Analisi di un corpus di titoli di giornale : un confronto tra strategie

Stefano Tartaglia, Raffaella Gonella, Chiara Rollero

Dipartimento di Psicologia, Università degli Studi di Torino

Via Verdi 10, 10124 Torino - Italia

tartagli@psych.unito.it

Abstract

The current research has been carried out to compare different techniques for the text analysis. The corpus analyzed was made up of front page titles published by the main Italian daily newspapers during the month before the last regional elections. We followed two aims : the content exploration and the comparison between newspapers. For the exploration we carried out a content analysis and two lexical analyses : a descendent cluster analysis and a correspondence analysis. To compare newspapers we carried out three ascendant hierarchical cluster analyses based on content and lexical analysis data. Theoretical and empirical implications are discussed.

Riassunto

La presente ricerca è stata condotta allo scopo di confrontare diverse tecniche utilizzabili per l'analisi di un testo. Sul medesimo corpus – costituito dai titoli presenti sulle prime pagine dei più venduti quotidiani italiani nel mese precedente le ultime elezioni regionali – abbiamo perseguito due obiettivi : l'esplorazione dei contenuti e il confronto tra le testate. Per l'esplorazione dei contenuti sono state effettuate un'analisi di contenuto classica e due analisi lessicali : una classificazione gerarchica discendente ed un'analisi delle corrispondenze lessicali. Per il confronto tra le testate sono state effettuate tre classificazioni gerarchiche ascendenti basate sui dati dell'analisi di contenuto e di due differenti analisi testuali. Attraverso i risultati ottenuti vengono discussi gli apporti forniti dalle varie tecniche in relazione agli obiettivi d'analisi.

Keywords : content analysis, lexical analysis, comparison between techniques, press

1. Introduzione

L'obiettivo di questa ricerca è il confronto tra differenti strategie di indagine dei dati testuali mediante l'applicazione di esse su un medesimo corpus. Nello specifico, i testi analizzati sono titoli di giornale, sono quindi costituiti da enunciati molto semplici e brevi, da un lessico abbastanza limitato e riguardano argomenti molto eterogenei (dalla politica allo sport, dalla cronaca all'economia).

Consapevoli del fatto che differenti tecniche permettono di centrare diversi obiettivi di indagine (Brugidou e Labbé, 2000 ; Brugidou et al., 2004 ; Desmarais e Moscarola, 2004) abbiamo però cercato di confrontare le potenzialità, i vantaggi e gli svantaggi di alcune strategie nel perseguire due obiettivi in particolare : l'esplorazione dei contenuti trattati nel corpus ; il confronto tra le testate e la loro classificazione.

2. La ricerca

Abbiamo scelto di analizzare il contenuto delle prime pagine di dieci testate a diffusione nazionale nel periodo compreso tra il 3 marzo e il 2 aprile 2005, ovvero il mese immediatamente precedente le ultime elezioni regionali italiane. Si è deciso di prendere in

considerazione gli otto quotidiani più venduti in Italia (escludendo quelli sportivi ed Il Sole 24 Ore, che è un quotidiano economico), ossia il Corriere della Sera, La Repubblica, La Stampa, Il Messaggero, Il Giornale, Il Resto del Carlino, Avvenire e Il Mattino, a cui sono stati aggiunti i due quotidiani politicamente schierati più venduti, uno di sinistra ed uno di destra : L'Unità e Libero.

Il corpus testuale analizzato è costituito dai titoli (compresi occhielli e sottotitoli) dei 3149 articoli apparsi sulle prime pagine dei quotidiani succitati. Su questo corpus sono state effettuate differenti analisi in relazione ai due obiettivi posti. A scopo esplorativo abbiamo condotto una analisi di contenuto e due analisi lessicali : una classificazione gerarchica discendente basata sulla cooccorrenza delle parole all'interno dei titoli (metodo Alceste, Reinert, 1986) e una analisi delle corrispondenze lessicali sulla matrice titoli per parole (Lebart e Salem, 1988). Per classificare le testate, invece, abbiamo effettuato tre classificazioni ascendenti gerarchiche : una basata sulle categorie dell'analisi di contenuto, una sulle parole caratteristiche di ciascun quotidiano (Alceste) ed una sulle coordinate fattoriali ottenute mediante analisi delle corrispondenze della matrice parole per testate (Lebart e Salem, 1988).

3. Esplorazione del testo

3.1. Analisi di contenuto

I titoli dei quotidiani sono stati classificati sulla base delle seguenti variabili : la testata, il tipo di articolo, l'argomento trattato ed il riferimento a personaggi politici o partiti di destra o di sinistra.

Come tipologia di articoli abbiamo utilizzato una classificazione a quattro modalità : *apertura* (9,1% del totale dei titoli), corrispondente alla notizia cui graficamente viene dato maggior risalto nella prima pagina ; *taglio* (26,4%), articoli che descrivono fatti ; *editoriale* (22,9%), articoli di commento spesso di opinionisti o esperti ; *rimandi* (41,5%), classe in cui abbiamo inserito tutti i titoli a cui non è associato un articolo sulla prima pagina ma che rimandano a servizi nelle pagine interne. L'argomento degli articoli è stato invece classificato in 8 categorie : *internazionali* (13%), *politica italiana* (27,1%), *politica internazionale* (5,7%), *cronaca interna* (11,3%), *cronaca estera* (5,2%), *economia* (7,5%), *sport* (4,4%) e *altro* (25,8%). Sono state inserite nelle categorie di cronaca le notizie di cronaca nera e giudiziaria mentre nella categoria internazionali sono state inserite le notizie estere varie non di cronaca né di politica (ad es. guerre e catastrofi naturali). Oltre all'argomento, all'interno dei titoli sono stati individuati anche alcuni eventi particolarmente rilevanti nel periodo analizzato. I cinque più trattati sono stati : la vicenda *Sgrena/Calipari* (275 titoli ; 8,7% del totale), la malattia del *Papa* (171 ; 5,4%), le *elezioni* regionali (136 ; 4,3%), la questione relativa alla lista *Mussolini* (110 ; 3,5%) e il caso *Terri Schiavo* (103 ; 3,3%)[1].

