4.
IL TRATTAMENTO DEL TESTO CON TALTAC

TALTAC (Trattamento Automatico Lessico-Testuale per l'Analisi del Contenuto) è un software per l'analisi testuale sviluppato da Sergio Bolasco, Francesco Baiocchi e Adolfo Morrone (http://www.taltac.it).

Seguendo le fasi principali di trattamento suggerite dal software è possibile delineare, a livello introduttivo, alcune strategie di base utili per qualsiasi indagine quali-quantitativa sui testi. TALTAC integra risorse e tecniche che fanno riferimento sia alla statistica che alla linguistica. A sua volta è predisposto per ricevere dati testuali che provengono da altri software, sia per esportare dati testuali disposti in matrici per l'analisi multidimensionale.

4. 1. LA BARRA DEGLI STRUMENTI

Le icone sulla barra degli strumenti permettono di identificare le fasi principali dell'analisi.

Normalizzazione: fase di pre-trattamento in cui si cerca di ridurre la variabilità del testo con l'applicazione di alcune procedure standard.

Misure lessicometriche: fase di analisi del vocabolario generato durante la fase di normalizzazione.

Segmentazione: fase di estrazione dei segmenti ripetuti.

Lessicalizzazione: fase di identificazione delle sequenze di forme (segmenti) definite dall'analista e di trasformazione di esse in forme grafiche semplici.

Tagging grammaticale: fase di riconoscimento delle forme grafiche e di applicazione delle categorie grammaticali.

Connessione lessicale: fase di confronto tra vocabolari al fine di valutarne l'intersezione

Risorse statistico-linguistiche: modulo in cui sono raccolte le tabelle/liste del DataBase di TALTAC e quelle generate durante la sessione di lavoro.

Text-Data Mining: modulo di gestione e ricerca sulle liste selezionate.

4. 2. PREPARAZIONE DEL CORPUS

Il primo passo da compiere è di preparare il file per la lettura da parte di TALTAC. La versione 1.6 permette di creare una sola partizione. La chiave della partizione è:

```
$P#
```

Pertanto il nostro corpus, che ora chiameremo *LEX1_TT*, si presenterà come segue:

```
$P#A_Statuto
Statuto del Regno di Sardegna
CARLO ALBERTO
Per la grazia di Dio Re di Sardegna, di Cipro e di Gerusalemme,
......
```

```
$P#B_Roma>
COSTITUZIONE DELLA REPUBBLICA ROMANA, 1849
PRINCIPII FONDAMENTALI
La sovranità è per diritto eterno nel popolo. Il popolo dello
......
$P#C_Italia
Costituzione della Repubblica italiana
Edizione del 1 gennaio 1948
Principî fondamentali
L'Italia è una Repubblica democratica, fondata sul lavoro.
La sovranità appartiene al popolo, che la esercita nelle forme e
......
$P#D_ONU
DICHIARAZIONE UNIVERSALE DEI DIRITTI DELL'UOMO
PREAMBOLO
Considerato che il riconoscimento della dignità inerente a tutti i
......
```

Il file *LEX1_TT* verrà copiato in una cartella che denominiamo *LEX1_TALTAC* e che sarà destinata ad accogliere tutti gli output della sessione di lavoro.

Dal menu **File** si apre una **Nuova Sessione** e le si assegna un nome (LEX1). Questa operazione costruisce un ambiente di lavoro che registrerà i riferimenti di percorso necessari nel Registro di Configurazione. In questo modo ogni volta che si vorrà ritornare alla sessione definita da quel nome basterà aprire la sessione indicata nell'elenco delle Sessioni aperte.

Dal menu **File** si seleziona il comando *Corpus e Liste esterne* e si apre la finestra di dialogo, dalla quale – tramite il pulsante di navigazione in corrispondenza della casella "Corpus" - si accede alle cartelle di lavoro per la selezione del corpus da caricare nel DataBase della sessione. Cliccando su OK il corpus selezionato appare nella barra di controllo dei file in uso in basso sullo schermo.

4. 3. FASE DI PRETRATTAMENTO: NORMALIZZAZIONE

Ora possiamo avviare la procedura di normalizzazione del corpus. La normalizzazione agisce sui caratteri "alfabetici". Il programma considera convenzionalmente come caratteri alfabetici tutti i caratteri che non vengono definiti come separatori.

I caratteri separatori di default (richiamabili dal Menu File – voce Separatori) sono i seguenti:

```
[TAB] [spazio] , . ; : ? ! ( ) [ ] { } \ / |
```

Dal menu **Moduli** – comando *Pretrattamento* selezioniamo la voce *Normalizzazione*. La finestra di dialogo della *Normalizzazione* (fig. 4.1) ci fornisce la selezione dei parametri di default per la **normalizzazione "leggera"** e per la **normalizzazione basata su liste**.

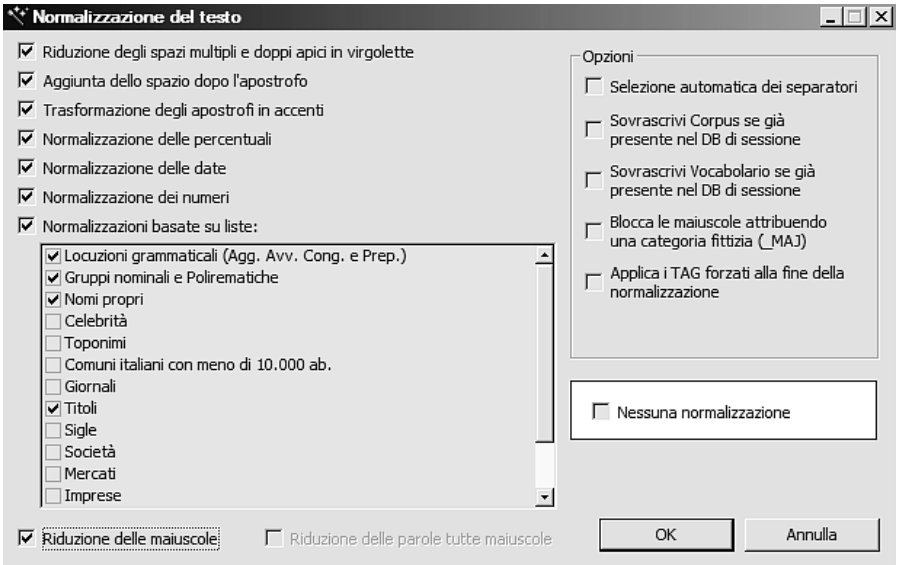

Fig. 4.1. – Normalizzazione del testo

La normalizzazione leggera è quella definita dai primi tre parametri:

1. Riduzione degli spazi multipli e dei doppi apici in virgolette.
2. Aggiunta dello spazio dopo l'apostrofo. Con questa opzione si aggiunge uno spazio vuoto dopo l'apostrofo.
3. Trasformazione degli apostrofi in accenti. Con questa opzione si trasformano le eventuali vocali seguite dall'apostrofo nelle corrispondenti vocali accentate.

La normalizzazione delle percentuali, delle date e dei numeri mira ad uniformare la grafia di queste forme per il loro corretto riconoscimento durante la fase di numerizzazione (codifica) del testo. In generale TALTAC introduce uno spazio intorno ai separatori (punteggiatura e parentesi).

La normalizzazione basata su liste si pone come obiettivo di categorizza-

re, già in questa fase di pretrattamento, le forme delle quali si vuole conservare la specificità. Le informazioni sulle liste sono inserite nel DataBase di normalizzazione di TALTAC.

Con la licenza "student" di TALTAC l'elaborazione è limitata a un corpus di 500 kb; durante l'esecuzione della fase di normalizzazione si apre una finestra che informa l'utente di questa limitazione.

Una delle funzioni principali in questa fase è il riconoscimento dei poliformi, locuzioni grammaticali (aggettivi, avverbi, congiunzioni e preposizioni), gruppi nominali e poliremathiche (tab. 4.1).

Tab. 4.1 – Esempi di poliformi (Bolasco, 1999, p. 195)

1) locuzioni grammaticali con funzioni di:	
- *avverbi:*	di più, non solo, per esempio, di nuovo, in realtà, più o meno, di fatto, del resto
	(luogo) a casa, in chiesa, al di là
	(tempo) di sera, un anno fa, al più presto
	(modo) in particolare, d'accordo, in piedi
- preposizioni:	fino a, da parte di, prima di, rispetto a, in modo da, per quanto riguarda
- *aggettivi*:*	in punto, di oggi, 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
2) idiomi e modi di dire:	io penso che, è vero che, non è che, per così dire, questo è tutto non c'è niente da fare, è un peccato
3) gruppi nominali polirematici:	buona fede, lavoro nero, mercato unico, punto di vista, cassa integrazione
4) verbi supporto e idiomatici:	si tratta di, tener conto, portare avanti, far fronte, far parte, prendere atto, dare vita, dare luogo, mettere a punto, venirne fuori, rendersi conto

* Alcuni aggettivi si possono anche trovare con funzione di avverbi

I **poliformi**, unità di senso (**lessie**) da considerare come unità minime del discorso non scomponibili, possono essere locuzioni grammaticali (*di_nuovo*, *di_fatto*, *del_resto*, *fino_a*, *in_modo_da*, *in_punto*, *il_fatto_che*, ecc.); oppure modi di dire (*io_penso_che*, *è_vero_che*, ecc.).

