguenti[2]:

- il file deve essere costituito da tante righe quante sono i "tipi" di parole da classificare (a seconda dei casi, tutte le parole presenti nel corpus o solo quelle considerate rilevanti);
- ogni sua riga deve riportare una label (ad es. un lemma) e una parola, separate dal carattere ";" (punto e virgola);
- ogni sua riga, senza spazi e ulteriori segni di punteggiatura, deve terminare con un "a capo" (ritorno di carrello);
- il formato del file deve essere del tipo "solo testo".

| Ad esempio: | Oppure: |
|---|---|
| (...) | (...) |
| influenza;influenzava | potere;influenzava |
| influenza;influenzavamo | potere;influenzavamo |
| influenza;influenzavano | potere;influenzavano |
| influenza;influenzavate | potere;influenzavate |
| influenza;influenzavi | potere;influenzavi |
| (...) | (...) |

Per le sue caratteristiche strutturali, questo tipo di file dizionario (*Dictio.diz*) può essere utilizzato anche per lemmatizzare e/o per classificare forme grafiche non riconosciute da T-LAB (ad es. parole appartenenti a lessici specialistici, a dialetti o a lingue straniere).

La Fig. 1 illustra la logica delle operazioni che fanno seguito alla sua importazione:

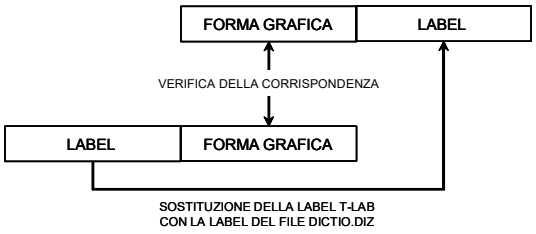



*Fig. 1 – Uso del file Dictio.diz*

[1] La possibilità di importare due tipi di dizionari è una prerogativa delle versioni T-LAB PRO (4.0 e seguenti).

[2] Per facilitare la costruzione e la modifica di questo tipo di dizionario, T-LAB consente di esportare un file *Dictio.diz* con tutte le parole presenti nel corpus in analisi, e in un formato già predisposto per ulteriori applicazioni.

Quanto al secondo tipo di file dizionario (*Dictionary.diz*), esso può essere importato e utilizzato solo nei casi in cui è stata scelta l'opzione della lemmatizzazione automatica. Infatti, la sua struttura – formalmente analoga a quella dei file *Dictio.diz* – prevede "accoppiamenti" tra label di primo e di secondo ordine: le prime, attribuite automaticamente da T-LAB (vedi colonna "label" di Tab. 1), sono generalmente lemmi; le seconde, definite dall'utilizzatore, possono essere lemmi e/o categorie tematiche derivate da specifici modelli di analisi.

Ad esempio:

```
(...)
potere;influenzare
potere;influenza
(...)
terapia;cura
terapia;curare
(...)
```

Oppure:

```
(...)
anomia;disorientamento
anomia;confusione
(...)
politica;governo
politica;parlamento
(...)
```

Anche se gli usi dei due tipi di dizionari sono spesso analoghi, tendenzialmente il primo (*Dictio.diz*) risulta più idoneo nei casi in cui la classificazione delle parole segue la logica del processo di lemmatizzazione; diversamente, il secondo (*Dictionary.diz*) risulta più adeguato quando si intendano applicare "griglie" per l'analisi di contenuto, non ultime quelle derivate da tradizioni di ricerca di origine anglosassone (ad es. i *dictionaries* implementati nel *General Inquire*).

La Fig. 2 illustra la logica delle operazioni che fanno seguito alla importazione di un file *Dictionary.diz*:

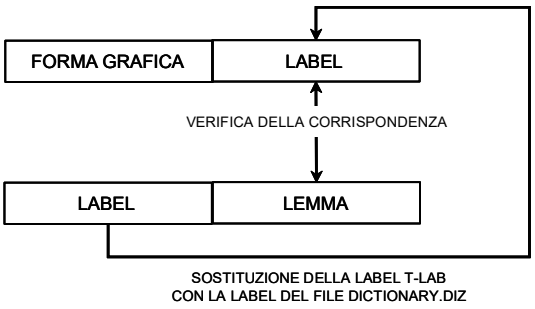



*Fig. 2 – Uso del file Dictio.diz*

In entrambi i casi (*Dictio.diz* e *Dictionary.diz*), le operazioni di modifica e/o costruzione possono essere realizzate mediante un qualunque editore di testi; mentre, per la importazione dei file è sufficiente aprire la finestra *Personalizzazione del dizionario* e cliccare sul pulsante "Importa dizionario" (vedi Fig. 5, paragrafo 2.2.2, prima parte del libro).