[1] Gli altri eventi considerati sono: gli avvenimenti in Iraq (*Iraq* - 2,7% del totale), il dibattito politico sulle modifiche alla Costituzione italiana (*Costituzione* - 1,9%), il patto di stabilità dell'Unione Europea (*Patto* - 1,8%), il referendum sulla fecondazione assistita (*referendum* - 1,6%), gli attentati di Unabomber (*Unabomber* - 1,5%), il dibattito sull'introduzione di dazi per le importazioni di merci (*dazi* - 1,2%), il terremoto in Indonesia (*sisma* - 1,1%), la questione dei rapporti tra Libano e Siria (*Libano* - 1%), il maltempo (*maltempo* - 0,8%), le azioni di matrice anarchica (*anarchici* - 0,7%), il processo ad esponenti delle nuove Brigate Rosse (*BR* - 0,5%), la candidatura del magistrato Casson alle elezioni comunali di Venezia (*Casson* - 0,5%), il caro petrolio (*petrolio* - 0,5%), il dibattito sul risarcimento sui decessi legati al fumo di sigaretta (*fumo* - 0,3%) e il congresso del Partito della Rifondazione Comunista (*Prc* - 0,3%).

La classificazione di alcuni titoli come temi particolari ci permetterà più avanti di verificare la bontà dell'esplorazione del testo effettuata da parte delle tecniche computerizzate. Il riferimento a politici o schieramenti appartenenti al Polo (coalizione di centrodestra) o all'Unione (coalizione di centrosinistra) nei titoli analizzati è stato riportato in due variabili a tre modalità per distinguere tra riferimenti positivi (titoli in cui chiaramente agli esponenti di una delle due coalizioni viene associata una caratteristica positiva), riferimenti negativi (titoli in cui chiaramente viene evidenziata una debolezza o un difetto di una coalizione) e riferimenti neutri, ovvero tutti i titoli che non suggeriscono chiaramente una valutazione positiva o negativa. I titoli che fanno riferimento al Polo sono in totale 547 (17,4% del totale) di cui 37 con riferimenti positivi, 151 negativi e 359 neutri. Per quanto riguarda l'Unione i riferimenti sono 306 (9,7%) : 16 positivi, 93 negativi e 197 neutri. Il Polo risulta quindi citato in misura maggiore rispetto all'Unione, ma questo è in parte dovuto al fatto che, essendo la coalizione al governo, viene menzionato in quasi tutti i casi in cui si parla di questioni istituzionali. Per vedere le relazioni tra le testate prese in esame, gli argomenti trattati e lo spazio dedicato alle due principali coalizioni politiche italiane abbiamo operato un'analisi delle corrispondenze multiple (procedura Homals) che ci ha permesso di estrarre due componenti principali tramite cui costruire uno spazio bidimensionale su cui rappresentare le relazioni tra le modalità delle variabili *testata, argomento, riferimento al Polo* e *riferimento all'Unione* (vedi fig. 1)

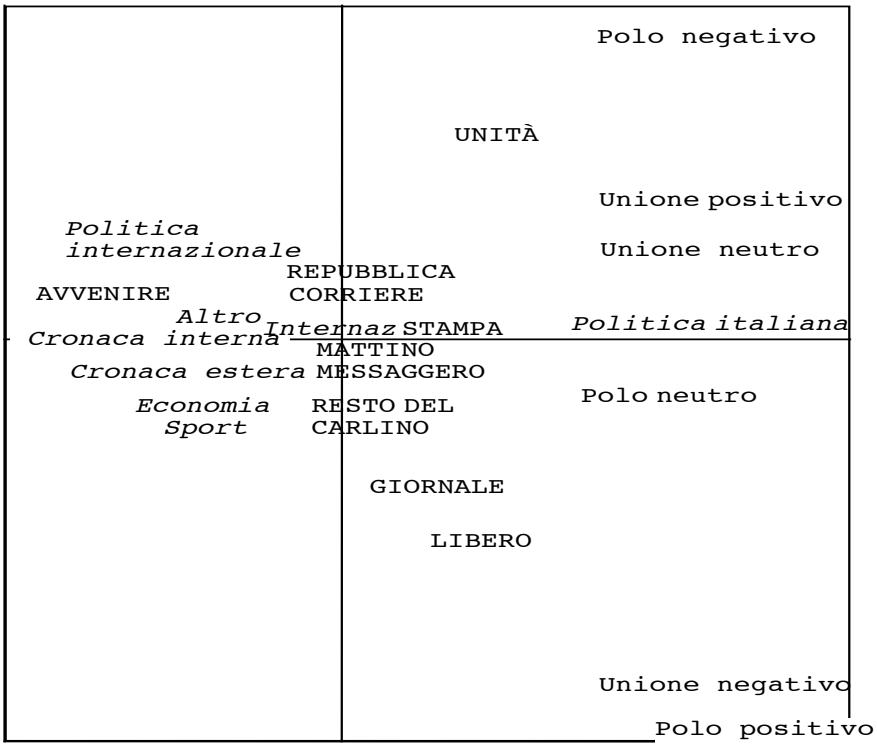

Figura 1 - Analisi delle corrispondenze multiple : proiezione sulle due componenti estratte delle variabili testata, argomento e riferimenti alle coalizioni.