I poliformi possono dare luogo a **polirematiche**, lessie complesse che hanno complessivamente un significato diverso dalle parole che li compongono (Bolasco, 1999, p. 196) e che quindi il software di analisi automatica dei testi deve saper riconoscere e conteggiare come occorrenze: *Presidente_della_Repubblica*, *Camera_dei_Deputati*, *Consiglio_Regionale*.

Alcune polirematiche possono essere valide solo all'interno di lessici specifici: *in_vigore*, per esempio, può essere trattato come una polirematica nel lessico giuridico, ma non nel lessico medico.

In generale il riconoscimento dei poliformi, e delle polirematiche in particolare, offre un contributo essenziale alla **disambiguazione** delle forme grafiche riducendo la **polisemia** delle parole per conseguire un livello accettabile di **monosemia**, che rimane comunque una meta molto difficile da conseguire se non attraverso linguaggi artificiali e simbolici (per esempio il linguaggio della matematica e della logica formale).

In questo primo passaggio di normalizzazione possiamo lasciare inalterate le opzioni, salvo che per la marcatura della casella di "Riduzione delle maiuscole". Questo ci consentirà di uniformare le forme grafiche con le maiuscole all'inizio del periodo (*Il / il*; *Un / un*; ecc.).

Al termine della prima fase di normalizzazione viene aperta una finestra di *Definizione dei caratteri* nella quale si può intervenire per modificare la distinzione tra **caratteri separatori** e **caratteri alfabetici** (fig. 4.2). Questa **tabella dei caratteri** è identica alla tabella dei codice carattere ANSI Windows che deriva dalla tabella ASCII - ISO 8859-1 (Latin-1). I caratteri con il marcatore sono trattati come separatori. Se non abbiamo ragioni per modificare questa scelta possiamo cliccare su Continua, proseguendo così con il caricamento del corpus nel DataBase della sessione e la creazione del vocabolario.

Il corpus normalizzato LEX1_TT_norm.txt appare nella barra di controllo dei file in basso sullo schermo. Il file viene salvato nella cartella di lavoro LEX1_TALTAC, pertanto selezionando il comando Apri nel menu File è possibile seguire il percorso e aprire il file per esaminarlo. Osserveremo subito alcuni cambiamenti avvenuti nella forma del testo: Statuto del Regno di Sardegna appare come statuto del Regno di Sardegna. Questo significa che il software ha ridotto a minuscola la lettera s di Statuto all'inizio del periodo (del "titolo", in questo caso) ma non la lettera R di Regno.

Carlo Alberto è diventato Carlo_*Alberto_NM*. Il software ha riconosciuto il nome proprio e lo ha trasformato in una polirematica aggiungendo il tag grammaticale NM che significa "Nome Proprio".

La stessa cosa è accaduta per la polirematica *grazia_di_Dio_N* cui è stato aggiunto il tag N che significa "Nome" (sostantivo). Da notare anche

in_mezzo_agli_PREPstar che identifica il poliforme con ruolo di una preposizione.

Durante questa operazione, per la natura stessa dei testi, si possono verificare degli "errori" involontari: ad esempio, le forme *Romana_NM* e *Regina_NM* sono state identificate come nomi propri.

Dalla versione 1.6.3, esiste anche la possibilità di non effettuare nessuna normalizzazione, laddove è il caso.

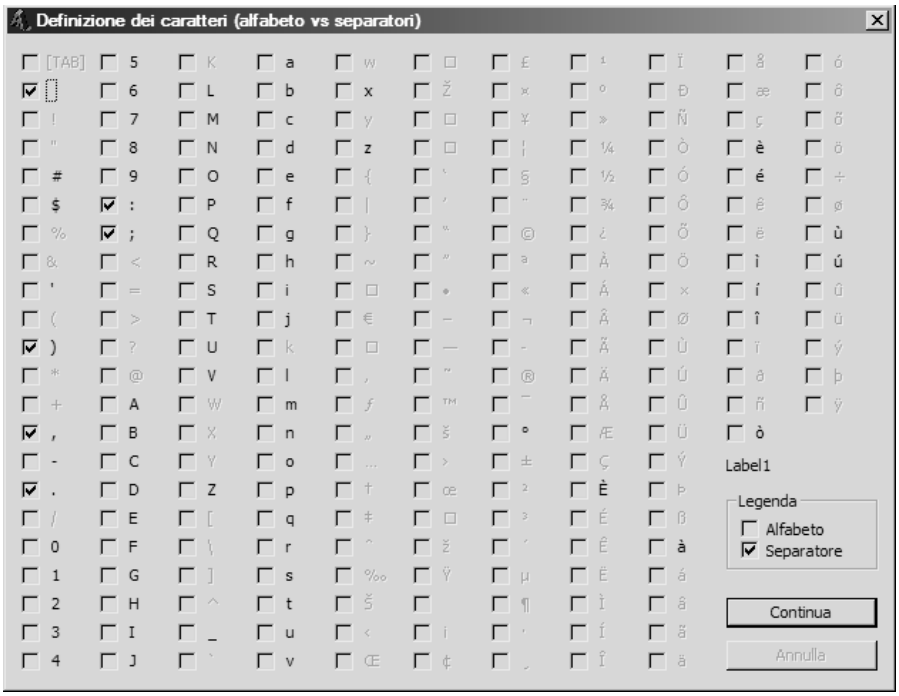

Fig. 4.2. – Finestra di definizione dei caratteri alfabetici e dei separatori

4. 4. ANALSI DEL VOCABOLARIO

A questo punto possiamo passare ad una prima analisi del vocabolario. Nel menu **Moduli**, selezioniamo il comando *Analisi del Vocabolario* e poi la voce Misure *Lessicometriche sul Voc. [TALTAC]* (fig. 4.3).

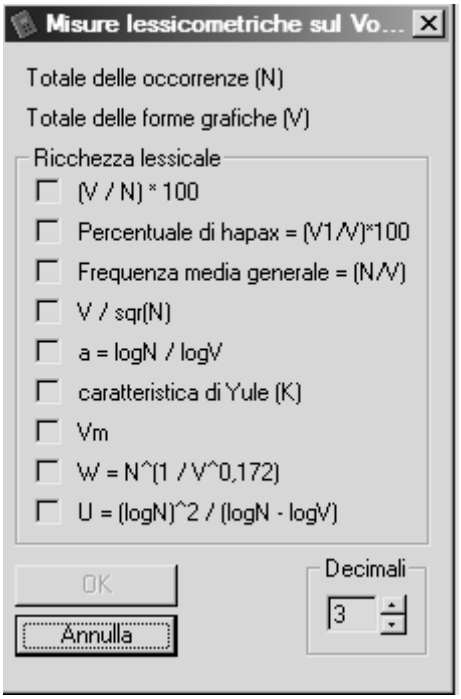

Fig. 4.3. – Misure lessicometriche sul Vocabolario

Il **totale delle occorrenze** o **dimensione del corpus** (N), come si è visto, è totale delle forme grafiche (parole, lessie, grafie) intese come "unità di conto" (*word token*).

Il **totale delle forme grafiche** o **ampiezza del vocabolario** (V) è il totale delle forme grafiche conteggiate come forme grafiche distinte (o parole distinte – *word type*).

La **ricchezza lessicale** viene misurata con alcuni indicatori in uso nella

statistica linguistica. I più rilevanti sono:

Estensione lessicale (*type/token ratio*)

$$\frac{V}{N}100$$

Percentuale di *hapax*:

$$\frac{V_1}{V}100$$

Frequenza media generale:

$$\frac{N}{V}$$

Coefficiente G (di Guiraud):

$$G = \frac{V}{\sqrt{N}}$$

Le misure di ricchezza del vocabolario sono sempre influenzate dal numero delle occorrenze. Ad esempio, la frequenza media generale è tanto più alta quanto più è esteso il corpus perché il totale delle occorrenze (per effetto delle alte frequenze delle parole forma) tende a crescere più rapidamente del totale delle parole distinte e, in ogni caso, le parole tendono a ripetersi con l'aumentare delle dimensione del corpus.

Il discorso inverso vale per la *type/token ratio* (che infatti è l'inverso della media generale delle parole): quanto più il corpus è grande tanto più il valore del rapporto è piccolo.

