UN ESEMPIO DI APPLICAZIONE DI ANALISI TESTUALE INFORMATICA:
la *Batracomiomachia* di Giacomo Leopardi
di Daniele Silvi

1.1 STATISTICHE DEL TESTO: T.A.C.T.

T.A.C.T. deve il suo nome al fatto di essere non un programma, bensì un sistema di sedici programmi riuniti in un unico pacchetto ed accessibili dallo stesso menu[1]. Progettato per operare sotto il sistema operativo MS-DOS[2], questo software è in grado di effettuare ricerche e analisi sulla base di testi letterari opportunamente codificati; in genere i ricercatori lo usano per elencare le occorrenze di una parola, combinazioni di parole o gruppi. Quello che il programma è in grado di generare sono liste, tabelle, grafici, concordanze, passando dalla semplice elencazione della frequenza di parole, lettere o frasi fino ad arrivare alle statistiche del rapporto *type-token*, creare anagrammi, dizionari lemmatizzati, collocazione delle parole in base ai loro rapporti con il resto del testo, ecc.

La prima operazione da compiere per eseguire ricerche con questi strumenti su di un testo è codificare il testo stesso: "codificare" significa che il testo deve acquisire le informazioni aggiuntive di cui abbiamo parlato da fornire alla macchina e quindi assumere un aspetto che al lettore umano non solo non dice nulla ma provoca anche un senso di inquietudine. Si tratta come prima cosa di riscrivere l'oggetto della nostra analisi in formato di testo *ASCII*[3]. Dopo aver fatto questo si passa ad inserire i *tag*, cioè le indicazioni che dicono alla macchina cosa una parola rappresenta, come deve apparire graficamente, che funzione logica ha all'interno del periodo, e così via. Questa fase si dice di "istruzione". Dobbiamo quindi utilizzare un linguaggio di marcatura appropriato che abbia particolari codici identificativi per ogni categoria morfo-sintattica precedentemente definita[4] e che sono racchiusi tra parentesi acute '< >', posti all'inizio e alla fine della parte testuale da segnalare,

[1] Questo software è prodotto dal *Centre for Computing in the Humanities* dell'Università di Toronto, ed è distribuito gratuitamente; si veda l'appendice per le indicazioni sull'URL e le sue modalità di uso e di reperimento.
[2] Attualmente lo si può far girare con tranquillità sotto Windows98, per i successivi sistemi operativi c'è qualche difficoltà per via dell'assenza di una modalità MS-DOS reale. Per i sistemi Windows2000 e WindowsXP si deve utilizzare il comando *forcedos*.
[3] *American Standard Code for Information Interchange*; per un totale di 256 caratteri rappresentabili col sistema binario ad 8 bit; tutti i caratteri all'interno della tabella Ascii sono chiamati "caratteri alfanumerici". Sul proprio personal computer è possibile accedervi premendo il tasto "Alt" e digitando la sequenza di numeri del carattere che si vuole ottenere.
[4] Per quanto riguarda la possibilità di costruire delle *D.T.D.* si consulti, ad esempio, K. Carey e S. Blatnik, *Guida a XML*, Milano, McGraw-Hill, 2002

o meglio da "marcare"[5]. Per fare degli esempi possiamo dire che in una sceneggiatura cose ovvie da marcare sono atti, scene e dialoghi, in un romanzo i capitoli, in un poema libri e stanze e così via fino a livelli sempre più specifici a seconda del tipo di analisi testuale che si intende compiere.

TACT è un sistema multilingue, il che significa che per supportare lingue straniere si serve della tabella Ascii estesa ed è in grado, con l'uso di opportuni *editor* di *font*, di coprire tutti i linguaggi europei moderni (francese, tedesco, greco, etc.).

Una volta che il testo è stato "marcato" l'*utility* "Makebase" lo converte in un *database* in modo da velocizzare e facilitare le operazioni di ricerca. In questa fase bisogna istruire il programma sull'alfabeto usato, sui caratteri speciali, i tag di riferimento e così via. Tutte queste specifiche vengo memorizzate in un *file* con estensione .MKS, che può anche essere riutilizzato per altri database testuali.

Le utilità "Mergebas" e "Buildbat" servono invece a gestire la fusione di più testi per creare dei file grandi partendo da parti di essi; la divisione in parti infatti è consigliata nel caso di testi lunghi per facilitare il *debugging*[6].

Dopo che il database testuale è stato creato (si tratta di un file con estensione TDB, Textual DataBase) il ricercatore è in grado di utilizzare sei programmi per operare ricerche all'interno di esso:

- *Usebase*, generalmente è il primo ad essere utilizzato, esso permette di selezionare una parola o un gruppo di parole (o anche una famiglia) e di visualizzarle in cinque forme: concordanza parola-contesto (KWIC), concordanza variante-contesto, l'intero testo, il grafico della distribuzione delle occorrenze, ed infine una tavola della distribuzione della parola all'interno del testo. Questi cinque *output* sono tra loro collegati in modo da poter accelerare i confronti e passare dall'uno all'altro immediatamente.
- *Collgen*, genera liste di gruppi di parole (due o più) che si ripetono più di una volta vicine tra loro nel testo in esame;
- *TactStat*, produce statistiche *type-token* per lunghezza-parola e frequenza-parola;
- *TactFreq*, produce liste di parole ordinabili alfabeticamente o in ordine inverso a seconda del numero di occorrenze che le caratterizzano;
- *Anagrams*, elabora tutti i possibili anagrammi di una parola di interesse.

[5] Sia essa una parola, una lettera, una frase, un intero paragrafo, etc.
[6] Per "debugging" si intende la fase di correzione degli errori.

1.2 UNA PROPOSTA DI LAVORO: LE TRADUZIONI DELLA *BATRACOMIOMACHIA*

Come dimostrazione delle potenzialità di *T.A.C.T.* abbiamo codificato ed analizzato le tre traduzioni del poemetto pseudo-omerico *Batracomiomachia* redatte da Giacomo Leopardi: il motivo della nostra scelta sta nel fatto che si tratta di un testo con varianti.