## 2. La manipolazione degli output

Gli output di T-LAB sono di due tipi: tabelle e grafici. In generale, per editarli e stamparli, bastano un browser[3] e due applicazioni disponibili nelle varie versioni di MS Windows: *Wordpad* e *Paint*, il primo per le tabelle e il secondo per i grafici.

Tuttavia, in molti casi, il ricercatore può essere interessato a modificare l'aspetto degli output, vuoi per costruire tabelle ad hoc, vuoi per predisporre grafici con un stile personalizzato.

Nel caso delle tabelle, quelle che T-LAB archivia in file con estensione .DAT[4], valgono i seguenti accorgimenti:

- aprire il file con un editore di testo (ad es. WordPad);
- selezionare i dati che interessano, copiarli e incollarli in un nuovo file;
- se il dati selezionati includono numeri con decimali, sostituire i punti con le virgole (opzione trova/sostituisci);
- salvare il nuovo file in formato solo testo con estensione .TXT.

Successivamente, se si intende utilizzare MS Excel, le operazioni da fare sono le seguenti:

- aprire Excel dal menu "Programmi";
- mediante Excel (menu File/Apri con opzione "tutti i file") cercare e aprire il file salvato;
- quindi, seguire i vari "passaggi" dell'importazione guidata: a) selezionare "campi delimitati" e cliccare su "avanti"; b) selezionare il delimitatore "spazio" e cliccare su "avanti"; c) selezionare il formato "testo" per le colonne contenenti parole e cliccare su "fine".

Se la tabella importata include misure che consentono la generazione di grafici, si può utilizzare la corrispondente funzione di MS Excel ("creazione guidata grafico").

Ad esempio, nel caso dell'analisi delle corrispondenze, T-LAB produce output grafici come quello in Fig. 1. Volendo, seguendo la procedura sopra descritta, l'utilizzatore, può generare grafici come quello in Fig. 2, sia selezionando le label da inserire, sia scegliendo il formato dei caratteri e dei colori.

[3] Caso degli output in formato HTML.

[4] La funzione "Tabelle" del menu T-LAB consente anche di salvare file in formato .XLS.

