UNIVERSITÀ DEGLI STUDI DI ROMA

"La Sapienza"

FACOLTÀ DI SCIENZE STATISTICHE

Diploma di Laurea di primo livello in Statistica per le Analisi Demografiche e Sociali

Il nome di Dio:

ricostruzione del campo semantico nei *newsgroups* di argomento religioso

Relatore: Prof. Luca C. Giuliano

Laureando: Alessandro Stabellini
Matricola N°04034234

Anno Accademico 2001/2002
Seduta di Laurea del 16/07/2002

Sommario

Introduzione

Il lavoro svolto tratta l'Analisi del Contenuto di alcuni messaggi inviati - in un determinato periodo temporale - in tre *Newsgroups* di argomento religioso ed ha come scopo quello di ricostruire il campo semantico della parola “Dio”, proponendo nel contempo una particolare metodologia operativa che consente di trattare dati testuali molto ampi ed articolati in tempi ridotti.
Nella Tabella 1 si riportano alcuni elementi utili per identificare i files dai quali ha tratto origine l’analisi.

Tabella 1: Quadro degli elementi identificativi dei files 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/ *Newsreader:* ***Netscape Communicator per Windows®*** *Parametri di salvataggio:* ***Seleziona tutto/salva***	17/09/99 al 05/10/99	2.356kB (2,29MB)
it.cultura.religioni **Server:** http://www.caspur.it/servizi/usenet/news/it.cultura.html news://news.caspur.it:it.cultura.religioni/ *Newsreader:* ***Netscape Communicator per Windows®*** *Parametri di salvataggio:* ***Seleziona tutto/salva***	18/09/99 al 06/10/99	555kB
it.cultura.ateismo **Server:** http://www.caspur.it/servizi/usenet/news/it.cultura.html news://news.caspur.it:it.cultura.ateismo/ *Newsreader:* ***Netscape Communicator per Windows®*** *Parametri di salvataggio:* ***Seleziona tutto/salva***	18/09/99 al 06/10/99	2.269kB (2,21MB)
	TOTALE	**5,06MB**

Sono stati usati diversi software, ognuno per un ben preciso scopo operativo e/o di analisi. In particolare:

- L’applicativo Textpad® (www.textpad.com) come editor di dati testuali per l’eliminazione del quoting (parte di testo citata) dai messaggi e per l’editing dei file di output e dei parametri dell’applicativo Spadt-T®;

- L’applicativo Lexico® per la generazione delle concordanze della parola Dio;
- L’applicativo Taltac® (www.taltac.it) per la Normalizzazione, la creazione del vocabolario la lemmatizzazione nonché per il tagging grammaticale (disambiguazione);
- Spad-T® per l’analisi della corrispondenze semplici parole_x_testi (*Newsgroups*);
- L’applicativo Excel® della Microsoft® per la selezione delle parole a partire dalla tabella con le coordinate, i contributi assoluti, le masse ed i cosen-quadri;
- L’applicativo S.P.S.S. per la generazione del grafico delle corrispondenze.

1.1. Definizioni e concetti di linguistica

Per ***corpus*** si intende un qualsiasi insieme di testi oggetto di studio. Tali testi possono essere letti secondo diversi punti di interesse come se si sfogliasse un libro in differenti modi (per capitoli, per paragrafi, ecc.) in funzione degli obiettivi prefissati.

Ogni differente lettura secondo diversi criteri genera - dal punto di vista della statistica testuale - un insieme di ***profili lessicali*** che costituiscono la base per l'analisi.

Questa definizione operativa di *corpus* è applicabile alle più diverse fonti testuali tra cui la trascrizione di discorsi orali e la traduzione di codici specifici.

Per ***testo*** si intende una fra le tante possibili partizioni del corpus. Il *testo* può essere considerato anche come un insieme di ***frammenti*** (frasi) di un ***discorso*** del ***parlante*** (colui che ha generato il testo) le cui ***parole*** (componenti elementari) sono denominate ***occorrenze*** (ovvero ogni sequenza di *parole* che appare in un *testo*). Le *parole* possono anche essere considerate come una sequenza di caratteri di un alfabeto predefinito delimitata da due separatori[1].

Ogni *frammento*[2] è definito da almeno una *proposizione* o *frase*, delimitata da due separatori "forti"[3], all'interno dei quali è possibile identificare uno o più

[1] Sono considerati, ad esempio, separatori **caratteri non appartenenti all'alfabeto** come lo spazio bianco (*blank*), la punteggiatura (, . : ; ? !), le virgolette, i "trattini" (- / |), le parentesi ({} () []) e ogni carattere speciale (#, @, $,£, $ ecc.) presente nel testo. Ma potrebbero considerarsi tali, *ad hoc*, i numeri o quant'altro.

[2] I frammenti possono essere naturalmente già definiti, come nel caso delle risposte libere in un questionario o dei titoli di articoli della stampa, o dei paragrafi e/o commi di un testo giuridico. Ma possono invece non esserlo, come nel caso dei testi letterari e di interviste non direttive. Allora si ricorre all'individuazione dei *separatori* [Vedi nota 1]

[3] In generale, tali separatori sono definiti da segni di punteggiatura, ma, per una buona segmentazione del testo devono essere adottati ulteriori criteri. Nella lettura automatica della seguente sequenza di testo "... in modo, tale che ... " saranno riconosciuti solo <in modo> e <tale che>. Occorre però evitare di scambiare ad esempio dei punti (punti di migliaia [1.15011] punti di abbreviazioni [Sig. Rossi, C.E.E.]) come falsi indicatori di fine frase o fine frammento. In alcuni *software* le frasi vengono identificate mediante algoritmi di riconoscimento di tali situazioni (grammatiche locali) capaci di risolvere tali ambiguità.

segmenti ripetuti. Un *segmento ripetuto* è una sequenza di *parole* (***sintagma***), tra tutte le disposizioni di 2,3,…,*q parole* che si ripetono più volte nel *corpus*, contenuta interamente in un *frammento*[4]. Tra i *segmenti* si considerano solo i *sintagmi* che costituiscono una *lessia* (*unità di senso* – Vedi pag.12), cioè tutte quelle sequenze di parole che esprimono un contenuto autonomo (***poliformi***). Queste sono di solito il 30% dei *segmenti*. Tra i *poliformi* vi sono le ***poliremat**iche*[5]**,** ovvero dei *sintagmi* che acquistano un significato diverso rispetto ai significati elementari delle parole semplici che li compongono (es. <disegno_di_legge> oppure i costrutti verbali <venir_meno, andare_al_creatore>). È facile che una *lessia* non semplice, in quanto unità semantica, sia spesso una *poliremat**ica*.

Si riporta di seguito un schema di esempio di *segmenti ripetuti, poliformi* e *poliremat**iche* [Bolasco, 1999: 195].

[4] Ad esempio una frase del tipo "il governo si propone di" contribuisce alle occorrenze di segmenti quali: <il governo>, <il governo si>, <il governo si propone>, <governo si>, <governo si propone>, <si propone>, <si propone di>, oltre che al segmento <il governo si propone di>. Bisogna notare che i *segmenti ripetuti* sono di per sé assai più numerosi delle stesse *forme grafiche* componenti un corpus. Per esempio, ad una soglia di 3 occorrenze, in un corpus molto ampio, si ottengono oltre 18.000 segmenti diversi, quando le *forme grafiche* distinte, a quella soglia, sono 9.400: un rapporto di 2 a 1. Al contrario in un corpus molto piccolo (minore di 5.000 occorrenze), a soglia di frequenza elevata (ad esempio 10) il numero di segmenti ripetuti può essere inferiore a quello delle parole alla stessa soglia. Nel parlato, o là dove il linguaggio è molto stereotipato, questo rapporto segmenti/parole può essere ancora più elevato: fino ad un valore pari a 4.

[5] La presenza di poliremat**iche** rende molto difficile il giusto riconoscimento di una frase da parte di tutti quegli algoritmi di traduzione automatica dei testi. Un esempio è il passo: <Portare avanti una casa è un'impresa> che equivale ad un modello generale di senso del tipo: <gestire una casa non è facile>.

Si ritiene che la presenza di poliremat**iche** in un testo possa arrivare a coprire il 60% dell'intero testo ove si considerano anche tutti i verbi–supporto compresi gli ausiliari essere, avere e fare.

Schema 1: *segmenti ripetuti, poliformi* e *polirematiche*

Segmenti

vuoti:	e di, con il, per la, non si, tra le, da tre, ma un, non c'
incompleti:	campo del, è il, quanto si, casa per
pieni:	buona volontà, programma di governo, politica industriale

Poliformi

locuzioni grammaticali con funzioni di:

avverbi:	di più, non solo, per esempio, di nuovo, in realtà, più o meno, di fatto, del resto
(luogo)	a casa, in chiesa, al di là
(tempo)	di sera, un anno fa, al più presto
(modo)	in particolare, d'accordo, in piedi
preposizioni:	fino a, da parte di, prima di, rispetto a, in modo da, per quanto riguarda
aggettivi:	in punto, di oggi, dei genere, in crisi, di cotone, in fiamme, alla mano
congiunzioni:	il fatto che, dal momento che, prima che, nel senso che, a patto che
interiezioni:	va bene!, grazie a Dio, mamma mia!, hai voglia!, punto e basta
idiomi e modi di dire:	io penso che, è vero che, non è che, per così dire, questo è tutto non c'è niente da fare, è un peccato
gruppi nominali polirematici:	buona fede, lavoro nero, mercato unico, punto di vista, cassa integrazione
verbi supporto e idiomatici:	si tratta di, tener conto, portare avanti, far fronte, far parte, prendere atto, dare vita, dare luogo, mettere a punto, venirne fuori, rendersi conto

Un *discorso* è caratterizzato da un ***linguaggio*** le cui componenti sono:

- l'**idioma** della comunità di appartenenza del *parlante* in un dato periodo storico (la lingua del parlante: italiano, inglese ecc.) che è la componente dovuta al *lessico* della lingua;
- il **contesto**, ossia l'ambito o il genere concettuale del discorso, l'aspetto tematico e/o settoriale della terminologia. Il linguaggio è diverso a seconda che tratti di politica, di letteratura, d informazione di sport;
- la specifica **condizione** di enunciazione del discorso (o di stesura del testo) che riflette la modalità d'interazione fra colui che emette (*E*) e colui che riceve (*R*) il messaggio, per cui si creerà una situazione diversa a seconda che il rapporto si stabilisca fra due soggetti ("uno a uno" o dialogo, lettura) o fra un soggetto e un gruppo ("uno a molti", manifesto, assemblea), oppure a seconda che il carattere del discorso sia formale o informale e si svolga, ad esempio, in pubblico o in privato oppure vi sia

co-presenza o meno fra *E* ed *R* (dialogo a vista o a distanza, via telefono o via mass-media) o che abbia carattere pedagogico/polemico.