Affascinato dalla materia e stimolato dall'alone di mistero che gravava attorno al suo autore, il poeta decide per una traduzione di propria mano: la *Batracomiomachia* era stata già più volte tradotta in Italia ma nessuna di queste traduzioni soddisfa Leopardi, che le giudica fredde e letterali e decide quindi che "una nuova traduzione della Batracomiomachia potesse non essere inutile all'Italia"[7].

La prima versione risale al 1815 ma la storia di questa traduzione inizia a farsi interessante nel 1821, quando il Poeta inizia un processo di revisione che durerà praticamente per tutta la sua vita fino ad approdare all'ultima revisione del 1826, ed è lecito chiedersi quindi perché e come; a queste domande tenteremo di dare una risposta servendoci appunto delle metodologie informatiche.

Leopardi apporterà significativi cambiamenti nel passare da una redazione all'altra delle sue traduzioni e possiamo evidenziare subito le varianti con l'aiuto di *T.A.C.T.* Solo a titolo di esempio riportiamo la parte iniziale della codifica della edizione del 1815 (si noti in particolare la struttura del *Tact header*):

[7] G. Leopardi, *Discorso sopra la Batracomiomachia*, in *Tutte le poesie e tutte le prose*, a cura di L. Felici e E. Trevi, Milano, Newton & Compton, 1997, p. 398-399.

<Note geninf>
<<TITOLO: La guerra dei topi e delle rane>>
<<(Poema 1815)>>
<<CURATORE: Walter Binni>>
<<AUTORE: Giacomo Leopardi>>
<<DIRITTI D' AUTORE: no>>
<<TRATTO DA: "Tutte le opere">>
<<a cura di Walter Binni, con la collaborazione di Enrico Ghidetti>>
<<prima edizione: Sansoni editore, 1969>>
<<edizione di riferimento per la codifica TACT:>>
<<LA GUERRA DEI TOPI E DELLE RANE in TUTTE LE POESIE E TUTTE LE PROSE, Newton & Compton, 1997>>
<<CODICE ISBN: 88-383-0875-6>>
<<Il testo in formato elettronico è stato scaricato dal seguente sito:>>
<<LIBER LIBER>>
<<http://www.liberliber.it>>
<<1a EDIZIONE ELETTRONICA DEL: 27 ottobre 1999>>
<<INDICE DI AFFIDABILITA': >>
<<0: il file è in attesa di revisione>>
<<1: prima edizione>>
<<2: affidabilità media (edizione normale)>>
<<3: affidabilità ottima (edizione critica)>>
<<ALLA EDIZIONE ELETTRONICA HANNO CONTRIBUITO:>>
<<Vittorio Volpi, volpi@galactica.it>>
<<REVISIONE:>>
<<Catia Righi, catia.righi@risorsei.it>>
<<Il testo elettronico è stato codificato in formato T.A.C.T. da >>
<<Daniele Silvi, d.silvi@libero.it>>
<<Un particolare ringraziamento all'indispensabile collaborazione di Chiara Colombo>>
<<Ultima revisione: 06 ottobre 2002.>>
<<Il testo non presenta errori rispetto all'edizione di riferimento>>
<<Batra15 = Batracomiomachia 1815 La guerra dei topi e delle rane>>
<Autore G.Leopardi><Titolo La guerra dei topi e delle rane>
<Batra15 I 1>
Grande impresa disegno, arduo lavoro: %
O Muse, voi dall'Eliconie cime %
A me scendete, il vostro aiuto imploro: %
Datemi vago stil, carme sublime: %
Antica lite io canto, opre lontane, %
La Battaglia dei topi e delle rane. %

<Batra15 I 2>

```
Sulle ginocchia ho le mie carte, or fate %
Che nota a ogni mortal sia l'opra mia, %
Che alla più lenta, alla più tarda etate %
Salva pur giunga, e che di quanto fia %
Che sulle carte a voi sacrate io scriva, %
La fama sempre e la memoria viva. %

<Batra15 I 3> …etc etc…
```

La codifica è abbastanza semplice, facciamo solamente notare che il segno % alla fine di ogni verso è un "contatore" e serve appunto a contare detti versi, mentre l'uso delle doppie parentesi acute "<< >>" indica righe di commento da ignorare nell'elaborazione; in grassetto sono stati invece evidenziati le variabili usate nei tag e seguite dai loro valori. In sostanza un *tag* di quelli usati nelle codifche per *T.A.C.T.* si chiama "reference tag" ed è costituito da due parti: una variabile (la prima parola contenuta nel *tag*) ed un valore (le rimanenti parole, tutte saparate da uno spazio). Quello che segue è il *file* .MKS che fornisce al programma le specifiche di codifica e di interpretazione:

```
[Alpha]
  a b c d e f g h i j k l m n o p q r s t u v w x y z
  0 1 2 3 4 5 6 7 8 9
[DiacRet]
  - \Acute\ \Grave\ \Circumflex\ \Cedilla\ \Umlaut\ \Tilde\
"'"
```

```
[Title]
  La guerra dei topi e delle rane [1815]
[RefTemplate]
  canto/stanza $CS, v. $verso
[ReferenceBracket]
  <       >       SupText   NoWordSep
[IgnoreBracket]
  <<      >>      SupText   NoWordSep
[WordCounter]
  WORD
      0
[LineCounter]
```

```
  LINE
      0
[Reference]
  Autore
  Autore                NonNumeric 9999
[Reference]
  Titolo
  Titolo                NonNumeric 9999
[Reference]
  Note
  Note                  NonNumeric 9999
[Reference]
  CS
  Batra15               NonNumeric 9999
[Counter]
  verso
      1 SupText   %

[Counter]
  conta
      1 NoSupText @CS
     verso
```

TABELLA 2. *File* *.MKS

Una volta ottenuto il database testuale possiamo operare le ricerche di cui *T.A.C.T.* è capace ed ottenere le informazioni ritenute più utili al fine di notare le eventuali differenze tra le tre versioni della *Batracomiomachia* leopardiana e trarne opportune conclusioni.