Alcuni autori hanno proposto delle misure indipendenti dall'ampiezza del vocabolario. Per un commento di questi indicatori occorre soffermarsi brevemente sulla relazione riscontrata tra rango e frequenza da un geniale linguista: George Kingsley Zipf (1902-1950). Se si ordinano secondo il rango le parole di un testo sufficientemente esteso (Zipf aveva preso come riferimento l'*Ulisse* di James Joyce, con 260.000 occorrenze), partendo dal rango più elevato, la frequenza della parole è ovviamente in relazione inversa. Questo non sorprende perché siamo stati noi ad assegnare il rango 1 alla parola con frequenza maggiore e così di seguito. Quello che sorprende è osservare che il prodotto della frequenza per il rango è approssimativamente costante:

Tab. 4.2. – Rapporto tra rango e frequenza delle parole nell'Ulisse di J. Joyce.

rango (*r*)	frequenza (*f*)	*f* × *r* = *c*
la parola 10ª	è usata 2.653 volte	26.530
la parola 100ª	265	25.500
la parola 1.000ª	26	26.000
la parola 10.000ª	2	20.000
la parola 29.000ª	1	29.000

Questa osservazione della **legge di Zipf** richiede naturalmente che a un certo rango (ad esempio, al rango 10) si assuma come valore di frequenza il valore medio delle occorrenze delle parole intorno al rango considerato (ad esempio da 1 a 20); infatti nel vocabolario di un corpus (come si è visto quando si è parlato delle fasce di frequenza) non vi sono tutte le classi di frequenza possibili e, nelle fasce di media e bassa frequenza, vi sono numerosi casi di parole che appartengono alla stessa classe di frequenza (Bolasco, 1999, p. 201). In seguito ad un'ampia discussione sulla validità di questa legge, i linguisti hanno ritenuto più opportuno esprimere l'equazione come:

$$f \times r^{a} = c$$

che in scala logaritmica si può esprimere anche come equazione della retta di regressione (fig. 4.4):

$$\log f = c + a \times \log r$$

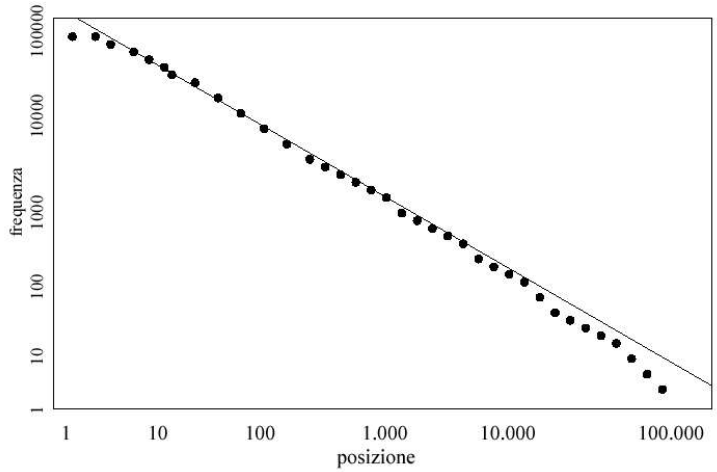

Fig. 4.4. – Legge di Zipf

per cui il coefficiente *a* indica l'angolo della retta (da intendersi con segno negativo) su un grafico a coordinate logaritmiche, in cui sull'asse *x* si riporta il logaritmo del rango e sull'asse *y* il logaritmo della frequenza; *c* è il punto in cui la retta interseca l'ordinata.

Il coefficiente *a* viene approssimato dal rapporto

$$\frac{\log N}{\log V}$$

e definisce, come si è detto, la pendenza della retta di regressione; in testi con un numero abbastanza elevato di occorrenze (50.000) il suo valore (con segno negativo) dovrebbe essere intorno a 1,15. Valori più elevati di 1,3 indicano che il vocabolario utilizzato non è particolarmente ricco (Tuzzi, 2003, p. 127). La G di Guiraud, per testi delle stesse dimensioni, assume un valore intorno a 22. Valori più grandi indicano una maggiore ricchezza lessicale, ma occorre tenere conto delle particolari tipologie di testi che vengono sottoposti a trattamento. La caratteristica di Yule nei testi più grandi è meno influenzata dalla presenza degli hapax (nella formula di K la i rappresenta la classe di frequenza e V_i il numero di forme grafiche appartenenti alla classe di frequenza i).

$$K = \frac{10^4 (\sum_{i=1}^{\infty} i^2 V_i - N)}{N^2}$$

Per approfondire l'interpretazione delle misure della ricchezza lessicale, molto complessa e molto discussa nella statistica linguistica, conviene riferirsi ai più recenti lavori di Cossette (1994) e di Labbé (1995).

Nella finestra di dialogo della figura 4.3 occorre spuntare le caselle delle misure lessicometriche richieste. Durante l'esecuzione di questa procedura si apre una finestra di dialogo che chiede di indicare l'ordine di grandezza delle frequenze normalizzate con un valore che approssima meglio il totale delle occorrenze (N) del corpus (nel nostro esempio con 14.733 occorrenze dovremo battere 10.000).

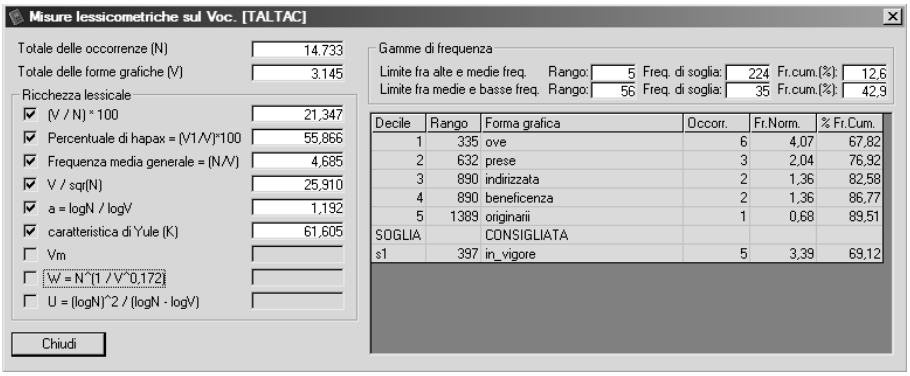

Fig. 4.5. – Misure lessicometriche sul vocabolario

Un confronto tra diversi corpora di dimensione crescente permette di apprezzare la validità e la sensibilità di alcuni indicatori di ricchezza lessicale.

Tab. 4.3. – Misure lessicometriche per diversi corpora secondo la dimensione

Corpora	**LEX1**	**Mongai**	**AnticoT**
Occorrenze N	14.733	142.485	627.325
Forme grafiche distinte	3.145	19.024	33.368
***type/token r.* = (V/N)*100**	21,35	13,32	5,32
% di hapax = (V1/V)*100	55,87	53,04	44,75
Frequenza media gen.= N/V	4,63	7,51	18,80
G di Guiraud	25,91	50,33	42,13
coefficiente *a*	1,19	1,20	1,28
Caratteristica di Yule (K)	61,60	43,05	53,48

Descrizione dei corpora:
Dalema98: corpus costituito da un unico testo tratto del discorso programmatico del primo governo D'Alema (programma "demo" di TALTAC).
LEX1: corpus costituito dai testi del Corpus LEX con eliminazione dei titoli privi di contenuto (Art., ART., Titolo, Sezione, ecc.) e delle rispettive numerazioni.
Mongai: corpus costituito dai testi di due romanzi di fantascienza di Massimo Mongai (*Il gioco degli immortali*, Mondadori, Milano, 1999; *Memorie di un cuoco d'astronave*, Mondandori, Milano, 1997).
AnticoT: corpus costituito dai testi della *Bibbia*, *Antico Testamento* (Pentateuco, Libri storici, Sapienziali, Profetici).
Tutti i testi sono stati scaricati dal sito LIBER LIBER del progetto Manuzio per la costituzione di una biblioteca telematica ad accesso gratuito (http://www.liberliber.it).

Aumentando le dimensioni del corpus la *type/token ratio* diminuisce. La frequenza media generale aumenta con il crescere delle dimensioni del corpus. La G è più alta nel corpus *Mongai* e nel corpus *Antico Testamento* che sono composti da testi letterari e poetici. La caratteristica di Yule è meno sensibile agli hapax nei testi di maggiori dimensioni, infatti un corpus dieci volte più grande di *LEX1* come *Mongai*, a parità di hapax, presenta un K sensibilmente inferiore.

In generale vale quanto già detto: i confronti di ricchezza lessicale si dovrebbero fare tra corpora o tra testi omogenei (tra testi giornalistici, tra testi di autori letterari, ecc.).

Esaminiamo il vocabolario dopo il trattamento di normalizzazione con le relative informazioni in questa fase dell'analisi. Nella colonna 3 (fig. 4.6) è riportata la lunghezza della forma grafica. Questa informazione – come si vedrà in seguito - è rilevante nella selezione delle parole. Le colonne 4-6 per il momento sono in gran parte vuote. Sono già allocate le categorie grammaticali e le lessie riconosciute automaticamente dal software nella fase di pretrattamento: i poliformi, i numeri, alcuni sostantivi chiaramente identificabili (*Stato*, *Governo*, *presidente della Repubblica*, ecc.).

Nella colonna 8 (fig. 4.6), accanto al rango viene indicata la gamma (**fascia**) di frequenza, nella colonna 9, il numero delle frequenze normalizzate su base 10.000 (da noi selezionata) e, nella colonna 10, le percentuali cumulate delle occorrenze sul totale.

