Tesi di diploma in Statistica

Il lessico dei Newsgroup di argomento religioso: lo studio di quattro casi esemplari con applicazione dello Spad-T®

- SINTESI -

Università degli studi di Roma "La Sapienza" Facoltà di Scienze Statistiche, Diploma Triennale Relatore: Prof. Luca C. Giuliano	Anno Accademico 1999/2000 Seduta di Diploma del 30 maggio 2000 **Voto: 110/110**	**Diplomato:Alessandro Stabellini** Matricola N° 04030550 Residente in Via G. Maspero, 6 00119-Ostia Antica-Roma Tel.: 065650327 03474861893

INTRODUZIONE

Il lavoro svolto eredita le esperienze maturate in campo s*tatistico-testuale* dall'Analisi del Contenuto (AdC)[1] e propone una particolare metodologia di analisi applicabile alla realtà del fenomeno *Usenet*.

Usenet[2], nella realtà sociale, può essere pensata come una grandissima banca dati le cui informazioni riguardano la gente, i suoi pensieri, le sue tendenze, le sue aspirazioni ed il suo sapere. Coglierne il senso o poterne rilevare dimensioni e caratteristiche, con gli strumenti messi a disposizione dall'AdC, può aiutare ad interpretare alcuni fenomeni sociali. Non solo: l'AdC fornisce i mezzi per poter riassumere le informazioni del grosso "contenitore" *Usenet* al fine di renderle più semplici sia in termini di accessibilità che di lettura del loro significato.

I dati di *Usenet*, infatti, altro non sono che l'insieme dei numerosi messaggi che ogni giorno vengono spediti in un determinato gruppo di discussione o *Newsgroup*. Questi messaggi seguono uno standard di composizione ma il loro formato è quello testuale. Appare allora semplice la messa a punto di una strategia che sfrutti le tecniche dell'AdC per analizzare il testo e produrre dei risultati in funzione di uno specifico obiettivo. In realtà vi sono diversi problemi proprio correlati al tipo d'informazione a disposizione, che richiedono, non solo particolari interventi operativi, ma anche determinate scelte di tipo metodologico.

---

[1] L'AdC ovvero la "Analisi del Contenuto", è un metodo scientifico in grado di effettuare una spiegazione totale ed oggettiva dei dati di informazione. Analizzare il contenuto di un documento o di una comunicazione significa cercare le informazioni, spiegarne il senso e classificare tutto ciò che esse contengono.

L'AdC è l'incontro tra la psicologia applicata, la statistica e la linguistica e si inserisce nel quadro di una logica della comunicazione intersoggettiva, partendo dal presupposto che l'analisi logico-semantica della produzione verbale di un individuo deve rivelare le sue opinioni ed i suoi atteggiamenti.

Il territorio dell'AdC è vasto: esso confina da una parte con la linguistica, in cui l'AdC è usata per scoprire lo stile, il lessico, le figure retoriche utilizzate da persone o gruppi e dall'altro con l'ermeneutica che usa l'AdC per ricercare i significati impliciti, le connotazioni di una parola o di una serie di parole (temi narrativi, generi culturali). Al centro del territorio trova posto un approccio dell'AdC volto alla classificazione logica del contenuto di un testo per cercare di comprendere il suo significato manifesto. Da qui l'inventario e la categorizzazione delle parole chiave di un testo, il riassunto di un articolo o di un volume, la classificazione delle risposte di un questionario e così via.

[2] "*Usenet* è un muro gigante di annotazioni *Post-It*, un inconscio collettivo. Tramite *Usenet* milioni di utenti di computer di nazionalità diverse condividono informazioni, presentano domande e risposte e conducono discussioni."

NetNews, più comunemente chiamata *Usenet*, è un sistema condiviso di messaggi che provengono da tutto il mondo in un formato standard. In poche parole *Usenet* è una comunità mondiale di bacheche elettroniche (o *BBS*), interconnesse fra di loro e strettamente associate ad Internet, le cui informazioni sono costituite da singoli messaggi, ciascuno dei quali può essere letto e condiviso da tutti gli utenti. Le *BBS*, a differenza della posta elettronica, consentono, infatti, di organizzare le informazioni come *risorse comuni condivise*, in *directory* (cartelle) che non appartengono al singolo utente: i messaggi arrivano in aree di proprietà del software della bacheca elettronica cosicché i fruitori possono leggere contemporaneamente la stessa copia, come se si trattasse di un giornale comune o di un avviso affisso ad un muro, mediante programmi software progettati per organizzare e leggere i numerosi messaggi. In pratica, invece di inviare un messaggio ad una persona, lo si invia, o "affigge", in un gruppo della *BBS*.

Chiunque può scrivere ciò che vuole, aspettarsi delle risposte o leggere fra milioni di altri messaggi, suddivisi per argomenti. Già, perché la bacheca *Usenet* è organizzata in gruppi di discussione o *Newsgroups*, ovvero in aree, ciascuna con un proprio "manifesto" o tema su cui basare la stesura dei messaggi da inviare.

OBIETTIVI

Il presente lavoro propone un percorso di ricerca in grado di analizzare il testo dei messaggi dei *Neswgroup*, con l'obiettivo finale di evidenziare, attraverso tecniche statistiche, i contenuti dei discorsi di un gruppo di scriventi avventori di un *forum* di discussione.

In realtà lo studio si spinge più in là, nell'intento di gettare le basi per il raggiungimento di una meta più ardua: una "lettura automatizzata" del *senso* contenuto nei messaggi *Usenet*, il tutto con un ridotto dispendio di risorse e con tempi operativi brevi.

La messa a punto del metodo ha cercato di non contravvenire alle caratteristiche dei metodi di ricerca quali la *quantificazione, la sistematicità, l'esaustività e l'oggettività*.

La perdita d'informazione, pagata a fronte di una maggiore velocità esecutiva nelle procedure adottate, è stata ritenuta accettabile per lo scopo proposto e per i risultati ottenuti.

LA METODOLOGIA

La tesi di ricerca si basa sull'Analisi del Contenuto di alcuni messaggi inviati - in un determinato periodo temporale - in quattro *Newsgroup* di argomento religioso. L'unione di questi messaggi ha originato un file di testo che ha rappresentato il ***corpus***[3] dal quale ha tratto origine lo studio. In Tabella 1 sono riportati alcuni elementi utili per identificare i file di origine.

Tabella 1: Quadro degli elementi identificativi dei file oggetto di studio

Newsgroup	**Riferim. Temporale**	**Lunghezza File di testo originale**
it.cultura.cattolica *Server:* http://www.caspur.it/servizi/usenet/news/it.cultura.html news://news.caspur.it:it.cultura.cattolica/	17/09/99 al 05/10/99 **19 giorni**	2.356kB (2,29MB)
it.cultura.ebraica *Server:* http://www.caspur.it/servizi/usenet/news/it.cultura.html news://news.caspur.it:it.cultura.ebraica/	18/09/99 al 06/10/99 **19 giorni**	152kB
it.cultura.religioni Server: http://www.caspur.it/servizi/usenet/news/it.cultura.html news://news.caspur.it:it.cultura.religioni/	18/09/99 al 06/10/99 **19 giorni**	555kB
it.cultura.ateismo Server**:** http://www.caspur.it/servizi/usenet/news/it.cultura.html news://news.caspur.it:it.cultura.ateismo/	18/09/99 al 06/10/99 **19 giorni**	2.269kB (2,21MB)
	TOTALE	**5,20MB**

[3] L'insieme dei testi oggetto di analisi.