Con il termine ***contesto locale*** si indica un determinato insieme di parole adiacenti ad un termine prefissato (di solito si considerano 5-10 parole prima e 5-10 parole dopo il termine) che funge da polo (*pivot*). Lo studio dei *contesti locali* di una parola viene detto ***analisi delle concordanze***.

L'***unità di contesto*** è un *frammento* di testo che ha generalmente una lunghezza variabile (da 120 o 200 parole). Spesso corrisponde ad un ***enunciato*** (proposizione con un senso compiuto) altre volte ad una ***frase*** (proposizione con una rilevanza sintattica).

La ***dimensione*** (**N**) o lunghezza di un *corpus* è data dal numero totale delle *occorrenze* (*parole*), mentre il ***vocabolario*** del *corpus* viene definito dall'insieme delle *parole* diverse distinte fra loro[6]. Il numero di *parole diverse* in un testo definisce 1'***ampiezza del vocabolario*** (**V**).

Vale la seguente relazione:

$$\mathbf{V}_1+\mathbf{V}_2+\mathbf{V}_3+\ldots+\mathbf{V}_i+\ldots+\mathbf{V}_{fmax}=\mathbf{V}$$

con $\mathbf{V}_i$ numero di parole diverse che appaiono (o ricorrono) *i* volte ($\mathbf{V}_1$ rappresenta quindi l'insieme delle parole che appaiono una sola volta $\mathbf{V}_2$ quelle che ricorrono due volte ecc.. L'insieme può essere costituito anche da una sola *parola*) e f_{max} valore delle occorrenze della parola con il maggior numero di occorrenze del vocabolario.

Il *vocabolario* di un *corpus* può essere espresso in ***forme grafiche*** (ossia parole tali e quali scritte nel testo) o in ***lemmi*** (ossia riconducendo le parole del testo al corrispondente vocabolo presente in un ***dizionario*** della lingua. Ove per ***dizionario*** si intende l'insieme dei lemmi di un idioma raccolti in un unico "inventario" o *database* lessicale che comprende non solo i *lemmi*, ma anche le ***forme flesse*** [le voci declinate dei sostantivi o aggettivi, o quelle coniugate dei

verbi], le ***forme composte**,* le ***locuzioni** e* le forme ***idiomatiche***. Per la definizione di *lemma* si veda più avanti).

È logico che il *vocabolario* di un *corpus* espresso in *forme grafiche* avrà un'ampiezza differente del vocabolario espresso in *lemmi*.

Si definisce ***lessico*** il discorso in *potenza*, ovvero un insieme virtuale di segni linguistici[7] - esistente nella memoria collettiva di una comunità o in quella di un individuo - da cui possono essere tratte tutte le parole di un *potenziale* discorso. È evidente che non tutte le forme possibili - sebbene siano conosciute - saranno di fatto attualizzate (*occorrenze*). Inoltre il *lessico* di un individuo è un riflesso delle sue appartenenze socio-antropologiche, ovvero delle sue origini, dell'esperienza e della cultura acquisita.

Si definisce ***rango*** il posto occupato da un termine in una graduatoria. Nei *vocabolari* per *occorrenze decrescenti*, ad esempio, un rango "elevato" è indicato da un numero piccolo.

Per ***lemma*** s'intende la *forma canonica* con cui una parola è presente in un dizionario della lingua (come entrata di una voce)[8]. Ad esempio le occorrenze <scrive> e <scrivevano> individuano due forme grafiche distinte, ovvero due *flessioni,* appartenenti ad uno stesso lemma: il verbo <scrivere>.

Secondo la definizione della grammatica tradizionale, ogni *parola* (in quanto ***monema***) può suddividersi in un radicale (***lessema***) e in una desinenza o un affisso (***morfema***): il primo individua l'aspetto lessicale e semantico del termine, il secondo quello grammaticale. La parola "scrivere", ad esempio, si articola in scriv-ere (Queste definizioni variano a seconda dei linguisti che ne discutono).

La *parola* di un *vocabolario* può essere distinta e categorizzata attraverso differenti criteri: rispetto al suo ruolo nella frase, alla sua categoria grammaticale o ad altri criteri.

---

[6] <casa> diversa da <case> o da <cane>

[7] ovvero uno *stock* mentale di radici lessicali (***lessemi***) da cui ricavare tutte le forme flesse dei corrispondenti *lemmi*. Se un individuo conosce il significato della radice "lavor-" può generare sia il sostantivo <lavoro>, sia il verbo <lavorare>, ma anche forme come <lavorante> o altre flessioni.

Molto spesso si fa distinzione tra *parole* ***vuote*** e *parole* ***piene***. Le prime sono le *parole* grammaticali o di legame (articoli, preposizioni, congiunzioni ed alcuni aggettivi) che non esprimono in sé un contenuto d'interesse ai fini dell'analisi, ma hanno una funzione *strumentale* in quanto cardini di costrutti lessico-grammaticali[9]. Le seconde sono portatrici di tutti quei *significati* oggetto di studio, delle parti "sostantive" del contenuto di un *discorso* (nomi e aggettivi), delle sue modalità di enunciazione (avverbi) o di azione (verbi) e per questo sono dette anche *parole principali*.

Inoltre due *parole* possono essere ***omografe*** e per questo avere lo stesso ***significante*** ma diverso ***significato*** (***polisemia***) (ad esempio <stato_S1> nell'accezione di "istituzione" distinto da stato <stato_S2> nell'accezione di "situazione/condizione"), oppure avere differente ***significante*** ma medesimo ***significato*** (***sinonimia***) (ad esempio <stupendo> e <splendido>, <abitanti> e <residenti>, <isolata> e <sperduta>).

[8] L'infinito per i verbi, il singolare per i sostantivi, il singolare-maschile per gli aggettivi.

[9] Anche le parole grammaticali possono avere importanza nell'interpretare un testo. Ad esempio, il sovrautilizzo di preposizioni come <in> o <di> sottolinea il carattere descrittivo del discorso; una prevalenza di <non>, <per> e <con> sottolinea particolari intenzionalità del parlante, mentre quella dei <ma> e <se> evidenzia elementi legati ad incertezza.

1.2. Dalla Linguistica alla Statistica

Base di partenza di ogni indagine statistica è la selezione del ***collettivo***. Il *collettivo* statistico è in questo caso rappresentato dal *corpus* ovvero da una raccolta di testi (*stock* di materiale testuale), omogenea sotto qualche punto di vista.

L'***unità statistica***, ovvero l'unità di osservazione, può essere di tre tipi a seconda degli obiettivi, del tipo e del livello di analisi del *corpus*.

- L'***unità di testo***, che corrisponde alla *parola* come *forma grafica* (grafia)[10]. È l'unità di analisi elementare per la lettura computerizzata di un testo: La *parola*, vista come <catena di caratteri di un alfabeto delimitata da due separatori che ne definiscono l'inizio e la fine>[11], diviene l'oggetto del ***riconoscimento automatico*** (*scansione*) di un testo[12]. Il risultato finale è la ***numerizzazione*** (ad ogni occorrenza diversa si associa un numero diverso) o l'***indicizzazione*** del *corpus* (ogni parola è identificata da un codice identificativo e da un indirizzo - pagina, riga - che indica la sua collocazione nel testo). In taluni studi la *forma grafica* può essere "riletta" successivamente come una specifica flessione di un *lemma* (Vedi pag.9) in quel processo che prende appunto il nome di ***lemmatizzazione***.
- L'***unità di contesto*** che corrisponde ad ogni *frammento* di testo (Vedi pag.5): sia esso una *frase* (proposizione sintatticamente indipendente), un *enunciato* (proposizione di senso compiuto), o una *risposta individuale* (al limite, costituita da una sola parola: <sì>) o quant'altro sia da considerarsi

[10] È evidente che una catena di caratteri non è necessariamente una parola di senso. Per cui <carta> e <catra> sono due occorrenze di *parole* diverse ma <catra> non è una parola dell'italiano: lo sarebbe dopo una correzione ortografica.

[11] Di fatto ogni *parola* di un testo è delimitata da spazi bianchi, mentre ogni frase dalla punteggiatura.

[12] Il computer infatti legge ogni informazione linguistica come una sequenza di *bytes* ciascuna delimitata dai separatori inizio e fine.

unitario sotto qualche punto d'interesse. Ai fini di un'analisi automatica di un *corpus*, può considerarsi come un'unica *unità di contesto* anche un intero testo (un libro, un discorso, un articolo di giornale), oppure una sua parte (un capitolo, un paragrafo, un titolo): ovvero un qualsiasi raggruppamento pertinente di frammenti.

L'*unità di contesto* è l'unità di analisi di studi basati sul confronto di *frammenti* del *discorso* al fine di individuare delle specificità o delle omogeneità rispetto a testi diversi[13].