Vocabolario [TALTAC] (sola lettura)

Forma grafica	Occorrenze totali	Lunghezza	Categoria grammaticale	Lemma/Lessia	Informazioni aggiuntive	Rango	Gamme di frequenza	cc.Tot.Norm. 10000	%Occ.Tot. Cum.
e	542	01				1	Alta	367,88	3,7
di	490	02				2	Alta	332,59	7,0
la	325	02				3	Alta	220,59	9,2
il	281	02				4	Alta	190,73	11,1
i	224	01				5	Media	152,04	12,6
della	224	05				5	Media	152,04	14,2
le	219	02				7	Media	148,65	15,6
per	178	03				8	Media	120,82	16,9
del	171	03				9	Media	116,07	18,0
è	169	01				10	Media	114,71	19,2
non	163	03				11	Media	110,64	20,3
legge	160	05				12	Media	108,60	21,4

Fig. 4.6. – Vocabolario di base LEX1

Le forme grafiche in ordine lessicometrico (ordinamento secondo la frequenza) si presentano anche in un ordinamento per **ranghi crescenti** (fig. 4.6, col. 7). Il rango è il posto occupato da una forma grafica nella graduatoria. La forma *e* (congiunzione) occupa il primo posto (rango 1) e appartiene alla **classe di occorrenze** *i*=542 come unica forma (nel senso che nessun'altra forma conta 542 occorrenze). La forma *legge* occupa il rango 12 ed è la prima

parola piena, cioè una parola che ci rivela qualche cosa della struttura semantica del corpus.

Al rango 5 troviamo due forme: *i* e *della*. Entrambe appartengono alla classe di occorrenze *i*=224 e definiscono il passaggio tra la fascia delle **alte frequenze** e la fascia delle **medie frequenze** (fig. 4.6, col. 8). La fascia delle medie frequenze inizia con la prima coppia di parole che hanno uno stesso numero di occorrenze.

Le parole che appartengono alla fascia delle alte frequenze di solito sono in massima parte parole vuote. Nei corpora di maggiori dimensioni nella fascia delle alte frequenze si riscontrano le **parole-chiave** che possono descrivere l'argomento principale dei testi in esame. E' rilevante, per la statistica testuale, osservare che le forme che appartengono ai primi 11 ranghi decrescenti sono parole vuote e rappresentano il 20,3% delle occorrenze.

Scorrendo rapidamente i ranghi decrescenti dal fondo della lista delle parole, partendo quindi dagli hapax per risalire verso l'alto, incontriamo classi di occorrenze crescenti consecutive: 1, 2, 3, ..., *i*, ... fino al rango 56, in corrispondenza della preposizione *nei*, cui segue (risalendo verso l'alto) una lacuna nelle classi di occorrenze crescenti (*i*=35). Dal rango 56 inizia quindi la fascia delle **basse frequenze**. Nella fascia delle basse frequenze, con classi di frequenze decrescenti fino a 1, si trova sempre la maggior parte delle parole distinte del vocabolario, in genere le parole principali. In questo caso sono 8.450 occorrenze, corrispondenti a 3.090 parole (pari al 98,25% delle parole distinte).

L'output delle misure lessicometriche presenta altre informazioni interessanti sulle gamme di frequenza (fig. 4.5, che si riporta in primo piano sul monitor cliccando sul segno "–" in alto a destra della finestra di Windows).

Le informazioni della figura 4.5 sono essenziali per compiere alcune valutazioni sulle **dimensioni minime** del corpus per una analisi automatica e sulla copertura del testo in funzione della scelta di determinate soglie alle quali collocare la selezione delle forme da analizzare (Bolasco, 1999 p. 203). Si tratta di scelte che non possono essere di natura esclusivamente quantitativa, ma che devono essere compiute con la massima attenzione tenendo conto anche di alcuni aspetti che riguardano le misure effettuate sul testo.

Quali sono le dimensioni minime che un corpus deve avere affinché sia adeguato per un'analisi statistica? Un testo troppo corto avrà ovviamente tutte parole diverse. Un criterio empirico suggerito dagli analisti (Bolasco, 1999, p. 203) è di osservare la *type/token ratio*: quando le parole distinte superano il 20% delle occorrenze il corpus non si può considerare sufficientemente esteso per un'analisi quantitativa. Nei corpora della tab. 4.3 possiamo notare come il corpus LEX1 sia appena adeguato. In generale un corpus di 15.000 occorrenze si

può considerare di "piccola dimensione"; un corpus di 50.000-100.000 occorrenze di media dimensione e un corpus maggiore di 200.000 occorrenze di grande dimensione. Un corpus sufficientemente grande (oltre le 500.000 occorrenze) può costituire una base per la costruzione di un lessico di frequenza rappresentativo di un linguaggio purché i testi siano abbastanza rappresentativi della sua eterogeneità. Le unità di un lessico di frequenza sono espresse in lemmi.

L'analisi testuale di un corpus, nelle applicazioni della statistica multidimensionale, non può prendere in esame l'intero vocabolario. Pertanto la copertura del testo non potrà mai essere del 100%. In genere, in tutte le indagini statistiche, l'obiettivo del ricercatore è di selezionare un piccolo numero di variabili che siano sufficientemente rappresentative dei caratteri essenziali del fenomeno oggetto di studio. Nell'analisi testuale questo obiettivo si consegue attraverso la scelta di una soglia di frequenza al di sotto della quale le parole possono essere abbandonate senza una significativa perdita di informazione.

Il **tasso di copertura del testo** (% COP) è pari alla percentuale di occorrenze che derivano dalle parole $V_{(s)}$ al di sopra della soglia s sul totale N di tutte le occorrenze del corpus.

Dalle misure lessicometriche sul vocabolario di *LEX1_TT* (fig. 4.5) apprendiamo che sul primo decile delle basse frequenze (il 10% delle parole distinte di bassa frequenza che sono al di sopra della soglia 6 indicata al rango 335 con la forma grafica *ove*) si ottiene una copertura del testo pari al 67,82%. Sulla mediana (5° decile) delle basse frequenze in corrispondenza della forma *originarii* si ottiene l'89,51% di copertura del testo. La soglia consigliata da TALTAC per l'analisi multimensionale è la soglia 5 che offre una copertura del testo del 69,12%. La copertura del testo è bassa, ma questo dipende dal fatto che il corpus è piccolo. Con il corpus AnticoT la soglia consigliata è 13 con una copertura del corpus dell'87,45%.

Ritornando all'analisi del vocabolario (la finestra con la figura 4.6 si riporta in primo piano cliccando su "ingrandisci" della barra "Vocabolario" in basso a sinistra del monitor), dal menu **Record** possiamo selezionare alcune operazioni per l'esplorazione del lessico. Ad esempio, selezioniamo la colonna "Lunghezza" cliccando sull'intestazione della tabella, poi dal menu **Record** selezioniamo il comando *Ordina in senso decrescente*: otteniamo un ordinamento delle forme grafiche che mette immediatamente in evidenza i poliformi che sono stati individuati da TALTAC (fig. 4.7). Nella figura 4.7 (e nelle figure seguenti) alcuni campi sono stati "nascosti" con il comando *Nascondi campo* del menu **Formato**. In qualsiasi momento i campi nascosti possono essere visualizzati con il comando *Scopri campo*.

Vocabolario [TALTAC] (sola lettura)

Forma grafica	Occorrenze totali	Lunghezza	Categoria grammaticale	Gamme di frequenza	cc.Tot.Norm. 10000
Consiglio_nazionale_dell_Economia_e_del_Lavoro	1	46	N	Bassa	0,68
Consiglio_superiore_della_Magistratura	6	38	N	Bassa	4,07
presidente_del_Consiglio_dei_ministri	5	37	N	Bassa	3,39
amministrazione_della_giustizia	2	31	N	Bassa	1,36
presidente_della_Repubblica	31	27	N	Bassa	21,04
Pubblica_Amministrazione	5	24	N	Bassa	3,39
giustizia_amministrativa	2	24	N	Bassa	1,36
giuridico-amministrativa	1	24		Bassa	0,68
amministrazioni_centrali	1	24	N	Bassa	0,68
Presidenza_del_Consiglio	1	24	N	Bassa	0,68
ministro_della_Giustizia	2	24	N	Bassa	1,36
presidente_del_Consiglio	1	24	N	Bassa	0,68
ordinamento_giudiziario	7	23	N	Bassa	4,75
a_suffragio_universale	3	22	AVV	Bassa	2,04
Consiglio_dei_Ministri	4	22	N	Bassa	2,71
ordinamento_giuridico	2	21	N	Bassa	1,36
procuratore_generale	2	20	N	Bassa	1,36
territorio_nazionale	4	20	N	Bassa	2,71
bilancio_dello_Stato	1	20	N	Bassa	0,68

Fig. 4.7. – Vocabolario LEX1: poliformi.