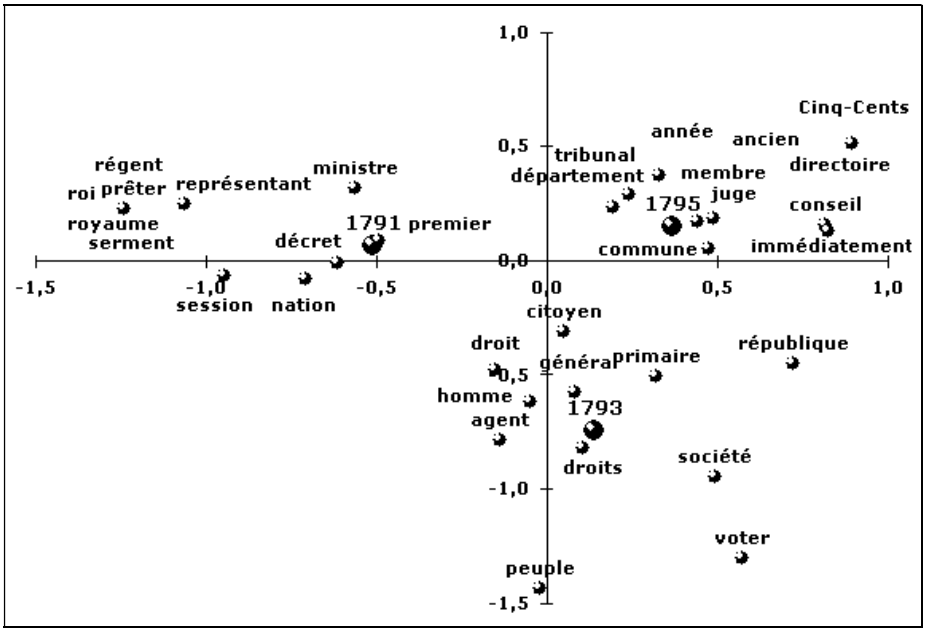



*Fig. 1 – Grafico T-LAB (Analisi delle corrispondenze)*

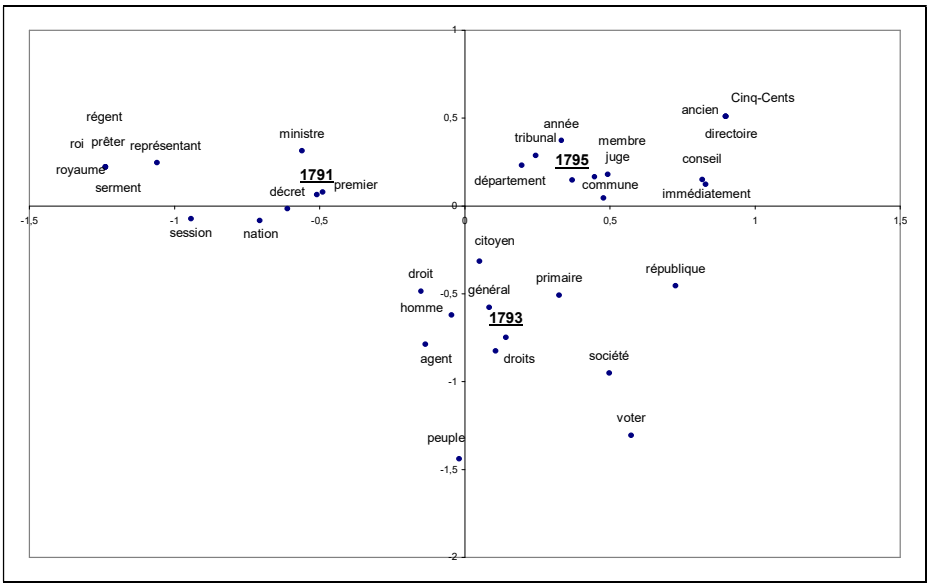



*Fig. 2 – Grafico ottenuto con MS Excel*

Tutti i grafici salvati da T-LAB sono in formato .BMP e possono essere agevolmente editati e modificati mediante vari tipi di software (ad es. Paint). Inoltre, mediante l'opzione "Inserisci/Immagine/Da file" possono essere acquisiti in tutti i programmi della suite MS Office (ad es. MS Word e MS PowerPoint).

In generale, nelle funzioni *Analisi e Mappe*, T-LAB consente di modificare la visualizzazione dei grafici, sia tramite lo spostamento di label all'interno dello spazio cartesiano, sia – in alcuni casi – tramite la ri-denominazione delle stesse label. Quando l'utilizzatore fa ricorso a questo tipo di opzione, deve tener conto del fatto che il salvataggio dell'immagine può presentare qualche anomalia[5]; quindi, se vuole evitarla, deve seguire un procedimento ad hoc.

Ad esempio, se il file generato da T-LAB è come quello mostrato in Fig. 3 e se l'utilizzatore non interviene in alcun modo, il suo salvataggio automatico produrrà un'immagine identica. Diversamente, se l'utilizzatore sposta qualche label del grafico (vedi Fig. 4), il suo salvataggio automatico produrrà un file come quello in Fig. 5 che, certamente, non è "elegante".

Per ovviare a questo problema, quando l'utilizzatore ha completato i suoi interventi e – in T-LAB – "vede" il grafico come in Fig. 4, deve copiare l'immagine dello schermo negli appunti di Windows (tasto "Stamp"), quindi, dopo aver aperto un programma grafico (ad es. Paint) deve incollare l'immagine in nuovo file e provvedere ad eventuali aggiustamenti.

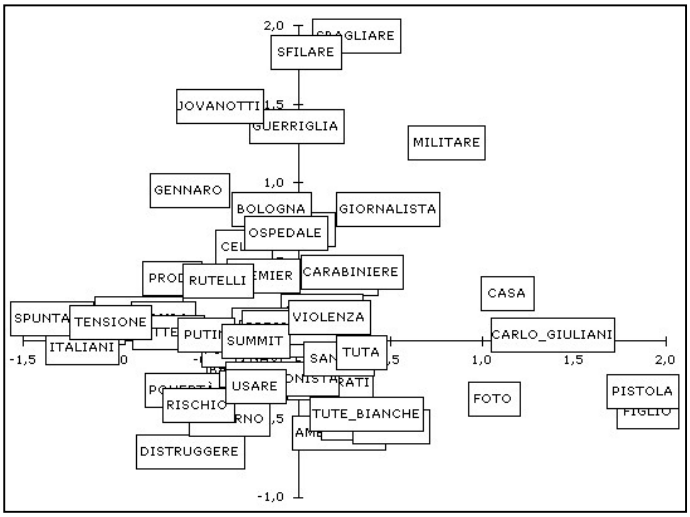



*Fig. 3 – Grafico mostrato da T-LAB*

[5] Anomalia causata dall'interazione tra T-LAB e l'ambiente Windows.