La prima componente estratta ha un autovalore pari a .462 (8,4% di inerzia spiegata) mentre la seconda ha un autovalore di .329 (6%). Come si può vedere dal grafico sull'asse orizzontale (prima componente) si posizionano differentemente gli argomenti trattati nei titoli analizzati : all'estremo positivo dell'asse si trova la politica italiana mentre tutti gli altri

argomenti si collocano sul semiasse negativo. Possiamo quindi interpretare questa dimensione come spazio dato alla politica italiana. L'asse verticale invece discrimina chiaramente in base all'orientamento politico : sul semiasse positivo si trovano le modalità riferimento negativo al Polo e riferimento positivo e neutro all'Unione mentre sul semiasse negativo si trovano le modalità riferimento positivo e neutro al Polo e negativo all'Unione.

A riguardo della collocazione delle testate possiamo notare come L'Unità e Libero siano i giornali che danno più spazio alla politica e lo fanno fornendo valutazioni coerenti con l'orientamento politico dichiarato. Un secondo gruppo di testate, che corrisponde ai quotidiani a più alta tiratura, si colloca vicino all'origine dell'asse orizzontale ma comunque sul semiasse della politica italiana. Queste testate sono Il Giornale, chiaramente posizionato in direzione della valutazione positiva del Polo, il Corriere della Sera e La Stampa, in posizione sostanzialmente uguale e tendenzialmente di equilibrio tra le due coalizioni, e La Repubblica, moderatamente spostata verso il favore per l'Unione.

I rimanenti giornali sono invece quelli che meno trattano la politica italiana. Nel terzo quadrante si collocano Il Messaggero, Il Mattino (la proprietà di queste due testate è la stessa) ed Il Resto del Carlino e come argomenti l'economia, la cronaca, sia interna che estera, e lo sport. Nel quarto invece si colloca Avvenire, che in assoluto è la testata che dà meno spazio alla politica italiana, come argomenti troviamo invece i fatti internazionali, la politica internazionale e la categoria residuale altro.

3.2. Lemmatizzazione

Per le analisi basate sulle cooccorrenze verbali è stato necessario operare una lemmatizzazione che riducesse la variabilità del testo accorpando le forme flesse dello stesso lemma (effettuata con Taltac). Il corpus non lemmatizzato è risultato composto da 48788 occorrenze di 11642 forme grafiche differenti (indice di ricchezza lessicale 23.86), di queste 6955 sono Hapax (59,74%). La lemmatizzazione (riduzione dei plurali e dei femminili al singolare maschile e riduzione di tutte le forme verbali all'infinito) ha ridotto il corpus a 7869 forme distinte di cui 4188 Hapax (53,22%) ; si è quindi deciso, per ridurre ulteriormente la complessità di questo corpus, di utilizzare per le analisi solo le parole con almeno 15 occorrenze e di eliminare articoli e preposizioni (parole funzionali alla costruzione del discorso ma considerate semanticamente poco rilevanti) ottenendo il corpus analizzato composto da 357 forme distinte per un totale di 13336 occorrenze.

3.3. Classificazione gerarchica discendente

La classificazione dei titoli di giornale è stata effettuata mediante il software Alceste 4.6 che, a partire dalla matrice unità di contesto (nel nostro caso i titoli) per parole, raggruppa le unità di contesto in classi progressivamente più piccole fino a quando per la creazione di nuove classi deve superare un numero minimo di unità di contesto per classe (Reinert, 1995). Abbiamo deciso di richiedere che la classificazione si arrestasse al momento in cui per creare una nuova classe si fosse dovuti scendere al di sotto dei 315 titoli per classe, corrispondenti al 10% del totale dei titoli.

La classificazione ottenuta secondo questi criteri è risultata composta da quattro classi che raggruppano complessivamente 2135 titoli pari al 67,8% del totale. Per ciascuna classe sono state identificate alcune parole caratteristiche, che ne permettono l'interpretazione, e le modalità delle variabili illustrative (testata, tipo di articolo, argomento, evento trattato, riferimenti politici) tipiche dei titoli appartenenti alla classe.

Le quattro classi identificate corrispondono a quattro mondi lessicali (Reinert, 1997) a cui si possono ricondurre i differenti temi trattati nelle prime pagine dei giornali : una classificazione degli argomenti sovrapponibile a quella da noi utilizzata nell'analisi di contenuto, ma meno dettagliata.

La ***classe 1*** raggruppa 578 titoli (28% dei titoli classificati) e corrisponde agli argomenti di politica interna. Le parole caratteristiche di questa classe sono tutte relative a temi politici. Nella tabella 1 sono riportate le trenta parole con i valori di Chi-quadro più elevati. Si possono notare numerosi riferimenti alla vicenda dell'esclusione e della successiva riammissione di Alternativa Sociale alle elezioni nella regione Lazio (Mussolini, Storace, firma, falso, Lazio, TAR, ricorso), alle elezioni (regionale, voto, lista, elezione, regione) ed a protagonisti della vita politica (Prodi, Unione, DS, Bertinotti).

Parole	^{2}	Parole	^{2}	Parole	^{2}
Mussolini	248.39	referendum	106.40	Senato	50.88
Storace	184.18	elezione	103.63	authority	49.91
regionale	168.53	leader	82.69	Ruini	47.18
firma	162.82	TAR	76.43	dimettere	46.16
Lazio	157.07	Mediaset	71.91	DS	45.88
Prodi	122.23	costituzione	64.92	RAI	44.89
falso	113.36	Unione	62.01	Bertinotti	44.46
voto	110.35	regione	56.73	sociale	43.43
riforma	107.79	ricorso	55.39	escluso	43.43
lista	107.01	vincere	51.40	alternativo	41.73

Tabella 1 - Classificazione gerarchica discendente : parole caratteristiche della classe 1 (prime 30 in ordine di Chi-quadro)

Per quanto riguarda le variabili illustrative, in questa classe sono significativamente presenti i titoli di Libero (Chi-quadro = 13.88) del Corriere della Sera (7.63) e de L'Unità (3.11), i titoli di tipo apertura (3.37), quelli che trattano di politica interna (346.92) e tutti i titoli che fanno riferimento alle due maggiori coalizioni politiche (Polo positivo, 18.12 ; Polo negativo, 34.22 ; Polo neutro, 52.47 ; Unione positivo, 6.45 ; Unione negativo, 394.26 ; Unione neutro, 175.90). Vengono associati a questa classe gli eventi Mussolini (230.96), Elezioni (164.15), Referendum (72.69), Costituzione (63.00), Libano (35.61), Congresso Prc (11.77) e Casson (9.38). L'unico di questi eventi che non riguarda la politica interna è Libano, gli altri sono tutti collocati dalla procedura in modo coerente con l'analisi di contenuto.