- La ***forma testuale*** che corrisponde alla ***lessia*****,** ovvero alla più piccola unità portatrice di *senso* - non ulteriormente decomponibile - rilevabile in un *corpus*. È l'unità minima significativa del discorso che può essere *semplice* (definita da una sola parola: <cane>, <tavola>), *composta* (costituita da più parole in via d'integrazione: <sangue freddo>) o *complessa* (individuata da una sequenza di parole fra loro connesse: <fare lo gnorri>, <dalla testa ai piedi>). Nell'applicazione dei metodi della *statistica testuale* e *di analisi del contenuto* si adottano unità di tipo misto, ora *semplici*, ora *complesse* che vengono appunto definite *forme testuali*. Così una *forma testuale* potrà essere sia un *lemma* (<scrivere>), sia una riduzione *lessematica* (Vedi pag.9) (ad esempio la radice <attual+> che fonderebbe <attuale>, <attualmente>, distinta da <attu+> frutto delle fusioni <attuare>, <attuazione>, <attuato>), sia un *significante* che rappresenti la fusione di *sinonimie* accertate nel *corpus* (Vedi pag. 10) (<accordo$>=accordo+alleanza+patto), ma anche un *poliforme* (locuzione grammaticale o poliremática di contenuto: <in_corso>, <bilancia_dei_pagamenti> od una frase fissa *idiomatica*, identificabile come un'entità (<andare_al_creatore>).

[13] L'obiettivo è quello di individuare differenti "universi lessicali" e lo studio ha interessi terminologici.

1.3. Il trattamento del testo

Vi sono diversi principi di ***normalizzazione*** attraverso i quali un testo viene trattato al fine di scegliere, nel sottoinsieme di parole scelte per l'analisi, quelle sulle quali intervenire per accrescerne il livello informativo.

In poche parole nel trattamento del dato testuale viene ridotta l'ambiguità e migliorata la ***monosemia*** (univocità dei significati), cercando, nel contempo, di lasciare intatto il contenuto del *testo* con il suo *sistema di variabilità dei significati*.

<<Più in generale, il criterio fondamentale che è alla base di ogni intervento sul testo si può così esprimere: *conservare distinte nel testo le variazioni significative in termini semantici e fondere le forme che costituiscono degli invarianti semantici.*>> [Bolasco, 1999: 213].

Nel paragrafo 1.2. si è già parlato di *lemmatizzazione*[14]. Essa è di fatto un processo di trattamento del testo attraverso cui vi è una trasformazione sistematica delle forme grafiche in *lemmi* (Vedi pag.9). Ciò risulta utile ed opportuno in alcuni casi (ad esempio per i verbi), in altri innocuo (aggettivi), mentre in altri ancora è addirittura dannoso (basti pensare ad alcuni sostantivi per cui il plurale spesso indica dei referenti concettuali diversi dal singolare: comunicazione ≠ comunicazioni; <scienze della comunicazione> e <scienza delle comunicazioni>. Linguaggi e multimedialità nel primo caso, ingegneria e progettazione dei servizi di trasporto nel secondo. O ancora: paese ≠ paesi: l'uno è "il nostro paese", l'altro sta per "le altre nazioni").

È evidente allora che sono necessari altri interventi capaci di massimizzare la ricerca del carattere *monosemico* delle parole di un *corpus* per valorizzarne l'accezione interna pur rimanendo il più possibile ancorati al contenuto del *testo*.

[14] Sono allo studio dei *lemmatizzatori* automatici basati su algoritmi di riconoscimento *markoviani* (Bolasco, dispense a.a. 1997-1998: 42).

La trasformazione delle *forme grafiche* in *forme testuali* (Vedi par.1.2.) è fra questi.

Nel passaggio dalle *forme grafiche* alle *forme testuali* (Vedi **Tabella 2**) si effettuano sostanzialmente due processi:

I. La ***disambiguazione***, ovvero la distinzione tra forme *omografe-polisemiche* (Vedi pag.10) che può essere:

- *grammaticale* (lemmi diversi: <posto_Verbo> e <posto_Sostantivo>);
- *semantica* (diverse accezioni di uno stesso lemma; <posto-Sostantivo> può significare “luogo”, “impiego”, “spazio”, “sedile”, “centro”);
- *lessico-grammaticale* isolando un *poliforme* (forze_politiche).

II. La ***fusione*** di forme che costituiscono degli invarianti semantici (*sinonimia.* Vedi pag.10) che può riguardare:

- equivalenti grammaticali (le diverse flessioni di un aggettivo, o di determinate voci di un verbo);
- equivalenze semantiche (i raggruppamenti di forme diverse, individuanti tratti semantici o insiemi di sinonimi).

Tabella 2
Esempio del processo di individuazione delle forme testuali

forme grafiche		*analisi lessico-grammaticale*			*lemmi*		*forme testuali*	
	occ.		occ.	cod.		occ.		occ.
POLITICA	1509				POLITICA_sost	1233		
		POLITICA_sost	866	1	(1+3)		POLITICA_sost	866
		POLITICA_ESTERA_polirern.	114				POLITICA_ESTERA_polirem.	114
		POLITICA_ECONOMICA_polirem.	140				POLITICA_ECONOMICA_polirem.	140
		POLITICA_agg	389	2				
POLITICHE	370							
		POLITICHE_sost	113	3			POLITICHE_sost	113
		POLITICHE_agg	132	4				
		FORZE-POLITICHE_polirem.	125	5			FORZE-POLITICHE_polirem.	125
POLITICO	241				POLITICO_agg	1003	POLITICO_agg.	878
		POLITICO_sost	0	6	(2+4+5+6+7+8+9)		(2+4+7+9)	
		POLITICO_agg	241	7				
POLITICI	116							
		POLITICI_sost	0	8				
		POLITICI_agg	116	9				
Totale occ.	2236		2236			2236		2236

Fonte: Bolasco, 1999

Queste *disambiguazioni o fusioni* possono essere messe in atto con differenti strumenti.

a) ***L'isofrequenza***

Definiamo *isofrequenza* la condizione di equilibrio o di stabilità - in numero di occorrenze - esistente fra alcune forme flesse di uno stesso *lemma*. Basterà scorrere un vocabolario di un *corpus* secondo l'ordine alfabetico e ci si renderà conto del fenomeno. L'ipotesi che è alla base di questo comportamento è che tanto più un termine è usato con funzioni, significati o forme diverse, tanto più è probabile che esso accumuli un numero maggiore di occorrenze tali da procurare il fenomeno contrapposto alla *isofrequenza*: la *non-isofrequenza*.

Quindi, se è pur vero che l'esistenza dell'isofrequenza non può, di per sé, costituire la prova di un'equivalenza di significato nei termini coinvolti, al contrario, il riscontrare una *non-isofrequenza* costituisce spesso l'indizio di un utilizzo plurimo della forma in questione.

Questa circostanza segnala pertanto l'opportunità, vuoi di una *disambiguazione*, ad esempio estraendo una locuzione, vuoi di una *fusione*.

Uno dei casi più evidenti è quello in cui una forma semplice è parte integrante e fondamentale di un *poliforme*. Ad esempio si può osservare in un *corpus* che le due flessioni <locali> e <locale> si presentano con occorrenze assai diverse in virtù della presenza della *poliremática* <enti_locali>. Una volta decodificate le occorrenze di tale composto, le due flessioni tornano in condizioni di *isofrequenza*.

b) ***La selezione dei poliformi***

Osservando un vocabolario di *forme grafiche* ci si accorge che molte parole comuni sono inspiegabilmente ai primi ranghi del vocabolario. Questa circostanza può essere il riflesso della presenza di *poliformi* (Vedi pag.6), in particolare quelli a contenuto prevalentemente grammaticale, che sono alla base della costruzione stessa del discorso. Si tratta soprattutto di *locuzioni* con funzione *avverbiale* (<in particolare>, <di nuovo>, <a casa>, <una volta>/<tempo fa>, <del tutto>, <alla fine>, <di fatto>), *aggettivale* (<a punto>, <a tempo determinato>, <in mano>, <alla mano>) o *prepositiva*

(<fino a>, <in modo da>, <da parte di>, <rispetto a>), o di *congiunzioni composte* (<dal momento che>, <certo che>, <come mai>, <a condizione che>), di *formule idiomatiche* (<tutte queste cose>, <è una cosa che>, <per così dire>, <io credo>) o infine di alcuni *verbi idiomatici* con funzione ausiliare di *verbi supporto* (<rendersi conto>, <andar fatto>, <portare avanti>, <venir meno>, <dare vita>, <va bene/male>, <far parte>, <far fronte>, <fare presto/tardi>, <mettere a punto>, <prendere atto>). Tutti i componenti queste espressioni, viste come frasi fisse, risultano avere nel vocabolario di un *corpus* in *forme grafiche*, un numero di occorrenze alterato rispetto al solo uso ordinario, come parole semplici con il loro significato elementare, diretto od originario. Disambiguare almeno alcune fra queste espressioni diventa necessario e, a volte, essenziale. Si riscontra, infatti, che tali *poliformi* hanno un comportamento (sotto il profilo semantico) assai diverso dalle parole semplici costituenti.
Sarebbe allora interessante poter valutare quanto le occorrenze di un *segmento* (Vedi pag.6) incidono sulle occorrenze delle forme semplici che lo compongono.

Per far questo esiste un indice, denominato ***IS***, costruito per selezionare alcune *poliremathiche* di contenuto (Vedi pag.6).

$$IS = \left(\sum_{i=1}^{L} \frac{f_{segm}}{f_{fg_i}} \right) \cdot P$$

ove, date le *L* forme grafiche componenti il *segmento*, si pone a rapporto *la* f_{segm} (n° occorrenze del segmento) a *ciascuna* f_{fg} (occorrenze delle forme grafiche componenti), moltiplicando poi la somma di tutti questi quozienti per *P*, quantità che esprime il numero di *parole piene* (Vedi pag.10) presenti nel *segmento*[15]. Tale indice è sempre positivo, si annulla quando il segmento è

[15] la ricerca automatica di tutti i segmenti ripetuti in un testo è per costruzione ridondante: avviene infatti cercando tutte le sequenze identiche di qualsiasi lunghezza 2, 3, 4 o 5 parole. Per cui ad esempio avremo <punto di> <di vista> <punto di vista> <dal punto di vista> <sotto il punto di vista> (vedi pag.6). Occorre quindi eliminare la ridondanza e selezionare solo quelli "pieni" (*poliremathica*) <punto di vista>, prendendo atto che l'ordine di grandezza della frequenza con cui essi occorrono è assai inferiore a quello delle corrispondenti *forme grafiche* elementari.

composto solo da *parole vuote*[16] ed ha il suo massimo pari a L^2. Condizione quest'ultima in cui tutte le occorrenze della *forma singola* sono date proprio dalla frequenza del *segmento*.