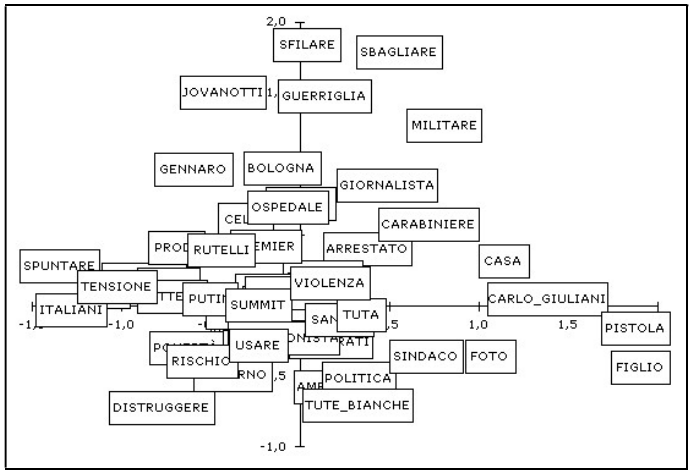



*Fig. 4 – Grafico modificato dall'utente*

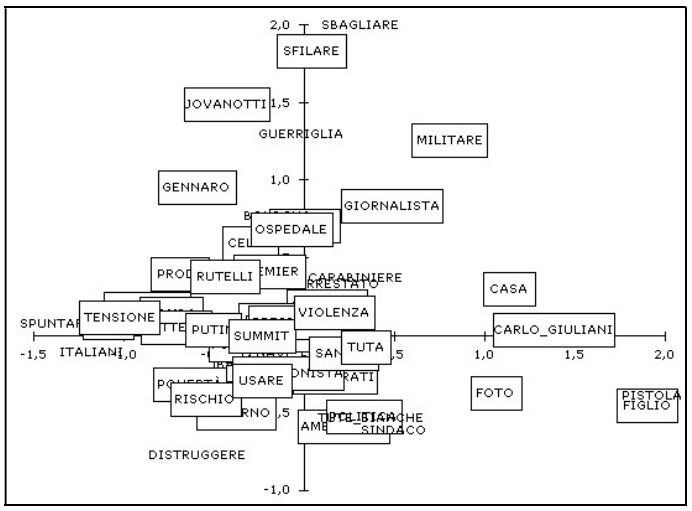



*Fig. 5 – Grafico di Fig. 4 come appare dopo il salvataggio automatico*

# Riferimenti bibliografici


ARISTOTELE (1962), *Etica Nicomachea*, Laterza, Bari.

- (1966), *Poetica*, Laterza, Bari.

AUSTIN J.L. (1962), *How to do Things with Words*, Oxford University Press, Oxford (tr. it. *Quando dire è fare*, Marietti, Torino, 1974).

BARDIN L. (1977), *L'analyse de contenu*, P.U.F., Paris.

BARTHES R. (1964), *Eléments de sémiologie*, Seuil, Paris (tr. it. *Elementi di semiologia*, Einaudi, Torino, 1966).

- (1966), *"*Introduction à l'analyse structurale des récit" in *Communications 8* (tr. it. in Aa.Vv. *L'analisi del racconto*, Bompiani, Milano, 1969).
- (1970b), *"*L'ancienne rhétorique" in *Communications 16* (tr. it. *La retorica antica*, Bompiani, Milano, 1972).
- (1980), *La chambre claire. Note sur la photographie*, Seuil, Paris (tr. it. *La camera chiara*, Einaudi, Torino, 1980).

BECCARIA G.L. (diretto da) (1994), *Dizionario di linguistica e di filologia, metrica, retorica*, Einaudi, Torino.

BENZECRI J.P & F. (1984), *Pratique de l'analyse des données. Analyse des correspondances & Classification*, Dunod, Paris.

BIORCIO R. (1993), *L'analisi dei gruppi*, Angeli, Milano.

BLACK M. (1962), *Models and Metaphors*, Cornell University Press, Itacha-London.

BLOOMFIELD L. (1933), *Language*, Holt, Reinehart & Winston, New York (tr. it. *Il linguaggio*, il Saggiatore, Milano, 1974).

BOBBIO N. (1980), *Norma,* in *Enciclopedia,* Einaudi, Torino, vol. IX, pp. 876-876.

