parte seconda

3. Introduzione metodologica

MARCO MARIN

Nella seconda parte di questo volume, che qui mi accingo a presentare, la sintesi dei risultati di ricerca lascia spazio all'esposizione di una parte dei dati ricavati mediante l'analisi lessicologica[1].

L'uso del mezzo informatico sta diventando, anche nel campo degli studi storici, fondamentale[2]. In numerosi ambienti le consuete reticenze verso le novità provenienti dalle discipline informatiche stanno sparendo ed una conoscenza più competente del mezzo informatico si sta facendo largo fra le nuove generazioni di studenti e docenti. L'attivazione di corsi universitari mirati e di dottorati di ricerca[3], volti all'integrazione dello strumento informatico nello studio della storia, può far avanzare la ricerca in maniera sostanziale.

La conoscenza degli studi dei linguisti[4] (più all'avanguardia, in questo campo, rispetto agli storici) risulta essere il punto di partenza per ogni ricercatore, che voglia utilizzare la linguistica computazionale (o forse con più precisione quel ramo della linguistica computazionale denominato «linguistica dei *corpora*»[5]) nello studio delle fonti. Al riguardo – però – l'interazione fra linguisti e storici è – a mio avviso – ancora allo stato embrionale, come lo è la formazione degli studenti.

La linguistica computazionale tende ad utilizzare l'elaboratore elettronico in modo da rendere più semplici e rapide le ricerche all'interno del testo[6]. L'utilizzo dei programmi di interrogazione dei testi velocizza e snellisce il lavoro: permette verifiche rapide e maggior libertà d'indagine. Mediante questi software si può controllare rapidamente un testo; stilare liste dei riferimenti[7] fruibili da chiun-

que non possa accedere ad un *corpus* informatizzato[8]; verificare – stilando liste di frequenza[9] – se un autore (o un gruppo di autori) usi una parola[10] o una lessia[11], che sia indice di una particolare teoria politica o filosofica; si può confrontare il lessico di più autori; si possono costruire *corpora* ipertestuali. Si può altresì verificare – mediante la ricerca delle cooccorrenze[12] – se due o più parole compaiano spesso affiancate formando locuzioni ricorrenti; si possono stilare delle liste di concordanze[13] di una o più parole, in modo da creare uno strumento facilmente consultabile da chiunque sia interessato allo studio della fonte in questione. In questo caso la lista delle concordanze funge anche da base filologica. Si possono – infine – comporre dei flussi temporali[14] che diano conto del variare dell'incidenza di alcune espressioni nel tempo; si possono ricercare le datazioni lessicografiche[15] di qualsiasi espressione.

Ma la linguistica computazionale non risolve la ricerca facendone una semplice questione di numeri (o di occorrenze). Il lavoro filologico rimane fondamentale, tanto più che l'utilizzo della linguistica computazionale nasconde alcune insidie.

Un esempio pratico del genere di trappole nascoste nell'uso della linguistica computazionale è, per rimanere nell'ambito delle analisi da noi condotte, il concetto di «terreur» in Robespierre. I risultati della ricerca delle occorrenze indicano come questa lessia ricorra 189 volte nelle *Œuvres*. 51 volte Robespierre la pronuncia nel periodo che va dal 27 luglio 1793 al 9 termidoro anno II. Il dato sembrerebbe indicare che, nel periodo in cui il Terrore si sviluppa e raggiunge il culmine, Robespierre si soffermi più volte ad analizzare le sue forme ed i suoi obiettivi. Ma se andiamo a verificare il testo, scopriamo che l'uso della parola *terreur*, in Robespierre, assume per lo più – anche durante dell'anno II – caratteristiche generiche ed il suo significato è quasi sempre quello di «paura».

Questo esempio può far intuire come solo la fantasia e la creatività del ricercatore, affiancate da una conoscenza profonda del materiale (fonti e storiografia) inerente all'oggetto della ricerca, possano permettere di utilizzare la linguistica computazionale con profitto. Creando categorie, schemi interpretativi e campi semantici adeguati, è possibile integrare i risultati dei programmi d'interrogazione dei testi all'interno degli studi storici. Programmi che risultano essere *anche* strumenti validi per verificare se gli schemi interpretativi creati sono corretti. Se lo storico sa interagire attivamente, la linguistica computazionale può risultare essa stessa una fonte d'idee nuove.

Per sviluppare uno studio dei testi che utilizzi la linguistica computazionale, la prima operazione da compiere è informatizzare lo scritto[16] (o gli scritti) che si vuole analizzare, cioè creare un *corpus* informatico.

Creare un *corpus* di testi sul quale sia possibile svolgere delle ricerche mediante l'utilizzo del *personal computer*, è un'operazione solitamente gravosa. Per svolgerla al meglio è auspicabile lavorare con un'equipe numerosa oltre che qualificata.

La prima operazione da compiere è l'acquisizione dello stampato cartaceo. L'operazione di acquisizione di pagine e volumi avviene attraverso una periferica che si collega al computer detta scanner (di qui il neologismo scannerizzazio-

ne, dall'inglese *to scan* = scrutare). Lo scanner acquisisce le pagine di testo come se fossero delle immagini. Terminata questa fase, i «file immagine» non possono ancora venire riconosciuti dai programmi di analisi computazionale in quanto non sono in formato testuale[17]. Per trasformare le immagini, che lo scanner ha elaborato, in file di testo, è necessario un software che compia un'operazione detta *OCR* (Optical Character Recognition). I documenti elaborati dall'*OCR* presentano comunemente le desinenze *doc*[18] e *txt*.

Una scelta importante da compiere – durante questa fase del lavoro – riguarda la suddivisione della fonte cartacea in porzioni di documento informatico. Si può procedere in quattro direzioni: a) creazione di un unico documento per tutto il testo acquisito; b) creazione di tanti documenti quante sono le pagine acquisite; c) creazione di tanti documenti quanti sono le fonti originali acquisite; d) creazione di documenti secondo segmenti temporali. È auspicabile che il testo acquisito sia convertito utilizzando tutti e quattro i metodi. Il primo sistema permette la ricerca delle occorrenze, delle cooccorrenze e delle espressioni di frequenza, il secondo ed il terzo sono utili nella ricerca dei riferimenti, delle concordanze e delle datazioni lessicografiche. Il quarto metodo permette di creare i flussi temporali.

Globalmente, le opere dei tre autori, di cui abbiamo creato i *corpora*, sono state informatizzate utilizzando tutti i sistemi esposti. Le *Œuvres* di Robespierre sono state acquisite tenendo come punto di riferimento l'edizione cartacea della *Phénix Éditions* del 2000[19]. Ciò mi ha permesso di creare, con semplicità, degli indici dei riferimenti rispetto all'indicazione di pagina di questa edizione[20]. Grazie al *Concordance*[21], che svolge questo lavoro in automatico, è stato ricostruito, poi, un unico documento che comprende tutto il testo delle *Œuvres*. Il *corpus* delle *Œuvres complètes* di Saint-Just presenta le stesse caratteristiche[22]. La sola differenza è che il *corpus* delle *Œuvres* di Robespierre, a causa del formato originale cartaceo in 8°, comprende due pagine dell'edizione stampata per ogni file[23]. Il *corpus* delle *Œuvres politiques* di Marat[24] – invece – rispecchia, nella divisione in file, la fonte edita originale da cui proviene[25]. Anche per Marat, poi, è stato creato un file unico comprensivo di tutto il testo delle *Œuvres politiques*. Infine, i file del *corpus* delle *Œuvres politiques* sono stati divisi anche secondo segmenti temporali. Questa operazione è stata resa possibile dalla struttura delle *Œuvres politiques*, all'interno delle quali i testi si presentano ordinati cronologicamente.