Sempre a titolo di esempio aggiungiamo una seconda codifica che è stata effettuata; detta codifica nasce dall'esigenza di analizzare in maniera più specifica le varianti, pertanto i tre testi (come si nota dall'esempio che segue) sono stati "fusi" in uno solo e ogni terna di versi è stata etichettata con la sigla dell'anno di edizione compresa tra asterischi, inoltre sono stati utilizzati dei marcatori speciali per effettuare ricerche sulle varianti. Nella tabella che segue viene data spiegazione di tali marcatori speciali:

Sono state inserite le seguenti prolettere nel testo elettronico come codici di identificazione di tutte le diverse versioni:

1815: _

1821: #

1826: ^

1815/21: °

1821/26: +

1815/26 =

TABELLA 3. Caratteri speciali utilizzati nella codifica "Varianti"

Inoltre si è reso necessario modificare di conseguenza il *file* .MKS aggiungendo questi caratteri in fondo all'alfabeto ed eliminando l'accento circonflesso dai diacritici (in quanto non è presente come carattere alfanumerico nel testo in esame). Quello che segue è il *file* risultante:

```
[Alpha]
  a b c d e f g h i j k l m n o p q r s t u v w x y z _ # ^ ø
+ =
  0 1 2 3 4 5 6 7 8 9
[DiacRet]
  - \Acute\ \Grave\ \Cedilla\ \Umlaut\ \Tilde\ "'"
[RefTemplate]
  $Versione, canto/stanza $CS, v. $verso
[ReferenceBracket]
  <       >       SupText   NoWordSep
[IgnoreBracket]
  <<      >>      SupText   NoWordSep
[WordCounter]
  WORD
      0
[LineCounter]
  LINE
      0
[Reference]
  Autore
  Autore                NonNumeric 9999
[Reference]
```

```
  Titolo
  Titolo                NonNumeric 9999
[Reference]
  Note
  Note                  NonNumeric 9999
[Reference]
  CS
  15-21-26              NonNumeric 9999
[Counter]
  verso
      1 SupText   %

[Counter]
  conta
      1 NoSupText @CS
  verso
[Label]
  Versione
  *       *       NoSupText NoWordSep NonNumeric 9999
```

TABELLA 4. *File* .MKS utilizzato per la codifica "Varianti"

Una volta superata questa fase di "istruzione" del programma possiamo accedere alle interrogazioni, che svolgiamo mediante *query*, una funzione specifica del programma che permette di ricercare singole parole, porzioni di testo, segni diacritici e via discorrendo. Concentriamoci ora sulla codifica "Varianti" ed utilizziamo il *file* "3Vers.txt" che genera un omonimo *database*. Per mezzo del *Query Dialog Box* otteniamo quindi una *distribution list* per ciascun codice, impostando come unità di segmentazione la stanza e come sequenza di *output* la successione nel testo.

Vediamo come prima interrogazione le varianti nella versione del 1821, tenendo presente che per l'analisi ci si è riferiti al solo primo canto (per un totale di 25 stanze). Nella prima colonna viene indicato il canto in numeri romani, nella seconda la stanza in cifre arabe, nella terza il numero delle occorrenze dei sintagmi o delle categorie ricercate (in questo caso le "varianti") e nella quarta una rappresentazione grafica (modificabile) di queste occorrenze.

I 1	4	****
I 2	4	****
I 3	2	**
I 4	5	*****
I 5	6	******
I 6	2	**
I 7	5	*****
I 8	5	*****
I 9	2	**
I 10	1	*
I 11	2	**
I 12	3	***
I 13	3	***
I 14	3	***
I 15	5	*****
I 16	4	****
I 17	3	***
I 18	5	*****
I 19	4	****
I 20	3	***
I 21	3	***
I 22	4	****
I 23	5	*****
I 24	6	******
I 25	3	***

Total: 92.

TABELLA 5. Varianti 1821

Operiamo la stessa interrogazione sulla versione del 1826; si noti come, anche in questo caso, gli asterischi indichino il numero delle varianti presenti in ciascuna stanza di 6 versi (indicata ad inizio riga, assieme al canto che in questo caso è sempre – ovviamente – I).

Questo sistema permette di avere subito un impatto visivo esauriente della distribuzione delle varianti:

I 1	6	******
I 2	6	******
I 3	6	******
I 4	6	******
I 5	6	******
I 6	4	****
I 7	5	*****
I 8	6	******
I 9	3	***
I 10	4	****
I 11	6	******
I 12	5	*****
I 13	3	***
I 14	3	***
I 15	5	*****
I 16	4	****
I 17	4	****
I 18	6	******
I 19	5	*****
I 20	3	***
I 21	5	*****
I 22	6	******
I 23	5	*****
I 24	6	******
I 25	5	*****

Total: 123.

TABELLA 6. Varianti 1826

Si vede subito che il maggior numero di varianti si trova appunto nella versione del 1826, dal che si può già ipotizzare come i cambiamenti siano divenuti più sostanziali nell'ultima revisione leopardiana. A riprova di questo possiamo generare grafici che ci indichino quante e quali cose siano invece rimaste invariate nel passare da una redazione all'altra; vediamo (sempre solo in forma numerica) per prima la tabella di distribuzione relativa ai cambiamenti avvenuti nel passare dalla versione 1821 a quella 1826:

```
I 6        |     2|**
I 7        |     1|*
I 9        |     1|*
I 10       |     2|**
I 13       |     3|***
I 14       |     3|***
I 16       |     1|*
I 17       |     2|**
I 20       |     2|**
I 21       |     1|*
I 25       |     1|*

Total: 19.
```

TABELLA 7. Sintagmi e diacritici invariati 1821/1826

La distribuzione evidenziata dalla tabella mostra le parti di testo che non sono state modificate nel passare dalla redazione del 1821 a quella del 1826.