La ***classe 2*** raggruppa 552 titoli (26% dei titoli classificati) e corrisponde agli argomenti di cronaca (vedi tab. 2). Tra le parole caratteristiche di questa classe ve ne sono molte tipiche della strutturazione giornalistica delle notizie di cronaca in generale (De Piccoli et al., 2003 ; Van Dijk, 1988) : la descrizione delle persone coinvolte in termini di età e ruolo sociale (figlio, bimbo, ragazzo, donna, giovane, madre) la precisazione dell'età degli attori coinvolti oppure dell'entità delle condanne quando si parla di cronaca giudiziaria (anno, tre, sei), la collocazione geografica esplicita (Napoli, Bologna), il riferimento alla vittima. Altre parole rimandano invece a eventi specifici occorsi nel periodo di tempo preso in esame come ad esempio un attentato di Unabomber, il sisma in Indonesia (tsunami, terremoto) ed il processo alle nuove BR (processo, Biagi, BR).

Parole	^{2}	Parole	^{2}	Parole	^{2}
Anno	167.38	sei	49.07	vittima	26.80
morto	104.92	Bologna	43.32	diventare	26.34
figlio	102.04	storia	41.83	BR	25.92
Napoli	75.19	terremoto	40.41	polemica	24.44
tsunami	63.75	processo	39.64	allarme	23.86
scuola	60.63	Biagi	38.77	donna	23.76
Scala	60.63	fare	34.43	giovane	23.17
bimbo	59.09	ragazzo	34.05	madre	22.90
Unabomber	54.84	Moratti	30.13	oltre	22.88
tre	54.04	piccolo	28.19	altro	22.10

Tabella 2 – Classificazione gerarchica discendente : parole caratteristiche della classe 2 (prime 30 in ordine di Chi-quadro)

In questa classe sono significativamente presenti i titoli di tipo rimando (11.98) ; quelli pubblicati su Avvenire (6.41) e su Il Resto del Carlino (5.35) ; quelli di cronaca italiana (140.18), cronaca estera (38.09), sport (3.58) e altro (99.13). Gli eventi associati questa classe sono Sisma (53.03), Unabomber (46.30), BR (43.32), Maltempo (14.38), Fumo (4.16) e fanno tutti riferimento a fatti di cronaca.

La ***classe 3*** è composta da 380 titoli (17% del totale dei titoli classificati) e rappresenta il tema dell'economia, come si può facilmente osservare dalle parole caratteristiche (vedi tab. 3).

Parole	^{2}	Parole	^{2}	Parole	^{2}
Patto	252.73	Bruxelles	112.10	sviluppo	69.77
banca	239.62	Fazio	97.95	contratto	67.75
UE	162.77	Siniscalco	88.54	aumenti	60.41
BNL	150.04	benzina	86.93	governo	60.08
stabilita	138.85	Cina	85.21	miliardo	58.91
competitivita	136.56	Bankitalia	82.24	offerta	58.91
dazi	135.78	accordo	80.46	record	58.09
statale	130.35	Irap	79.14	Europa	50.22
Opa	116.83	gas	74.45	risparmio	49.63
economia	113.09	Antonveneta	69.77	nuovo	42.80

Tabella 3 – Classificazione gerarchica discendente : parole caratteristiche della classe 3 (prime 30 in ordine di Chi-quadro)

In questa classe sono significativamente sovrarappresentati i titoli di tipo taglio (Chi-quadro = 2.81), quelli pubblicati da La Repubblica (3.62), quelli di argomento economico (446.56) e di politica internazionale (40.66). Gli eventi associati a questa classe sono tutti economici : Patto (206.25), Dazi (122.45) e Petrolio (39.91).

La ***classe 4***, 625 titoli (29% del totale dei titoli classificati), tratta la vicenda Sgrena/Calipari ed in generale la questione irachena. Sono maggiormente presenti in questa classe i titoli di tipo editoriale (chi-quadro = 9.24) e apertura (8.94), quelli de La Stampa (9.39), di argomento internazionale (292.65) ed altro (6.12). Sono associati coerentemente gli eventi Sgrena/Calipari (402.69) e Iraq (131.11) mentre viene associato in maniera errata l'evento Papa (10.34).

Parole	^{2}	Parole	^{2}	Parole	^{2}
Sgrena	188.57	eroe	56.17	riscatto	37.24
Calipari	186.86	verita	55.40	ambasciatore	36.50
USA	175.03	Giuliana	53.71	funerale	36.50
Iraq	166.16	sparare	51.94	auto	35.55
Bush	109.96	morte	44.57	annuncio	34.99
Ciampi	104.03	rapitore	43.86	servizio	34.84
americano	81.66	agente	43.86	uomo	34.82
Bagdad	81.18	Nicola	42.14	guerra	34.82
ritiro	66.39	soldato	38.95	rispondere	34.05
avere	60.56	giornalista	38.93	uccidere	34.04

Tabella 4 – Classificazione gerarchica discendente : parole caratteristiche della classe 4 (prime 30 in ordine di Chi-quadro)

3. 4. Analisi delle corrispondenze lessicali

Abbiamo effettuato l'analisi delle corrispondenze sulla matrice titoli per parole per mezzo del software Spad.t. Questa analisi permette di esplorare le relazioni tra le parole (le colonne della matrice) mediante l'estrazione di componenti latenti allo scopo di rappresentare graficamente le relazioni in uno spazio bidimensionale. Nello specifico sono state estratte due componenti : la prima ha autovalore .60 (.78% di inerzia spiegata), la seconda ha un autovalore pari a .56 (.74% di inerzia spiegata)[2]. Dato l'alto numero di parole inserite nell'analisi, l'interpretazione delle singole componenti risulta difficoltosa e poco chiara, mentre è molto più semplice e chiarificatrice l'interpretazione della proiezione delle parole sul grafico determinato dalle due componenti (vedi fig. 2).