BOLASCO S. (1999), *Analisi Multidimensionale dei dati. Metodi, strategie e criteri di interpretazione*, Carocci, Roma.

BORGES J. (1952), *El idioma analitico de John Wilkins*, Émecé Editores, Buenos Aires (tr. it. in *Tutte le opere*, Einaudi, Torino, vol I, pp. 1004-5).

BOURDIEU P., PASSERON J.C. (1970), *La reproduction. Eléments pour une théorie du système d'enseignement*, Minuit, Paris.

BREMOND R. (1966), *"*La logique des possibles narratifs" in *Communications 8* (tr. it. in Aa.Vv. *L'analisi del racconto*, Bompiani, Milano, 1969).

BRUNER J.S., BROWN R.W. (1956), *A Study of Thinking*, J. Wiley & Sons Inc., New York (tr. it. in *Il pensiero. Strategie e categorie*, Roma, Armando, 1973).
BURT C.L. (1940), *Factors of the Mind*, University of London Press, London.
CALVINO I. (1993), *Lezioni americane*, Einaudi, Torino.
CAMPELLI E. (1996), *Metodi qualitativi e teoria sociale*, in CIPOLLA C., DE LILLO A. (a cura di) (1996) pp. 17-36.
CARLI R., PANICCIA R.M., LANCIA F. (1988), *Il gruppo in psicologia clinica*, La Nuova Italia Scientifica, Roma.
CICOGANI E. (2002), *L'approccio qualitativo della Grounded Theory in psicologia sociale: potenzialità, ambiti di applicazione e limiti*, in MAZZARA B.M. (a cura di) (2002) pp. 43-60.
CINANNI V. (1990), *Dimensioni di somiglianza*, il Mulino, Bologna,.
CIPOLLA C., DE LILLO A.M. (a cura di) (1966), *Il sociologo e le sirene. La sfida dei metodi qualitativi*, Angeli, Milano.
DRAY W. (1957), *Laws and Explanation in History*, Clarendon Press, Oxford (tr. it. *Leggi e spiegazione in storia*, il Saggiatore, Milano, 1974).
DI BERNARDO G. (1983), *Le regole dell'azione sociale*, il Saggiatore, Milano.
DUBOIS J., GIACOMO M., GUESPIN L., MARCELLESI C. E J.B, MEVEL J.P.. (1973), *Dictionnaire de linguistique*, Larousse, Paris (tr. it. *Dizionario di linguistica*, Zanichelli, Bologna, 1979).
ECO U. (1973), *Segno*, ISEDI, Milano.
- (1979), *Lector in fabula. La cooperazione interpretativa nei testi narrativi*, Bompiani, Milano.
- (1981), *Significato,* in *Enciclopedia,* Einaudi, Torino, vol. XII, pp. 830-876.
- (1983), "Corna, zoccoli, scarpe. Alcune Ipotesi su tre tipi di abduzione", in Eco U., Sebeok T.A., *Il segno dei tre*, Bompiani, Milano.

ELKANA Y. (1981), *A programmatic Attempt at an Anthropology of Knowledge*, in E. Mendelssohn and Y. Elkana (eds.), *Sciences and Cultures*, D. Reidel Publishing Company, Dordrecht (tr. it. *Antropologia della conoscenza*, Laterza, Bari, 1989).
FILIASI CARCANO P. (1969), *La metodologia nel rinnovarsi del pensiero contemporaneo*, Libreria Scientifica Editrice, Napoli.
FORNARI F. (1977), *Il Minotauro. Psicoanalisi dell'ideologia*, Rizzoli, Milano.
FOUCAULT M. (1966), *Les mots et les choses,* Gallimard, Paris (tr. it. *Le parole e le cose*, Rizzoli, Milano, 1980).
- (1969), *L'archéologie du savoir,* Gallimard, Paris (tr. it. *L'archeologia del sapere*, Rizzoli, Milano, 1967).

GADAMER H.G. (1960), *Wahreit und Methode*, J.B.C. Mohr, Tübingen (r. it. *Verità e Metodo*, Fabbri, Milano, 1972).
- (1961), *Kleine Schriften*, J.B.C. Mohr, Tübingen (tr. it. *Ermeneutica e metodica universale*, Marietti, Torino, 1973).

GENETTE G. (1966), "Frontières du récit" in *Communications 8* (tr. it. in Aa.Vv. *L'analisi del racconto*, Bompiani, Milano, 1969).
- (1972), *Figures III*, Paris, Seiul (tr. it. *Figure III*, Einaudi, Torino, 1976).