In definitiva i file che si riferiscono a Robespierre o a Saint-Just danno l'indicazione del volume[26] e della pagina delle edizioni recenti in cui sono state pubblicate, mentre i file delle *Œuvres politiques* di Marat indicano la pubblicazione maratiana da cui provengono e l'anno ed il mese in cui sono state stampate. Ovviamente in entrambi i casi è possibile risalire, mediante un raffronto fra i risultati dei programmi di interrogazione dei testi e le edizioni originali, alle informazioni mancanti.

Un problema che deve essere affrontato in fase di acquisizione del testo riguarda la sua pulizia: la presenza di eventuali errori grafici o di porzioni di testo inutili o dannose al fine delle ricerche lessicologiche. Infatti – anche se i programmi

che svolgono l'*OCR* sono molto sofisticati – spesso lo stampato presenta imperfezioni, sbavature, lettere parzialmente cancellate, soprattutto nelle edizioni datate e nelle ristampe anastatiche. Il 99% degli errori che si verificano dopo aver *riconosciuto* un testo tramite l'*OCR* sono dovuti proprio alla cattiva qualità di stampa. Nell'ambito delle ricerche che stiamo svolgendo è il caso delle *Œuvres* di Robespierre, la cui riproduzione anastatica presenta – in alcuni punti e soprattutto nei volumi dal VI al IX – numerose imperfezioni, arrivando – nei casi limite – alla quasi impossibilità di lettura del testo. Per questo motivo, seppure io abbia svolto un'approfondita correzione del testo, stimo che ci sia, nel *corpus* delle *Œuvres* di Robespierre, una percentuale di errori quantificabile al massimo al 4% (circa 3,5%). Sono giunto a questo dato mediante stime e calcoli fatti su un campione casuale di 150 documenti (circa 300 pagine cartacee).

I volumi dal I al V ed il X, la cui qualità di stampa è migliore, coprono circa la metà (49,86%) dei 2577 file in cui è divisa l'opera informatizzata di Robespierre. Sono 1285 documenti, in cui ho riscontrato, sempre mediante una ricerca a campione, un indice d'errore inferiore all'1%. Dei restanti documenti, seppure in buona parte corretti, quelli provenienti dai volumi VII,VIII,IX hanno percentuali d'errore ancora considerevoli, a causa della pessima qualità di stampa degli originali cartacei. Sono 927 documenti, circa 1/3 del totale (35,97%). Qui la percentuale d'errore si attesta attorno all'7%.

Per minimizzare la percentuale d'errore nei risultati delle ricerche linguistiche basate sul *corpus* a mia disposizione, sono stato aiutato dalla presenza, su Internet, degli archivi informatici dei volumi VI, VII, VIII, IX delle *Œuvres* di Robespierre presenti sul sito dell'*ATILF*[27] gestito dal *CNRS*[28] francese. Il *CNRS* permette – previo abbonamento – di svolgere ricerche sui *corpora* messi in linea tramite il programma di interrogazione dei testi *STELLA*[29]. Ho potuto così confrontare i risultati di ricerca ottenuti attraverso il *Concordance*, il *DBT*, il *Bruco*[30], con le verifiche effettuate presso il sito dell'*ATILF* e correggere i punti in cui il *corpus* in mio possesso risultava essere più carente.

S'incorre in un ulteriore problema metodologico nel momento in cui i testi acquisiti sono edizioni critiche e/o annotate di fonti originali. Il problema da risolvere riguarda l'attendibilità dei risultati di un *corpus* che contenga porzioni di testo che non provengono dal pugno dell'autore, come possono essere le note, i titoli ed i paragrafi esplicativi. Lavorando a livello di liste di frequenza assoluta[31] il problema non è affatto secondario. Per avere delle liste di frequenza assoluta contenenti unicamente parole provenienti dal lessico del personaggio che si sta studiando, i nostri *corpora* non dovrebbero contenere le note, i titoli ed i paragrafi esplicativi. Questa scelta – però – toglierebbe al ricercatore un supporto prezioso e, in alcuni casi, impoverirebbe il testo da eventuali varianti stilistiche. A tale proposito un chiaro esempio (ma se ne potrebbero citare moltissimi) si può trovare nel *Discours sur les peines infamantes*[32], all'interno del quale è presente – in nota – una variante. Nelle due edizioni del *Discours sur les peines infamantes* (1784, 1785), Robespierre utilizza prima *félicité publique*, poi *bonheur public*.

I *corpora* dei tre autori trattati in questo volume, sono comprensivi di note, paragrafi esplicativi e titoli. Rispecchiano, in maniera puntuale, lo stampato dal quale sono stati acquisiti[33].

Una scelta metodologica valida può essere quella di lavorare, soprattutto nel caso di confronti fra più autori, sui dati della frequenza relativa[34]. Operazione che viene compiuta, in questo volume, per mettere a confronto il lessico di Robespierre, Marat e Saint-Just[35]. L'idea di preferire l'indice di frequenza relativa all'indice di frequenza assoluta, risente della considerazione che, all'aumento del totale delle parole in un testo, corrisponda un aumento proporzionale delle occorrenze della parola ricercata nell'apparato di supporto alla lettura. È probabile, infatti, che al di fuori del testo di un autore (nelle note, nei titoli e nei paragrafi esplicativi) compaia qualche occorrenza delle lessie che sono oggetto di ricerca. Questo fa sì che la variazione dell'indice di frequenza relativa non sia considerevole in base alla presenza o meno dell'apparato di supporto alla lettura. La scelta che massimalizza le possibilità offerte dalla linguistica computazionale (pur essendo la più complicata e dispendiosa) è, comunque, quella di avere a disposizione più *corpora* della medesima opera, che servano a svolgere operazioni diverse.

In merito a Robespierre, un'ulteriore questione che deve essere messa in luce, riguarda le fonti primarie dalle quali sono stati raccolti gli interventi per essere inseriti nei volumi delle *Œuvres*. Molti dei passi raccolti nelle *Œuvres* – infatti – sono stati tratti da giornali, i quali, come è ovvio, non sempre hanno l'interesse (o lo spazio) per riportare le parole esatte dell'oratore, ma ne trascrivono il senso generale. Se si dovesse fare uno studio sul lessico di Robespierre (e non sul pensiero o le teorie politiche) tutti gli interventi riportati dai giornali dovrebbero essere passati attentamente al vaglio. Dovrebbero essere presi in considerazione solo i passi in cui il giornale riporta le parole dell'oratore. Il *corpus* risulterebbe, per questo motivo, molto più scarno. Non bisogna dimenticare – inoltre – che la maggior parte degli interventi raccolti dai curatori delle *Œuvres*, vengono presentati secondo le numerose varianti presenti nelle differenti testate giornalistiche. Questo crea – e deve essere tenuto nella dovuta considerazione – numerose ripetizioni[36].