Il *Newsgroup it.cultura.cattolica* raccoglie messaggi dedicati alla tradizione, nonché alle esperienze che riguardano la dottrina cattolica; il gruppo *it.cultura.ebraica* è dedicato alla storia, alla cultura e alle tradizioni ebraiche; *it.cultura.religioni*, invece, riguarda le religioni e la religiosità in genere, mentre *it.cultura.ateismo* è dedicato agli atei[4].

Il primo elemento che salta subito all'occhio, leggendo la Tabella 1, è la quantità di dati a disposizione. Per dare un'idea della corrispondenza tra testo e byte, basti pensare che il file in oggetto, di 5,20MB, unione dei messaggi inviati nei rispettivi *Newsgroup* nel periodo di rilevazione, contiene un numero di parole che si aggira intorno alle 750.000 occorrenze[5].

L'estensione del file di dati, il periodo temporale in cui tali dati sono stati raccolti (di soli 19 giorni), i tempi e le risorse stimate per l'elaborazione hanno suggerito l'uso, per tale ricerca, di una metodologia diversa rispetto a quella adottata da altri studi effettuati nello stesso campo.

Alla base del metodo, infatti, vi è un processo di automatizzazione, diviso in due fasi, che ha consentito di snellire e velocizzare diverse operazioni. L'intento era quello di selezionare gli elementi pertinenti all'analisi cercando di non contravvenire a quelle che sono le caratteristiche fondamentali dei metodi di ricerca. E per far questo si è cercato di costruire uno strumento di analisi automatizzato, il più possibile indipendente dal *set* di dati usato per la sua *standardizzazione*, semplicemente ricercando le *regolarità* che sono alla base del fenomeno oggetto di studio.

Nella **prima fase** del processo automatizzato, infatti, sono state dapprima individuate, in alcuni documenti ufficiali reperibili in Internet, le regole che stabiliscono i formati, le caratteristiche e la struttura dei messaggi *Usenet* allo scopo di individuare quali elementi del messaggio dovevano essere tenuti e quali potevano essere scartati perché di scarso interesse per lo studio. Successivamente si sono sfruttati gli automatismi che alcuni *Newsreader-client* offrono all'utente, ovvero il salvataggio su supporto elettronico di determinate parti, ben precise e selezionabili, di un insieme di messaggi.

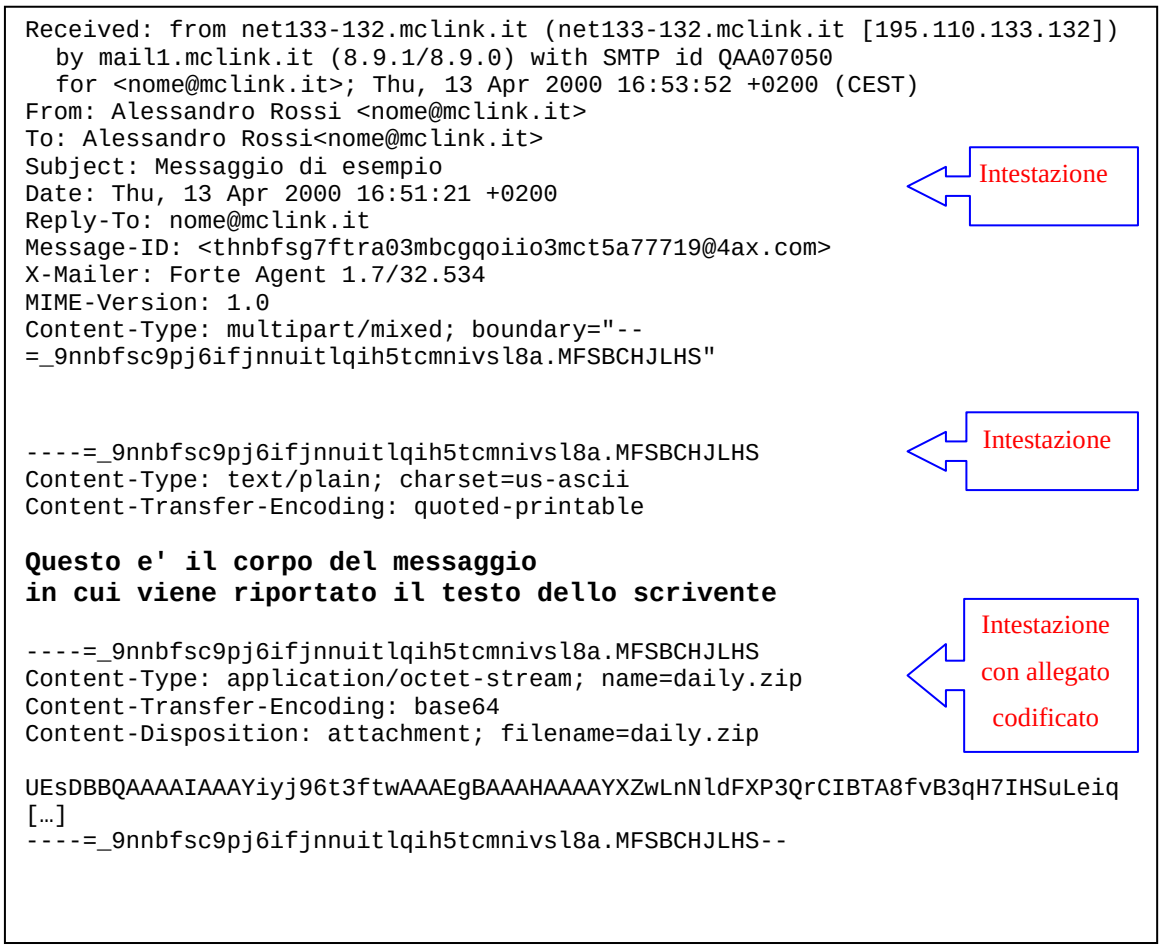
Ogni messaggio *Usenet* è costituito da due parti: una *intestazione* seguita da un *corpo*. L'*intestazione* contiene le istruzioni in formato testo che servono più che altro al *server* e al *newsreader-client* (programma per la stesura e lettura dei messaggi) per poter individuare instradare e formattare il messaggio, mentre il *corpo* è rappresentato dal contenuto vero e proprio del messaggio stesso.

[4] Nonostante la diversa area tematica dei *Newsgroups* in esame è frequente trovare persone, ad esempio di fede cattolica, che scrivono e conducono discussioni sia su *it.cultura.ebraica* che su *it.cultura.ateismo*, proponendo spunti di riflessione e punti di vista differenti.

[5] Un antecedente studio, simile a questo, ma con obiettivi differenti, utilizzava un file di circa 600KB contenente al suo interno poco più di 90.000 parole.

In Figura 1 si riporta un esempio di un messaggio *Usenet* con allegato.