GEERTZ C. (1973), *The Interpretation of Cultures*, Basic Books, New York (tr. it. *Interpretazione di culture*, il Mulino, Bologna, 1987).

GINZBURG C. (1979), *Spie. Radici di un paradigma indiziario*, in A. Gargani (a cura di), *Crisi della ragione*, Einaudi, Torino.
GLASER B.G., STRAUSS A.L. (1967), *The Discovery of Grounded Theory*, Aline, Chicago.
GREENACRE M.J. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press, New York.
GREIMAS A.J. (1966), *Sémantique structurale*, Larousse, Paris (tr. it. *La semantica strutturale*, Rizzoli, Milano, 1968).
GREIMAS A.J., COURTÉS J. (1979), *Sémiotique. Dictionnaire raisonné de la théorie du language*, Haschette, Paris.
HAIR J.F., ANDERSON R.E., TATHAM R.L., BLACK W.C. (1995), *Multivariate Data Analysis with Readings*, Prentice-Hall International, London.
HANSON N. R. (1958), *Patterns of Discovery. An Inquiry into the Conceptual Foundations of Science*, Syndics of the Cambridge University Press, England (tr. it. *I modelli della ricerca scientifica. Ricerca sui fondamenti concettuali della scienza*, Feltrinelli, Milano, 1978).
HARRÉ R. (1974), *Some Remarks on "Rule" as a Scientific Concept*, in T. Mischel (eds.), *Understanding other Persons,* Basil Blackwell, Oxford, 1974: 143-184.
HARRÉ R., SECORD P.F. (1972), *The Explanation of Social Behaviour*, Basil Blackwell, Oxford (tr. it. *La spiegazione del comportamento sociale*, il Mulino, Bologna, 1974).
HEMPEL C.G. (1952), *Fundamentals of Concept Formation in Empirical Science*, The University of Chicago Press, Chicago (tr. it. *La formazione dei concetti nella scienza empirica*, Feltrinelli, Milano, 1961).
HESSE M. B. (1980*), Models and Analogies in Science*, University Notre Dame Press, Indiana (tr. it. *Modelli e analogie nella scienza*, Feltrinelli, Milano, 1980).
HINTIKKA J. (1973*), Logic, language games and information*, Oxford University Press, London (tr. it. *Logica, giochi linguistici e informazione*, il Saggiatore, Milano, 1975).
HJELMSLEV L. (1943), *Omkring sprogteoriens grundloeggelse*, Kobenhavn, Munksgaard (tr. it. *I fondamenti della teoria del linguaggio*, Einaudi, Torino, 1968).
- (1959), Essais linguistiques (*Travaux du Cercle Linguistique de Copenhague*, vol. XII), Nordisk Sprong-og Kulturforlag, Copenhagen.
HUGHES G.E., CRESSWELL M.J. (1968), *An Introduction to Modal Logic*, Methuen, London (tr. it. *Introduzione alla logica modale*, il Saggiatore, Milano, 1973).
JAKOBSON R. (1963), *Essais de linguistique générale*, Editions de Minuit, Paris (tr. it. *Saggi di linguistica generale*, Feltrinelli, Milano, 1966).
KUHN T.S. (1962), *The Structure of Scientific Revolutions*, University of Chicago, Chicago (tr. it. *La struttura delle rivoluzioni scientifiche*, Einaudi, Torino, 1969).
- (1977), *The Essential Tension*, University of Chicago, Chicago (tr. it. *La tensione essenziale. Cambiamenti e continuità nella scienza,* Einaudi, Torino, 1985).
KRIPKE S. (1980), *Naming and Necessity*, Basil Blackwell, Oxford (tr. it. *Nome e necessità*, Boringhieri, Torino, 1982).