In futuro (tempo, forze e possibilità permettendo) conto di poter creare *corpora* più raffinati, che rispondano alle diverse esigenze di ricerca.

3.1 – Software di interrogazione dei testi utilizzati Concordance, DBT, Bruco

I risultati delle ricerche di linguistica computazionale, esposti nella seconda parte di questo volume, sono stati ricavati utilizzando tre software, i quali, studiati con l'intento di compiere pressoché i medesimi compiti, sono stati realizzati con interfaccia, database e motori di ricerca abbastanza diversi. Intrecciando i risultati ottenuti mediante questi tre software si può essere sicuri della correttezza

numerica del proprio lavoro. Come ho già indicato in precedenza, dove possibile, si è proceduto ad un'ulteriore verifica intrecciando i dati già acquisiti con quelli ricavabili mediante il programma *STELLA* sul sito dell'*ATILF*.

Il primo software utilizzato, il *Concordance*[37](ultima versione 3.1), è un programma anglo-americano. I suoi punti di forza sono la rapidità di elaborazione e la semplicità d'uso. Anche l'utente poco competente può utilizzare questo programma con discreto profitto.

Le sue peculiarità positive sono:

1) un'interfaccia molto semplice ed intuitiva;

2) la possibilità di utilizzare i file *txt* grezzi – che sono stati prodotti dall'OCR – senza alcuna etichettatura preliminare del testo[38];

3) la possibilità di creare database complessi formati da molteplici file *txt*; il programma raccoglie assieme i *txt*, creando un unico database/*corpus*[39];

4) la possibilità di personalizzare la lista degli indicatori di confine di parola (separatori);

5) la possibilità di creare dei database mirati che rispondano a delle esigenze precise. Per esempio, il *Concordance* può ricercare le cooccorrenze, oppure limitare la ricerca delle occorrenze solo ad alcune parole (tramite l'opzione *pick list*[40]). Queste funzioni danno la possibilità di evitare la creazione, se non è necessaria, della lista esaustiva delle occorrenze/concordanze, in cui sono presenti tutte le occorrenze/concordanze di tutte le parole. I vantaggi di queste funzioni si possono valutare in tempo ed in quantità di spazio logico occupato.

6) Il *Concordance* presenta – in maniera del tutto automatica – la lista della frequenza delle collocazioni di una parola rispetto alle quattro parole che la precedono e che la seguono. Questa opzione facilita notevolmente la ricerca delle lessie complesse.

La lacuna principale del *Concordance* è di non permettere l'estrapolazione dei riferimenti rispetto alle pagine dell'opera cartacea d'origine. I riferimenti, infatti, vengono espressi attraverso il numero della riga del documento *txt* sul quale abbiamo svolto la ricerca delle occorrenze/concordanze.

Anche l'impossibilità di creare database con una quantità illimitata di file può essere considerato un limite di questo programma. Il numero massimo di documenti *txt* diversi che un solo database può contenere, infatti, è 387. Per creare un unico database contenente i 2577 documenti acquisiti dall'edizione cartacea delle *Œuvres* di Robespierre, per esempio, ho dovuto lavorare per passaggi successivi raccogliendo il materiale in *txt* più grandi (contenenti alcune centinaia di *txt* sorgente) per poi poter raccogliere questi ultimi in un unico database.

Un'indicazione importante, per chiunque non abbia mai usato questo programma, riguarda il formato dei file in entrata, il quale, oltre ad essere obbligatoriamente *txt* (il che non è un problema), deve contenere «l'interruzione di linea» altrimenti il programma, pur funzionando, diviene considerevolmente più lento. Nel mio caso ho dovuto aprire i tre file *txt*, contenenti le opere degli autori, mediante *Microsoft Word* e salvarli come «file di testo con l'interruzione di linea[41]».

Il secondo programma che ho utilizzato è il *DBT* (ultima versione *DBT 2000*). Il *DBT* è un programma italiano sviluppato a Pisa[42]. I suoi limiti principali sono la difficoltà di utilizzo da parte di un utente non esperto ed il grosso lavoro preliminare sul *txt* sorgente. Anche per riconoscere semplicemente un testo, il *DBT* deve avere una sigla all'inizio del documento che lo configura come documento acquisibile. Questa sigla è di tipo %NOME.

Un altro problema in cui sono incappato nell'utilizzo del *DBT* riguarda il nome della cartella in cui il programma deve venir installato. Se una qualsiasi delle cartelle del percorso in cui il programma viene installato ha un nome più lungo di 8 caratteri, il programma non funziona. Questo inconveniente dipende dalle condizioni di gestione della memoria dei sistemi hardware/software negli anni in cui *DBT* è stato concepito e richiederebbe, per essere eliminato, di riscrivere completamente il programma.

Il vantaggio più grande, che ho riscontrato in questo programma rispetto agli altri, si riferisce alle accentazioni. Il *DBT* non richiede che una parola venga ricercata seguendo la corretta grafia degli accenti ma, in output, espone tutte le parole che corrispondano all'input senza preoccuparsi degli accenti. Questo è un grosso vantaggio soprattutto nei casi, come quello delle *Œuvres* di Robespierre, in cui (a causa dei diversi periodi di pubblicazione, delle scelte dei curatori o a causa degli originali da cui è stato tratto il testo) alcune parole vengono stampate con diverse forme grafiche degli accenti. Per citare solo un esempio, la parola «intérêt» è riscontrabile in questa veste nei tomi delle *Œuvres* contenenti i discorsi (tomi VI-X), mentre, nei primi volumi (tomi I-V delle *Œuvres*), è presente la forma «intérét». Questo comporta (utilizzando *Concordance* e *Bruco*) l'obbligo, per l'utente, di una ricerca che copra più parole; ricerca che, oltre a rallentare tutto il lavoro, aumenta le possibilità di errori.

Una possibilità offerta dagli ideatori del *DBT* è quella di integrare il programma ad un sito Internet, per rendere possibili le ricerche su *corpora* direttamente dalla rete[43]. La licenza – in questo caso – risulta essere abbastanza costosa.

Il *DBT* prevede l'installazione di una versione di *Microsoft Word* per l'esportazione di porzioni di testo.

Mi rendo conto di aver parlato principalmente delle lacune del *DBT*, ma sinceramente il suo utilizzo non è per nulla intuitivo. Si pensi che il manuale si compone di più di 350 pagine. Intuisco che il *DBT* abbia, conoscendolo a fondo, grosse possibilità di utilizzo, ma rimango scettico su alcune scelte fatte dai programmatori. Il *DBT*, in ogni caso, esprime il massimo delle sue potenzialità su file altamente etichettati[44]. Mediante un'etichettatura preliminare – infatti – il programma può rispondere a ricerche complesse come individuare le parole che sono alla fine di un verso (in un testo di poesia) o se una forma è usata da uno specifico personaggio (in un testo teatrale). Mediante un *tag* specifico è possibile verificare le porzioni di testo in corsivo[45].