4. 5. IL RICONOSCIMENTO DELLE FORME GRAMMATICALI

Il riconoscimento delle forme grammaticali (tagging grammaticale) consiste nella attribuzione di ciascuna forma ad una categoria grammaticale. Questa operazione è fondamentale per la procedura di disambiguazione delle parole. Le categorie grammaticali presenti nel DataBase di TALTAC sono le seguenti:

N = sostantivo
A = aggettivo
V = verbo
AVV = avverbio
DET = determinante
PREP = preposizione
CONG = congiunzione
PRON = pronome
ESC = interiezione
J = ambigua
FORM = forma idiomatica

NM = nome proprio
DAT = data
NUM = numerale
O = parola straniera

La categoria "ambigua" (J) identifica le parole compatibili con più categorie. Ad esempio, *legge* (sostantivo femminile) e *legge* (terza persona singolare dell'indicativo presente del verbo *leggere*). Le forme non riconosciute (parole rare, parole straniere, parole di un lessico specialistico, errori di ortografia, ecc.) non vengono attribuite ad alcuna categoria (il campo di attribuzione rimane vuoto).

Dal menu **Moduli** – *Analisi lessicale* selezioniamo *Tagging grammaticale – Vocabolario [TALTAC]*; oppure dalla barra degli strumenti selezioniamo l'icona corrispondente. La finestra di dialogo ci offre delle scelte. Per il momento possiamo lasciare inalterato il tagging di base:

Fig. 4.8. – Finestra di dialogo del tagging grammaticale

	Forma grafica	Occorrenze totali	Lunghezza	Categoria grammaticale	Entrate nel Lemmario	Gamme di frequenza	cc.Tot.Norm. 10000
▶	e	542	01	J	2	Alta	367,88
	di	490	02	J	2	Alta	332,59
	la	325	02	J	3	Alta	220,59
	il	281	02	DET	1	Alta	190,73
	i	224	01	J	2	Media	152,04
	della	224	05	PREP	1	Media	152,04
	le	219	02	J	2	Media	148,65
	per	178	03	PREP	1	Media	120,82
	del	171	03	PREP	1	Media	116,07
	è	169	01	V	1	Media	114,71
	non	163	03	J	2	Media	110,64
	legge	160	05	J	2	Media	108,60

Fig. 4.9. – Vocabolario LEX1 dopo la fase di tagging grammaticale

L'output del tagging grammaticale ci restituisce la tabella *Vocabolario* con ordinamento lessicometrico (fig. 4.9). Tuttavia, selezionando la colonna desiderata sulla tabella *Vocabolario*, possiamo sempre effettuare, dal menu **Record**, un riordinamento in senso crescente o decrescente.

Ora abbiamo altre informazioni sul vocabolario. La colonna "Categoria grammaticale" contiene per ciascuna forma una etichetta di classificazione. L'articolo determinativo il è classificato come DET. L'aggettivo indefinito ogni è classificato A.

Se clicchiamo con il tasto destro del mouse su una forma otteniamo tutte le informazioni contenute nel DataBase del lessico di riferimento di TALTAC (fig. 4.10).

	Forma grafica	Occorrenze totali	Categoria grammaticale	Entrate nel Lemmario	Gamme di frequenza	cc.Tot.Norm. 10000
	legge	160	J	2	Media	108,60
	l'	157	DET	1	Media	106,56
	o	149	J	3	Media	101,13
	dei	141	J	2	Media	95,70
▶	sono	125	J	3	Media	84,84
	che					80,77
	essere					74,66
	delle					73,30
	in					71,27
	può					69,23
	con	99	PREP	1	Media	67,20
	alla	97	J	2	Media	65,84
	a	93	J	2	Media	63,12
	diritto	85	J	3	Media	57,69

Forma	CAT	Lemma potenziale	OrigTAG	Imprinting	InfoAgg
sono	N	sono	1	s_m	
sono	V	essere	1	indic_pres_s/pl_1/3	
sono	V	sonare	1	indic_pres_s_1	

Fig. 4.10. – Dettaglio del tagging grammaticale

Molte forme grafiche hanno un'etichetta J. Questa etichetta indica che non è stato possibile attribuire una categoria grammaticale perché la forma grafica è ambigua. Scoprendo il campo CAT_AC se ne conoscono tutte le possibili categorie che in teoria la forma può assumere.

La colonna 6 indica le "entrate del lemmario" e cioè quante sono le forme grafiche che, per ciascuna voce, possono essere identificate come lessie "autonome"

Per la forma *sono* troviamo tre lessie:

- *sono*, sostantivo maschile per una forma letteraria di *suono*;
- *sono*, indicativo presente, prima persona singolare e terza persona plurale del verbo *essere*;
- *sono*, indicativo presente, prima persona singolare del verbo *sonare*.

Se la forma grafica è classificata in una lessia univoca allora la forma è contrassegnata da un lemma. Ad esempio, le forme *applica* ed *applicarsi* vengono classificate nel verbo *applicare*; *anni* e *anno* nel sostantivo *anno*.

La lemmatizzazione è parzialmente indipendente dal tagging. Una forma ambigua può essere ricondotta ad un lemma senza modificare la sua classificazione come ambigua. Ad esempio, *altra* e *altre*, pur essendo forme ambigue (J) vengono classificate nel lemma *altro* poiché la forma canonica delle differenti CAT di queste forme è sempre la stessa.

Come si è detto, nella fase di tagging grammaticale alcune forme grafiche non vengono riconosciute dal software. Lo possiamo vedere evidenziando nella tabella del vocabolario "Categoria grammaticale" e selezionando *Ordina in senso crescente* dal menu **Record**. Così vengono visualizzati i record con il "campo vuoto". Ad esempio, la forma *giuridico-amministrativa* ha una grafia non identificabile nel DataBase; le forma *dovario* è un termine tecnico che indica la dote regale; la forma *sempreché* è una forma desueta.

4. 6. LA LEMMATIZZAZIONE

Nella fase di tagging grammaticale il software ha riconosciuto i lemmi, pertanto è possibile generare una lista di forme grafiche classificate secondo il lemma corrispondente. Dal menu **Calcola**, selezioniamo la voce *Fusioni di – Lemma/Lessia*: ci viene suggerito un nome per la tabella da salvare "Fusioni di Lemma/Lessia di Vocabolario [TALTAC] (con TAG grammaticale)" che sarà inserita nel DataBase della sessione.

	Lemma/Lessia	Numero di unità lessicali	Occorrenze totali	Categoria grammaticale	Dispersione	Uso
	espressa	1	1,00	J	0,00	0,00
	espressamente	1	2,00	AVV	0,00	0,00
	espresse	1	2,00	J	0,00	0,00
	espressi	1	1,00	J	0,00	0,00
	espressione	1	1,00	N	0,00	0,00
	espropriare	1	1,00	V	0,00	0,00
	espropriazione	1	1,00	N	0,00	0,00
	essenziale	2	2,00	J	0,00	0,00
	essere	1	110,00	J	0,83	91,38
	essere	1	2,00	N	0,00	0,00
	essere	6	204,00	V	0,79	160,98
	esserlo	1	1,00		0,00	0,00
	esso	4	25,00	J	0,73	18,13
	estendere	1	1,00	V	0,00	0,00
	estensione	1	1,00	N	0,00	0,00
	esteri	1	1,00	J	0,00	0,00
	estero	1	1,00	A	0,00	0,00

Fig. 4.11. – Fusioni di lemma/lessia

La tabella presenta 2.619 record con tutte le forme grafiche (comprese quelle non disambiguate "J" e le forme non riconosciute "campo vuoto") classificate (fuse) nel lemma corrispondente. Ad esempio, dove è stato possibile, tutte le coniugazioni del verbo *essere* sono state classificate nel lemma corrispondente (V, 204 occorrenze), 110 occorrenze sono classificate come grammaticalmente ambigue (J) e 2 occorrenze come sostantivo (N).
Tuttavia questa lemmatizzazione "grezza", effettuata su tutto l'insieme delle categorie grammaticali, non è attendibile: il 42,5% dei lemmi è classificato come ambiguo e 151 forme non sono riconosciute. Le classificazioni ambigue sono eccessive e per ridurle sarebbero necessari degli interventi puntuali di disambiguazione che possono essere effettuati solo con un attento esame delle concordanze. Vedremo nel cap. 6 come è possibile ovviare parzialmente a questo problema con il tagging avanzato. In generale, la procedura di lemmatizzazione, essendo sostanzialmente un'operazione di misura tramite classificazione, va condotta con cura e con la piena consapevolezza del ricercatore rispetto agli scopi che intende perseguire. L'attenzione alla qualità del dato deve sempre prevalere rispetto alla quantità.

5.
SEGMENTAZIONE E ANALISI DI SPECIFICITÀ CON TALTAC

Le fasi di normalizzazione, di costruzione del vocabolario e di tagging grammaticale sono fondamentali per un esame completo del corpus in vista di qualsiasi strategia di analisi. Quando ci troviamo di fronte a vocabolari molto ampi, composti di oltre 10.000 forme grafiche distinte, l'interpretazione automatica dei testi richiede necessariamente la selezione di un sottoinsieme di parole con un alto contenuto di informazione che sia rappresentativo del contenuto del corpus. L'estrazione dei segmenti ripetuti, la lessicalizzazione e l'individuazione delle forme peculiari rappresentano momenti significativi in vista di questo obiettivo.