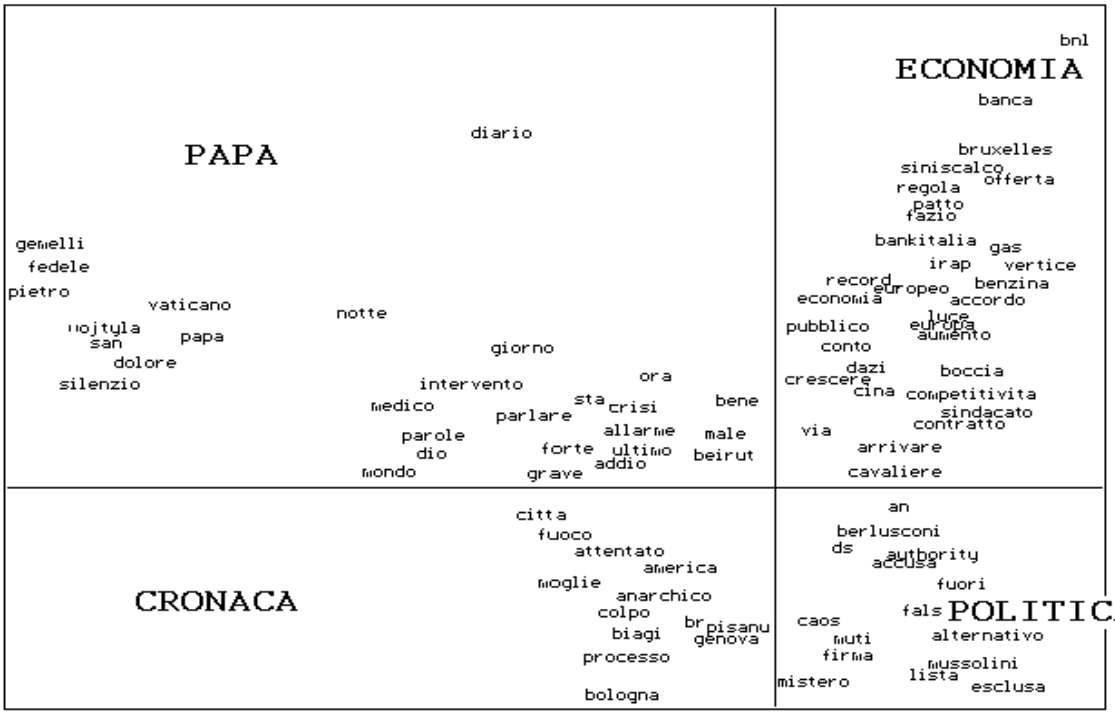

Figura 2 - Analisi delle corrispondenze lessicali : proiezione delle parole sulle prime due componenti estratte.

[2] In questo tipo di analisi è normale spiegare delle quote di inerzia molto basse, dipende dal fatto di operare su matrici con un elevato numero di colonne (nel nostro caso più di 350).

Le parole sovrapposte non sono state riportate nel grafico, tuttavia bisogna tenere presente che le parole più utili all'interpretazione sono quelle che si distanziano dalla nuvola di punti centrale (quelle che nel grafico compaiono). Si può notare come le parole dei quattro quadranti appartengono chiaramente a universi semantici differenti : nel primo quadrante (in alto a destra) troviamo i termini che definiscono il discorso ed i temi economici (es. banca, economia, aumento, conto, offerta, benzina), i nomi dei protagonisti dell'economia e della politica economica (Fazio, Siniscalco, Bankitalia, Bruxelles, sindacati) e i temi di discussione economica del marzo 2005 (dazi, Cina, competitività).

Nel secondo quadrante (in basso a destra) si trova invece la politica italiana e le sue polemiche : si vedano in particolare i riferimenti alla vicenda Mussolini (Mussolini, escluso, lista, firma, falso) ed ai protagonisti della politica nazionale (Berlusconi, Ds, An).

Il terzo quadrante (in basso a sinistra) è quello della cronaca, si vedano i riferimenti al processo alle nuove Br (processo, Biagi, Br, Bologna), agli attentati anarchici (anarchici, attentato, Genova) e la citazione del Ministro dell'Interno (Pisanu).

Il quarto quadrante (in alto a sinistra) è infine occupato principalmente dalle parole che descrivono l'agonia del Papa : Papa, Wojtila, San, Pietro, Vaticano, Gemelli (il policlinico), intervento, medico, fedele, etc…

Questa interpretazione è supportata anche dalla proiezione sugli assi delle variabili illustrative argomento ed eventi (vedi fig. 3).