L'indice ***IS***[17] appena visto ci dà un aiuto per valutare l'opportunità della ***lessicalizzazione*** consentendo di valutare l'impatto della frequenza su alcune delle parole chiave coinvolte nel processo di trasformazione delle unità.

La *lessicalizzazione* è quel processo che porta a considerare un *sintagma* (o un qualunque raggruppamento di *parole*) come un solo elemento lessicale. In altri termini la lessicalizzazione consiste nella trasformazione del testo, dovuta al riconoscimento di una sequenza di forme grafiche, come una sola unità di senso o *lessia*. Ad esempio < capo dello stato> verrà modificata in un'unica unità lessicale <capo_dello_stato>[18].

Si prenda come esempio il seguente *segmento*.

"teste rasate" con frequenza *F* =18 ed *IS* = 3,636, con numero di occorrenze della parola "teste" nel *corpus* pari ad f_1 = 22 e con numero di occorrenze della parola "rasate" pari ad f_2 = 18. Il valore di *IS* vicino al suo massimo rivela in effetti che vi è un buon assorbimento della *forma singola* ("pivot") da parte del *poliforme* che la contiene. Infatti l'81% della forma <teste> ed il 100% della forma <rasate> viene "catturato" dal segmento.

Se la sequenza avesse assorbito circa il 50% della frequenza della parola, entrambi avrebbero apportato informazioni utili e differenti.

c) ***L'individuazione del linguaggio peculiare***

L'individuazione del linguaggio peculiare può essere vista come la ricerca dell'insieme minimo di parole massimamente rappresentativo del *vocabolario* che consente di ridurre le ambiguità presenti nel *corpus*.

Comunque si apprezza la presenza di *segmenti* quando la loro frequenza è superiore o uguale a 3 o 4 occorrenze.

[16] Presupponendo di avere definito una lista di parole vuote, l'indice consente di scartare i segmenti vuoti o irrilevanti in termini di grado d'assorbimento; questi, generalmente, sono oltre l'80% dell'intero inventario

[17] Esiste anche un altro metodo per isolare sistematicamente i *poliformi* di un testo. Esso si basa sul confronto dell'inventario dei segmenti ripetuti di un *corpus* con una qualche lista significativa di *poliformi*, specifica di un settore o di un genere di linguaggio. L'intersezione delle due liste - quella del *corpus* e quella specifica - consente di isolare i segmenti pieni.

[18] Il carattere "_" *underscore* dovrà essere cancellato dall'elenco dei separatori (Vedi pag. 5).

Si parte dal presupposto che le parole più frequenti in un *corpus* (anche dette ***parole tema***) non sempre sono ***parole chiave*** (*peculiari, tipiche*) del *corpus* stesso. Ove per *parola chiave* si intende una parola *sovra/sotto-utilizzata* rispetto alla sua frequenza standard nei normali contesti d'uso. E si sceglie un modello di riferimento (rappresentato da un tipo di lessico[19]) rispetto al quale calcolare il sovra/sotto-uso delle *parole chiave*. Così facendo ci si affida ad un criterio che consente di selezionare le parole di un corpus non soltanto sulla base del loro più o meno elevato numero assoluto di occorrenze.

Tale criterio misura la *peculiarità* in termini di *specificità* sia positiva che negativa. La prima correlata con le parole più frequenti, mentre la seconda con quelle così rare da essere "quasi assenti", forse perché volutamente evitate dal locutore.

La misura di specificità, per ciascuna parola, è allora data, ad esempio, dal seguente rapporto:

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

che costituisce uno scarto standardizzato della frequenza relativa, dove f_i è il numero di occorrenze normalizzate della *i-esima* parola nel corpus ed f_i^* il corrispondente valore nel lessico assunto come modello[20], mentre la quantità al denominatore è lo scarto quadratico medio della frequenza relativa (Vedi pag. 25). Come è facile notare, questo rapporto è pari alla radice quadrata dell'i-esimo contributo ad un *chi-quadrato*.

In assenza di un modello di linguaggio di riferimento, si potrebbe lo stesso giungere all'individuazione delle parole chiave del corpus effettuando, in via

[19] Con la crescita delle potenzialità informatiche di calcolo, attualmente non è difficile effettuare raccolte di testi per la messa a punto di liste di frequenza, assemblando stock sempre più ampi (anche milioni di occorrenze) di materiali riguardanti periodi, generi e situazioni differenti (un tempo solo testi scritti, più recentemente anche testi parlati). Queste liste permettono di costruire i cosiddetti *lessici di frequenza:* in pratica, dei vocabolari ordinati per numero decrescente di occorrenze, o meglio, secondo il loro rango in termini di frequenza d'uso. I lessici di frequenza possono essere utilizzati come modelli di riferimento per la valutazione del sovra/sottouso delle parole nel corpus oggetto di studio.

[20] Tali quantità possono essere espresse anche in termini d'indice d'uso (Vedi pag.17). Tale confronto è tanto più valido quanto più il corpus è connesso al lessico.

preliminare, un'analisi delle corrispondenze sul *corpus* in *forme grafiche*, a soglia di frequenza elevata.

Così facendo si evidenziano sul primo piano fattoriale alcuni "punti cardinali" della struttura del contenuto. Si procede poi ad altre analisi, con soglie di frequenza via via decrescenti, per scoprire quali siano le parole che restano stabili in queste simulazioni e quali siano i contenuti che si definiscono come "sottocampi" o dettagli semantici di tali punti cardinali. Così facendo, si identificano i termini sui quali è opportuno concentrare gli interventi di disambiguazione o di fusione.

Nonostante siano stati appena descritti metodi e criteri per intervenire sul testo con una certa sistematicità, l'atteggiamento che occorrerebbe comunque assumere è quello di procedere con parsimonia: alcuni interventi, infatti, potrebbero procurare più danni che vantaggi (caduta di frequenza, frammentazione delle occorrenze e perdita della *forma* perché al di sotto della *soglia di frequenza* ecc..).

Uno dei criteri fondamentali che guidano gli interventi sul testo consiste da un lato nel tendere a ridurre il numero delle unità lessicali da considerare per l'analisi e dall'altro nel cercare di aumentare il tasso di copertura del testo[21], a parità di numero di unità considerate[22].

Esiste un test in grado di legittimare la scelta di fondere/non fondere o disambiguare più termini basato sulla ricostruzione - mediante simulazione - delle "regioni di confidenza" sul piano fattoriale[23].

Si considerano le parole che sarebbero oggetto di pretrattamento.

- Si effettuano le disambiguazioni e si analizza come quest'ultime si comportano sul piano fattoriale. Se i loro punti producono regioni disgiunte, la loro disambiguazione è legittima;

[21] Il t*asso di copertura del testo (%cop)* è dato dal valore percentuale $\boldsymbol{N_{(s)}/N}$ (dove $N_{(s)}$ esprime l'ampiezza del *corpus* sopra il *livello di soglia*, mentre *N* il numero delle occorrenze del *corpus* stesso).

[22] Si ricorda che gli interventi riguardano circa un 10% delle forme del vocabolario da analizzare (che a sua volta potrebbe aggirarsi intorno al 12 % di *V*.

[23] Si sfrutta in pratica la proprietà vicinanza=somiglianza dei punti sui piani. Le matrici sono matrici di frequenza <parole x subtesti>.

- si effettuano le fusioni e, come sopra, si studia il loro comportamento sul piano fattoriale. Se esse hanno regioni di confidenza fortemente incluse, è evidente che una loro fusione sotto un unico lemma non inficerebbe l'analisi.

In Figura 2 ed in Figura 1 si riportano alcuni esempi tratti dalle dispense del Prof. Bolasco da cui si evince come sia legittimo fondere le quattro voci del participio passato del verbo essere, mentre come sia inopportuno mischiare il singolare ed il plurale del nome politica.

Figura 2: Participio passato del verbo essere

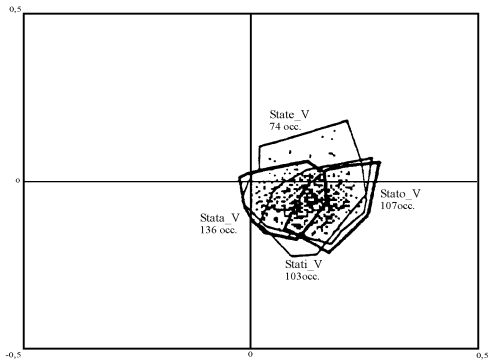
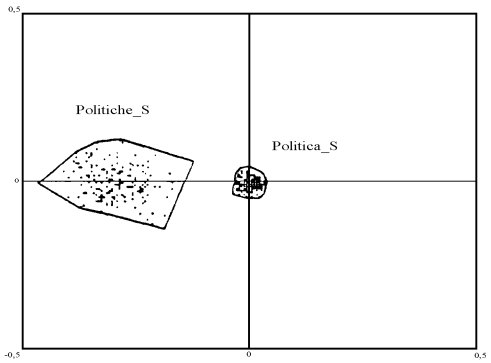


Figura 1: Singolare/plurale del nome politica



1.4. Tipi di matrici di dati testuali

Al fine di un'analisi statistica del contenuto, ogni *corpus* deve tradursi in un'opportuna matrice di dati.

Fondamentalmente, si possono avere tre diversi tipi di matrici di dati testuali[24]:

- {frammenti x forme}, in cui in riga si hanno i *frammenti* di testo, da considerarsi come unità statistiche d'analisi (od "osservazioni") e in colonna si hanno le *forme* selezionate per lo studio, da considerarsi come variabili relative a ciascuna unità.