Figura 1: tipico messaggio *Usenet*

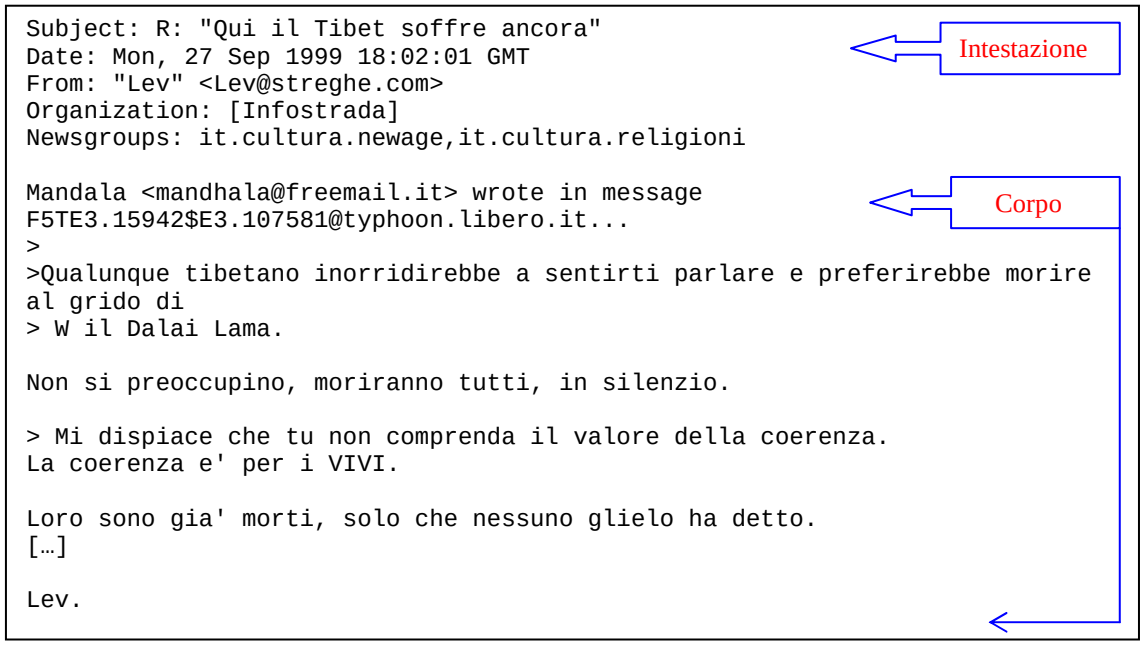


Scartare a priori le *intestazioni* e considerare il solo *corpo* come elemento di analisi poteva essere un criterio di selezione automatizzato, rapido, ma "pericoloso" per diversi aspetti.

Un primo aspetto riguardava la perdita d'informazione che si sarebbe ottenuta: l'intestazione contiene infatti, oltre allo "oggetto", dati circa la provenienza del messaggio e l'identità dello scrivente. Un secondo aspetto riguardava l'impossibilità di correggere alcuni errori di formattazione ed interpretazione del *newsreader-client*, o l'incapacità di porre rimedio all'uso incorretto del *newsreader* da parte dell'utente stesso.

Ecco come compare uno dei messaggi, tra quelli considerati per l'analisi, dopo il salvataggio automatizzato

Figura 2: messaggio *Usenet* dopo il salvataggio automatizzato



Si nota dalla **Figura 2** che oltre al *corpo* sono state mantenute alcune righe dell'*intestazione* che contengono informazioni circa l'identità dello scrivente, l'oggetto, la data e la provenienza del messaggio utili per eventuali "incroci" statistici di natura descrittiva.

La **seconda fase** del processo di automatizzazione è più complessa da un punto di vista operativo. Lo scopo è sempre quello di individuare gli elementi pertinenti all'analisi attraverso strumenti di selezione il più possibile precisi ed oggettivi con la difficoltà maggiore, in questo caso, dell'individuazione di regole ben precise alle quali riferirsi, così come era avvenuto nella prima fase dell'automatizzazione.

L'attenzione è rivolta esclusivamente al *corpo* del messaggio, rappresentato, di fatto, da uno scritto contenente un linguaggio molto vicino al parlato sia per la struttura del mezzo di comunicazione (*Internet*), che per la modalità intrinseca del tipo di colloquio tenuto con gli altri "avventori" (botta e risposta)[6]. Ciò fa sì che molte regole sintattico-grammaticali del linguaggio scritto vengano, nei messaggi *Usenet*, stravolte. A questo si aggiunge, non solo la difficoltà di cogliere il significato che un singolo emettitore dà ad un insieme di parole, ma anche l'impossibilità di attribuire un linguaggio ad una situazione/condizione particolare a tal punto da non poter essere poi generalizzato ad altri casi ad esso comparabili.

[6] Ne messaggi *Usenet* è sovente l'uso del *qouting* (dall'inglese *to quote*: citare). In pratica per commentare o rispondere ad un messaggio si citano i passi del messaggio contrassegnandoli in qualche modo per indicare che sono parole altrui e specificando la fonte originale. Questa funzione consente di inserire il messaggio desiderato, solitamente rientrato e con un carattere quale ">" oppure "|" sul margine sinistro di ciascuna riga, anteponendo una riga al messaggio stesso per indicare l'autore e la data d'affissione.

Tutto questo ha suggerito una linea operativa basata sostanzialmente sulla ricerca di regole che stabiliscono la composizione di quegli elementi, contenuti nel *corpo* del messaggio, che possano essere rintracciati in tutti i messaggi inviati nei *Newsgroup* e, allo stesso tempo, superflui ai fini dell'analisi. Tali elementi sono rappresentati dagli indirizzi di posta e di Internet, dalle firme digitali e da quant'altro identificabile da una ben precisa struttura di forma all'interno del file di testo.

Ovviamente la procedura di individuazione di eliminazione e/o sostituzione di questi elementi nei files non poteva avvenire manualmente: ciò avrebbe vanificato il tentativo di ridurre i tempi di elaborazione.

Per l'occasione è stato allora scritto, compilato e sviluppato un **software** che sotto forma di tool applicativo permettesse in automatico quanto appena descritto e che per caratteristiche e potenzialità potesse essere applicato a differenti *set* di dati. Non solo, gli automatismi del software dovevano essere in grado, quasi contemporaneamente, di effettuare un intervento di *normalizzazione* sul testo dei messaggi.

In Tabella 2 sono riportati, in termini di estensione in byte, gli effetti dell'automatizzazione sul file di dati originario.

Tabella 2: Effetti dell'automatizzazione sui file originali

Prima	File	Dopo
2,21MB	atei.txt	1,87MB
2,29MB	catto.txt	2,01MB
152KB	ebraica.txt	125KB
555KB	religio.txt	506KB
5,20MB	**Totale**	**4,50MB**

Una maggiore riduzione in termini di byte la si sarebbe ottenuta eliminando le parti del testo citate (vedere Nota 6) così come è stato effettuato in uno studio simile a questo, contenuto negli atti della quinta giornata internazionale d'Analisi statistica dei dati testuali tenutasi nel mese di marzo del 2000 a Losanna. Si è deciso, invece, di mantenere il testo ripetuto per l'impossibilità di stabilire una legge di composizione del fenomeno su cui basare gli algoritmi di intervento automatico sul testo.

Il passo successivo ha riguardato la formattazione e la **preparazione del testo** al fine di sottoporlo alle procedure dello Spad-T®. Applicativo di origine francese che sfrutta l'interfaccia DOS® per eseguire diverse analisi di natura statistico-testuale.

Importantissimo da riportare il fatto che il testo non è stato sottoposto ad alcun intervento di *lemmatizzazione* o di *disambiguazione* propri dei dettami dell'AdC, la cui applicazione avrebbe comportato la lettura analitica dell'intero documento e aumentato, così, in maniera considerevole i tempi di elaborazione. Gli unici interventi correttivi sul vocabolario - apportati con la procedura

CORTE dello Spad-T® - si sono basati su di un criterio che prevedeva l'eliminazione senza discriminazione dei pronomi, degli articoli, dei verbi ausiliari, di alcuni aggettivi e di alcune parole di uso comune in Internet (come E-mail, wrote ecc…), nonché il riporto a forma canonica solo di quelle forme che, "palesemente", non davano adito a significati ambigui (facciamo = fare; chiese = chiesa).

In Tabella 4 è riportato un quadro riassuntivo di tutte le procedure effettuate nello Spad-T®, nonché dei software usati per elaborare i dati e delle strategie adottate.

Nota: Nell'analisi delle corrispondenze semplici è stata adottata una matrice il cui esempio è riportato in Tabella 3.