Questi due programmi hanno un grosso vantaggio sul terzo, cioè la possibilità di configurare una lista standard di parole (detta *pick list*). Questo permette di

limitare le ricerche alle parole cui si è interessati, velocizzando notevolmente il lavoro. Ma se per il *Concordance* la *pick list* è integrata con l'interfaccia, cioè l'utente la può modificare dall'interno del programma, per il *DBT* questo non è previsto; la *pick list* – infatti – deve essere modificata editando un file del tipo "LISTA.wrd". Il terzo programma non ha ancora sviluppato l'opzione *pick list*.

L'ultimo programma che ho utilizzato è un software non ancora in commercio sviluppato da mio padre, Bruno Marin. È denominato *Bruco* (*Brevetto di Ricerca dell'Ubicazione delle Concordanze, Occorrenze, Cooccorrenze*). Il suo difetto principale è la lentezza nella creazione del database e nella ricerca delle cooccorrenze[46]. Ovvia, di contro, ad alcune carenze – soprattutto nella ricerca dei riferimenti – degli altri due software.

Il *Bruco*, diversamente dagli altri programmi, lavora su file *doc*, conformi al 100% al testo cartaceo (anche come formattazione). Questa scelta permette di avere due vantaggi: 1) in fase di correzione del testo è possibile trovare, con facilità, i riferimenti all'opera cartacea e controllare così, immediatamente, il testo originale[47]; 2) si può estrarre la lista dei riferimenti relativi al volume (o ai volumi) da cui i documenti sono stati acquisiti (indicazione del volume, della pagina e della riga).

Al pari del *Concordance*, il *Bruco* permette di utilizzare il carattere Jolly (*) per ricercare tutte le parole che presentino un gruppo di lettere[48]. Diversamente dal *Concordance* non permette di personalizzare la lista degli indicatori di confine di parola (separatori), rendendo così alcune ricerche molto più complicate[49].

Bisogna sottolineare che il Bruco, come il *DBT*, prevede la presenza sul *pc* di un software di videoscrittura. Nel caso non si abbia a disposizione una versione di *Microsoft Word* è possibile scaricare – dal sito www.openoffice.org – un software (*open source*) che svolge le stesse funzioni. Questo programma si chiama *OpenOffice*. La versione più recente è la 2.0.

L'analisi di questi software ha cercato di essere il più possibile precisa e puntuale ma risente, in ultima analisi, del tipo di lavoro di cui mi sono occupato. Non pretende, quindi, di essere esaustiva.

In prospettiva, conto di poter utilizzare pienamente anche il programma di interrogazione dei testi *STELLA* (che il Dipartimento di Storia dell'Università di Trieste si accinge ad acquistare) e di acquisire familiarità con gli strumenti e le metodologie della lessicometria.

3.2 – Considerazioni metodologiche sulla grafia delle parole

Visto che le ricerche lessicologiche si basano sull'unità fondamentale «parola», non è questione di secondaria importanza accennare alla forma grafica che assumono alcune parole nei diversi testi, che abbiamo preso in considerazione durante le nostre ricerche.

I curatori delle *Œuvres* di Robespierre annotano di aver mantenuto l'ortografia degli originali settecenteschi[50]. Per quanto riguarda le *Œuvres complètes* di Saint-Just e le *Œuvres politiques* di Marat, i rispettivi curatori avvertono – invece – che l'ortografia è stata aggiornata all'uso contemporaneo[51]. A questo proposito non posso che essere d'accordo con Cesare Vetter, il quale, nel suo *Il dispotismo della libertà*, esprime scetticismo nei confronti della scelta, dei curatori delle *Œuvres politiques* di Marat, di «ammodernare grafia e punteggiatura e di correggere gli errori grammaticali[52]».

L'ortografia delle parole pone problemi alla linguistica computazionale[53]. Un testo in cui una parola compaia con molteplici forme grafiche, rende più complicato il lavoro di ricerca delle occorrenze, delle concordanze ma soprattutto delle cooccorrenze. Due possono essere le strade percorribili. Mantenendo fissa la necessità di conoscere a fondo il lessico dell'autore studiato, è possibile, da un lato, uniformare il *corpus* secondo degli standard grafici[54]. In questo caso rimane la necessità di verificare il testo per le varianti ortografiche. Dall'altro lato, mantenendo il testo nella sua forma originaria, la ricerca di alcune cooccorrenze (come possono essere «Être – suprême» o «faibles – lois») non è sempre un lavoro semplice[55]. Forse la seconda strada è la migliore ma prevede, comunque, un lavoro capillare e attento. I *corpora* che abbiamo prodotto rispettano l'ortografia e la grammatica presenti nelle edizioni da cui sono stati acquisiti.

3.3 – Guida alla lettura e scelte metodologiche inerenti alle liste di frequenza[56]

I primi risultati della ricerca lessicologica che presento sono le liste di frequenza assoluta e di frequenza relativa delle *Œuvres* di Robespierre, delle *Œuvres politiques* di Marat e delle *Œuvres complètes* di Saint-Just. È necessario illustrare qualche scelta metodologica effettuata.

Innanzitutto devo indicare quali motivazioni mi hanno portato a dividere le liste di frequenza dei tre autori in «liste di frequenza di parole[57]», «liste di frequenza di lessie composte e complesse» e «liste di frequenza di nomi propri».

Sulle *liste di frequenza di parole* (*lessie semplici* e, in alcuni casi, *lessie composte*) non è necessario fare alcuna considerazione particolare essendo, questo, uno degli studi lessicologici più semplici e consueti.

La considerazione di non includere in questi indici le *liste di frequenza delle lessie composte e complesse*, deriva – principalmente – da questioni inerenti all'indice di frequenza relativa. Le parole e le lessie composte e complesse fanno parte di due insiemi numericamente diversi ed incommensurabili.

Se da un lato è valida l'affermazione che la frequenza relativa normalizzata di «peuple» è 0,26953%, visto che questo dato si calcola dividendo il numero delle occorrenze di *peuple* per il totale delle parole presenti nel testo, è corretto d'altro canto affermare che la frequenza relativa normalizzata di «amis de la liberté» è

0,01070%? La lessia *amis de la liberté* è composta da quattro parole. L'eventuale dato della frequenza relativa normalizzata si basa su un insieme (il totale delle parole) che conta, per ogni occorrenza di questa espressione, quattro parole. Mi chiedo: per avere un dato valido bisogna moltiplicare l'insieme «totale delle parole» per il numero delle parole che compongono l'espressione (in questo caso quattro)? Non credo, visto queste considerazioni, che l'indice di frequenza relativa sia un indice valido per le lessie costituite da più parole.

Infine l'idea di creare delle *liste di frequenza di nomi propri* deriva dalla considerazione che i nomi propri sono, nel lessico, elementi diversi rispetto agli altri. Una motivazione subordinata risente della volontà di rendere più rapida (e quindi più semplice) la consultazione delle liste di frequenza dei nomi propri.