Un'altra interrogazione molto utile è quella di evidenziare tutti i versi varianti per gruppi, mantenendo ad inizio riga l'indicazione della versione e segnalando per ciascun gruppo il canto e la stanza di provenienza. Quello che segue è, come sempre, solo un esempio:

```
  15, canto/stanza I 1, v. 1
  *15*_          Grande impresa disegno, arduo lavoro:
```

21# Mentre a novo m' accingo arduo lavoro,
26^ Sul cominciar del mio novello canto,

15, canto/stanza I 1, v. 2
*15*_ O Muse, voi dall' Eliconie cime
21# O Muse, voi da l' Eliconie cime
26^ Voi che tenete l' eliconie cime

15, canto/stanza I 1, v. 3
*15*_ A me scendete, il vostro aiuto imploro:
21# Scendete a me ch' il vostro aiuto imploro:
26^ Prego, vergini Dee, concilio santo,

15, canto/stanza I 1, v. 6
*15*_ La Battaglia dei topi e delle rane.
21# La Battaglia de' topi e de le rane.
26^ Segno insolito a i carmi, io prendo a dire.

15, canto/stanza I 2, v. 1
*15*_ Sulle ginocchia ho le mie carte, or fate
21# Su le ginocchia ho le mie carte; or fate
26^ La cetra ho in man, le carte in grembo: or date

TABELLA 8. Varianti rispetto alla versione 1815

Nell’esempio della tabella n. 8 si è utilizzata la versione 1815 come riferimento e si sono quindi chieste tutte le varianti a questa relative. Si potrebbe poi eseguire un’interrogazione chiedendo le varianti rispetto alla versione 1821, o solamente tra la 1821 e la 1826. In pratica il testo dell’esempio precedente non considera i casi in cui le tre versioni coincidono, quelli in cui la versione 1815 è identica alla 1821 ed infine quelli in cui la versione 1815 è identica alla 1826. Mescolando i vari tipi di interrogazione si possono avere tutte le varianti ed i loro contesti.

Un’altra interrogazione interessante è la “KWIC”, acronimo di *Key-Word in context*; questa operazione permette di cercare una data parola ed individuarne le occorrenze con la collocazione nel testo. Abbiamo eseguito questa interrogazione sulla parola “topo”. Nel caso specifico è stata utilizzata la codifica “Varianti”, pertanto insieme alle occorrenze vengono indicate anche le versioni di dette occorrenze:

```
15, canto/stanza I 4, v. 1                Un topo un dì, fra' topi il
21, canto/stanza I 4, v. 2                Un topo un dì, fra' topi il
26, canto/stanza I 4, v. 2                Un topo, de le membra il più ben
15, canto/stanza I 8, v. 1       Rispose il topo: Amico, e che mai brami?
21, canto/stanza I 8, v. 2         disse il topo, "e che mai brami? |
26, canto/stanza I 8, v. 2              E 'l topo a lui: "Quel che saper tu
15, canto/stanza I 8, v. 6                 Topo di raro cor, d' anima
```

21, canto/stanza I 8, v. 7 **Topo** di fino pel, d' anima
21, canto/stanza I 14, v. 4 che il **topo** cada in quell' ordigno |
26, canto/stanza I 14, v. 4 che 'l **topo** incorra in quell'
15, canto/stanza I 18, v. 2 Saltovvi il **topo**, e colle mani il collo |
21, canto/stanza I 18, v. 6 da prima il **topo**, malaccorto, |
26, canto/stanza I 18, v. 6 Rideva il **topo**, e rise il malaccorto |
15/21, canto/stanza I 22, v. 3 Il **topo** inorridì, gelò la rana; |
15/21, canto/stanza I 22, v. 5 e il **topo** sventurato |
26, canto/stanza I 22, v. 6 celarsi, e 'l **topo** sventurato |

TABELLA 9. Risultati di una interrogazione ***Kwic*** – parola "topo"

Notiamo anche che questo tipo di interrogazione si può fare anche per sintagmi, laddove per "sintagma" si intenda una sequenza di più forme consecutive. Utilizzando appositi operatori[8] si può ulteriormente estendere il campo di indagine; per comunicare al programma che stiamo fornendo un sintagma si deve usare la barra verticale (|) tra un termine e l'altro. Potremmo così cercare, ad esempio, tutte le parole seguite o precedute da un certo aggettivo, e

[8] L'asterisco (*) rappresenta zero o più ripetizioni del carattere precedente. In combinazione con ANY (.*) significa: qualsiasi stringa di caratteri (anche zero). Le parentesi quadre ([ ]) racchiudono una classe di caratteri, es. [aeiuo] trova qualunque vocale semplice, accentata o apostrofata. Abbiamo poi il carattere "negate" (~) che esclude dalla ricerca i caratteri specificati; mentre l'operatore "range" specifica una sequenza di caratteri nell'alfabeto del testo, es.: [a:c] trova qualunque lettera compresa tra a e c, estremi compresi. Infine l'operatore "escare" rappresentato dal *backslash* (\) che ha due funzioni: definisce il carattere seguente come elemento del testo e identifica la stringa seguente come formula. La prima delle due funzioni elencate è inerente proprio alle codifiche in esame in questa tesi, in quanto ci siamo serviti di diverse promettere per "etichettare" il testo, e questo operatore (nelle ricerche) segnala che determinati segni sono prolettere e non metacaratteri.

contestualizzarle producendo la stampa della riga, del contesto ed, eventualmente, della versione (come nel caso dei testi in esame).