Tabella 3: Matrice utilizzata dallo Spad-T® per l'analisi delle specificità e per l'Analisi delle corrispondenze semplici

Testo / **Forma**		ATEI	CATT.	RELIGIO	EBRAICA
1	DIO	82	35	40	145
2	SAN	56	77	19	70
3	PER	49	62	33	12
…		…	…	…	…
i	…	29	10	25	56
…		…	…	…	…
$V_{(s)}$	***CHIESA***	3	1	5	7

Tabella 4: Quadro riassuntivo delle procedure effettuate nello Spad-T®, trattamento dei dati per la generazione del grafico delle analisi delle corrisp.

Procedura NUMER: numerizzazione del testo

Soglia	Risposte	Occorrenze	N.ro par. dist.	Par. distinte	%COP
0	117.529	747.859	40.852	5,5%	100%
2	″	716.975	17.644	2,5%	96%
5	″	687.370	9.683	1,4%	92%
10	″	658.034	5.805	0,9%	88%
20	″	622.453	3.350	0,54%	83,2%
30	″	597.520	2.353	0,4%	80%
50	″	565.393	1.533	0,3%	75,6%

Procedura Corte: correzione del vocabolario

Eliminazione senza discriminazione di pronomi, articoli, verbi ausiliari e di alcuni aggettivi e parole di uso comune in Internet. Riporto a forma canonica solo delle forme non ambigue. **Criterio della massima velocità operativa.**

Procedura SETEX1: modifica dei parametri di lavoro

Innalzamento della **soglia** di frequenza a **50**, **esclusione** delle parole inferiori a **3 caratteri**.
Risultato:
- Occorrenze: 161.261
- Parole distinte: 764

(fonte Excel®)

Procedura MOCAR: analisi delle specificità, generazione delle prime **50** parole caratteristiche

Cattolici	Ebraica	Atei	Religioni
1. TRADUZIONE	TORAH	1. ATEO	1. MORTE
2. GESÙ	SUKKOT	2. CREDERE	2. CADUTA
3. TDG	EBRAICO	3. CREDENTE	3. EVA
4. MUTAZIONE	ISRAELE	4. RELIGIONE	4. MEDITAZIONE
5. GEOVA	EBREI	5. ATEISMO	5. MAESTRO
6. MARIA	LIBRO	6. ESISTERE	6. THOMAS
7. TNM	DIVINITÀ	7. SCIENTIFICO	7. ADAMO
8. SIGNORE	HOME	8. SCIENZA	8. PRINCIPIO
9. SCRITTURA	ITALIA	9. ESPERIENZA	9. ALBERO
10. EFISIO	MEMORIA	10. LOGICA	10. ATTENZIONE
11. SELEZIONE	CORSO	11. DIOS	11. GIORNO
12. SERMONTI	VECCHIO	12. UNIVERSO	12. VITA
13. PAOLO	GRAZIE	13. WILDE	13. NASCITA
14. SANTO	STORIA	14. SCHIANTO	14. SCOPO
15. CRISTO	UMANITÀ	15. CERVELLO	15. CREAZIONE
16. PREGHIERA	BIBLICO	16. LAMENTO	16. UOMO
17. SPECIE	GOVERNO	17. IPOTESI	17. FRUTTO
18. GRECO	RISPOSTA	18. MATEMATICA	18. SIRITUALE
19. […]	[…]	19. […]	19. […]

Procedura SEGME: ***costruzione dei*** **segmenti**

Calcolo delle frequenze dei differenti segmenti ripetuti e costruzione della tabella di contingenza che incrocia le risposte (in riga) ed i segmenti (in colonna).

Procedura MOCAR: generazione dei primi **50** *segmenti* caratteristici

Cattolici	Ebraica	Atei	Religioni
1. DIRE OGGI	1. SOLO PAROLA	1. PRE MORTE	1. BASE COMUNE
2. NUOVO MONDO	2. POTER RISPONDERE	2. UNO SCHIANTO	2. SAN GIOVANNI
3. BUONO FEDE	3. VENIRE DARE	3. LIBERO ARBITRIO	3. TRE GIORNO
4. GESÙ CRISTO	4. MAGGIORE PARTE	4. PECCATO ORIGINALE	4. SITO INTERNET
5. AMARE DARE	5. DUEMILA ANNO	5. CULTURA GENERALE	5. OTTOBRE 1999
6. TESTI GRECO	6. NESSUNO PARTE	6. DIO ESISTERE	6. CHIESA CRISTIANO
7. BUONO NOTIZIA	7. BUONO LETTURA	7. POTER VOLARE	7. UNO STATO
8. TRADUZIONE LETTERALE	8. DOVERE RISPONDERE	8. SPAZIO TEMPO	8. LIBERTÀ RELIGIOSO
9. TERZO MILLENNIO	9. OTTOBRE 1999	9. IPOTESI DIVINA	9. SAN PAOLO
10. UOMO NUOVO	10. DIRITTO UMANO	10. METODO SCIENTIFICO	10. CHIESA CATTOLICO
11. CRISTO GESÙ	11. SAN PIETRO	11. PADRE FIGLIO	11. GESÙ CRISTO
12. LINGUA ORIGINALE	12. FORSE DOVERE	12. LIBRO SACRO	12. CRISTO STORICO
13. GRAZIE GRAZIE	13. MONDO CATTOLICO	13. HOME PAGE	13. POTER VEDERE
14. TESTI BIBLICO	14. CONSCERE BENE	14. NORMALE PADRE	14. BUONO LETTURA
15. BIBBIA TDG	15. CAPIRE MEGLIO	15. AFFERMAZIONE DIO	15. UOMO DIO
16. FARE BENE	16. DUE ANNO	16. […]	16. GIOVANNI PAOLO
17. UNO SOLO	17. […]		17. […]
18. GESÙ DIRE			
19. […]			

Procedura APLUM: analisi delle corrispondenze semplici parole x testi

- Calcolo della massa
- Calcolo dei contributi assoluti e relativi
- Coordinate delle parole sui tre assi

Procedura APLUM: analisi delle corrispondenze semplici segmenti x testi

- Calcolo della massa
- Calcolo dei contributi assoluti e relativi
- Coordinate dei segmenti sui tre assi

Trattamento dell'Output dell'APLUM: uso di un particolare editor di testi (Textpad®), porting dei dati in Excel® e nell'SPSS® (parole)

1. Eliminazione parole con contributi assoluti ≤ 0,05 (Textpad®)
2. Ordinamento parole secondo contr.ass.asse_1, contr.ass.asse_2, contr.ass.asse_3 (Excel®)
3. Eliminazione parole dal basso contributo e dal significato troppo generico per la valutazione del"senso" (Excel®)
4. Selezione delle parole con una condizione filtro (contr.ass.asse_1≥0,4|contr.ass.asse_2≥0,4| contr.ass.asse_3≥0,4) (SPSS®)
5. **Rappresentazione grafica su (F_1, F_2)** delle parole di cui al punto 4. (SPSS)
6. Selezione delle parole con una condizione filtro (SPSS®)
   [(contr.ass.asse_1≥0,1| contr.ass.asse_2≥0,1| contr.ass.asse_3≥0,1) AND NOT (contr.ass.asse_1≥0,4|contr.ass.asse_2≥0,4| contr.ass.asse_3≥0,4)]
7. **Rappresentazione grafica su (F_1, F_2)** delle parole di cui al punto 6. (SPSS®)

Trattamento dell'Output dell'APLUM: uso di un particolare editor di testi (Textpad®), porting dei dati in Excel® e nell'SPSS® (*segmenti*)

1. Eliminazione *segmenti* con contributi assoluti ≤ 0,03 (Textpad®)
2. Ordinamento *segmenti* secondo contr.ass.asse_1, contr.ass.asse_2, contr.ass.asse_3 (Excel®)
3. Eliminazione *segmenti* dal basso contributo e dal significato troppo generico per la valutazione del"senso" (Excel®)
4. **Rappresentazione grafica su (F_1, F_2)** dei *segmenti* di cui al punto 3. (SPSS®)

Si noti dalla Tabella 4 la bassa percentuale di copertura del testo già a soglia zero (procedura NUMER). Ciò a denotare la connotazione del linguaggio adottato nei *Newsgroups*: molto vicino al parlato.