	F_{max}						$F_{(s)}$
	1	2	3	…	j	…	$V_{(s)}$
Forma	*di*	*è*	*il*		…		*era*
Framm.							
1	1	1	0	…	0	…	0
2	1	0	1	…	1	…	0
…				…		…	
i	0	1	1	…	1	…	1
…							
n	1	1	0	…	0	…	0

Dove le righe possono essere: singole risposte degli intervistati, le singole proposizioni, i singoli versi, i commi. Mentre le colonne sono costituite dalle *unità lessicali* (Vedi pag.18) selezionate dal vocabolario del *corpus*, espresso ora in *forme grafiche*, *lemmi* o *segmenti*, ora in forme miste (*lessie*, *forme testuali*). Le colonne della matrice costituiscono le informazioni elementari di ogni unità e sono quindi le componenti di un *profilo lessicale*, che sarà oggetto dello studio. In ogni casella (*i,j*)
della matrice è riportato il numero di occorrenze dell'unità lessicale *j-esima* presente nell'osservazione *i-esima*. Di fatto questa matrice è in prevalenza una tabella booleana (1 = presenza, 0 = assenza) poiché solo raramente vi sono più occorrenze di una stessa parola in un frammento. Sarà comunque

[24] Le matrici conterranno solo le unità lessicali, V_{fmax},...,V_s, che superano una prescelta soglia di frequenza *s*, dal momento che si analizza non l'intero vocabolario di un corpo, ma un sua parte (Vedi pag. **Errore. Il segnalibro non è definito.**).

una matrice di tipo sparso, cioè con oltre il 95% delle caselle nulle, poiché ciascun frammento è composto di qualche decina di parole, mentre l'ampiezza del vocabolario è solitamente di diverse centinaia - se non migliaia - di unità lessicali. In questa matrice, si perde l'informazione relativa alla disposizione dei termini all'interno di ciascun frammento, mentre se ne conosce la loro combinazione. Ad esempio, il frammento <una casa bella veramente> coinciderebbe con <veramente una bella casa>. Tali "perdite d'informazione" sono tuttavia trascurabili rispetto al contenuto informativo dovuto alla combinazione dei termini stessi; se, al contrario, dovessero ritenersi fondamentali, non si dovrà fare ricorso a questo genere di matrici.

A questa matrice può venire associata, per ciascun frammento, una serie di categorizzazioni (A, B, C, ...) che registrano le modalità di altrettante variabili qualitative numeriche. Ad esempio, nel caso di un corpus di una raccolta di articoli di giornale (ogni articolo costituisce un frammento), è possibile associare all'articolo categorizzazioni sull'autore (A), sulla posizione nella pagina (B), sull'argomento trattato (C) ecc.

Si avrà così la seguente matrice delle variabili categoriali:

A	B	C	D
1	3	5	2
2	1	3	1
…			
2	3	2	2
…			
1	3	1	2

- {forme x testi}, in cui si ha in riga il vocabolario selezionato allo scopo e in colonna i testi (o parti) secondo cui si considera suddiviso il *corpus*. *L'unità lessicale* è quindi l'unità statistica d'analisi e il *testo* costituisce la variabile di studio**.** L'informazione statistica interna alla matrice è la frequenza (numero di occorrenze assolute) con cui una "parola" (*forma, segmento o lessia*) ricorre in ciascun testo. Il profilo lessicale d'interesse è spesso il profilo colonna, dal momento che si confronteranno i diversi testi sulla base della differente presenza (frequenza) delle parole. In ogni caso, la lettura diretta di

tali profili richiederebbe di trasformare le occorrenze assolute in occorrenze normalizzate.

	Testo	1	2	…	j	…	*T*
Forma.							
1	*di*	82	35	…	40	…	145
2	*è*	56	77	…	19	…	70
3	*il*	49	62		33	…	12
…		…	…	…	…		…
i	…	29	10		25	…	56
…		…	…		…		…
$V_{(s)}$	*era*	3	1	…	5	…	7

- {forme x forme}, in cui sia le righe che le colonne descrivono il "vocabolario" prescelto. L'informazione statistica interna alla matrice è una misura di similarità (espressa come sola presenza/assenza o come grado di correlazione) che registra il tipo o livello di co-occorrenza fra le forme, all'interno dei testi. Questa matrice può venire utilizzata ad esempio per ponderare i profili colonna nelle matrici del tipo {frammenti x forme}.

	cosa	madre	casa	vita	dare	…
cosa	1	0	1	0	0	…
madre	0	1	1	1	1	…
casa	1	1	1	1	0	…
vita	0	1	1	1	1	…
dare	0	1	0	1	1	…
…	…	…	…	…	…	…

1.5. L’analisi multidimensionale del contenuto

Una volta raggiunto un buon livello del dato linguistico come dato statistico, ossia una volta ricercata una qualche approssimazione della *monosemia* per le parole significative ai fini dell’analisi, si procede all’applicazione delle tecniche statistiche per differenti livelli di analisi.

Si possono individuare due livelli di studio. In un primo livello si affronta per così dire uno studio esterno o “verticale” del *testo*: l’analisi è di tipo *lessicale* perché l’interesse si basa sulla terminologia utilizzata (*vocabolario*). Il risultato finale è l’*analisi delle specificità*.

In un secondo livello si fa ricorso a tecniche statistiche multidimensionali (ad esempio di tipo fattoriale) capaci di studiare il contesto generale delle varie co-occorrenze delle parole attraverso lo studio dei profili lessicali descritti dalle matrici dei dati, fino alla ricostruzione dei sintagmi fondamentali presenti nel *corpus*. Il risultato finale è la ricostruzione dei principali modelli di comportamento del *senso*.

- **Specificità di forme e frasi in un testo**

  Con il termine *specificità* si intende indicare se e quanto una parola sia tipica o specifica di un sub-testo, nell’ambito di uno stesso *corpus*, o - più in generale - quanto una forma sia specifica rispetto ad un qualche modello di linguaggio di riferimento.

  Una misura di *specificità* di una parola in un testo, di solito ottenuta a livello di forme testuali, viene calcolata a partire dalla tabella di frequenza che ripartisce le occorrenze totali di una forma del corpus nei vari sub-testi in cui essa occorre.

  Essa può esser data semplicemente da:

$$z = \frac{(x - x_{teor})}{\sigma_x}$$

  Che è del tutto simile alla formula già vista a pag.19.

Si arriva a tale formula per il calcolo della *specificità* partendo da alcune semplici considerazioni.

<<Come è noto si possono indicare con:

$$E(x) = n \cdot p \quad \text{e} \quad \sigma_x = \sqrt{n \cdot p \cdot q}$$

rispettivamente la media del numero assoluto di occorrenze di una parola e il suo scarto quadratico medio, ove *p* (e *q)* è la probabilità, come frequenza relativa, dell'apparire della parola (e rispettivamente del suo non apparire) in un testo, ed *n* è il numero di "prove" che si immagina di effettuare per ottenere la parola in oggetto. Nel nostro caso *n* è pari al numero totale di occorrenze nel sub-testo: ipotizzando ogni *tranche* di corpus della stessa dimensione, *n* è costante in tutto il corpus. Questo schema teorico sottintende, nell'ipotesi di indipendenza fra eventi, che l'apparire delle occorrenze di una parola in ciascun sub-testo possa essere immaginato come un evento aleatorio, ove *p* appunto è la probabilità di ottenere quella parola ogni *n* "prove".

Ogni qualvolta si ottiene una proporzione di occorrenze di molto superiore (o inferiore) a questa quantità *n·p* si può supporre che ciò non sia dovuto al caso ma piuttosto sia l'espressione di una qualche "causa" specifica. Ha senso allora voler misurare in termini di uno scarto relativo questa differenza. Tale scarto prenderà la forma seguente:

$$z = \frac{(x - x_{teor})}{\sigma_x}$$

Ora in ambito linguistico, la frequenza relativa *p* di una parola in un testo è di fatto sempre bassissima, per cui, volendo semplificare il calcolo, possiamo esprimere σ_x come $\sigma_x = \sqrt{n \cdot p}$, in quanto il prodotto di *p·q* è praticamente sempre uguale a *p*. Ma il tal modo lo s.q.m. della frequenza assoluta di una parola è pari alla radice quadrata della frequenza assoluta teorica. In questo senso lo scarto standardizzato […] assume la forma

$$z = \frac{(x - x_{teor})}{\sqrt{x_{teor}}}$$ >> [Bolasco,1999: 227]

Questo rapporto può essere valutato utilizzando i criteri classici della significatività statistica con alcune considerazioni.

<<[…] Assumendo il *corpus* come una popolazione e ogni sua parte (sub-testo) come un campione, il modello distributivo di riferimento - per valutare in termini probabilistici il numero di occorrenze di una parola presenti in questo campione - è quello di una legge *ipergeometrica*, legge vicina alla distribuzione *multinomiale* quando le frequenze relative sono molto piccole rispetto alla popolazione. Sotto particolari condizioni (frequenze assolute osservate non inferiori ad una certa frequenza) a sua volta quest'ultima è ben approssimata da una variabile casuale normale. In pratica quando si stabilisce un livello di soglia minimo sul numero di occorrenze di una parola, per considerarla "in analisi", il calcolo delle specificità avviene attraverso un *valore-test* che confronta la frequenza relativa di una parola nella parte, con la corrispondente frequenza relativa nel corpus totale. Questo test è effettuato sotto l'ipotesi di un'approssimazione normale, per cui è possibile assumere i classici limiti degli intervalli di confidenza di una variabile standardizzata *z* e assumere le regole ben conosciute della distribuzione di Gauss.

Quando *z* è all'incirca intorno allo zero ciò significa che la parola è presente nel sub-testo in proporzioni puramente aleatorie, ossia tanto quanto in media ci si può aspettare. In tal caso la parola non è significativa, quindi in qualche modo è "banale", come dire che fa parte del vocabolario di base (necessario alla costruzione) del testo.

Se *z* è superiore, in valore assoluto, a 2 sappiamo che la sua presenza è significativamente diversa da quella attesa (sotto una ben determinata ipotesi teorica, che è quella dell'equidistribuzione e quindi di indipendenza, all'interno di un certo schema di estrazione e di un modello probabilistico di riferimento). Quindi il numero delle sue occorrenze è significativo, sia in termini positivi che negativi.

Nel primo di questi due ultimi casi si dirà che il numero di occorrenze della parola in esame nel sub-testo supera largamente il valore atteso per puro effetto del caso e che la parola è *caratteristica* del testo (specificità positiva).