Bisogna sottolineare che le liste di frequenza consultabili in questo volume non sono esaustive. Questa scelta (obbligata) deriva dalla necessità di comprimere la pubblicazione in costi sostenibili.

Le lessie semplici, composte e complesse ed i nomi, inseriti nelle liste di frequenza ragionate, sono stati scelti in base all'importanza che rivestono nella produzione dei tre autori, nel dibattito storiografico e nel lessico dell'epoca. I criteri di scelta non riguardano la loro frequenza nel testo[58]. Scorrendo le liste si possono incontrare oltre a parole, come «liberté», che segnano – nelle *Œuvres* di Robespierre – più di 5000 occorrenze nel testo, anche parole come «perfectibilité» che – nella stessa sede – contano una sola occorrenza (*hapax*[59]). Ciò che può risultare strano è che – nelle liste di frequenza inserite nelle pagine seguenti – compaiono anche parole che nel testo degli autori non sono presenti, come, in Robespierre, «autonomie». Anche l'assenza di una parola è un risultato di ricerca. È ovvio che se potessi presentare delle liste di frequenza esaustive non ci sarebbe bisogno di indicare le parole che non compaiono nelle opere dei tre autori.

Ciò che non viene mai preso in considerazione (se non nelle liste delle cento parole più frequenti) sono le parole vuote[60]. Gli studi che abbiamo condotto sui tre autori riguardano – infatti – solamente il lessico e tralasciano altri aspetti fondamentali della dimensione linguistica (ma non solo) come lo stile e la struttura sintattica e grammaticale del discorso.

Il criterio di presentazione delle liste di frequenza ragionate è alfabetico.

Concludo dando ancora tre indicazioni, utili come guida di lettura alle liste di frequenza: 1) nei casi in cui sia possibile ed il senso non cambi, le occorrenze delle lessie costituite da due parole (solitamente un sostantivo ed un aggettivo), non tengono conto della rispettiva posizione delle parole. Ad esempio, gli indici di frequenza di «bon citoyen» comprendono anche le occorrenze di «citoyen bon». Ovviamente questo non è possibile con espressioni del tipo di «amour maternel», poiché non si troverà mai «maternel amour».

2) In tutte le liste di frequenza alcune lessie sono state accorpate ad altre; a volte perché si tratta di variazioni grafiche della stessa lessia; a volte per accorpare singolare e plurale della stessa lessia[61]. Esempi: «âme» e «ame»; «apocalyptique» e «apocalyptiques». Ovunque, scorrendo le liste di frequenza, si trovi una

lettera (o una parola) fra parentesi come nel caso di «A(â)me», la lettera (o la parola) fra parentesi va *sostituita* a quella che la precede. Nel caso si riscontri, all'interno della parentesi, una lettera preceduta da una barra come in «apocalyptique(/s)», la lettera fra parentesi va *aggiunta* alla parola. Il numero delle occorrenze – in questi casi – si riferisce alla somma delle occorrenze delle due parole. Come per l'occorrenza di «A(â)me» = 264, la quale è la somma dell'occorrenza di «ame» = 48 e dell'occorrenza di «âme» = 216.

3) Le definizioni tratte dalla lingua inglese *word types* e *(word) tokens* riflettono la terminologia degli studi di linguistica computazionale più recenti: «Ogni occorrenza delle parole testuali è detta *token*. Un testo è costituito da un certo numero di *word tokens*, ossia da un certo numero di parole (che possono anche ripetersi nel testo stesso), costituito dalla somma di tutte le occorrenze di qualunque tipo di parola nel testo. Le forme delle parole diverse sono invece dette TIPI DI PAROLE (*word types*)[62]».

3.4 – Guida alla lettura e scelte metodologiche inerenti agli indici delle concordanze

Innanzitutto indico – in maniera sintetica – gli argomenti dei tomi delle *Œuvres* di Robespierre da cui sono stati tratti gli indici delle concordanze di «bonheur», «félicité», «heureux», «démocratie», «démocratique», «terreur», «liberté civile», «liberté politique», «liberté publique» e gli indici delle concordanze delle cooccorrenze[63] «monarchie» – «république», «terreur» – «bonheur», «terreur» – «heureux»[64]:

Tomo I: *Œuvres littéraires*.
Tomo II: *Œuvres judiciaires*.
Tomo III: *Correspondance*[65].
Tomo IV: *Le Défenseur de la Constitution*.
Tomo V: *Lettres de Maximilien Robespierre, membre de la Convention nationale de France, à ses commettants*[66].
Tomo VI: *Discours* (1789-1790)[67].
Tomo VII: *Discours* (janvier-septembre 1791)[68].
Tomo VIII: *Discours* (octobre 1791-septembre 1792)[69].
Tomo IX: *Discours* (septembre 1792-juillet 1793)[70].
Tomo X: *Discours* (27 juillet 1793-27 juillet 1794)[71].
Come ho già accennato sopra, il testo delle *Œuvres* di Robespierre non contiene solamente materiale scritto, edito, stampato o pronunciato da M. Robespierre. Una parte del testo è composto da: 1) passi di giornali che si riferiscono a Robespierre; 2) corrispondenza ricevuta da Maximilien e da Augustin Robespierre o inviata da quest'ultimo; 3) note, paragrafi esplicativi e titoli inseriti dai curatori dell'opera.

Sicuramente le concordanze di *bonheur*, *félicité*, *heureux*, ecc. presenti in queste sezioni di testo non possono essere conteggiate ed inserite negli indici delle concordanze come se fossero pronunciate da Robespierre. Per non perdere la ricchezza di questa casistica e per mantenere una corrispondenza fra il dato della frequenza assoluta ed il numero delle concordanze presentate, ho pensato di inserire i riferimenti, che si riferiscono a queste concordanze «esterne», in uno speciale indice che ho denominato «indice dei riferimenti esterni».

Un altro indice «speciale» delle concordanze è l'«indice delle concordanze delle cooccorrenze». L'*indice delle concordanze delle cooccorrenze* è un indice in cui sono inseriti i contesti in cui è presente la cooccorrenza ricercata. In questo primo volume presento gli indici delle concordanze delle cooccorrenze a distanza 50[72] di «monarchie» – «république», di «terreur» – «bonheur» e di «terreur» – «heureux».

3.4.1 – Contesti

Per quanto riguarda i contesti delle concordanze, si è scelto di non limitarli – come comunemente avviene in pubblicazioni analoghe – alla riga in cui compare la parola cercata ma – per mantenere un significato comprensibile – tutti i contesti presentati sono compresi fra due segni d'interpunzione forti (due punti; un punto ed un punto e virgola; due punto e virgola[73]).

Segnalo che i contesti in cui compaiono due (o più) delle parole di cui vengono stilati gli indici delle concordanze, sono stati ripetuti. Per esempio, il seguente contesto: «Heureux de la félicité de mes concitoyens, je passerais des jours paisibles dans les délices d'une douce et sainte intimité», è presente sia nell'indice delle concordanze di *heureux* che in quello di *félicité*.