L'estensione del testo e quindi l'elevato numero di occorrenze ha comandato la scelta di una frequenza di soglia pari a 50 e l'esclusione delle parole inferiori a 3 caratteri. In queste condizioni si è ottenuto un numero di occorrenze pari a 161.261 unità con 764 parole distinte. Su questi dati è stato effettuato successivamente uno studio basato sull'analisi delle corrispondenze (semplici) tra parole e *Newsgroups* e tra *segmenti* e *Newsgroups*, ove per *segmento*[7] si intende una sequenza di parole contenuta in ciascuna riga del documento.

L'interpretazione dei risultati è scaturita da una ben precisa strategia d'intervento, diversificata nel caso delle parole e dei segmenti, mirata più che altro alla semplificazione della lettura dell'output dell'analisi. Gli elementi dei grafici delle due procedure erano, infatti, così numerosi da rendere molto difficile l'attribuzione di un significato al loro ordinamento.

Gli interventi sull'output hanno riguardato l'eliminazione delle parole e dei segmenti che meno hanno contribuito - da un punto di vista analitico - alla determinazione degli assi fattoriali e che possedevano un significato ritenuto troppo generico per la valutazione del "senso" (fare, andare, andare a fare, ecc…). Solamente per le parole, infine, è stata necessaria una condizione filtro che consentisse di ottenere un grafico con gli elementi più significativi.

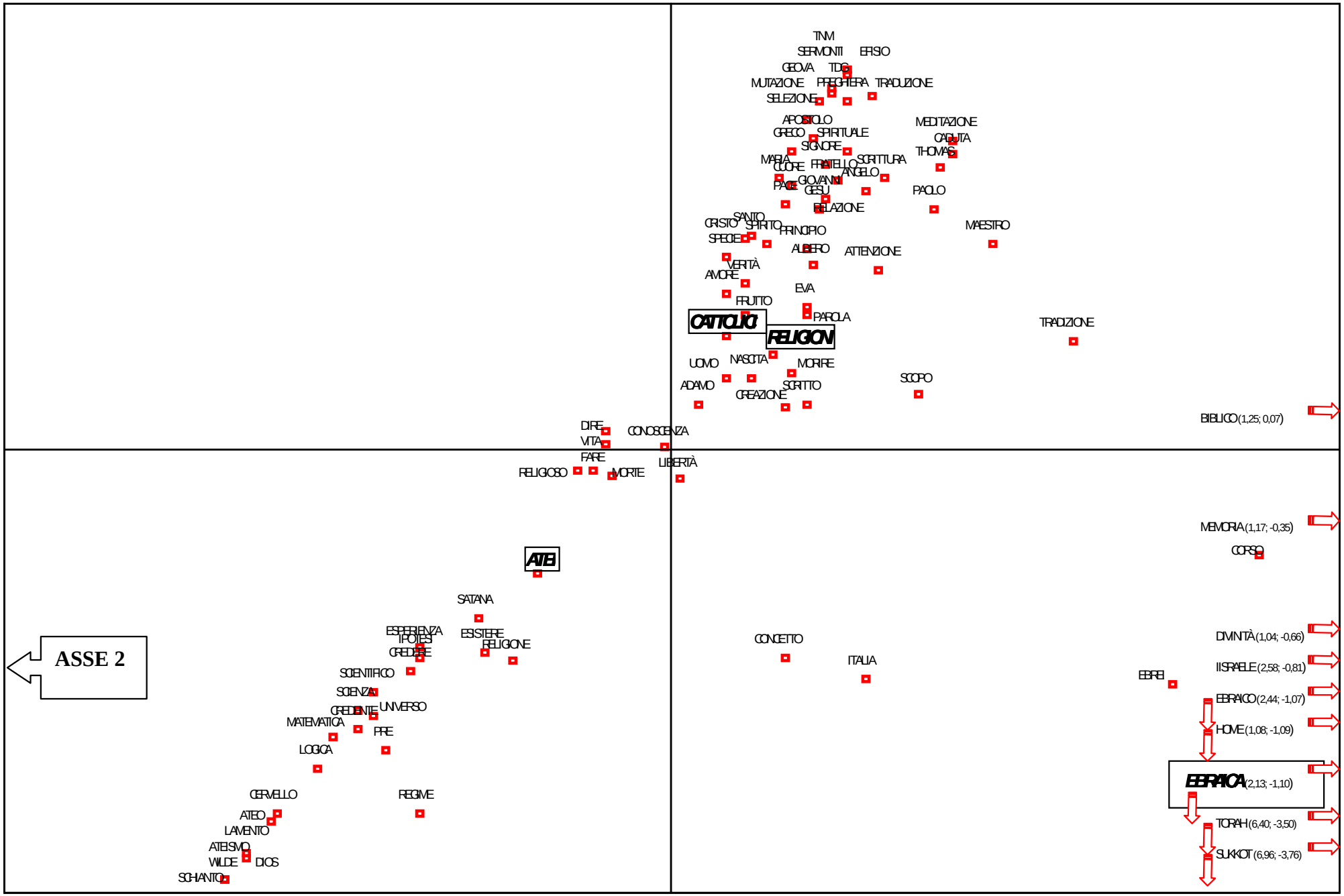
Il livello di interpretazione dei risultati si è basato sulla valutazione della differente disposizione (corrispondenza) sul piano fattoriale delle parole o dei segmenti rispetto ai *Newsgroups.* Una vicinanza tra una determinata parola ed un *Newsgroup* o tra un segmento ed un *Newsgroup* si traduceva, per costruzione, in una ricorrenza significativa della parola o del segmento per quel particolare *Newsgroup*. Al contrario una lontananza di un elemento da un *Newsgroup* corrispondeva ad una sua assenza significativa all'interno del *Newsgroup* stesso. Un'attrazione, invece, tra parole o tra segmenti sul piano fattoriale lasciava intendere una similarità nella struttura distributiva delle loro frequenze nei gruppi di discussione. E simmetricamente una opposizione degli elementi era dovuta ad una disuguaglianza dei corrispettivi profili distributivi.

Viene posta inoltre particolare attenzione agli elementi del piano fattoriale più distanti dall'origine, ovvero alle parole o ai segmenti la cui più bassa frequenza contribuisce a differenziare le relative distribuzioni.

Qui di seguito sono riportati alcuni **grafici** dai quali sono state avanzate le ipotesi circa il contenuto dei discorsi dei partecipanti ai *Newsgroups*.

[7] La generazione dei segmenti ripetuti avviene grazie ad una procedura dello Spad-T® che prende il nome di SEGME.

Grafico 1 Analisi delle corrispondenze lessicali: parole ad alto contributo assoluto



Dal **Grafico 1** si osserva che parole come "TNM" (Traduzione Nuovo Mondo Sacre scritture), "TDG" (Testimoni Di Geova), "GEOVA", addensate nel quadrante positivo, testimoniano una concentrazione di argomenti riguardanti le varie traduzioni della Bibbia. Sullo stesso quadrante, lo scienziato "SERMONTI" (teorico dell'evoluzionismo)[8] insieme alle parole "MUTAZIONE" e "SELEZIONE"[9] polarizzano, invece, l'attenzione sulle teorie Darwiniane dello sviluppo della vita sulla terra.