Nel secondo caso si dirà che la sua così bassa frequenza è anch'essa significativa, per cui vi sarà una qualche causa per la quale la parola non è

presente nel testo quanto ci si potesse aspettare. La parola si dice allora anti-caratteristica o "rara" o anche mal rappresentata.

Una selezione di parole con specificità positive S^+ o negative S^- consente di individuare alcuni tratti salienti del sub-testo, in modo da identificarne i principali contenuti>> [Bolasco, dispense anno accademico 1997-1998: 52].

Una estensione del criterio di selezione delle *forme caratteristiche* è quello dell'**estrazione di frasi significative** che consiste nell'identificare alcuni contesti locali che appunto contengono tali *forme*. Ciò si basa sul principio che una frase è tanto più caratteristica quante più parole ad alta specificità essa contiene. Pertanto se si considera di calcolare il valor medio dei *valori-test* delle parole che formano la frase, più è elevata questa quantità, più significativa è la frase[25].

Con una semplificazione si può immaginare di adottare come informazione il rango associato alle forme caratteristiche per ciascun sub-testo. Ovvero data la lista delle forme di un sub-testo, secondo la loro specificità positiva decrescente, e consideratone il rango (ranghi bassi = alta specificità) si calcola il rango medio delle parole della frase. Se il rango medio è piccolo vuol dire che essa contiene solo parole caratteristiche.

- **Gli assi fattoriali: una interpretazione in chiave linguistica**

Un insieme di unità lessicali - ordinate lungo un asse fattoriale - può essere concepito e interpretato alla stregua di un *sintagma* (Vedi pag.6). Il concetto che c'è alla base è che l'ordinamento degli elementi su di un asse crea significato. La lettura di quest'ordine genera il senso (d'un enunciato) da attribuire all'asse. Sarà così possibile trarre il senso latente e globale del sistema di significati elementari messi in gioco dalle forme oggetto di analisi su un fattore, limitandosi alla interpretazione delle sole "parole" significative per l'asse considerato.

[25] Naturalmente questa misura è influenzata dal numero di parole, in quanto tende a privilegiare le frasi corte. Infatti, a parità di forme caratteristiche, ogni parola banale che si aggiunge nel calcolo tende ad abbassare la media

Il differente disporsi degli elementi su di un nuovo asse genera un diverso *sintagma* e conseguentemente un differente punto di vista dal quale leggere i contenuti del *corpus*.

2.1. Cos'è *USENET*

<< *Usenet* riguarda le persone […] *Usenet* è l'insieme delle persone che sanno cos'è *Usenet* >>. [Dern, 1995: 197]

NetNews, più comunemente chiamata *Usenet*, è un sistema condiviso (*sharing system*) di messaggi che provengono da tutto il mondo in un formato standard. In poche parole *Usenet* è una comunità mondiale di bacheche elettroniche (*BBS*)[26], strettamente associate ad Internet, le cui informazioni sono costituite da singoli messaggi, ciascuno dei quali può essere letto e condiviso da molti utenti. I messaggi sono organizzati in gruppi di argomenti o "*Newsgroup*".

Usenet comprende molti computer in molte organizzazioni e sedi diverse.

Gli utenti di *Usenet* leggono e forniscono contributi (affiggono messaggi) nella sede *Usenet* locale. Ogni sede *Usenet* distribuisce le affissioni dei suoi clienti alle altre sedi *Usenet* in base ai vari parametri di configurazione impliciti ed espliciti e a sua volta riceve affissioni dalle altre sedi.

Tramite *Usenet* milioni di utenti di computer in tutto il mondo condividono informazioni, presentano domande e risposte, conducono discussioni tra più utenti.

Esistono pacchetti software speciali per gli utenti di *Usenet* (*Newsreader* per la lettura e l'affissione di messaggi).

<< *Usenet* non è un'organizzazione. Nessuna persona o gruppo ha autorità su *Usenet* nel suo insieme. Nessuno controlla chi riceve un contributo, quali articoli sono propagati nelle varie sedi, chi può affiggere gli articoli o qualsiasi altra cosa. Non esiste un'*azienda Usenet*, né un gruppo di utenti *Usenet*: ogni utente è solo >>. [Dern, 1995: 198].

---

[26] Le *BBS*, a differenza della posta elettronica, consentono di organizzare le informazioni come *risorse comuni condivise*, in directory che non appartengono al singolo utente. I messaggi arrivano in aree di proprietà del software della bacheca elettronica: molti utenti 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. Invece di inviare un messaggio ad una persona, lo si invia, o "affigge", in un gruppo della *BBS*.

I messaggi *Usenet* viaggiano con qualsiasi mezzo, che può essere *UUCP* (*Unix to Unix copy*), in cui i computer effettuano chiamate da modem a modem e inoltrano le copie di tutti i messaggi appropriati, oppure attraverso *Internet*[27] e anche via radio, dischetto, nastro di *backup* e *CD-Rom.*

[27] Nel settore delle reti, il termine tecnico per designare il collegamento di reti è *internetworking*, mentre per indicare una rete di reti viene impiegato il termine *internetwork o internet.* La nuova internet work, incentrata su ARPAnet fu soprannominata Internet, con la "I" maiuscola.

3.1. Il caso di studio: ricostruzione del campo semantico della parola DIO

I. Introduzione all'analisi

Il materiale raccolto comprende i messaggi inviati, in un determinato periodo temporale, su 3 *Newsgroups* (Vedi il capitolo su *Usenet*) di argomento religioso*: it.cultura.cattolica, it.cultura.ateismo, it.cultura.religioni*.
Nel capitolo introduttivo (pag. 3) è riportato uno schema dal quale si desume la provenienza dei messaggi ed il mezzo usato per scaricarli dalla Rete.
L'unione di questi messaggi ha generato un *corpus* di circa 5,06 Mega Byte di dati dal quale è stato creato il file contenente le concordanze (vedi pag. 8) del termine "Dio": quest'ultimo oggetto di un particolare approccio metodologico volto all'analisi del contenuto per la ricostruzione e/o la determinazione del campo semantico,appunto, della parola "Dio".

II. Fasi operative

- La generazione delle concordanze è stata possibile tramite l'uso dell'applicativo Lexico® che ha dato origine ad un file in cui ciascuna riga costituisce un "contesto locale" (appunto concordanza) della parola *Dio*. Di fatto, un insieme di righe in cui sono stati presi 40 caratteri prima e 40 caratteri dopo - fino alla parola completa - del termine_polo (pivot) Dio.
  Esempio: una delle righe generate da Lexico®

enza considerare il tempo storico in cui Dio cominciava a manifestare , il suo modo

  Nota: in questa fase, sul file trattato da Lexico®, è stata necessaria una prima normalizzazione del testo (vedi pag. 13) in cui è stato sostituita la Stringa < ‘> (spazio + accento) con <’> (accento); questo per recuperare le parole accentate attraverso gli automatismi del Taltac®.

- Successivamente al Lexico®, è stato usato l'applicativo Taltac® per normalizzare il testo e per effettuare il tagging grammaticale, ovvero la disambiguazione (Vedi pag.14) dei termini.

  Alla fine del processo di cui sopra, sempre tramite il Taltac®, è stato ricostruito il *corpus* delle concordanze con i termini lemmatizzati (vedi pag. 11) e disambiguati.

  Nella stessa sede è stata usata una procedura del Taltac® per il confronto del dizionario lemmatizzato del corpus delle concordanze con un dizionario dei poliformi base. Ciò, sostanzialmente, al fine di rivelare le specificità nel linguaggio adottato all'interno dei tre Newsgroups durante il periodo di rilevazione. I risultati del confronto sono riportati più avanti in questo capitolo.
- A questo punto, il corpus delle concordanze trattato dal Taltac® è stato editato da un editor di testi in maniera tale da poter essere processato dall'applicativo Spad-T®. Alla fine, il corpus originale presentava una struttura di seguito riportata:

  ****NEWSGROUPS n. 1
  qui comincia il testo del primo frammento di testo (concordanza), appartenente al NEWGROUPS 1, contenente la parola "Dio".

  I trattini che precedono segnano la fine del periodo, ovvero del frammento di testo di cui sopra. Questa accortezza è necessaria in quanto, in loro assenza, lo Spad-T troncherebbe il periodo ad 80 colonne video.
  ****NEWSGROUPS n.2
  segue come sopra con le stesse regole sullo stesso file.
  ****NEWSGROUPS n.3
  segue come sopra con le stesse regole sullo stesso file.
  ====
  I quattro uguali servono a segnare la fine del Corpus da analizzare

- Infine si è utilizzato lo Spad-T®.

  L' applicazione adottata nello Spad-T® è stata la "Texte" o testo unico (usata, appunto, per quei casi in cui non vi è bisogno di associare di alcun file numerico per la codifica delle variabili).

  All'interno dello Spad-T® si è proceduto come segue:

  - attraverso la procedura "Numer" si è numerizzato il testo. Di seguito sono riportati alcuni risultati della procedura:
    numero dei frammenti di testo (concordanze) = 954

[Newsgroups Atei = 220; Cattolici = 324; Religioni = 410]

numero totale delle parole =14.579

numero delle parole distinte = 2.659

percentuale delle parole distinte =18.2

- attraverso la procedura “Setex” si è innalzata la soglia di frequenza delle parole considerate fino al valore 5 e si è impostato a 3 il valore di lunghezza delle parole in analisi.

  Ciò ha consentito di ridurre il numero delle parole distinte e di portarle a 285;

- applicando infine la procedura “Aplumm” che ha permesso l’analisi delle corrispondenze semplici "parole x testi" nonché il calcolo di alcuni valori utili alla sua interpretazione quali la massa, i contributi assoluti e relativi degli elementi. L’output della “Aplumm” è stato editato attraverso un editor di testi e successivamente con l’applicativo Excel per far sì che fosse importato all’interno dell’S.P.S.S. per la tracciatura dei grafici.