3.4.2 – Guida alla lettura delle concordanze

Tomi I-V delle *Œuvres* di Robespierre.
Quella di seguito è la prima concordanza di «bonheur» che compare nelle *Œuvres*:
1) Tomo I
2) *Discours sur les peines infamantes, couronné par l'académie de Metz en 1784*, pp. 5-77[74].
3) p. 20 (1)
4) C'est un sublime spectacle de voir les compagnies sçavantes, sans cesse occupées d'objets utiles à l'interet public, inviter le génie, par l'appas des plus flatteuses récompenses à combattre les abus qui troublent le bonheur de la société.
Vado ad illustrare gli elementi da cui è composta:

1) Indicazione del volume delle *Œuvres* a cui si riferisce la concordanza. Questa indicazione compare una sola volta per ogni lista di concordanze, le successive concordanze si considerano appartenenti al volume in questione fino all'indicazione di volume successiva.

2) Indicazione dell'opera robespierriana (opera edita, giudiziaria, giornale o missiva) in cui compare la concordanza. I numeri di pagina si riferiscono alla numerazione dei tomi delle *Œuvres* dove è contenuta l'opera in questione.

3) Indicazione della pagina in cui compare la concordanza. L'eventuale numero fra parentesi identifica quante concordanze della parola in questione compaiono nella pagina. In caso di nessuna indicazione s'intende 1.

4) Contesto della concordanza. In maiuscoletto la parola (o le parole) di cui si espone la concordanza. Vengono mantenuti i segni d'interpunzione presenti nel testo originale; anche quelli che chiudono il contesto presentato.

In caso sia presente, prima di un contesto, unicamente un'indicazione di pagina, si deve considerare la concordanza come facente parte dello stesso testo della concordanza precedente. Il numero fra parentesi si riferisce sempre alla quantità di occorrenze presenti nella pagina.

Tomi VI-X delle *Œuvres*.
Le indicazioni che si riferiscono ai volumi dei discorsi sono leggermente più complicate:
1) Tomo VII
2) Société des Amis de la Constitution
3) Séance du 11 mars 1791, 1er intervention
4) *Sur la responsabilité des ministres*, p. 121.
5) p. 121 (1), *Cicéron à Paris*, n° 39, p. 5
6) Or, des ministres doivent être responsables de fait et de droit envers la nation, puisqu'ils tiendront dans leurs mains le bonheur et la tranquillité de l'empire. Je conclus donc à ce que les ministres soient électifs.

1) Stesso discorso fatto poco sopra al numero 1).

2) Indicazione del luogo in cui il discorso è stato pronunciato[75].

3) Indicazione della data in cui è stato pronunciato il discorso. In caso Robespierre intervenga più di una volta nella medesima seduta, indico a quale intervento mi riferisco.

4) Titolo dato dai curatori delle *Œuvres* agli interventi di Robespierre. I numeri di pagina si riferiscono alla numerazione dei tomi delle *Œuvres* dove è contenuto il discorso in questione.

5) Indicazione della pagina in cui occorre la concordanza; fonte originale (giornale, discorso stampato) da cui è stata tratta. Il numero fra parentesi – come sopra – si riferisce al numero di occorrenze della parola nella pagina.

6) Contesto della concordanza. In maiuscoletto la parola (o le parole) di cui si espone la concordanza. Vengono mantenuti i segni d'interpunzione presenti nel testo originale; anche quelli che chiudono il contesto presentato.

3.5 – Scelte metodologiche inerenti alle cooccorrenze

Una questione che mi si è imposta, e che ha rivestito particolare importanza, è inerente al lavoro sulle cooccorrenze. Si definisce «cooccorrenza» una ricerca volta a trovare due parole in un medesimo contesto[76]. Il parametro principale che deve essere settato in caso di ricerca di cooccorrenze, è la distanza fra le parole da cercare.

Tutte le ricerche effettuate, sono state svolte utilizzando tre misure di distanza:

1) Distanza 1. Mediante questo criterio s'intende ricercare non delle vere e proprie cooccorrenze, bensì delle lessie costituite da due parole, come possono essere «être suprême» o «bien public».

2) Distanza 10. Usato per verificare la presenza di due parole nella medesima frase o in periodi attigui. Spesso se due parole compaiono nella stessa frase, fanno parte del medesimo discorso e sono intimamente legate.

3) Distanza 50. Si tratta d'individuare parole che probabilmente non fanno parte della medesima frase o discorso, ma comparendo nella stessa pagina (e non troppo lontane) presentano comunque un legame logico, seppure alle volte non molto forte.

Nella ricerca delle cooccorrenze non viene mai valutata la rispettiva posizione delle parole nel testo. Esempio: le cooccorrenze a «distanza 1» di «bonheur» nel contesto sottostante sono sia l'articolo «le» (che precede «bonheur»), sia l'aggettivo «public» (che segue «bonheur»)[77]:

> Toujours convaincu que la liberté et le BONHEUR public sont attachés à la propagation des principes, je vous envoie plusieurs ouvrages que je confie à votre patriotisme[78].

3.6 – Riferimenti

L'ultima indicazione riguarda la lista dei riferimenti[79] di alcune lessie presenti nelle *Œuvres* di Robespierre. A causa del metodo usato nell'acquisizione delle *Œuvres*, il software che utilizzo per la ricerca dei riferimenti (*Bruco*) non mi permette di sapere se la parola cercata si trova sulla facciata sinistra o su quella destra di due pagine attigue[80]. Esempio: se un'occorrenza di *bonheur* compare a pagina 407 del X tomo, io non posso sapere – in maniera automatica – se in effetti l'occorrenza è a pagina 407 o in quella che immediatamente la precede (p. 406). Nella lista dei riferimenti – e, in alcuni casi, in altri indici lessicologici – ho indicato con un asterisco (*) le parole di cui ho verificato manualmente i riferimenti.

NOTE ALL'INTRODUZIONE METODOLOGICA

1 Per la bibliografia sulla linguistica computazionale e sull'approccio linguistico (analisi del discorso, lessicografia, lessicologia, lessicometria) alla rivoluzione francese cfr. – oltre alle note dell'*Introduzione*; della parte prima, capitolo primo e alle note della presente *Introduzione metodologica* – i lavori di Jacques Guilhaumou (http://revel.unice.fr/corpus/document.html?id=8; http://www.cavi.univ-paris3.fr/lexicometrica/article/numero0/jgadlex.htm; per la bibliografia a tutto il 2004 cfr. http://dispol.ens-lsh.fr/IMG/pdf/biblioguijac.pdf; per i lavori successivi cfr. http://publiens.ens-lsh.fr/), di Bernard Quemada (http://www.u-cergy.fr/dictionnaires/auteurs/quemada.html; per la bibliografia cfr. http://www.udc.es/grupos/lexicografia/q.htm), di Jean Pruvost (per la bibliografia cfr. http://www.udc.es/grupos/lexicografia/p.htm), di Étienne Brunet (http://www.cavi.univ-paris3.fr/lexicometrica/article/numero1/hypertexte.htm; per la bibliografia completa e per alcuni *links* a lavori di lessicografia computazionale cfr. http://ancilla.unice.fr/-brunet/pub/brunet.html#ouvrages); di André Salem (per la bibliografia cfr. http://www.cavi.univ-paris3.fr/ilpga/ED/dr/asdr/pubcomplete.htm). Cfr. inoltre i seguenti siti: http://www.atilf.fr/; http://www.cnrs.fr/; http://www.ens-lyon.fr/web/nav/; http://www.atala.org/; il sito dei *Cahiers de Lexicologie*: http://atilf.atilf.fr/jykervei/cahlex.htm. Per gli aggiornamenti sulla bibliografia linguistica cfr. *Bibliographie linguistique/Linguistic Bibliography*: http://wwwo.kb.nl/blonline/; http://publiens.ens-lsh.fr/; http://orbita.bib.ub.es/lexic/. Per un indice bibliografico molto vasto di testi inerenti alla lessicografia cfr. http://www.udc.es/grupos/lexicografia/bibliografia.htm.