Volendo analizzare le *specificità*[10], ovvero le parole caratteristiche dei quattro *Newsgroup*, generate dalla procedura MOCAR dello Spad-T® ed esplicitate nel **Grafico 2** per i *Newsgroup* "CATTOLICI" e "RELIGIONI", si vede come le parole appena descritte siano tipiche del gruppo dei "CATTOLICI". D'altronde la vicinanza tra il gruppo "RELIGIONI" e sue *specifiche* parole come "UOMO", "NASCITA", "MORIRE", "FRUTTO", "AMORE", "ALBERO", "PRINCIPIO" lascerebbero presupporre la trattazione di argomenti legati alla figura biblica dell'albero della vita il cui frutto comunica l'immortalità .

Sul quadrante negativo del **Grafico 1**, invece, si nota come nel *Newsgroup* dedicato agli "ATEI", le tematiche riguardano la "MATEMATICA" la "LOGICA" la "ESPERIENZA" e la "SCIENZA". Parole queste, che si trovano molto vicine al termine "RELIGIONE" che risulta essere particolarmente caratteristico del gruppo degli "ATEI". Nel gruppo "EBRAICA" (quarto quadrante del **Grafico 1**) si parla, invece, di "TORAH" della festa delle capanne (SUKKOT)[11] nonché di "ISRAELE".

[8] Sermonti Giuseppe, da non confondere con Vittorio (scrittore e traduttore), docente universitario e biologo di fama internazionale, è stato autore di ricerche all'avanguardia nel campo della genetica dei microrganismi. Ha scritto importanti testi scientifici tra cui ricordiamo *Genetics of antibiotics producing microorganisms* (Wiley & Sons) e *Genetica generale* (Boringhieri). È autore anche di numerosi libri e saggi di riflessione critica sulla scienza moderna in generale o su alcuni aspetti particolari. Ricordiamo: *La mela di Adamo e la mela di Newton* (Rusconi, 1974), *Dopo Darwin* (Rusconi, 1980), *Le forme della vita* (Armando, 1981*)* e il recentissimo *Dimenticare Darwin* (Rusconi, 1999).

[9] Le parole TNM, TDG, GEOVA, MUTAZIONE, SERMONTI sono quelle che rispetto all'ASSE 2 sono più distanti dall'origine.

[10] Una parola è tanto più specifica per un *Newsgroup* quanto più la sua frequenza totale nel *corpus* è assorbita dalla sua stessa frequenza all'interno del *Newsgroup*. Alla base del calcolo delle specificità vi è un articolato test di natura statistica.

[11] Da "Sukkah" o "Succot" località ad est del Giordano dove Giacobbe costruì una casa circondata da capanne per il bestiame dalle quali questo luogo prese il nome (Cfr. Bibbia Gen 33, 17; Gdc 8, 4-5). È un termine che indica anche una località dove si accamparono gli Ebrei dopo aver lasciato Ramesse (Cfr. Bibbia Es 12, 37; Nm 33, 5-6). Durante la festa dei Tabernacoli o delle Capanne (che nell'anno 1999 si festeggia il 25 settembre) gli Ebrei dovevano risiedere in una capanna per una settimana sia per festeggiare la vendemmia, sia per ricordare a se stessi di essere stati un popolo nomade sotto la guida di Dio (Cfr. Bibbia Lv 23, 39-43; Ne 8,14).

Grafico 2 Analisi delle corrispondenze lessicali: parole ad alto contributo assoluto, quadrante positivo e specificità