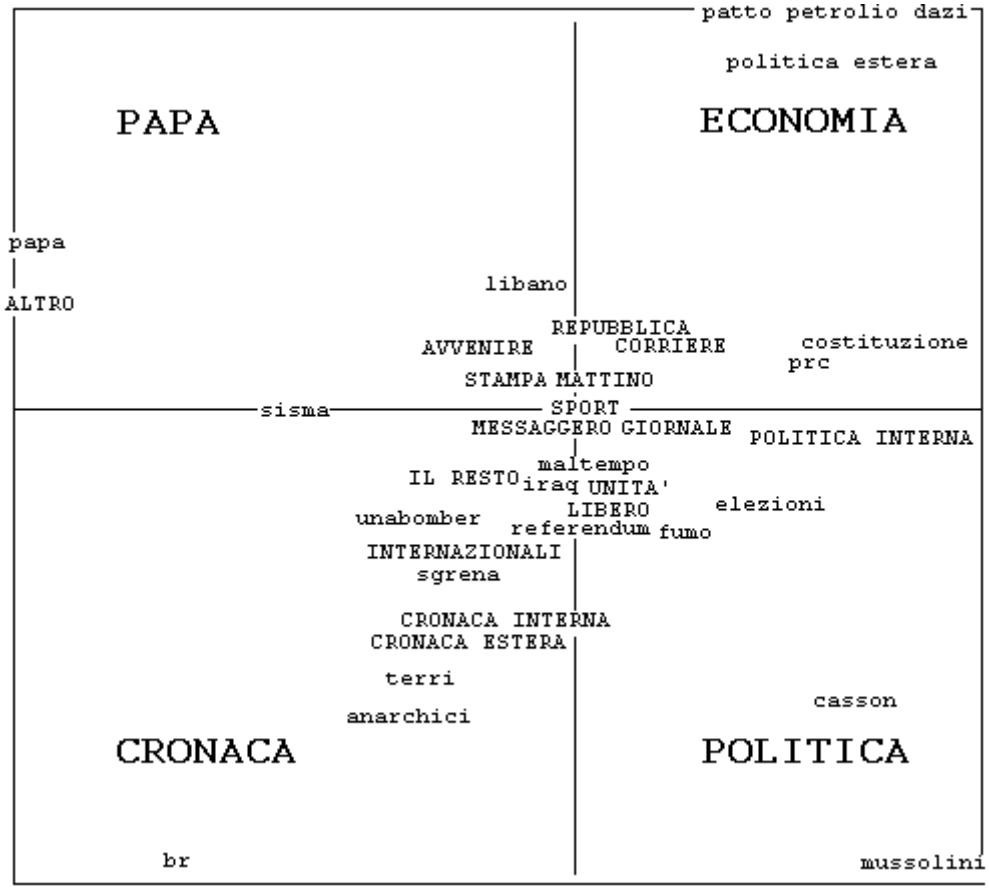

Figura 3 - Analisi delle corrispondenze lessicali : proiezione delle variabili illustrative testata, argomento ed eventi sulle componenti estratte.

Nel primo quadrante si trovano infatti gli argomenti economia e politica estera e gli eventi collegati *patto*, *petrolio* e *dazi* ; nel secondo, politica interna e gli eventi politici *elezioni*, *referendum*, *Mussolini* e *Casson* ; nel terzo, gli argomenti cronaca interna, cronaca estera e internazionali e gli eventi *Terri Schiavo*, *Unabomber*, *Br*, *anarchici*, *Sgrena/Calipari* e *sisma* ; nel quarto quadrante c'è la categoria argomento altro e l'evento *Papa*. Lo sport si colloca all'origine degli assi in posizione di equidistanza rispetto ai macro argomenti evidenziati dai quattro quadranti. Possiamo vedere come alcuni eventi non si collocano esattamente nella zona del grafico in cui ci saremmo aspettati di trovarli : gli eventi politici *costituzione* e *prc* sono vicini al quadrante della politica ma si trovano in quello dell'economia ; gli eventi di cronaca *fumo* e *maltempo* sono nel quadrante politico ; infine gli eventi *Libano* e *Iraq* non sono riconducibili a nessuno dei quattro universi semantici evidenziati dalla nostra interpretazione dell'analisi delle corrispondenze lessicali. Per quanto riguarda le testate, è interessante notare come le due testate politicamente schierate (L'Unità e Libero) risultano esattamente sovrapposte nel quadrante politico e molto vicine a quello della cronaca. Anche Il Giornale è nel quadrante politico mentre il Corriere della Sera e La Repubblica sono collocate nel quadrante economico. Il Resto del Carlino è l'unica testata collocata nel quadrante della cronaca mentre Avvenire, più distante dall'origine, e La Stampa si trovano nel quadrante in cui si trova il lessico legato alla malattia del papa. Infine, Il Mattino e Il Messaggero sono le due testate più prossime all'origine.

3.5. Confronto tra le tecniche esplorative

In generale possiamo affermare che le analisi testuali computerizzate utilizzate permettono di esplorare il contenuto del corpus testuale di titoli di giornale in maniera appropriata, anche se meno in profondità di quanto sia possibile fare con una analisi di contenuto condotta manualmente dai ricercatori. Sia la classificazione gerarchica discendente che l'analisi delle corrispondenze lessicali, infatti, identificano tre uguali mondi lessicali : quello della politica interna, quello dell'economia e quello della cronaca. Le categorie *argomento* ed *evento* dell'analisi di contenuto vengono associate abbastanza correttamente a questi mondi lessicali ricostruiti automaticamente. Potremmo quindi dire che le procedure computerizzate permettono una esplorazione meno dettagliata del contenuto ma decisamente più economica (in termini di tempo e risorse) su un corpus così vasto.

Oltre però ai tre mondi lessicali evidenziati da entrambe le tecniche di analisi testuale, la classificazione gerarchica discendente e l'analisi delle corrispondenze lessicali danno risultati differenti a riguardo di un quarto tema rilevante nei titoli analizzati. La classificazione operata con Alceste distingue all'interno del corpus una classe di titoli legati alla questione Sgrena/Calipari ed all'Iraq mentre l'analisi delle corrispondenze distingue il lessico dei titoli riguardanti la malattia del Papa. Entrambi questi mondi lessicali fanno riferimento, a differenza degli altri tre, a due eventi specifici e non a degli argomenti generali. L'analisi di contenuto ci permette di affermare che questi due eventi sono effettivamente i due più trattati dai giornali nel periodo preso in esame. Resta però di difficile interpretazione il motivo per cui, partendo da un'identica matrice, le due tecniche portino a conclusioni in parte differenti.