Infine dobbiamo parlare di un'altra importante funzione del pacchetto *Tact*: *TactStat*. Essa fornisce un complesso quadro numerico di statistiche del testo, generando tabelle, istogrammi e misurazioni statistiche sui *token*, i *type*, la lunghezza delle parole, la prima e l'ultima lettera delle parole e su tutte le lettere che compongono le parole. Per *default* queste informazioni statistiche vengono memorizzate in un *file* che ha lo stesso nome del *database* e l'estensione .STA. Una volta lanciato il programma si deve inserire il nome completo del *database* testuale da utilizzare, selezionare l'*output* dei risultati (file o stampante) e specificare (se si vuole) il nome del *file* di *output*, che altrimenti prenderà automaticamente lo stesso nome del *database* testuale. L'*output* di TACTstat si compone di cinque parti: (1) tabella e statistiche della frequenza delle parole, (2) un istogramma della lunghezza delle parole, misurate in lettere, (3) istogramma alfabetico e a frequenza discendente e statistiche per le prime lettere di tutte le parole, (4) istogramma alfabetico e a frequenza discendente e statistiche per l'ultima lettera di tutte le parole, (5) istogramma alfabetico e a frequenza discendente e statistiche per tutte le lettere che compongono le parole. Queste statistiche si sarebbero potute costruire anche manualmente, ma con enorme fatica, tempi lunghissimi ed elevate probabilità di errore. Prima di passare all'analisi di dette tabelle è opportuno chiarire però il significato e le differenze delle parole *token* e *type*. In sostanza quando si vuole valutare la ricchezza del linguaggio in un testo si deve valutare il rapporto tra le parole e la classe a cui queste parole appartengono: il rapporto parola/classe è appunto il rapporto *token/type*. Il *token* quindi è l'occorrenza (insomma quante volte la singola parola si ripete nel testo) ed il *type* è una via di mezzo tra l'occorenza e il lemma, sostanzialmente la forma grafica. Il numero totale dei *types* ci fornisce quindi il numero totale di parole (forme grafiche) diverse usate nel testo. Vediamo ora la tabella delle statistiche per quanto riguarda la versione della *Batracomiomachia* del 1815 e spieghiamo tutte le voci presenti:

Frequency Rank	Observed Freq. of Rank	Words in Frequency	Types Total	Tokens Total	% of Types	% of Tokens	% of word in freq.
1	906	906	906	906	68.53	29.03	29.03

2	197	394	1103	1300	83.43	41.65	12.62
3	63	189	1166	1489	88.20	47.71	6.06
4	52	208	1218	1697	92.13	54.37	6.66
5	29	145	1247	1842	94.33	59.02	4.65
6	10	60	1257	1902	95.08	60.94	1.92
7	12	84	1269	1986	95.99	63.63	2.69
8	12	96	1281	2082	96.90	66.71	3.08
9	6	54	1287	2136	97.35	68.44	1.73
10	4	40	1291	2176	97.66	69.72	1.28
11	3	33	1294	2209	97.88	70.78	1.06
12	3	36	1297	2245	98.11	71.93	1.15
13	1	13	1298	2258	98.18	72.35	0.42
14	1	14	1299	2272	98.26	72.80	0.45
15	1	15	1300	2287	98.34	73.28	0.48
16	1	16	1301	2303	98.41	73.79	0.51
17	2	34	1303	2337	98.56	74.88	1.09
18	2	36	1305	2373	98.71	76.03	1.15

20	1	20	1306	2393	98.79	76.67	0.64
21	1	21	1307	2414	98.87	77.35	0.67
22	1	22	1308	2436	98.94	78.05	0.70
24	1	24	1309	2460	99.02	78.82	0.77
25	2	50	1311	2510	99.17	80.42	1.60
27	2	54	1313	2564	99.32	82.15	1.73
28	1	28	1314	2592	99.39	83.05	0.90
33	1	33	1315	2625	99.47	84.11	1.06
35	1	35	1316	2660	99.55	85.23	1.12
49	1	49	1317	2709	99.62	86.80	1.57
52	1	52	1318	2761	99.70	88.47	1.67
53	1	53	1319	2814	99.77	90.16	1.70
61	1	61	1320	2875	99.85	92.12	1.95
96	1	96	1321	2971	99.92	95.19	3.08
150	1	150	1322	3121	100.00	100.00	4.81

Number of Types = 1322

Number of Tokens = 3121

```
Type/Token ratio  =        0.424
Token/Type ratio  =        2.361
Hapax Legomena    =      906
Hapax Dislegomena =      197
Hapax Legomena/Dislegomena ratio   =     4.5990
Hapax Legomena/Number of Types     =     0.6853
Hapax Legomena/Number of Tokens    =     0.2903
Hapax Legomena cubed/Types squared =   425.5217
Variance ( S.D. squared )          =    40.1036
Standard Deviation (S.D.)          =     6.3327
Coefficient of skewness            =    14.2758
Coefficient of kurtosis            =   272.7742
Herdan's characteristic            =     0.0738
Yule's characteristic              =   619.5178
Carroll TTR (Types / Sqrt of 2 X Tokens) =    16.7328
Most Frequent word "e" occurred 150 times
repeat rate (Tokens / frequency most frequent word) =
20.8067
```

TABELLA 10. Statistica delle frequenze nella versione 1815

- *Frequency Rank*, fornisce in ordine crescente le frequenze dei *token* del testo. Nella tabella non c'è, ad esempio, il numero 19, ciò significa che non ci sono *token* che hanno frequenza 19. In pratica non ci sono parole che compaiano 19 volte.
- *Observed Frequency of Rank*, fornisce, per ogni riga, la somma totale dei *types* che contengono i *tokens* con la stessa frequenza, riportata dalla prima colonna. Si tratta della somma totale delle parole che compaiono un *tot* di volte. Nell'esempio in esame ci sono, in altri termini, 906 parole che compaiono una sola volta, 197 che compaiono 2 volte, e così via.
- *Words in Frequency*, fornisce, per ogni riga, il numero complessivo dei *tokens* contenuti nei *types* riportati dalla seconda colonna. Per ottenere tale numero è sufficiente moltiplicare i dati numerici delle prime due colonne.