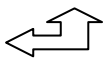
  Prima della tracciatura del grafico, però sono state eliminate manualmente alcune parole “vuote” ritenute superflue ai fini dell’analisi. Di seguito la lista delle parole eliminate:

1	così	noi
2	cui	non
5	da	né
a	dei	per
ai	di	proprio
alla	di	qualche
allora	ed	qualcosa
almeno	egli	quale
altrimenti	in	quando
anche	io	quanto
anche_se	la	quello
che	le	quello
chi	lo	questi
ché	lo	questo
ci	ma	quindi
ciò	me	se
come	mi	si
con	ne	www

III. Commenti e conclusioni

Da una prima analisi delle specificità tratta da un confronto con un dizionario base dei poliformi (cfr. Taltac® pag. 33) di cui, qui di seguito, si riportano i primi 50 termini con la più alta percentuale d'uso nel vocabolario sperimentale e con uno scarto positivo (differenza tra le occorrenze nel corpus sperimentale con quelle nel dizionario base), si evidenzia come il *pivot* "Dio" attragga su di sè parole tipiche di un determinato contesto religioso. A testimonianza di ciò i termini : "miracoli", "profeta", "biblico", "immacolata", "divinità", "onnipotente", "ateo", "sommo".

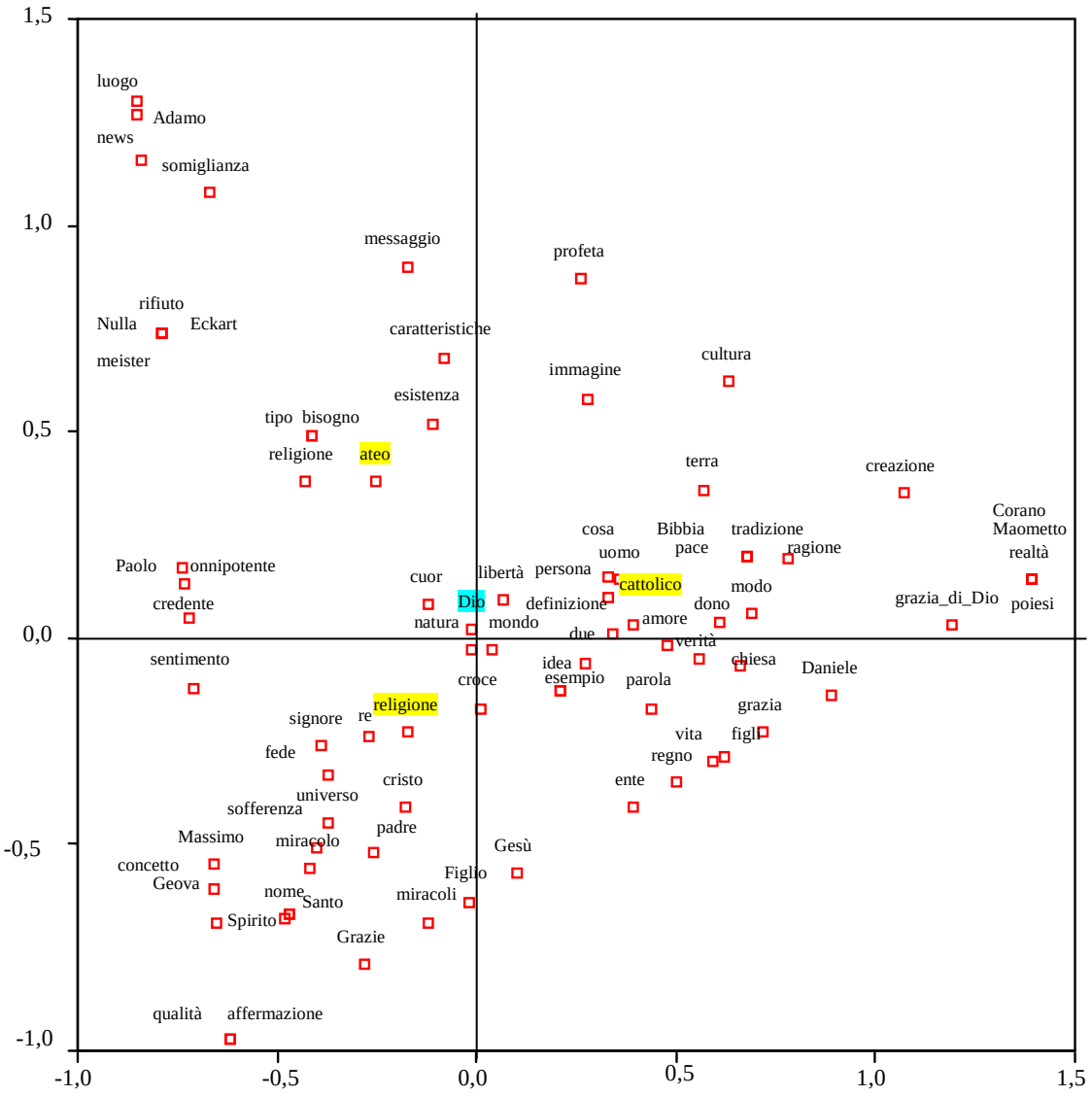
Forma grafica	**Uso**	**Scarto sulle Occorrenze**
esista	42,06	0,06
soffre	36,02	0,45
estraneo	32,85	0,35
miracoli	25,78	0,22
somiglianza	16,24	3,27
muta	14,65	0,88
profeta	14,6	1,46
credente	13,25	2,52
onnipotente	9,94	3,11
sommo	9,03	0,45
ateo	8,95	6,29
creò	8,29	1,64
voluti	7,34	0,54
chirurgo	5,6	0,89
mela	5,43	0,32
inventando	5,42	0,32
biblico	5,41	3,24
presuppongono	5,01	0,32
insensibile	4,73	1,11
colte	4,73	0,51
oppositore	4,61	0,89
Immacolata	3,96	0,73
pertinenza	3,87	0,51
assimilabile	3,6	0,51
apparisse	3,54	0,51
Segui	3,41	0,15
cavoli	3,16	0,32
divinità	3,14	1,9
dona	3,12	1
diametralmente	3,03	1
bla	2,73	4,38
svanisce	2,67	1
affermi	2,52	1
invocazione	2,52	1
sottomesso	2,46	0,73
demonio	2,3	4
credendo	2,27	1,64
divengono	2,24	0,22
millenaria	2,22	0,22
infallibile	2,21	0,22
pregate	2,2	2
scandalizza	2,2	1
sentirti	2,18	0,22
versetto	2,18	0,22
indelebile	2,16	0,22
santoni	2,13	0,22
adorano	2,09	0,22
fattosi	2,08	1
sentiranno	1,81	1,34



Nelle pagine seguenti i grafici delle analisi delle corrispondenze la cui interpretazione permette di cogliere meglio tali specificità legandole anche al tipo di Newsgroups da cui le parole sono state tratte.

Grafico 1: Analisi delle corrispondenze parole x testi. NOMI

(Newsgroups in giallo sul grafico)



Il Grafico 1 riporta l'analisi delle corrispondenze parole x testi (ovvero Newsgroups), dove le parole sono solo nomi.

Come era logico aspettarsi, la parola "Dio", *pivot* del corpus delle concordanze, è posizionata nel *centroide,* mentre i Newsgroups Atei (nel grafico "ateo" in giallo), Religioni (nel grafico "religione" in giallo) e Cattolici (nel grafico "cattolico" in giallo), occupano posizioni ben distinte l'uno rispetto all'altro.

In particolare il Newsgroups dei Cattolici risulta essere un forte "attrattore" per parole come "uomo", "amore", "definizione", "persona" e "pace" e caratterizzante per parole come "creazione", "poiesi", e "grazia_di_Dio". Queste ultime molto distanti dall'origine degli assi e quindi, in virtù della metrica del chi-quadrato, elementi in media con masse più piccole (parole + rare presenti solo in determinati contesti).

Il Newsgroups degli Atei, nel I quadrante negativo, risulta essere, invece un attrattore per i termini "esistenza" , "caratteristiche", "religione" e "bisogno" e caratterizzante per "Meister Eckart" (un mistico medievale) e per le parole "rifiuto" e "Nulla", quasi a voler testimoniare delle argomentazioni, nel periodo di rilevazione dei messaggi, intorno proprio alla figura di Eickart ed ai suoi pensieri nel gruppo degli Atei.

Infine il Newsgroups Religioni attrae su di sé termini molto particolari come: "signore", "re", "padre", "miracolo" e "sofferenza".

Sembrerebbe quindi, secondo una interpretazione del tutto personale, che il Dio dei Cattolici sia un Dio di "tradizione", di "amore" e di "dono", quelli degli atei di "bisogno" e di "religione", mentre quello delle Religioni si di "sofferenza", di "concetto" e di "affermazione" (quest'ultima parola molto caratterizzata da tale Newsgroups)

Grafico 2: Analisi delle corrispondenze parole x testi. VERBI (Newsgroups in giallo sul grafico)