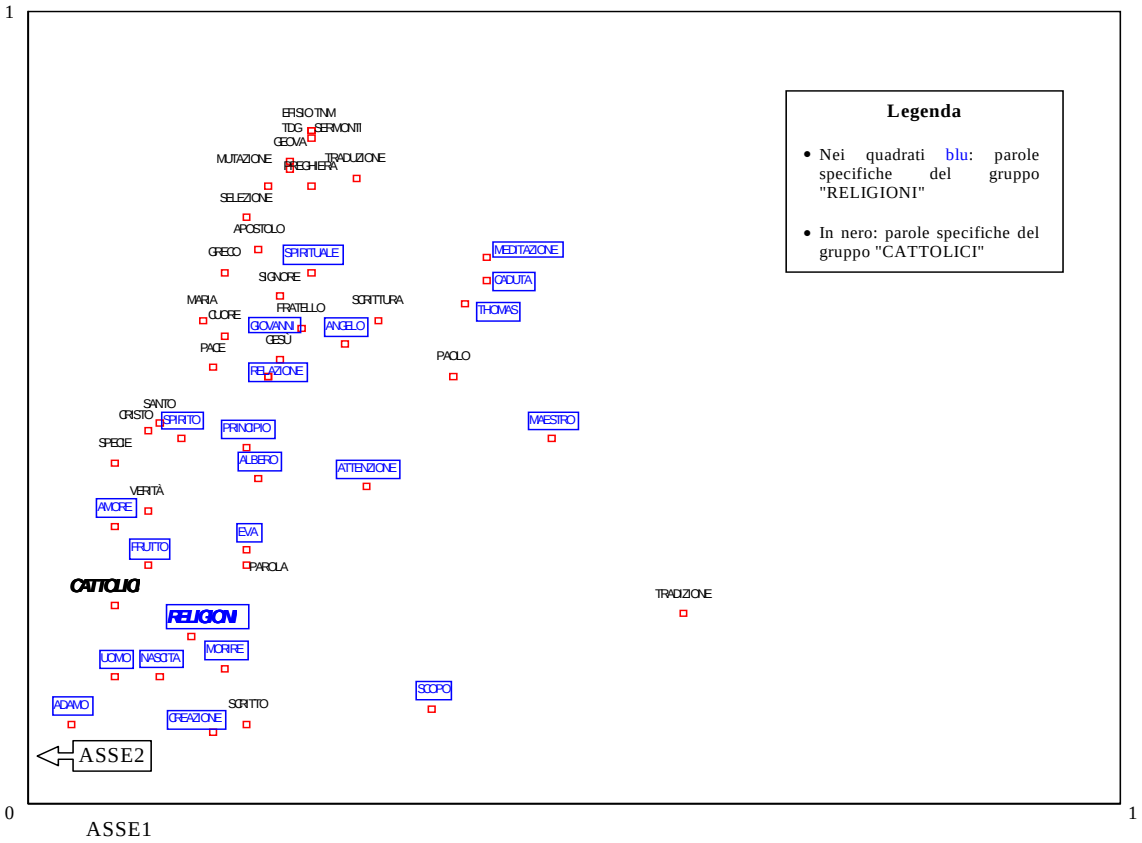
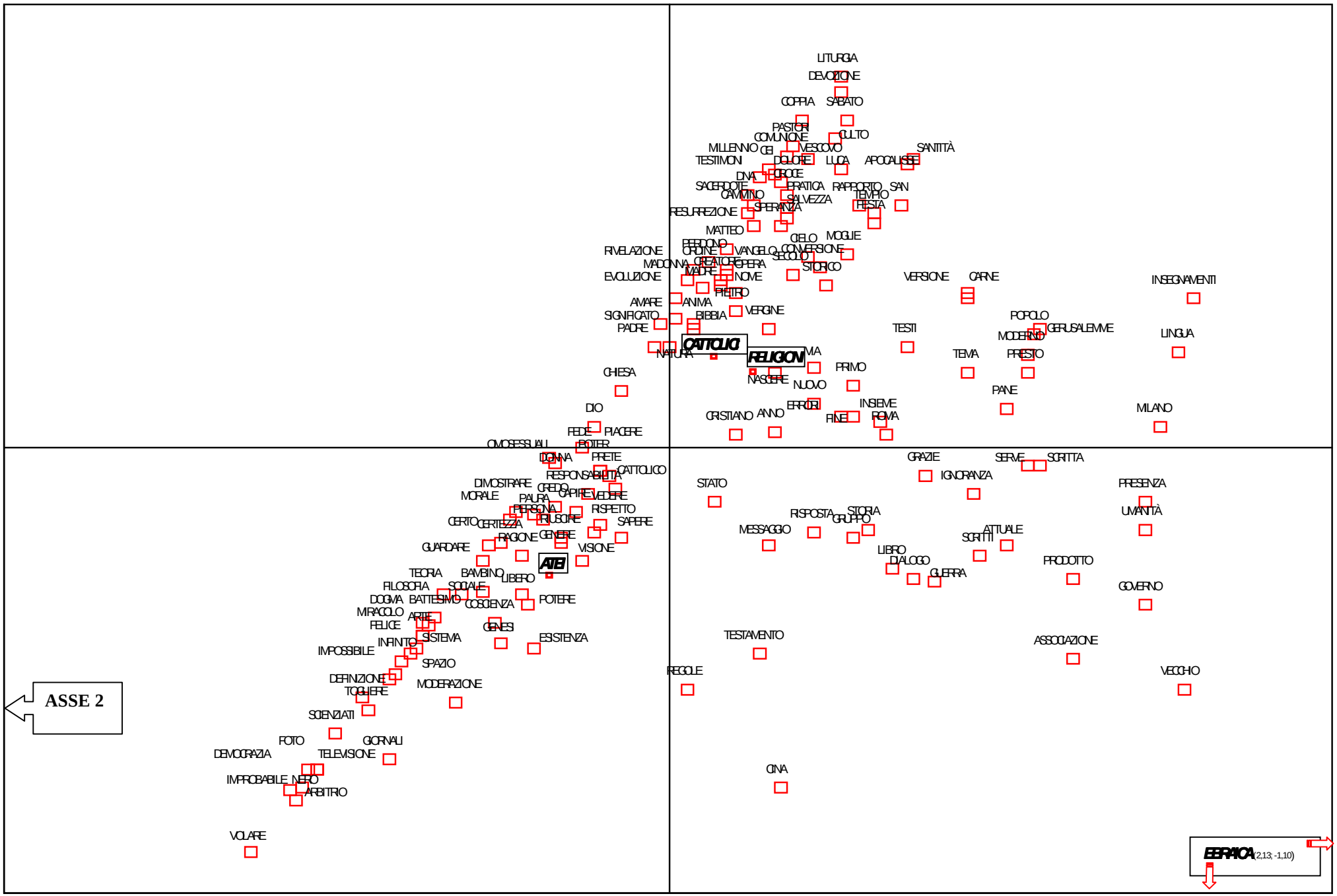
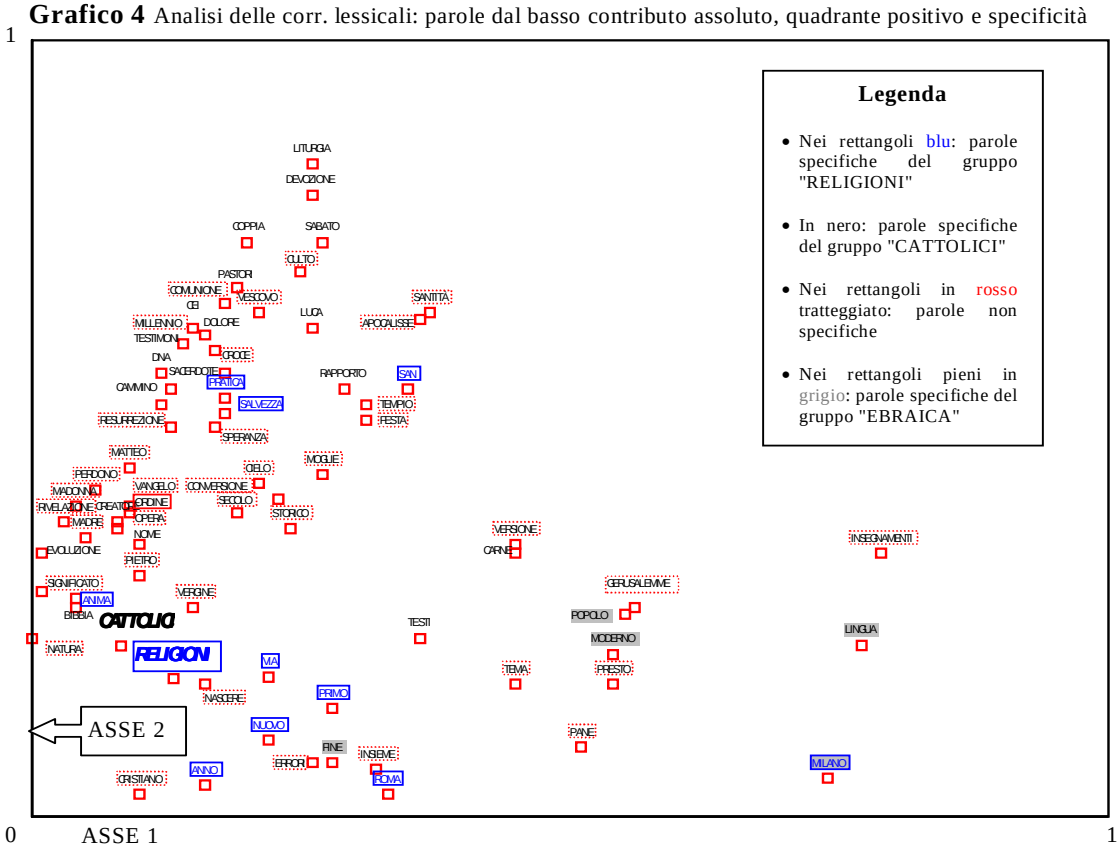


Grafico 3 Analisi delle corrispondenze lessicali: parole a basso contributo assoluto



Se si osserva il **Grafico 3** e la relativa discriminazione tra gruppi e specificità del quadrante positivo (**Grafico 4**), si noterà che anche le parole "CEI" (una edizione della Bibbia), "DNA" e "TESTIMONI" oltre ad essere specifiche del gruppo dei "CATTOLICI", confermano le teorie su di una trattazione in esso di argomenti inerenti alla Bibbia dei Testimoni Di Geova e allo sviluppo della vita. Per quanto , invece, riguarda il gruppo degli "ATEI" è interessante notare la vicinanza di termini quali "FILOSOFIA", "DOGMA", "BATTESIMO", "BAMBINO", "LIBERO" che farebbero pensare a discorsi sul valore dei sacramenti o addirittura sul senso di talune scelte in campo religioso. Sullo stesso quadrante le parole "GIORNALI", "TELEVISIONE", "DEMOCRAZIA" vicine l'una all'altra testimonierebbero, sempre nel gruppo degli "ATEI", alcuni argomenti di attualità come base di alcuni discorsi.