- *Types Total*, fornisce, in ordine crescente, il totale dei *types* presenti nel testo. Ogni riga si ottiene sommando progressivamente i dati della seconda colonna.
- *Tokens Total*, fornisce, in ordine crescente, il numero totale dei *tokens* presenti nel testo. Ogni riga si ottiene sommando progressivamente i dati della terza colonna.
- *Percentage of Types*, fornisce, in ordine crescente, la percentuale dei *types* presenti nel testo.
- *Percentage of Tokens*, come la precedente ma per quanto riguarda i *tokens*.
- *Percentage of word in Frequency*, fornisce la percentuale dei *tokens* rispetto alla totalità dei medesimi presenti nel testo.

Per avere un quadro esaustivo della situazione dovremmo generare una lista completa di concordanze (operazione oltremodo semplice con *T.A.C.T.*), cosa che – per ragioni di spazio – è stata qui omessa; tuttavia dalla sola analisi della tabella delle statistiche si possono ricavare preziose informazioni.

Per rendere significativa questa analisi facciamo lo stesso spoglio per le rimanenti versioni della *Batracomiomachia* e confrontiamo i risultati:

Frequency Rank	Observed Freq. of Rank	Words in Frequency	Types Total	Tokens Total	% of Types	% of Tokens	% of word in freq.
1	955	955	955	955	71.59	29.99	29.99
2	172	344	1127	1299	84.48	40.80	10.80
3	85	255	1212	1554	90.85	48.81	8.01
4	33	132	1245	1686	93.33	52.95	4.15
5	23	115	1268	1801	95.05	56.56	3.61
6	10	60	1278	1861	95.80	58.45	1.88

7	8	56	1286	1917	96.40	60.21	1.76
8	7	56	1293	1973	96.93	61.97	1.76
9	4	36	1297	2009	97.23	63.10	1.13
10	3	30	1300	2039	97.45	64.04	0.94
11	3	33	1303	2072	97.68	65.08	1.04
13	1	13	1304	2085	97.75	65.48	0.41
14	2	28	1306	2113	97.90	66.36	0.88
15	3	45	1309	2158	98.13	67.78	1.41
17	1	17	1310	2175	98.20	68.31	0.53
18	1	18	1311	2193	98.28	68.88	0.57
19	1	19	1312	2212	98.35	69.47	0.60
20	2	40	1314	2252	98.50	70.73	1.26
21	4	84	1318	2336	98.80	73.37	2.64
22	1	22	1319	2358	98.88	74.06	0.69
25	2	50	1321	2408	99.03	75.63	1.57
29	1	29	1322	2437	99.10	76.54	0.91
30	1	30	1323	2467	99.18	77.48	0.94

31	1	31	1324	2498	99.25	78.45	0.97
32	1	32	1325	2530	99.33	79.46	1.01
33	1	33	1326	2563	99.40	80.50	1.04
38	1	38	1327	2601	99.48	81.69	1.19
56	1	56	1328	2657	99.55	83.45	1.76
61	2	122	1330	2779	99.70	87.28	3.83
82	1	82	1331	2861	99.78	89.86	2.58
86	2	172	1333	3033	99.93	95.26	5.40
151	1	151	1334	3184	100.00	100.00	4.74

Number of Types = 1334
Number of Tokens = 3184
Type/Token ratio = 0.419
Token/Type ratio = 2.387
Hapax Legomena = 955
Hapax Dislegomena = 172
Hapax Legomena/Dislegomena ratio = 5.5523
Hapax Legomena/Number of Types = 0.7159
Hapax Legomena/Number of Tokens = 0.2999
Hapax Legomena cubed/Types squared = 489.4389
Variance (S.D. squared) = 49.3326
Standard Deviation (S.D.) = 7.0237
Coefficient of skewness = 12.4926
Coefficient of kurtosis = 201.7666

Herdan's characteristic = 0.0806
Yule's characteristic = 723.6244
Carroll TTR (Types / Sqrt of 2 X Tokens) = 16.7168
Most Frequent word "e" occurred 151 times
repeat rate (Tokens / frequency most frequent word) = 21.0861

TABELLA 11. Statistica delle frequenze nella versione 1821

Frequency Rank	Observed Freq. of Rank	Words in Frequency	Types Total	Tokens Total	% of Types	% of Tokens	% of word in freq.
1	987	987	987	987	72.04	30.68	30.68
2	196	392	1183	1379	86.35	42.87	12.19
3	71	213	1254	1592	91.53	49.49	6.62
4	27	108	1281	1700	93.50	52.84	3.36
5	16	80	1297	1780	94.67	55.33	2.49
6	17	102	1314	1882	95.91	58.50	3.17
7	9	63	1323	1945	96.57	60.46	1.96
8	8	64	1331	2009	97.15	62.45	1.99
10	3	30	1334	2039	97.37	63.38	0.93

11		2	22	1336	2061	97.52
64.07	0.68					
12		3	36	1339	2097	97.74
65.18	1.12					
13		1	13	1340	2110	97.81
65.59	0.40					
14		1	14	1341	2124	97.88
66.02	0.44					
15		2	30	1343	2154	98.03
66.96	0.93					
16		1	16	1344	2170	98.10
67.45	0.50					
17		2	34	1346	2204	98.25
68.51	1.06					
18		1	18	1347	2222	98.32
69.07	0.56					
19		3	57	1350	2279	98.54
70.84	1.77					
21		3	63	1353	2342	98.76
72.80	1.96					
22		1	22	1354	2364	98.83
73.48	0.68					
23		1	23	1355	2387	98.91
74.20	0.71					
24		1	24	1356	2411	98.98
74.95	0.75					
25		1	25	1357	2436	99.05
75.72	0.78					
28		1	28	1358	2464	99.12
76.59	0.87					
35		1	35	1359	2499	99.20
77.68	1.09					
37		1	37	1360	2536	99.27
78.83	1.15					

38	1	38	1361	2574	99.34	80.01 1.18
45	1	45	1362	2619	99.42	81.41 1.40
46	1	46	1363	2665	99.49	82.84 1.43
48	1	48	1364	2713	99.56	84.33 1.49
56	1	56	1365	2769	99.64	86.07 1.74
61	1	61	1366	2830	99.71	87.97 1.90
77	1	77	1367	2907	99.78	90.36 2.39
80	1	80	1368	2987	99.85	92.85 2.49
90	1	90	1369	3077	99.93	95.65 2.80
140	1	140	1370	3217	100.00	100.00 4.35