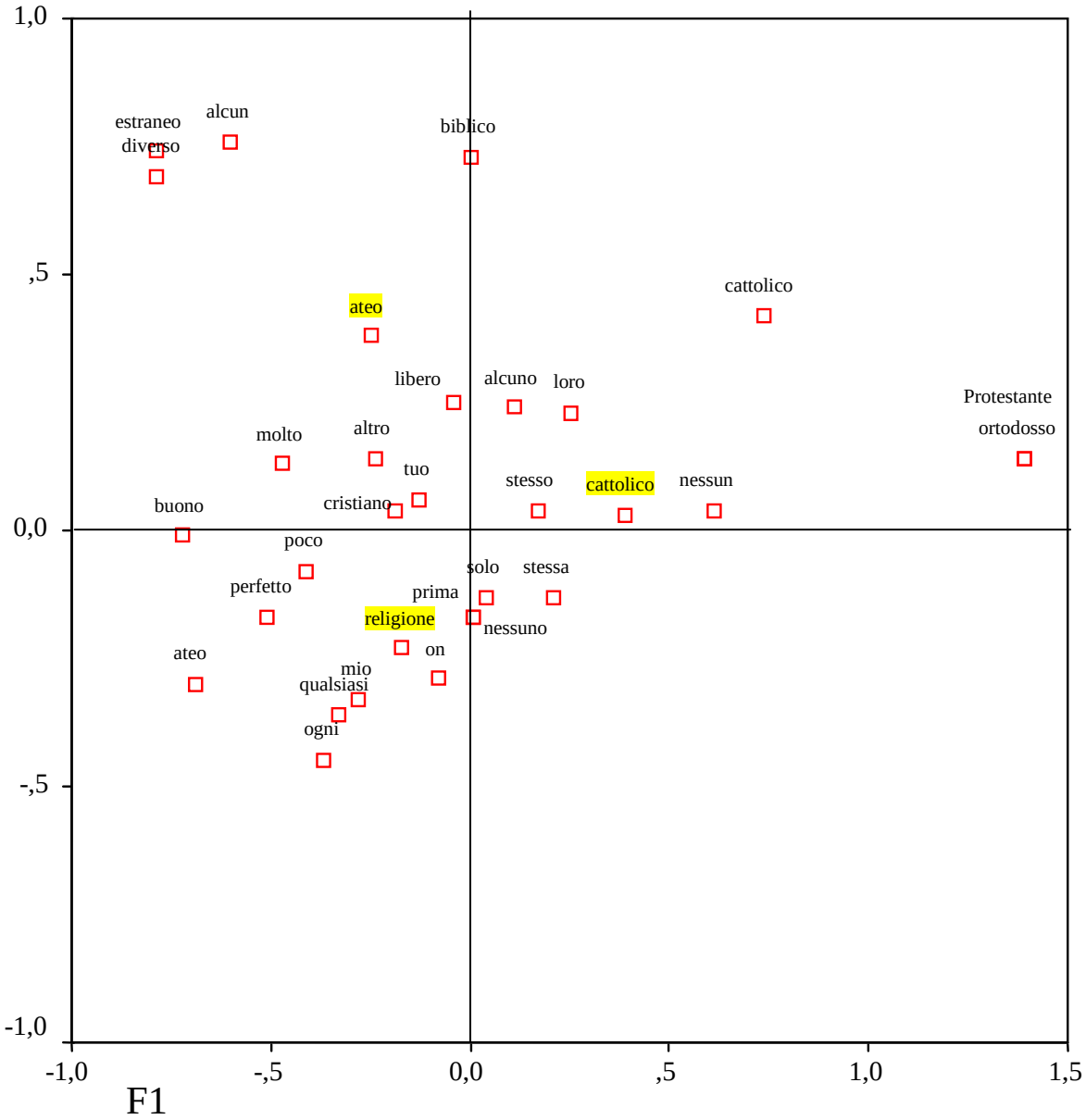
Il Grafico 2 riporta l'analisi delle corrispondenze parole x testi (ovvero Newsgroups), dove le parole sono solo verbi.

Usando la stessa logica interpretativa adottata per i soli nomi (cfr. commenti al Grafico 1), si può affermare che il Newsgroups dei Cattolici è un attrattore per i verbi quali "parlare" e "dire_che" e caratterizzante per il verbo "chiedere".

Per il Newsgroups degli Atei si conferma quanto detto per i nomi: la co-presenza di verbi quali "rifiuto", "negare" vicino a "meister" (volutamente riportato nel grafico) e "detto", lascia pensare ad alcune tematiche che riguardano Eickart.

Per il Newgroups religione, invece, è significativa la sua presenza insieme a verbi quali "significare", "conoscere", "confondere" e "seguire".

Grafico 3: Analisi delle corrispondenze parole x testi. AGGETTIVI (Newsgroups in giallo sul grafico)



Per quanto riguarda il Grafico 3, non si ritiene quest'ultimo particolarmente esplicativo da poter permettere delle analisi interpretative: sono troppo pochi e generici gli elementi in esso rappresentati.

In conclusione, si può affermare che la metodologia adottata si è rivelata efficace. Essa ha consentito, infatti, di trattare dati testuali molto articolati in tempi ridotti fornendo - nel contempo - al ricercatore gli elementi necessari al raggiungimento dello scopo proposto: evidenziare il contesto semantico della parola Dio attraverso gli strumenti messi a disposizione dalla Statistica.

BIBLIOGRAFIA

Alvisi F. Baldacci A. Cornacchini A.	*Perl*, Università degli Studi di Bologna Facoltà di Scienze Matematiche, Fisiche e Naturali, http://antares.csr.unibo.it/~alvisi/
Amaturo E. (1993)	*Messaggio, Simbolo, Comunicazione. Introduzione all'Analisi del Contenuto*, La Nuova Italia Scientifica, Roma
Giuliano L.C.	*Appunti del corso di Metodi Quantitativi per le Scienze Sociali*, a.a. 96-97
Benedetti C. (1989)	*Istituzioni di Statistica*, Veschi, Milano
Bolasco S. (1999)	*Analisi Multidimensionale dei dati*, Carocci, Roma
Bolasco S. (1997-1998)	*Metodi per l'analisi statistica dei dati testuali*, dispense a.a. 1997-1998 Corso di Statistica III – Facoltà di Economia, Scuola di Specializzazione di Metodi e Tecniche della Ricerca Sociale.
Cipriani R. Bolasco S. (1995)	*Ricerca qualitativa e computer*, Franco Angeli, Milano
Crescimanni A.	*Appunti di analisi dei dati*, dispense del corso di Analisi dei Dati del D.U. in Statistica a.a. 1997/98
Cutillo Enrica A. (1996)	*Lezioni di Statistica Sociale I,II*, CISU, Roma
Dern D.D. (1995)	*Alla scoperta di Internet*, McGraw-Hill, Milano
Fraire M. (1994)	*Metodi di Analisi Mulidimensionale dei Dati*, CISU, Roma
Frosini, Montinaro, Nicolini (1994)	*Il Campionamento da popolazioni finite*, UTET, Torino
JADT (2000)	*Actes des 5es Journées internationales d'Analyses statisque des Données Textuelles 9-11 Mars2000*, M. Rajman & J.C. Chappelier, École Polytechnique fédérale de Lausanne
Leti G. (1983)	*Statistica descrittiva*, Il Mulino, Bologna
Liverani M. (1996)	*Introduzione al Perl*, Dispense del corso di introduzione alla programmazione in Perl, http://www.faqit.to/docs/perl/perl.html
Memoli R. – Saporiti A. (1995)	*Disegno della Ricerca e Analisi dei Dati*, Euroma, Roma

Orsi R. (1985) *Probabilità ed inferenza statistica*, Il Mulino, Bologna

RFC 783 (1981) *The TFTP protocol (revision 2)*, http://www.pasteur.fr/infosci/RFC/

RFC 821 (1982) *Simple mail transfer protocol*, http://www.pasteur.fr/infosci/RFC/

RFC 822 (1982) *Standard for the format of Arpa Internet text messages*, http://www.pasteur.fr/infosci/RFC

RFC 959 (1985) *File Transfer Protocol (FTP),* http://www.pasteur.fr/infosci/RFC/

RFC 1036 (1987) *Standard for Interchange of USENET Messages*, http://www.pasteur.fr/infosci/RFC/

RFC 1123 (1989) *Requirements for Internet Hosts -- Application and Support*, http://www.pasteur.fr/infosci/RFC/

RFC 1341 (1989) *MIME (Multipurpose Internet Mail Extensions): Mechanisms for Specifying and Describing the Format of Internet Message Bodies*, http://www.pasteur.fr/infosci/RFC/

RFC 1421 (1993) *Privacy Enhancement for Internet Electronic Mail: Part I: Message Encryption and Authentication Procedures*, http://www.pasteur.fr/infosci/RFC/

RFC 1422 (1993) *Privacy Enhancement for Internet Electronic Mail: Part II: Certificate-Based Key Management*, http://www.pasteur.fr/infosci/RFC/

RFC 1423 (1993) *Privacy Enhancement for Internet Electronic Mail: Part III: Algorithms, Modes, and Identifiers*, http://www.pasteur.fr/infosci/RFC/

RFC 1424 (1993) *Privacy Enhancement for Internet Electronic Mail: Part IV: Key Certification and Related Services*, http://www.pasteur.fr/infosci/RFC/

RFC 1426 (1993) *SMTP Service Extension for 8bit-MIMEtransport*, http://www.pasteur.fr/infosci/RFC/

RFC 1521 (1993) *MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies*, http://www.pasteur.fr/infosci/RFC/

RFC 1563 (1994) *The text/enriched MIME Content-type*, http://www.pasteur.fr/infosci/RFC/

RFC 1740 (1994) *MIME Encapsulation of Macintosh files – MacMIME*, http://www.pasteur.fr/infosci/RFC/

RFC 1847 (1995) *Security Multiparts for MIME: Multipart/Signed and Multipart/Encrypted*, http://www.pasteur.fr/infosci/RFC/

RFC 1866 (1995) *Hypertext Markup Language - 2.0*, http://www.pasteur.fr/infosci/RFC/

RFC 1991 (1996) *PGP Message Exchange Formats*, http://www.pasteur.fr/infosci/RFC/

RFC 2015 (1996) *MIME Security with Pretty Good Privacy (PGP)*, http://www.pasteur.fr/infosci/RFC/

RFC 2183 (1997) *Communicating Presentation Information in Internet Messages:The Content-Disposition Header Field*, http://www.pasteur.fr/infosci/RFC/

RFC 2387 (1998) *The MIME Multipart/Related Content-type*, http://www.pasteur.fr/infosci/RFC/

Ricolfi L. (1997) *La ricerca qualitativa,* NIS, Roma

Smelser Neil J. (1995) *Manuale di Sociologia*, il Mulino, Bologna

Spad-T® *Introduction à SPAD-T intégré Version 1.5 P.C,* CISIA, Saint –Mandé (France)

Stabellini A. (2000) *Tesi di diploma Universitario, Università di Roma La Sapienza, facoltà di Scienze Statistiche*

What is SGML? *Extract from the OII Technology Handbook*, http://www.techapps.co.uk/iihb_sgml.html

Crittografia http://www.netzapping.com/banks/tesi/pesavento/crittog.htm http://www.newtech.it/utenti/netzapping/banks/tesi/pesavento/crittog.htm

What is the UUencode Compression Program? http://www.ualberta.ca/CNS/HELP/filetran/uuencode.html