5. 1. ESTRAZIONE DEI SEGMENTI RIPETUTI E LESSICALIZZAZIONE

Come si è visto nella procedura di Lexico3 (vedi 3.9), i segmenti ripetuti sono sequenze di forme grafiche formate da tutte le disposizioni a 2,3, ..., *q* forme che si ripetono per un certo numero di volte nel corpus (Bolasco, 1999, p. 194). TALTAC permette di eliminare i segmenti ridondanti e i segmenti vuoti utilizzando, durante la procedura di estrazione, due file (adeguatamente modificabile dall'utente) di parole vuote all'inizio (VuoteI.txt) e alla fine (VuoteF.txt) del segmento.

Dal menu **Moduli** – *Analisi dei segmenti*, selezionando la voce *Individuazione dei segmenti* (o dalla corrispondente icona sulla barra degli strumenti) si accede alla finestra di selezione delle modalità di esecuzione della procedura:

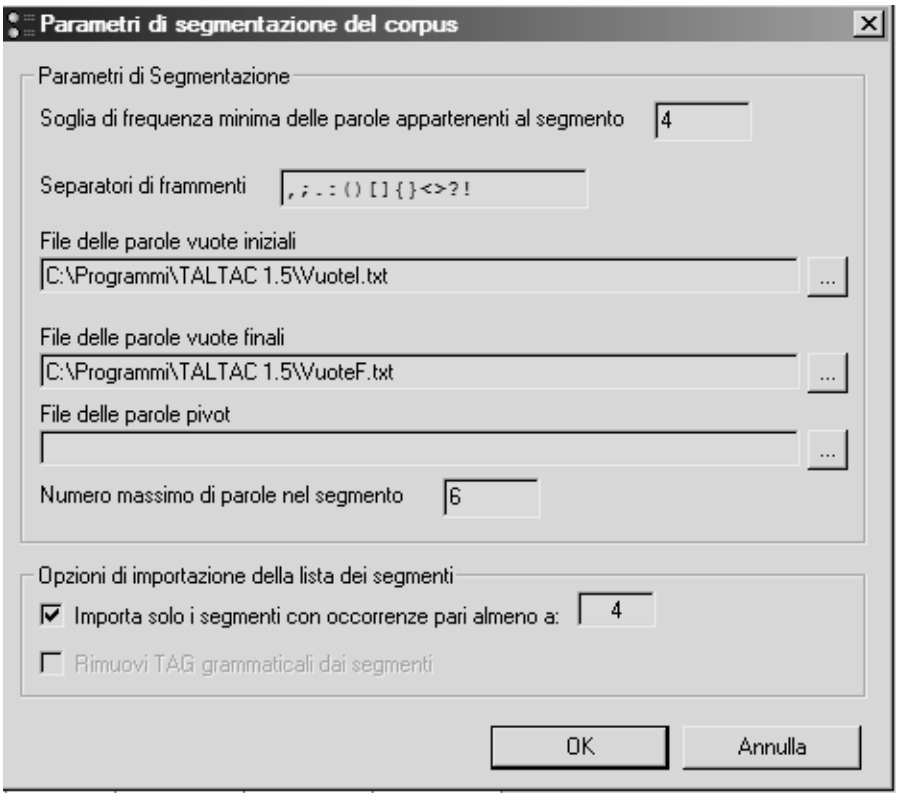

Fig. 5.1. – Parametri di segmentazione del corpus

L'individuazione dei segmenti ripetuti, soprattutto nei file molto grandi, può essere una procedura *time expensive*. E' sempre opportuno scegliere una soglia minima di occorrenze per i segmenti da individuare e importare nella lista. I segmenti ripetuti vengono individuati come tutte le disposizioni a 2, 3, ..., *q* parole (numero massimo di parole nel segmento) con classe di frequenza *i* (soglia di frequenza minima delle parole appartenenti al segmento) che si ripetono *n* volte nel corpus (importa solo i segmenti con occorrenze pari almeno a...); i segmenti individuati dipendono da *i* , ma i segmenti visualizzati nella lista dipendono dalla soglia *n*. E' evidente che se si vogliono visualizzare i segmenti che si ripetono almeno *n* volte è del tutto inutile fissare $i < n$ (un segmento costituito da almeno una parola che ha 3 occorrenze non può presentarsi con 4 occorrenze). Nel nostro esempio scegliamo di selezionare 4 come

soglia minima di frequenza delle parole e 4 come soglia minima di occorrenze dei segmenti da importare nella lista. Con questi parametri vengono individuati 225 segmenti (fig. 5.2).

Segmento	Occorrenze totali	Lunghezza	Indice IS	Indice IS relativo	Informazioni aggiuntive
delle Camere	14	2			
dell' Assemblea	8	2			
di cui	5	2			
di ciascuna Camera	4	3			
di Cassazione	4	2			
di approvazione	4	2			
devono essere	4	2			
deve essere	12	2			
dello Stato_N	32	2			
delle Regioni	4	2			
delle loro	6	2			
delle libertà	4	2			
di due terzi	5	3			
delle due Camere	6	3			
di esercitare	4	2			
della sua	10	2			
della società	4	2			
della Repubblica	53	2			
della Regione	8	2			
della libertà	7	2			
della famiglia	4	2			
della Costituzione	15	2			
della Corte	7	2			
della Camera_dei_Deputati_N	10	2			
della Camera	7	2			
dell' uomo	8	2			
a referendum	4	2			
delle leggi	12	2			
di Stato_N	4	2			
è stato	4	2			
è sempre	4	2			
è promulgata	4	2			
è libera	5	2			
è inviolabile	5	2			
due terzi	7	2			
due Camere	10	2			

Records visibili: 225 su 225

Fig. 5.2. – Segmenti ripetuti

Come di consueto, i segmenti possono essere ordinati per occorrenze o per lunghezza dal menu **Record**.

Sulla selezione dei segmenti ripetuti è possibile calcolare un **indice di significatività** (IS) dei segmenti per valutare la loro rilevanza nel corpus. Dal menu **Moduli** – *Analisi dei segmenti* selezioniamo il comando *Calcolo indice IS su – Lista dei segmenti (TALTAC)* ; alla richiesta di calcolare l'indice IS anche sugli hapax rispondiamo "No". L'output della lista dei segmenti si arricchisce di due colonne: l'indice IS e l'indice IS relativo (fig. 5.3).

Lista dei segmenti [TALTAC] (con indice IS)

Segmento	Occorrenze totali	Lunghezza	Indice IS	Indice IS relativo	Informazioni aggiuntive
ogni individuo ha diritto	18	4	4,71	0,29	
nessun individuo potrà essere	6	4	4,62	0,29	
Parlamento in seduta comune	6	4	4,09	0,26	
in seduta comune	9	3	3,31	0,37	
nessuno può essere	10	3	3,07	0,34	
diritti civili e politici	5	4	3,07	0,19	
dal Parlamento in seduta comune	4	5	3,01	0,12	
a_maggioranza_AVV assoluta dei suoi	4	4	3,01	0,19	
autorità giudiziaria	7	2	2,83	0,71	
può essere arrestato	5	3	2,78	0,31	
Consigli regionali	7	2	2,75	0,69	
due terzi	7	2	2,56	0,64	
trattati internazionali	6	2	2,42	0,61	
a_maggioranza_AVV assoluta	6	2	2,42	0,61	
potere legislativo	4	2	2,40	0,60	
Assemblea Costituente	6	2	2,34	0,59	
entrata_in_vigore_N della Costituzione	9	3	2,32	0,26	

Fig. 5.3. – Segmenti ripetuti: indice di significatività (IS)

Come si interpreta l'indice assoluto IS ? L'indice mostra il grado di assorbimento del segmento ripetuto rispetto alle parole che lo costituiscono (Morrone, 1993; Bolasco, 1999, p. 221).

$$IS = \left[\sum_{i=1}^{L} \frac{f_{segm}}{f_{fg_i}} \right] \times P$$

Per ciascuna delle forme (*L*) che compongono il segmento si considera il rapporto tra le occorrenze del segmento (f_{segm}) sulla forma grafica che ne fa parte (f_{fg}). La somma da 1 a *L* di questi rapporti viene moltiplicata per le parole piene (*P*) che costituiscono il segmento.

Il segmento *Parlamento in seduta comune* (6) è composto da *Parlamento* (23), *in* (105), *seduta* (12) e *comune* (11), pertanto:

$$IS = \left[\frac{6}{23} + \frac{6}{105} + \frac{6}{12} + \frac{6}{11} \right] \times 3 = 4{,}09$$

L'**indice IS relativo** viene rapportato al suo massimo (L^2) e quindi varia tra 0 e 1. I due indicatori offrono informazioni diverse. L'indice IS assoluto è fortemente condizionato dal numero di parole piene che costituiscono il segmento, pertanto mette in evidenza i segmenti più lunghi, costituiti da un maggior numero di parole, ma anche meno frequenti. L'indice IS relativo mette ai primi ranghi i segmenti più corti che spesso rappresentano i termini specialistici del lessico. Infatti chiedendo l'ordinamento dei segmenti secondo il valore decrescente dell'indice IS relativo troveremo nei primi ranghi *autorità giudiziaria*, *Consigli regionali*, *due terzi*, *a maggioranza assoluta*, *trattati internazionali*. I segmenti ripetuti con un grado di assorbimento più elevato sono evidentemente dei poliformi che conviene trattare come una sola occorrenza (un'unica parola) piuttosto che attraverso le forme grafiche che li compongono.