2 Cfr. J. Guilhaumou, *A propos de l'analyse de discours: les historiens et le «tournant linguistique»*, in «Langage et société», n. 65, septembre 1993, pp. 5-38. Non siamo riusciti a reperire in tempo utile il recente saggio di Jacques Guilhaumou: *La langue politique et la Révolution française*, in «Langage et société», n. 113, septembre 2005, pp. 63-92.

3 Un esempio importante è il dottorato in «Storia ed informatica» attivato presso l'Università degli Studi di Bologna: cfr. http://www.unibo.it/NR/ exeres/2C642B00-0715-4498-A0E1-9F02193A5D04.htm?WBCMODE=PresentationUnpublished.

4 Cfr. – per quanto riguarda l'ambito italiano – l'attività di Antonio Zampolli, che promosse e coordinò a lungo l'istituto di linguistica computazionale (*ILC*) di Pisa: http://www.ilc.cnr.it/ e in particolare http://www.ilc.cnr.it /AZ/ultimointrod-web.pdf. Tra le pubblicazioni in cartaceo cfr. – in particolare – G. Adamo, *Analisi informatica di testi: problemi e prospettive*, in *Calcolatori e scienze umane*, Milano, 1992, pp. 350-365; M. Lana, *L'uso del computer nell'analisi dei testi*, Milano, 1994; Idem, *Il testo nel computer. Dal web all'analisi dei testi*, Torino, 2004; S. Spina, *Fare i conti con le parole. Introduzione alla linguistica dei corpora*, Perugia, 2001; I. Chiari, *Informatica e lingue naturali. Teorie e applicazioni computazionali per la ricerca sulle lingue*, Roma, 2004; I. Chiari, T. De Mauro (a cura di), *Parole e numeri. Analisi quantitative dei fatti di lingua*, Roma, 2005; A. Lehmann, F. Martin-Berthet, *Introduction à la lexicologie, sémantique et morphologie*, Paris, 2005; Aa. Vv., *La linguistique de corpus (actes des deuxièmes journées de la linguistique de corpus, Lorient, 12-14 septembre 2002)*, Rennes, 2005; A. Niklas-Salminen, *La lexicologie*, Paris, 2005.

5 Cfr. I. Chiari, *Informatica e lingue naturali*, cit., pp. 10-11: «La linguistica dei corpora [...] esamina grandi quantità di produzioni linguistiche, scritte o parlate, osservandone le caratteristiche: il lessico, la sintassi, le cosiddette 'collocazioni', la catena fonica, le strutture morfologiche. [...] Tenta di portare alla luce caratteristiche delle lingue altrimenti non rilevabili». Per definire con più precisione il tipo di lavoro effettuato, bisogna indicare che il nostro è uno *studio storico* che si avvale: 1) dei mezzi informatici – principalmente *corpora* informatizzati e programmi d'interrogazione dei testi – necessari per portare alla luce le caratteristiche del *lessico* degli autori; 2) delle definizioni, delle categorie, dell'apparato teorico della «linguistica dei *corpora*». D'ora in avanti utilizzo «linguistica computazionale» nel senso indicato in questa nota.

6 Cfr. più avanti, nota 16.

7 www.diseur.unict.it/ddi/html/definizioni.html: «Il *riferimento* è l'indice del contesto, ed è relativo all'opera, al componimento, al verso (o ai titoli, alle dediche, ecc.)».

8 Ciò che secondo il linguaggio informatico si dice database di un testo, secondo la linguistica è un *corpus*. Secondo il gruppo EAGLES (*Text Corpora Working Group Reading Guide*. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica computazionale), un *corpus* è: «A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language». Un *corpus elettronico* è: «A corpus which is encoded in a standardized and homogeneous way for open-ended retrieval tasks». Cfr. I. Chiari, *Informatica e lingue naturali*, cit., p. 32.

9 La lista di frequenza assoluta, esaustiva o parziale, è un indice contenente il numero di occorrenze, di tutte o di una parte, delle parole presenti in un testo. La definizione più comune di «occorrenza» è: ricorrenza di una parola in un testo concordato. Per una definizione più articolata cfr. il sito del Centro d'Informatica Letteraria Italiana dell'Istituto di Letteratura italiana dell'Università di Catania all'indirizzo www.diseur.unict.it/ddi/html/definizioni.html.

10 www.diseur.unict.it/ddi/html/definizioni.html: «Per *parola* s'intende, in termini d'informatica linguistica, l'unità o forma grafica isolabile tra due spazi bianchi o due separatori (come uno spazio bianco e un segno d'interpunzione, due interpunzioni, ecc.). I lemmi sono quasi tutti formati da una sola parola, ma possono essere formati anche da due o più parole». Cfr. nota 11. Cfr. anche I. Chiari, *Informatica e lingue naturali*, cit.

11 Per la definizione di lessia (*lexie*) cfr. B. Pottier, *Linguistique générale, théorie et description*, Paris, 1974; http://www.tlab.it/it/allegati/help__it__online/glos__fr__def.html; http://atilf.atilf.fr/dendien/scripts/fast.exe?mot=lexie; http://perso.wanadoo.fr/ldelafosse/Glossaire/L.htm#lexie. La lessia – come è noto – può essere semplice, composta e complessa. Isabella Chiari (*Informatica e lingue naturali*, cit., pp. 53-56), sulla scorta del linguista inglese J. R. Firth, propone il termine «collocazione»: «Le collocazioni sono particolari espressioni composte da più di una parola grafica, che tuttavia si comportano semanticamente e spesso morfosintatticamente come un solo lessema. Sono anche genericamente definite come gruppi di parole grafiche che co-occorrono con una probabilità maggiore che se fossero indipendenti». Seppure l'uso del termine «collocazione» semplificherebbe il linguaggio, non ho ancora deciso di adottarlo a causa della possibilità di confondere la «collocazione» (intesa in questo senso) con il «riferimento».

12 Per cooccorrenza (o co-occorrenza) s'intende la presenza, nel medesimo contesto, di due parole. Come specificato più avanti (§ 3.5 – Scelte metodologiche inerenti alle cooccorrenze) le tipologie di cooccorrenza si differenziano in base alla distanza che separa le due (o più) parole.