KRIPPENDORFF K. (2004), *Content Analysis. An Introduction to its Methodology* (Second Edition), Sage Publications, London.
LANCIA F. (1984), *Il linguaggio in prospettiva psicosociale*, in Aa.Vv. *La Psicologia nella società moderna*, Claire, Milano, pp. 77-86.
- (1990), "Prassi e resoconto in psicologia clinica", *Rivista di Psicologia Clinica*, 1: 52-64.
- (1996), *La valutazione delle risorse umane: contesti culturali e qualità del servizio*, in Borgogni L. (a cura di) *Valutazione e motivazione delle risorse umane nelle organizzazioni*, Angeli, Milano, 63-83.
- (2002), *La logica di un testoscopio,* www.tlab.it.
LEBART L. MORINEAU A. PIRON M. (1995), *Statistique exploratoire multidimensionnelle*, Dunod, Paris.
LEBART L. SALEM A. (1994), *Statistique textuelle*, Dunod, Paris.
LIPSCHUTZ S. (1965), *Probability*, McGraw-Hill, New York (tr. it. *Il calcolo delle probabilità*, Etas, Milano, 1975).
LÉVI-STRAUSS C. (1958), *Anthropologie structurale*, Plon, Paris (tr. it. *Antropologia strutturale*, Bompiani, Milano, 1971).
LORENZ K. (1989), *Hier bin ich – who bist du?*, R. Piper GmbH & Co., München (tr. it. *Io sono qui. Tu dove sei? Etologia dell'oca selvatica*, Mondadori, Milano, 1990.
LOTMAN J.M. (1992), *Kul'tura i Vzryv*, Gnosis, Moskva (tr. it. *La cultura e l'esplosione. Prevedibilità e imprevedibilità*, Feltrinelli, Milano, 1993).
LOTMAN J.M. USPENSKIJ B.A. (1973), *Tipologia della cultura*, Bompiani, Milano.
MARANDA P. (1995), *DiscAn: un programma di analisi reticolare per la costruzione delle carte semantiche*, in R. Cipriani, S. Bolasco (a cura di), *Ricerca qualitativa e computer*, Angeli, Milano, 1995, pp. 171-183.
MARRADI A. (1996), *Due famiglie e un insieme*, in CIPOLLA C., DE LILLO A. (a cura di) (1996) pp. 167-178.
MARTINET A. (1960), *Eléments de linguistique générale,* Colin, Paris (tr. it., *Elementi di linguistica generale*, Laterza, Bari, 1971).
MATURANA H. VARELA F. (1980), *Autopoiesi and Cognition. The Realization of the Living*, D. Reidel, Dordrecht (tr. it. *Autopoiesi e cognizione. La realizzazione del vivente*, Marsilio, Venezia, 1985).
MAZZARA B.M. (a cura di) (2002), *Metodi qualitativi in psicologia sociale,* Carocci, Roma.
MICHELET B. (1988), *L'analyse des associations,* Thèse de doctorat*,* Université Paris VII*, Paris*.
MILLER G.A., GALANTER E., PRIBRAM K.H., (1960), *Plans and Structure of Behaviour,* Rinehart and Winston, New York (tr. it. *Piani e struttura del comportamento*, Angeli, Milano, 1979).
MOORE S., MYERHOFF B. (1977), "Secular Ritual: Forms and Meanings" in MOORE S., MYERHOFF B. (eds.), *Secular Ritual,* Van Gorcum: Assen, Netherlands, 1977, pp. 3-24.
ORE O. (1963), *Graphs and Their Uses*, Yale University (tr. it. *I grafi e le loro applicazioni*, Zanichelli, Bologna, 1976).