Il trattamento dei poliformi avviene attraverso la procedura di **lessicalizzazione**, attraverso la quale il software viene istruito a riconoscere i segmenti che vogliamo trasformare in lessie complesse fino a modificare il vocabolario di base per le operazioni successive. Il segmento *autorità giudiziaria*, ad esempio, con la lessicalizzazione viene modificato in *autorità_giudiziaria*.

La procedura di lessicalizzazione inizia con la marcatura dei segmenti nella colonna delle "Informazioni aggiuntive" scrivendo nel campo in corrispondenza del segmento scelto un codice a scelta, ad esempio "S". Per rendere possibile la scrittura nei campi della lista dei segmenti il file deve essere aperto dall'icona "Risorse statisco-linguistiche" (sulla barra degli strumenti) smarcando la casella "Sola lettura" in basso a destra della finestra di dialogo. Ultimata la fase di marcatura dei segmenti prescelti:

1) Selezionare l'intestazione di colonna "Informazioni aggiuntive".
2) Cliccare sull'icona "Text-Data Mining" sulla barra degli strumenti per aprire la corrispondente finestra di dialogo.
3) Marcare l'opzione di ricerca "Records LIKE" inserendo nella finestra un codice qualisasi; per es. "s".
4) Cliccare su OK.

A questo punto la lista selezionata contiene esclusivamente i segmenti da lessicalizzare. Pertanto la lista dovrà essere salvata dal menu **File** selezionando *Esporta in un file di testo* e poi la voce *Lista di lessicalizzazione*. La lista verrà salvata nella cartella di lavoro con il nome *LEX1_ldl.txt*.

La procedura prosegue con la lessicalizzazione dei segmenti selezionati e

la successiva rinumerizzazione che riporta i segmenti selezionati nel vocabolario della sessione.

Dal menu **Moduli**, selezionare il comando *Analisi dei Segmenti* e poi la voce *Lessicalizzazione* viene richiamata una finestra di dialogo con l'indicazione del "Testo da lessicalizzare", del file con la "Lista dei segmenti da unire nel testo" (che di default è quello salvato nella fase precedentemente, ma potrebbe essere anche un file creato appositamente dall'utente) e del "File di output con il testo lessicalizzato". A questo punto il software chiede che il testo da lessicalizzare sia in formato record. Salvo casi particolari il file è già in formato record, pertanto la risposta sarà NO. Ad operazione conclusa nel testo vengono apportare le modifiche richieste, controllabile facilmente aprendo la lista "Vocabolario [TALTAC]" ed effettuando una esplorazione delle forme grafiche attraverso lo strumento "Text-Data Mining".

5. 2. ESTRAZIONE DELLE FORME PECULIARI

Per un approfondimento dell'analisi del corpus possiamo individuare quali sono le parole caratteristiche per ciascuna della partizione (per ciascun testo, in questo caso).

L'analisi di specificità rivolta alla individuazione delle forme peculiari è sempre basata sulla sovra o sotto utilizzazione delle forme rispetto a un modello di riferimento (Bolasco, 1999, p. 223). I modelli di riferimento possono essere i lessici di frequenza oppure l'intero corpus rispetto a una sua partizione, come nel nostro caso. La misura di specificità è data dal seguente scarto standardizzato della frequenza relativa,

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

dove f_i è il numero delle occorrenze normalizzate della *i*-esima forma grafica nella partizione e f_i^* è il valore corrispondente nel modello di riferimento (corpus o lessico di frequenza). Il valore al denominatore è lo scarto quadratico medio della frequenza relativa; poiché la frequenza relativa di una parola, nell'analisi di un corpus, è sempre bassissima, di fatto lo s.q.m. equivale alla radice quadrata della frequenza teorica (Bolasco, 1999, p. 227). Come già si è visto per quanto riguarda il software Lexico3, la stima della significatività dello

scarto standardizzato viene calcolata in base alla legge di distribuzione ipergeometrica che è la legge della variabile *N* = "numero di oggetti del tipo V che trovo raccogliendo a caso *n* oggetti tra una quantità *q* di oggetti", dato come noto il numero *v* degli oggetti di tipo V. Per l'analisi delle forme grafiche si assume che V siano le forme grafiche distinte, *v* siano le rispettive occorrenze, *n* siano le occorrenze delle parole che costituiscono il corpus e *q* siano le occorrenze delle parole che costituiscono il lessico (Tuzzi, 2003, pp. 131-134).

Dal menu **Moduli** selezioniamo il comando *Analisi della Specificità – Estrazione delle parole caratteristiche*. Ci viene chiesto di fissare un livello alfa di probabilità (di default è fissato a 0,025) e una soglia di frequenza delle parole (di default è fissata a 5). Nella finestra successiva ci viene chiesto se calcolare la dispersione con il metodo Montecarlo. Rispondiamo "No" e proseguiamo (per informazioni dettagliate su questo indice di dispersione vedi Tuzzi, 2003, p. 134 sg.). Al termine dell'elaborazione compaiono due finestre di informazioni tecniche che possiamo chiudere, passando ad esaminare l'output della lista di Specificità (fig. 5.4).

Normalizzazione a lettura)

	Forma grafica	Occorrenze totali	A_Statuto	B_Roma	C_Italia	D_ONU	Parole caratteristiche	p-value (A_Statuto)	Speci (A_St	p-value (B_Roma)	Speci (B_Rc	p-value (C_Italia)	Speci (C_Ita	p-value (D_ONU)	Speci (D_Of
▶	e	542	83	45	331	83	spec			0,02118	neg			0,01393	pos
	di	490	94	56	264	76	spec					0,00183	neg	0,01360	pos
	la	325	45	34	220	26	spec					0,00307	pos	0,00987	neg
	il	281	62	27	163	29	spec	0,00934	pos						
	i	224	40	27	142	15	spec							0,00508	neg
	della	224	28	29	139	28	banale		ban		ban		ban		ban
	le	219	43	16	150	10	spec					0,00686	pos	0,00009	neg
	per	178	28	27	111	12	spec							0,01289	neg
	del	171	38	23	101	9	spec							0,00189	neg
	è	169	35	30	96	8	spec			0,00529	pos			0,00083	neg
	non	163	34	23	98	8	spec							0,00138	neg
	legge	160	17	15	123	5	spec	0,02229	neg			0,00001	pos	0,00005	neg

Fig. 5.4. - Specificità

La lista di Specificità ci dà le occorrenze sul totale del corpus e le occorrenze per ciascun testo. Nella colonna 7 ("Parole caratteristiche") vengono indicate le forme che hanno una specificità positiva o negativa. Nelle colonne successive per ciascun testo è indicato il valore di probabilità alfa ≤ 0,025 e inoltre viene indicato se si tratta di specificità positiva o negativa. La lettura della specificità diventa più agevole se selezioniamo una colonna (per esempio la colonna 14: "p-value del testo ONU") e selezioniamo nel menu **Record**: *Ordina in senso crescente*. In questo modo otterremo le parole caratteristiche a partire dal valore di probabilità più basso (fig. 5.5), ossia le più specifiche (sia positive che negative).

	Forma grafica	Occorrenze totali	A_Statuto	B_Roma	C_Italia	D_ONU	Parole caratteristiche	p-value (A_Statuto)	Speci (A_St	p-value (B_Roma)	Speci (B_Ro	p-value (C_Italia)	Speci (C_Ita	p-value (D_ONU)	Speci (D_O
	individuo	39	0	0	1	38	spec					0,00000	neg	0,00000	pos
	diritto	85	8	3	34	40	spec			0,01289	neg	0,00012	neg	0,00000	pos
	ogni	84	8	10	28	38	spec					0,00000	neg	0,00000	pos
	1	16	0	1	1	14	spec					0,00001	neg	0,00000	pos
	ha	62	6	4	23	29	spec					0,00018	neg	0,00000	pos
	2	17	0	1	2	14	spec					0,00005	neg	0,00000	pos
	libertà	37	1	2	13	21	spec	0,00967	neg			0,00172	neg	0,00000	pos
	sua	34	2	1	14	17	spec					0,01875	neg	0,00000	pos
	diritti	39	3	3	15	18	spec					0,00476	neg	0,00000	pos
	potrà	11	2	0	0	9	spec							0,00000	pos
	uomo	9	0	0	1	8	spec					0,00363	neg	0,00000	pos
	Dichiarazione	6	0	0	0	6	spec_orig							0,00000	pos
	ad	45	5	2	21	17	spec							0,00001	pos
	considerato	9	0	0	2	7	spec					0,02488	neg	0,00001	pos
	Nazioni_Unite	5	0	0	0	5	spec_orig							0,00003	pos

Fig. 5.5. – Parole specifiche del testo ONU

Individuo, *diritto*, *ogni*, *libertà*, *uomo*, *Dichiarazione* sono le parole più caratteristiche del testo ONU. Da notare che nella colonna 7 la modalità "spec_orig" indica il fatto che la parola Dichiarazione è presente con 6 occorrenze esclusivamente in questo testo. La parola *legge* ha un'alta specificità negativa (un valore basso di probabilità) e questo ci indica che essa è usata raramente (5/160) in questo testo rispetto al corpus.