Number of Types = 1370
Number of Tokens = 3217
Type/Token ratio = 0.426
Token/Type ratio = 2.348
Hapax Legomena = 987
Hapax Dislegomena = 196
Hapax Legomena/Dislegomena ratio = 5.0357
Hapax Legomena/Number of Types = 0.7204
Hapax Legomena/Number of Tokens = 0.3068
Hapax Legomena cubed/Types squared = 512.2834
Variance (S.D. squared) = 45.9992
Standard Deviation (S.D.) = 6.7823

```
Coefficient of skewness            =    11.7400
Coefficient of kurtosis            =   177.6344
Herdan's characteristic            =     0.0780
Yule's characteristic              =   681.4803
Carroll TTR (Types / Sqrt of 2 X Tokens) =    17.0797
Most Frequent word "e" occurred 140 times
repeat rate (Tokens / frequency most frequent word) =
22.9786
```

TABELLA 12. Statistica delle frequenze nella versione 1826

Per spiegare a cosa si riferiscono tutti questi numeri possiamo, ad esempio, leggere la prima riga di quest'ultima tabella (n. 12) nella seguente maniera: 987 parole differenti (col. 2) ricorrono una volta (col. 1) per un totale di 987 volte (col. 2 moltiplicato col. 1=col. 3). Queste ripetizioni coprono il 30,68 % di tutte le parole presenti nel testo (col. 8). Inoltre questo primo grado di frequenza (il grado di frequenza "1") contiene in tutto 987 parole (col. 4) che rappresentano il 72,04 % del vocabolario totale (col. 6) di 1.370 parole (ultima riga della colonna 4 oppure prima voce della didascalia, "number of types"). Sempre in questo primo grado di frequenza possiamo quindi contare un totale (tra ripetizioni e hapax) di 987 parole (ovviamente il numero è sempre lo stesso perché ci troviamo sul primo grado di frequenza. L'esempio sarà più chiaro con il secondo grado) che corrisponde al 30,68 % del totale[9] (col. 7) espresso nell'ultima riga della colonna 5, cioè di 3.217 parole totali (comprese di ripetizioni).

Facciamo, per maggiore chiarezza, la stessa lettura sulla seconda riga, o secondo grado di frequenza: la frequenza 2; si tratta insomma delle parole che ricorrono due volte nel testo[10]. Dunque abbiamo che 192 parole differenti (col. 2) ricorrono due volte (col. 1) per un totale di 392 occorrenze (col. 2 moltiplicato col. 1=col. 3). Queste ripetizioni coprono il 12,19 % di tutte le parole presenti nel testo (col. 8). Inoltre questo secondo grado di frequenza contiene in tutto 1.183 parole (col. 4, che rappresenta la somma dei valori della colonna 2, fin qui incontrati) che rappresentano l'86,35 % del vocabolario totale (col. 6) di 1.370 parole (ultima riga della colonna 4 oppure prima voce della didascalia, "number of types"); in sostanza

[9] Vale anche in questo caso la considerazione fatta poco prima per il dato numerico non percentualizzato: il numero rimane lo stesso perché ci troviamo nel primo grado delle frequenze.
[10] Che gli ideatori del programma, con un curioso ossimoro, chiamano "hapax dislegomena".

stiamo dicendo che, tra parole che ricorrono una sola volta e quelle che ricorrono due volte abbiamo già coperto un buon 86% del testo. Già da questo dato si può dedurre che si tratta di un testo dal vocabolario ricco e variato e molto poco ripetitivo. Sempre in questo grado di frequenza possiamo quindi contare un totale (tra ripetizioni e hapax) di 1.379 parole (ora il numero rappresenta la somma dei valori del primo grado con quelli del secondo) che corrisponde al 42,87 % del totale (col. 7) espresso nell'ultima riga della colonna 5, cioè di 3.217 parole totali (comprese di ripetizioni).

Dal confronto di tutte e tre le tabelle mettiamo in evidenza solo alcuni aspetti, a titolo di esempio:

1. Nei prospetti che sono in calce ad ogni tabella figura la voce *Hapax Legomena* che indica quante parole compaiono una volta sola nel testo: nella versione 1815 sono 906, nella versione 1821 sono 955 e nella versione 1826 sono 987. Tale numero è andato sempre crescendo, e in maniera evidente se confrontiamo la prima con l'ultima versione, segno chiaro che il testo nel corso degli anni si è arricchito di nuovi termini ed è stato sfrondato di ripetizioni. L'andamento di questi risultati numerici denuncia quindi un arricchimento lessicale.
2. Sempre nelle voci a fine tabelle si nota la *Type/Token ratio* che esplicita, per l'appunto, il rapporto tra il numero dei *types* ed il numero dei *tokens*. Questo dato numerico ci fornisce un'idea ancora migliore della ricchezza del testo:

> Una delle indicazioni alle quali aggrapparci per interpretare la ricchezza del linguaggio di un determinato testo è il rapporto tra le parole e le classi nelle quali possono essere catalogate (in inglese *token* e *type*). […] Una conseguenza abbastanza immediata (ma da prendere con le molle, come tutte le misurazioni) è che un testo in cui il rapporto *token/type* sia alto deve essere un testo in cui vi sono pochi *type* rispetto ai *token* e quindi potrà essere considerato un testo ripetitivo, con un vocabolario non molto ricco e tendenzialmente poco faticoso da leggere.[11]

Anche in questo caso il dato è significativo, poiché nel passaggio dalla versione 1821 alla versione 1826 il rapporto in questione si è abbassato di 0,039, passando da 2,387 della

[11] Giuseppe Gigliozzi, *Il testo e il computer*, Milano, Mondadori, 1977, p. 193-194.

versione 1821 a 2,348 della versione 1826. Per di più dobbiamo tenere conto del fatto che questo decremento è avvenuto in un testo abbastanza breve, il che ne aumenta l'importanza:

> …notare come una cosa sia l'aggiunta di un *token* in un testo di 10 parole, tutt'altra faccenda l'inserimento di un termine in un lavoro di 1000 parole. L'incremento dei termini diverrebbe significativo nel primo caso e di scarso interesse nel secondo.[12]

Tralasciando altre considerazioni "numeriche", vogliamo ora dare qualche esempio di come effettivamente questo linguaggio si sia arricchito attraverso lo studio più ravvicinato di alcune delle varianti, a nostro giudizio, più significative.