ORTONY A. (1979) (eds.), *Metaphor and Thought*, Cambridge University Press, Cambridge.
OSGOOD C.E. (1959), *The Representation Model and Relevant Research Methods*, in Ithiel de Sola Pool (ed.), *Trends in Content Analysis* (pp. 33-88). University of Illinois Press, Urbana.
PANOFSKY E. (1955), *Meaning in the Visual Arts. Papers in and on Art History*, Doubleday, New York (tr. it. *Il significato delle arti visive*, Einaudi, Torino, 1996.
PEIRCE C.S. (1868), "Questions Concerning Certain Faculties Claimed for Man", *Journal of Speculative Philosophy*, 2: pp. 103-114, tr. it. in C.S. Peirce, *Le leggi dell'ipotesi Antologia dai Collected Papers* (a cura di M.A. Bonfantini, R. Grazia, G. P. Proni), Bompiani, Milano, 1984.
- (1984), *Le leggi dell'ipotesi. Antologia dai Collected Papers* (a cura di M.A. Bonfantini, R. Grazia, G. P. Proni), Bompiani, Milano.
- (1980), *Semiotica*, testi tratti da *Collected Papers*, Einaudi, Torino.
PERELMANN C. (1977), *L'empire rhétorique. Rhétorique et argumentation*, Vrin, Paris (tr. it. *Il dominio della retorica*, Einaudi, Torino, 1981).
POTTIER B. (1974), *Linguistique générale, théorie et description*, Klincksieck, Paris.
POTTIER B. (1992), *Sémantique générale*, P.U.F., Paris.
POPPER K.R. (1959), *The Logic of Scientific Discovery*, Hutchinson, London (tr. it. *La logica della scoperta scientifica*, Einaudi, Torino, 1970).
PRIETO J.L. (1966), *Messages et signaux*, P.U.F., Paris (tr. it. *Lineamenti di semiologia. Messaggi e segnali*, Laterza, Bari, 1971).
PROPP V.JA. (1928), *Morfologija skazski*, Academia, Lelingrad (tr. it. *Morfologia della fiaba*, Einaudi, Torino, 1966).
RASTIER F. (1987), *Sémantique interprétative*, P.U.F., Paris,.
- (1995), "La sémantique des thèmes ou le voyage sentimental", in F. Rastier (dir.), *L'analyse thématique des données textuelles,* Didier, Paris, 1995, pp. 223-249.
RICOEUR P. (1969), *Le conflict des interprétation*, Seuil, Paris (tr. it. *Il conflitto delle interpretazioni*, Jaca Book, Milano, 1972).
- (1970), "Qu'est-ce qu'un texte ? Expliquer et comprendre", in Bubner R. et alii, *Hermeneutik und Dialektik*, Mohr, Tübingen (tr. it. in P. Ricoeur, *Dal testo all'azione. Saggi di ermeneutica*, Jaca Book, Milano, 1983).
- (1975), *La métaphore vive*, Seuil, Paris.
- (1984), *Temps et récit II. La configuration dans le récit de fintion*, Seuil, Paris (tr. it. *Tempo e racconto 2. La configurazione del racconto di finzione*, Jaca Book, Milano, 1985).
- (1987), "Logica ermeneutica?", *AUT AUT*, 217-218: pp. 64-110.
RICOLFI L. (a cura di) (1997), *La ricerca qualitativa,* La Nuova Italia Scientifica, Roma.
RIZZI A. (1990), *Analisi dei dati. Applicazioni dell'informatica alla statistica,* La Nuova Italia Scientifica, Roma.
SALTON G. McGILL M.J. (1984), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.
SAPIR E. (1949), *Culture, Language and Personality*, University of California, Berkeley (tr. it. *Cultura, linguaggio e personalità*, Einaudi, Torino, 1972).

SAUSSURE (de) F.(1922), *Cours de Linguistique générale*, Payot, Lusanne-Paris (tr. it. *Corso di linguistica generale*, Laterza, Bari, 1967).
SCHMIDT S.J. (1973), *Texttheorie/Pragmalinguistik,* in Althaus H.P. et al. *Lexikon der Germanistischen Linguistik*, Verlag, Niemeyer, Tübingen (tr. it. *Teoria del testo e Pragmalinguistica*, in Conte M.E. (a cura di), *La linguistica testuale*, Feltrinelli, Milano, 1977).
- (1973), *Texttheorie. Probleme einer Linguistick der Sprachlichen Kommunikation*, München, Verlag (tr. it. *Teoria del testo*, il Mulino, Bologna, 1982).
SEARLE J.R. (1969), *Speech Acts. An Essay in the Philosophy of Language*, Cambridge University Press, London (tr. it. *Atti linguistici*, Boringhieri, Torino, 1976).
STRATI A. (1997), *La Grounded Theory,* in RICOLFI L. (a cura di) (1997) pp. 125-163.
TODOROV T. (1966), "Les catégories du récit littéraire" in *Communications 8* (tr. it. in Aa.Vv. *L'analisi del racconto*, Bompiani, Milano, 1969).
WATZLAWICK P., BEAVIN J.H., JACKSON D.D. (1967), *Pragmatic of Human Communication. A Study of Interactional Patterns, Pathologies, and Paradoxes*, W.W. Norton & Co, New York (tr. it. *Pragmatica della comunicazione umana. Studio dei modelli interattivi, delle patologie e dei paradossi*, Astrolabio-Ubaldini, Milano, 1971).
WINDELBAND W. (1894), *Geschhite und Naturwissenschaft*, Kaiser-Wilhelms-Universitat, Strasburg.
WHORF B. (1956), *Language, Thought and Reality*, Mass. MIT Press, Cambridge (tr. it. *Linguaggio, pensiero e realtà*, Boringhieri, Torino, 1970).
WITTGENSTEIN L. (1961) *Tractatus Logico-philosophicus*, Routledge and Kegan, London (tr. it. *Tractatus logico-philosophicus e Quaderni 1914-1916*, Einaudi, Torino, 1964*)*.
WRIGHT VON G.H (1971), *Explanation and Understanding*, Cornell University, Ithaca New York (tr. it. *Spiegazione e comprensione*, il Mulino, Bologna, 1977).

# Indice analitico (luoghi delle definizioni)