5. 3. CONFRONTO CON UN LESSICO DI FREQUENZA

L'analisi di specificità per l'individuazione delle forme peculiari può essere effettuato, come si è detto, anche tra un corpus e un lessico di riferimento. Vediamo, prima di tutto, quali sono i lessici presenti nel TALTAC. Dal menu **Moduli** – *Estrazione di informazione* selezioniamo la voce *Visualizza risorse statistico-linguistiche*.

- Il *Lessico di Poliformi (FDP)* è stato costruito sulla base di un campione di oltre 4 milioni di occorrenze (121.786 forme grafiche diverse) del linguaggio contemporaneo (scritto e parlato) ed elenca 3.925 poliformi con le rispettive frequenze.
- *Italiano standard – FG>1* è un lessico basato sul campione precedente (stampa, discorsi parlamentari, documenti ufficiali, saggistica, biografie, interviste, dialoghi, composizioni scolastiche) costituito di forme grafiche con indice d'uso >1 (50.464 forme grafiche distinte).
- *Italiano standard – lemmi dei verbi* elenca 2.605 lemmi di verbi tratti da forme grafiche non ambigue.
- *Linguaggio comune – FG con uso >50 (REP-90)* è una lista costituita da 60.489

forme grafiche, con indice d'uso > 50 tratte da una raccolta di 270 milioni di occorrenze di 10 annate del quotidiano "La Repubblica" (1990-1999).

- *Linguaggio comune – lemmi dei verbi (REP-90)* è una lista costituita da 4.907 lemmi di verbi tratti da forme non ambigue.
- *Lessico del discorso programmatico di governo (TPG_F)* è il lessico costituito da 3.000 lemmi tratti da un'analisi dei discorsi programmatici dei Presidenti del Consiglio della Prima Repubblica dal 1948 al 1994 (Bolasco 1996).
- La *Tavola di confronto dei lessici di frequenza* elenca 14.493 lemmi con le rispettive misure di frequenza tratte da:

  - *Vocabolario di Base della lingua italiana (VdB)* di Thornton, Iacobini e Burani (1997): per questi lemmi si conosce solo la fascia di frequenza.

  - *Vocabolario fondamentale (Vfond)* costituito di 2.739 lemmi di massima disponibilità tratti del LIF.

  - *Lessico Italiano di Frequenza (LIF) dell'italiano scritto*, di Bortolini e altri (1972) costituito di 5.360 lemmi.

  - *Lessico di frequenza dell'italiano parlato (LIP)* di De Mauro, Vedovelli, Voghera, Mancini (1993).

  - *Vocabolario elettronico della lingua italiana (VELI, 1989)*, costituito da 9.994 lemmi.

  - *Lessico dei bambini (LE)*, 2029 lemmi.

Possiamo effettuare un confronto tra il vocabolario del corpus *LEX1* con Tag grammaticale e *Italiano standard* che riporta le forme grafiche del lessico contemporaneo.

Dal menu **Moduli** – *Analisi lessicale*, selezioniamo il comando *Confronto con un lessico di frequenza*. La finestra di dialogo (fig. 5.6) ci offre alcune opzioni. Questo primo confronto è effettuato sulla lista di **intersezione**: ciò che vogliamo ottenere è un tabella con le forme grafiche comuni che ci permetta di identificare le forme sovra-utilizzate e sotto-utilizzate nel nostro lessico (*LEX1*) rispetto al lessico di confronto (*Italiano standard – FG>1*). Per le due liste, opportunamente selezionate dalle risorse interne di TALTAC (per la lista "modello") e dalle liste della sessione (per la lista da confrontare), deve essere indicato il campo sul quale effettuare il confronto. In questo caso si tratta del campo "Forma grafica". I campi da inserire per la visualizzazione dell'output dipendono dagli scopi del confronto. Il campo su cui calcolare lo scarto standardizzato è quello delle occorrenze totali in entrambe le liste (le frequenze d'uso sono le occorrenze ponderate con una misura di dispersione delle forme nel testo). La marcatura della casella "Maiuscole/minuscole" comporta, come al solito, il tener conto della presenza delle maiuscole (*case sensitive*).

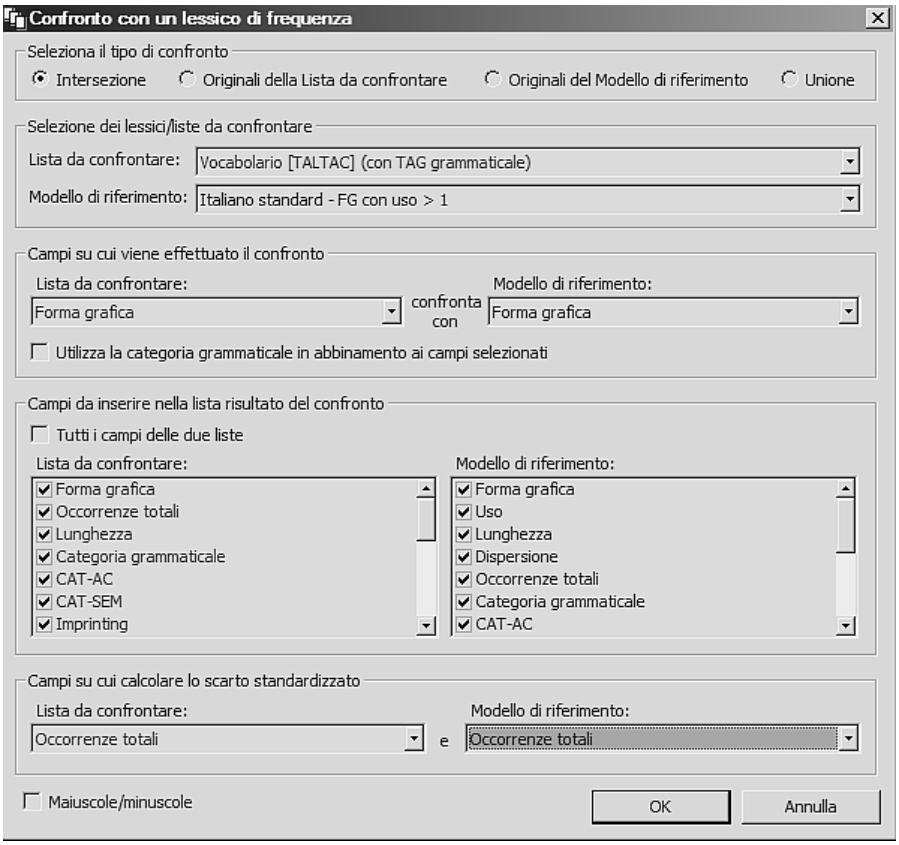

Fig. 5.6. – Finestra di dialogo del confronto con un lessico di frequenza

Cliccando su OK, nella finestra successiva ci viene indicato il file di intersezione delle due liste che sarà salvato nel DataBase della sessione (fig. 5.7). Il file, indipendentemente dalle opzioni di visualizzazione nella finestra di TALTAC, contiene tutte le informazioni del DataBase e potrà essere utilizzato con altri software.

Nella prima colonna sono riportati i valori dello scarto standardizzato sulle occorrenze normalizzate dai quali possiamo trarre informazioni sulle forme peculiari in questo corpus rispetto al lessico standard. Evidentemente, trattandosi di un corpus di testi giuridici, troviamo essenzialmente termini tecnici del lessico giuridico.

	Scarto sulle Occorrenze	Forma grafica	Occorrenze totali	Lungh	Categor gramma	CAT-AC
▶	386,900	Repubblica	77	10	N	N
	344,318	Camere	60	06	N	N
	262,294	stabiliti	21	09	J	A+N+V
	247,312	nominati	14	08	J	A+V
	239,965	individuo	39	09	J	A+N+V
	228,519	legge	160	05	J	N+V
	210,282	Reggente	5	08	J	A+N+V
	206,037	Reggenza	4	08	N	N
	205,998	giurisdizione	12	13	N	N
	192,712	inviolabile	7	11	A	A
	182,102	promulgata	5	10	V	V
	182,102	elegge	5	06	V	V
	181,893	leggi	66	05	J	A+N+V

Fig. 5.7. – Confronto di LEX1 con il lessico di frequenza Italiano Standard

La seconda opzione "Originali della lista da confrontare" permette di estrarre le forme grafiche peculiari che sono presenti nel file di confronto ma non nel file di modello. In questo caso, evidentemente, non sarà necessario calcolare lo scarto in quanto le forme visualizzate sono forme "uniche", assolutamente originali. Nel nostro esempio l'output con questa opzione visualizza 596 forme, tra le quali: *Re*, *presidente_della_Repubblica*, *Camera_dei_Deputati*, *Consiglio_regionale*, *consoli*, ecc.).

Dal Menu **Estrazione d'informazione** con l'opzione *Inserisci scarto nel vocabolario* è possibile riportare i valori dello scarto ottenuto nella tabella vocabolario; di conseguenza i record del vocabolario che hanno il campo scarto senza un valore corrispondono alle forme "originali".