Siamo nel Canto II e viene presentato Giove che parla al concilio degli Dei. È un vero e proprio *topos* letterario, e acquista sempre maggiore spessore via via che da una versione si passa all'altra, subendo peraltro l'influenza delle letture del giovane Leopardi, in particolar modo della *Secchia rapita*. Alla stanza 20, verso 3, del canto II leggiamo nel 1815 "Mai li soccorrerò…", che nel '21 diventa "Per me non fiaterei…" per giungere infine, nel 1826, a un perentorio "Vadan, per conto mio, tutti a Plutone". Se passiamo alla stanza 24 il risultato è ancora più evidente, soprattutto dal punto di vista della conformazione al lessico tassioniano. Il testo del 1815 recita "Vegliar dovei con fiero duol di testa / Fino a quel tempo, in cui spunta la luce, / Allor che il gallo svegliasi e fa festa. / Orsù, nessun di noi si faccia duce / de' combattenti che a pugnar sen vanno, / abbiasi chicchessia vittoria, o danno". Nel 1821 il cambiamento è minimo, anzi si tratta quasi di un ulteriore raffinamento del registro linguistico: "Vegliar dovei con fiero duol di testa / Fin quando spunta la diurna luce, / Allor che il gallo svegliasi e fa festa. / Orsù verun di noi schermo né duce / Si faccia di costor che in guerra vanno: / Abbiasi chicchessia vittoria o danno". Nel 1826 il cambiamento è radicale, il linguaggio, ora pungente e corposo, si è spolverato di dosso l'alone di "purismo" che opprimeva le versioni precedenti: "Postami per dormire un pocolino, / Ecco un crocchiare eterno di ranocchi / M'introna in guisa tal, ch'era il mattino / Già chiaro quando prima iochiusi gli occhi. / Or quanto a questa guerra, il mio parere / E' lasciar fare e starcela a vedere".

Possiamo ora brevemente tirare le somme per individuare un sistema di varianti:

1. Tutto il registro linguistico subisce, nel passaggio dal '21 al '26, un cambiamento radicale, in direzione di un linguaggio più "forte", più ficcante, pungente e "realistico". Parole come "suggon" diventano "succian", "forato e

[12] *Ibidem*, p. 194.

guasto” diviene “trasformato in un cencio”, “rosero il mio velo” diviene “me l’han rotto”, “la loro armata” muta in “quella marmaglia”, “distruggerem l’esercito nemico, né fia chi dal pantan faccia ritorno” che diventa “tutto quanto l’esercito nemico manderem senza sangue a la malora”, etc. L’elenco potrebbe continuare per intere pagine, ma ci basta qui aver dato un’idea dell’andamento generale.

2. Le preposizioni articolate subiscono una dissociazione nel passare dalla lezione del ’15 a quella del ’21: “sulla” diventa “su la”, “colla” diventa “con la”, “all’” diventa “a l’”, e così via. Un vero e proprio procedimento di revisione linguistica che si dimostra così essere non solo proprio dell’”ultimo” Leopardi ma anche del Leopardi più giovane.
3. Tutta l’ultima versione è più “leopardiana”, sia nello stile che nella scelta dei termini, ed è il segno di come il Poeta fosse divenuto, da traduttore, imitatore.

1.3 CONCLUSIONI

Siamo in una fase di transizione, fra “innovazioni e coincidenze con la tradizione filologica”[13] e pertanto molte pratiche risultano essere un ibrido tra le due tradizioni. Forse molti problemi si risolveranno se il futuro del libro si evolverà del tutto nella direzione elettronica, quando cioè l’*e-book* sostituirà definitivamente il libro cartaceo e di esso non si avrà più alcuna traccia, come in un romanzo di Ray Bradbury[14], tuttavia per ora dobbiamo muoverci in questo presente e cercare di far collaborare le metodologie: l’esperimento di analisi che è stato propostovuole solo mostrare una piccola parte delle potenzialità di un computer riguardo l’analisi di un testo. Molte cose inoltre sono state – volutamente – tralasciate, come ad esempio il confronto dei rapporti *token/type* e della ricchezza e distribuzione del lessico tra la *Batracomimachia* ed altre opere, presupposte ispiratrici, come ad esempio la *Secchia rapita*; oppure la ripartizione delle battute tra i diversi personaggi e la ricchezza lessicale di ognuno; o ancora un confronto più impegnativo con il lessico di tutta la produzione leopardiana e l’apparizione o la differente frequenza di un termine all’interno dell’esperienza poetica del Poeta.

A conclusione del nostro discorso possiamo senz’altro dire che, a fronte degli enormi vantaggi che si possono trarre da un’analisi informatizzata del testo, queste tecniche devono

[13] Raul Mordenti, *Informatica e critica dei testi*, Roma, Bulzoni, 2001, p. 29.
[14] Mi riferisco in particolare a *Fahrenheit 451*.

superare un naturale sbarramento ideologico che ne mostra un'immagine troppo tecnica o, in alcuni casi, snaturante della sostanza artistica dell'opera:

> Nonostante i loro tratti esteticamente rilevanti, tuttavia, gli strumenti digitali sono comunemente considerati al servizio del discorso scientifico. Inoltre, la denominazione pià comune, "tecnologie digitali", impone loro un'etihetta dal tono minacciosamente tecnicistico.[15]

Daniele Silvi

[15] Jerome Mcgann, *La letteratura dopo il eorld wide web*, Bologna, B.U.P., 2002, p. 252