4. Classificazione dei testi

Anche per la classificazione delle testate sono state utilizzate tre tecniche differenti, una basata sulla analisi del contenuto e due sulle analisi testuali. Per la classificazione basata sull'analisi di contenuto è stata creata una nuova matrice con in riga le dieci testate ed in colonna le modalità delle variabili *argomento*, *evento*, *riferimento al Polo* e *riferimento*

all'Unione. Per avere delle misure confrontabili tra le testate, nelle celle di questa matrice non sono state inserite le frequenze ma le percentuali sul totale dei titoli della testata. Su questa matrice è stata effettuata una classificazione gerarchica ascendente per indagare le similitudini tra le varie testate. Questa strategia ci è sembrata la migliore per l'obiettivo che ci siamo posti e per il fatto che è la stessa utilizzata da Alceste e Spad.t (i software utilizzati per effettuare le altre due classificazioni). Consapevoli del fatto che il metodo di classificazione può influenzare fortemente il risultato della stessa (Aldenderfer e Blashfield, 1984 ; Bolasco, 1999) abbiamo optato, anche sulla base delle indicazioni di un anonimo revisore che per questo ringraziamo, per il metodo di Ward, basato sulla minimizzazione della varianza. Questo metodo *space-dilating* tende a creare più clusters di dimensioni simili piuttosto che unire i casi progressivamente ai grappoli principali. Il nostro interesse era di riunire le testate in raggruppamenti basati sui contenuti. Il dendrogramma ottenuto è riportato in figura 4.

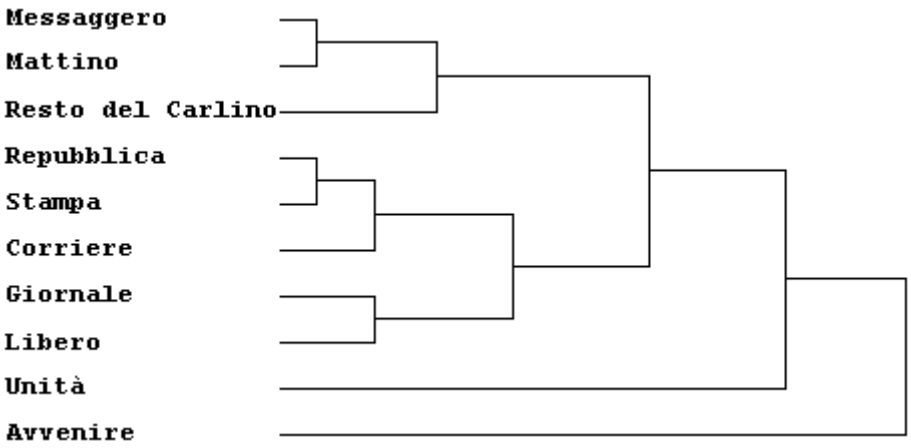

Figura 4 – Classificazione delle testate, dati analisi di contenuto (metodo di Ward)

Secondo questa classificazione, operata sulla base degli argomenti ed eventi trattati e sul tipo di riferimenti alle forze politiche, le testate più simili tra di loro sono Il Messaggero ed Il Mattino seguite dalla coppia Stampa e Repubblica. A queste prime due coppie si aggiungono rispettivamente Il Resto del Carlino alla prima ed Il Corriere della Sera alla seconda. Il Giornale e Libero costituiscono un terzo raggruppamento. Infine, appaiono maggiormente differenziati L'Unità e Avvenire che si uniscono all'albero principale nelle ultime aggregazioni.

La seconda classificazione è stata invece operata mediante Alceste sulla base delle specificità dei lessici delle dieci testate (quella che in Alceste è chiamata la procedura Tri-croisé). Il dendrogramma prodotto è riportato in figura 5.

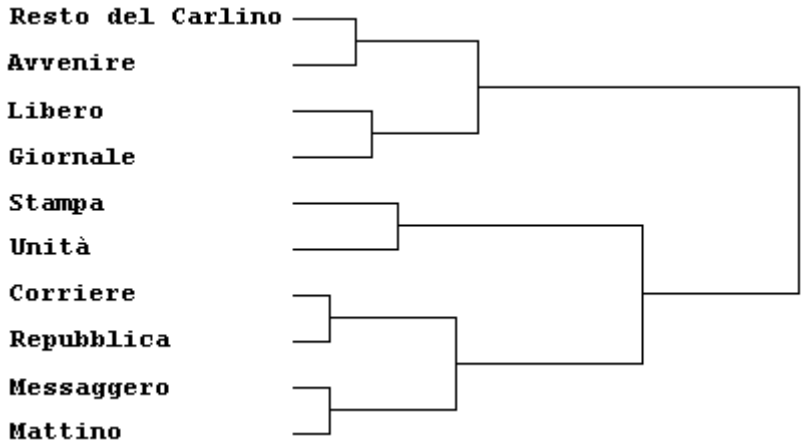

Figura 5 – Classificazione delle testate operata da Alceste (procedura Tri-croisé)

Anche in questo caso le testate che risultano più simili sono Il Messaggero ed Il Mattino. In generale però la classificazione dà risultati abbastanza differenti dalla precedente. In questo caso il dendrogramma è composto da due grappoli separati, in quello in alto sono riuniti i giornali di destra (Libero ed Il Giornale) e le due testate che danno meno spazio alla politica (Il Resto del Carlino e Avvenire) mentre in basso vi sono tutte le altre testate. Rispetto alla precedente classificazione in questo caso non rimangono testate isolate ma tutte (comprese L'Unità e Avvenire) vengono inserite in cluster nelle prime aggregazioni.

L'ultima classificazione è stata invece condotta operando una analisi delle corrispondenze lessicali sulla matrice parole per testate ; successivamente le testate sono state classificate utilizzando le coordinate fattoriali delle prime sei componenti estratte. Queste analisi sono state effettuate per mezzo di Spad.t. Il dendrogramma ottenuto è riportato nella figura 6.