Grafico 4 Analisi delle corr. lessicali: parole dal basso contributo assoluto, quadrante positivo e specificità



Se si passa al **Grafico 5**, dal quale si legge l'ordinamento lungo gli assi fattoriali dei *segmenti* caratteristici, non solo si trovano alcune conferme delle ipotesi appena formulate, ma si riescono ad aggiungere ulteriori elementi di valutazione del fenomeno. Nel quadrante negativo, infatti si trovano *segmenti* quali "BIBBIA TDG", "NUOVO MONDO", "TRADUZIONE LETTERALE", "TESTI GRECO", "SELEZIONE NATURALE" ma anche "CHIEDERE SCUSA" e "GIOVANNI PAOLO" la cui specificità , appartenendo al gruppo "EBRAICA", lascerebbe

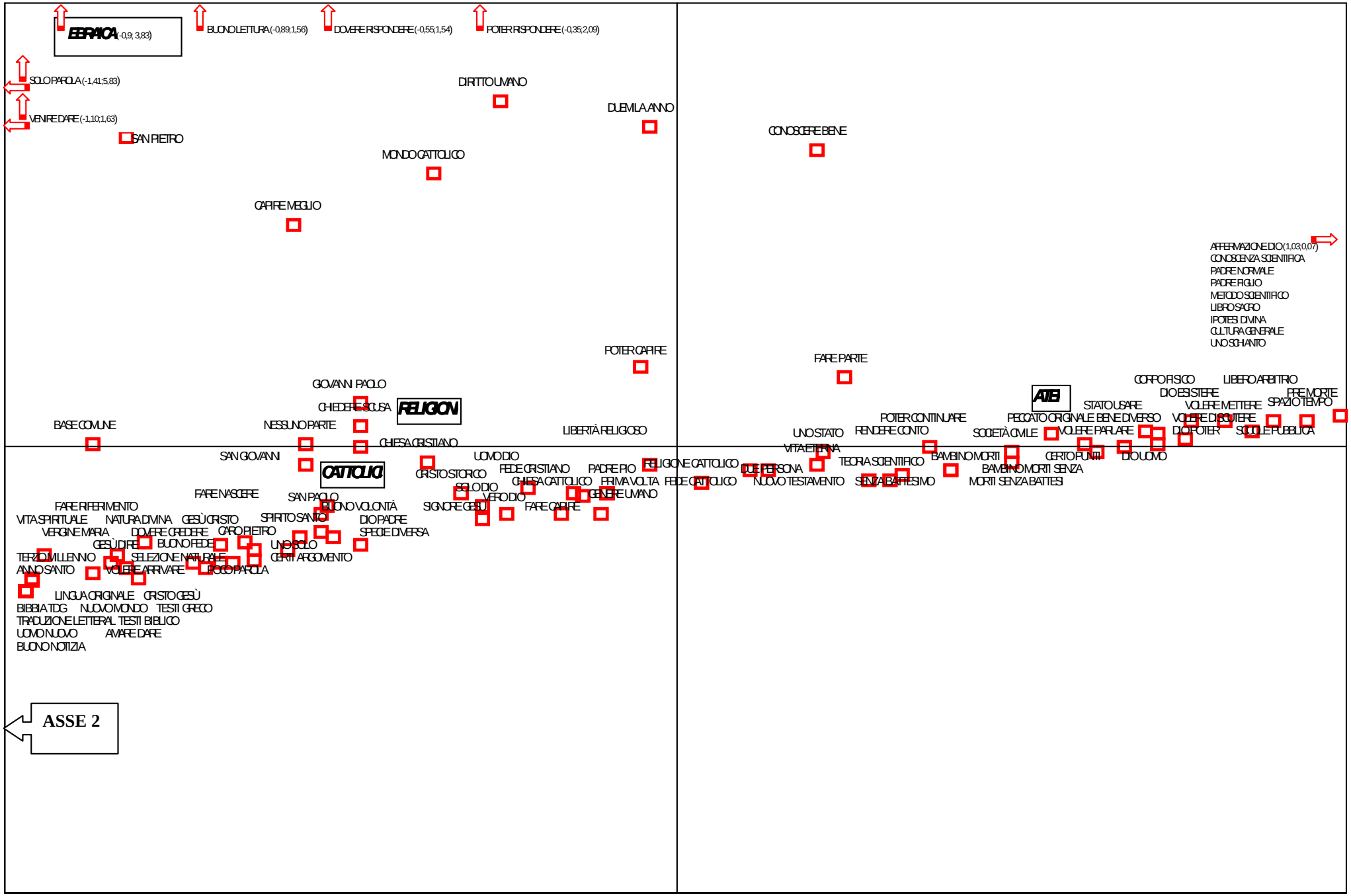
intendere il viaggio del Papa in Israele[12] e le sue intenzioni diplomatiche come temi di alcuni discorsi dei parlanti[13].

Nel secondo/quarto quadrante, sempre del **Grafico 5**, i *segmenti* "PECCATO ORIGINALE" e "BAMBINI MORTI SENZA BATTESIMO", addensati attorno al gruppo "ATEI", confermano anch'essi quanto detto in relazione al **Grafico 3**, oltre alla teoria, derivante dalla lettura del **Grafico 1**, che riguarda le tematiche scientifiche di alcuni discorsi.

[12] Viaggio che si è tenuto nei primi mesi del 2000. Periodo, questo, successivo alla data di rilevazione dei messaggi sui *Newsgroups.*

[13] Una conferma potrebbe anche venire dalla parola "MEMORIA", che oltre ad essere specifica del gruppo "EBRAICA" è posizionata vicino ad esso, nel **Grafico 1**, ai limiti esterni del quarto quadrante.

Grafico 5 Analisi delle corrispondenze lessicali: segmenti caratteristici



CONCLUSIONI

In conclusione, si può affermare che la metodologia adottata si è rivelata efficace. Essa ha consentito, infatti, di trattare dati testuali molto ampi e articolati in tempi ridotti fornendo - nel contempo - al ricercatore gli elementi necessari al raggiungimento dello scopo proposto: evidenziare gli argomenti e le tematiche trattate da un gruppo di parlanti attraverso gli strumenti messi a disposizione dalla Statistica.

E' comunque necessaria una puntualizzazione.

La metodologia adottata e le strategie operative intraprese derivano sostanzialmente dalla natura e dalle caratteristiche dei dati a disposizione. Ma, quando si parla di natura dei dati, non si fa riferimento solo alla loro estensione o formattazione, che pur influenzano scelte importanti in ambito di valutazione delle risorse impiegate. Il fenomeno *Usenet* crea un "indotto" comunicativo, che riguarda non solo il linguaggio, ma anche i comportamenti. Nei *Newsgroups*, i formalismi della vita quotidiana vengono in parte annullati: ci si rivolge all'altro con un linguaggio confidenziale e molto vicino al parlato; si scrive a "botta e risposta" con un modo ed una terminologia del tutto particolare; nell'affissione dei messaggi si favorisce il rispetto di una *Netiquettes*, ovvero di un insieme di norme di buona educazione e di comportamento in Rete, ma in realtà ciascuno può discutere liberamente sui temi proposti. L'interattività del mezzo Internet fa sì che moltissimi "avventori" - abituali e non - possano dialogare sullo stesso argomento e ciò può contribuire allo svolgimento di discussioni molto lunghe con il pericolo dell'*off-topic* (ovvero del fuori tema) oppure del *flame* (scontro verbale). Non solo, ma - grazie alla struttura di *Usenet* - chiunque può iniziare a far parte di una discussione, venendosi così a modificare alcune dinamiche proprie di un colloquio.

L'alias (ovvero lo pseudonimo o il *nick-name* che la maggior parte degli utenti usano per mascherare il proprio nome reale su *Usenet*) è anch'esso un fattore che influenza la comunicazione: l'anonimato, infatti, può costituire un mezzo attraverso cui esprimersi senza costrizioni.

E' chiara quindi la complessità del fenomeno *Usenet*, per cui risulta difficile determinare delle "regolarità" o riscontrare degli atteggiamenti già noti nella vita quotidiana.

Uno studio sistematico del contenitore *Usenet*, come abbiamo visto, non è impossibile, ma deve partire dai presupposti appena visti e ritornarvi nel momento in cui è necessario confrontare le risorse impiegate con la precisione dei risultati ottenuti.