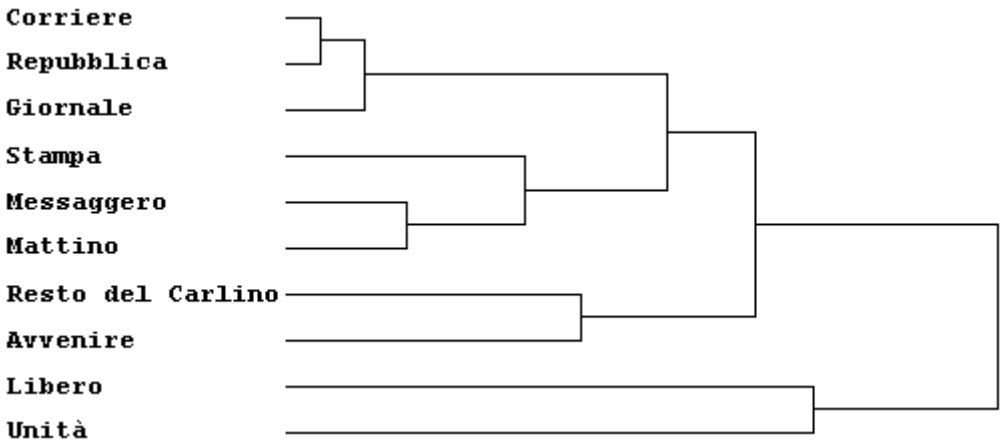

Figura 6 – Classificazione delle testate in base alle coordinate fattoriali ottenute mediante analisi delle corrispondenze lessicali sulla matrice parole per testata.

Anche in questo caso Il Corriere della Sera e La Repubblica sono associati strettamente ed a loro viene aggiunto Il Giornale. Come nelle precedenti due analisi anche Il Messaggero ed Il Mattino sono considerati affini. A questa coppia viene aggiunta successivamente La Stampa.

Le testate più differenti dall'insieme sono Il Resto del Carlino e Avvenire (raggruppate come nella precedente classificazione) e le due politiche (Libero e L'Unità).

Dal confronto dei grafici le tre classificazioni sono risultate convergenti solo in parte. Alcune coppie di quotidiani sono infatti associati in modo analogo dalle differenti strategie di analisi. Prendendo in considerazione livelli più generali di raggruppamento invece le strutture evidenziate differiscono maggiormente.

Secondo un criterio di interpretabilità dei risultati riteniamo preferibili le classificazioni basate sulle categorie dell'analisi di contenuto manuale e quella basata sulla analisi delle corrispondenze lessicali della matrice parole per testate.

5. Conclusioni

Sulla base del lavoro svolto pensiamo che per l'analisi di un vasto corpus testuale di titoli di giornale, o avente le stesse caratteristiche (enunciati brevi e argomenti trattati eterogenei), sia sicuramente indicato l'utilizzo di tecniche di analisi testuali automatizzate che danno buone garanzie di affidabilità e comportano un notevole risparmio di tempo e risorse. In particolare per quanto riguarda l'esplorazione del testo, la scelta tra una classificazione ed una analisi delle corrispondenze lessicali deve essere fatta sulla base del tipo di obiettivi che ci si è posti. L'analisi delle corrispondenze fornisce dei risultati meno facili da interpretare univocamente ma che hanno il vantaggio di permettere una lettura multidimensionale del testo. Al contrario la classificazione, forzando le parole e le unità di contesto all'interno dei clusters, permette di ottenere dei risultati più facilmente leggibili. In altre parole, l'analisi delle corrispondenze è più indicata per una esplorazione del corpus in vista di ulteriori analisi mentre la classificazione è funzionale ad una sintesi del testo.

Per la classificazione dei testi sulla base di una variabile illustrativa (nel nostro caso la testata) i nostri risultati indicano come più affidabile la classificazione effettuata sulla base delle coordinate fattoriali ottenute dall'analisi delle corrispondenze lessicali della matrice parole per modalità della variabili di selezione.

Bibliografia

Aldenderfer M.S. e Blashfield R.K. (1984). *Cluster analysis*. Sage.

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

Brugidou M. e Labbé D. (2000). Le vocabulaire syndical français à la lumière de l'analyse des données textuelles et de la statistique lexicale. *Actes des Journées Internationales d'Analyse Statistique des Données Textuelles 2000,* http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/38/38.pdf.

Brugidou M., Mandran N., Moine M. e Salomon A. (2004). Les apports de l'analyse textuelle pour l'analyse électorale : les questions ouvertes du panel électoral de 2002. *Actes des Journées Internationales d'Analyse Statistique des Données Textuelles 2004*, http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_019.pdf

De Piccoli N., Colombo M., Mosso C. e Tartaglia S. (2003). Stampa quotidiana e sentimento di insicurezza urbana. In B. Zani (a cura di), *Sentirsi in/sicuri in città*. Il Mulino.

Desmarais C. e Moscarola J. (2004). Analyse de contenu et analyse lexicale, le cas d'une étude en *management public*. Lexicometrica, *Actes du colloque "L'analyse de données textuelles : De l'enquête aux corpus littéraires"*. *http://www.cavi.univ-paris3.fr/lexicometrica/archives.html*.

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

Reiner M. (1986). Un logiciel d'analyse lexicale : Alceste. *Les Cahiers d'Analyse des Données*, 9 (4).

Reinert M. (1995). I mondi lessicali di un corpus di 304 racconti di incubi attraverso il metodo « Alceste ». In R. Cipriani e S. Bolasco (a cura di), *Ricerca qualitativa e computer. Teorie, metodi e applicazioni*. Franco Angeli.

Reinert M. (1997). Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique de divers corpus. *Lexicometrica*, 0. http://www.cavi.univ-paris3.fr/lexicometrica/archives.html.

Van Dijk T. A. (1988). *News as discourse*. Erlbaum.