13 Cfr. www.sapere.it: «La concordanza è la lista di tutte le parole presenti in un testo, elaborata di solito con l'ausilio di apparecchiature elettroniche: *concordanze alfabetiche*, in cui le singole parole sono registrate in ordine alfabetico, seguite solo dall'indicazione del luogo dove esse compaiono; *concordanze di frequenza*, in cui le singole parole, accompagnate da un numero che indica quante volte compaiono nel testo, sono registrate, in ordine crescente o decrescente, secondo la loro frequenza di impiego nel testo stesso; *concordanze delle* o *per forme*, in cui ogni parola, registrata sempre secondo la forma in cui compare, è seguita da un breve contesto tratto da ogni passo dove essa ricorre; *concordanze per lemmi*, in cui tutte le forme di una parola (p.e. i diversi tempi e modi di un verbo) vengono raggruppate secondo un unico "lemma", cioè secondo la forma fondamentale di quella parola (p.e. l'infinito presente): di solito vi è aggiunto anche un breve contesto». Cfr. www.diseur.unict.it/ddi/html/definizioni.html: «Dizionario relativo a uno o più testi, in genere di un solo autore, e contenente, di norma in ordine alfabetico (e senza definizione semantica), i lemmi ai quali si possono ricondurre tutte le parole del *corpus* concordato. La concordanza normalmente riproduce i contesti nei quali sono realizzati i lemmi nelle loro varie forme e occorrenze. Una concordanza si dice esaustiva (o integrale o totale) quando non esclude nessuna parola

dalla lemmatizzazione». In questo volume vengono stilate liste di *concordanze per forme* di alcune lessie (*bonheur*, *félicité*, *liberté politique*, …), corredate dal riferimento e dal contesto in cui compare ogni occorrenza della lessia in esame.

14 Per flusso temporale intendo un indice (affiancato da relativo grafico) che dia conto del numero delle occorrenze di una o più parole, presenti in un *corpus*, in diversi segmenti significativi di tempo (giorni, mesi, anni). In questo primo volume non presento – seppure il lavoro sia già stato impostato – alcun flusso temporale. Rimando al secondo volume la presentazione di questo tipo di lavori lessicologici.

15 Per le datazioni lessicografiche della lingua francese cfr. http://atilf.atilf.fr/jykervei/ddl.htm: *Base Historique du Vocabulaire Français* (*Datations et Documents Lexicographiques*).

16 Visto la definizione forzatamente estensiva di «testo» che comprende necessariamente testi scritti, testi stampati e testi in formato digitale e vista la necessità di chiarezza e distinzione fra «testo scritto o stampato» e «testo digitale», d'ora in poi utilizzo le parole «scritto» e «stampato» intendendo «testo scritto» e «testo stampato». Riservo la parola «testo» (senza aggettivazione) ai casi in cui il significato sia generico: «insieme di parole contenute in uno scritto, uno stampato o un documento in formato digitale». La parola «documento» deve essere intesa sempre – in questa sede – come «documento informatico di testo», «archivio digitale in cui sono contenute informazioni di testo», «file di testo». La parola inglese «file» (intesa in questa introduzione, in mancanza di alcuna aggettivazione, come «file di testo») risulta così essere un sinonimo di «documento».

17 Il formato testuale di gran lunga più utilizzato sui comuni *personal computer* presenta l'estensione *txt*. Questa desinenza indica un file riconoscibile da qualsiasi programma di editor testuale di qualsiasi sistema operativo.

18 Questa desinenza si riferisce ai documenti compatibili con il software *Microsoft Word*.

19 Si tratta di una riedizione anastatica (Paris, 2000) dell'edizione delle *Œuvres de Maximilien Robespierre*, a cura della *Société des études robespierristes* (1912-1967). Ho usufruito dei volumi I-X; di prossima pubblicazione il vol. XI.

20 Mediante il *Bruco* (produttore e proprietario Bruno Marin, mio padre), unico programma – a nostra disposizione – che stilasse liste dei riferimenti.

21 Cfr. § 3.1 – Software di interrogazione dei testi utilizzati.

22 Il lavoro si è basato sull'edizione pubblicata dalle Éditions Gérard Lebovici: L.-A. Saint-Just, *Œuvres Complètes*, édition établie par Michèle Duval, Paris, 1984.

23 A questo proposito, non posso, al momento attuale, dare liste di riferimenti perfettamente rispondenti alle pagine dell'edizione delle *Œuvres* di Robespierre. L'approssimazione è di una facciata. Il programma che uso per l'estrapolazione dei riferimenti (il *Bruco*) presenta, però, l'indicazione della riga – progressivamente per tutto il documento *doc* – in cui compare la concordanza. Mediante quest'informazione è possibile dare un'indicazione di massima più precisa. Sapendo che ogni pagina presenta circa 50 righe nella pagina pari e 50 in quella dispari, se l'indicazione di riga è superiore a 50 probabilmente la concordanza si riferisce alla pagina dispari. Il mio prossimo obiettivo sarà di creare un *corpus* delle *Œuvres* di Robespierre adeguato, su cui il *Bruco* possa funzionare al meglio.

24 L'edizione utilizzata è quella stampata dalla Pôle Nord: J.-P. Marat, *Œuvres politiques 1789-1793*, 10 voll., texte et guide de lecture préparés par J. De Cock et Ch. Goëtz, Bruxelles, 1989-1995.

25 Ad esempio se si trova ne *Les chaînes de l'esclavage*, nel *Plan de législation criminelle*, nel numero 166 de *L'Ami du peuple*, …

26 Questo avviene solamente per Robespierre visto che le *Œuvres complètes* di Saint-Just sono state pubblicate in volume unico.

27 Analyse et Traitement Informatique de la Langue Française.

28 Centre National de la Recherche Scientifique.

29 Per la bibliografia messa a disposizione dall'*ATILF* cfr. http://atilf.atilf.fr/artis/nvlbiblio.htm. Il *corpus* è denominato *FRANTEXT* e comprende opere in lingua francese dal XVI al XX secolo.

30 Cfr. più avanti.

31 Nelle pagine del sito www.frantext.fr dedicate agli abbonati viene data la seguente definizione di frequenza assoluta: «La fréquence absolue d'une forme graphique (nous dirons plus simplement "mot") dans un corpus est le nombre d'occurrences de cette forme dans le corpus».

32 Cfr. M. Robespierre, *Discours sur les peines infamantes couronné par l'Académie de Metz en 1784*, in *Œuvres*, cit., t. I, pp. 5-76, a p. 37.

33 Colgo l'occasione per rimandare alle introduzioni dei curatori dei diversi tomi delle *Œuvres* per le questioni riguardanti la completezza delle stesse, i testi mancanti e le integrazioni prospettate: cfr. in particolare C. Mazauric, *Présentation*, in M. Robespierre, *Œuvres*, cit., pp. I-XXIX, alle pp. XIV-XVI.

34 www.frantext.it (zona abbonati): «La fréquence relative d'une forme graphique dans un corpus est égale à la fréquence absolue de cette forme divisée par la somme des fréquences absolue de toutes le formes graphiques du corpus. Ainsi, si le mot "maison" a 2 occurrences dans un corpus contenant un million d'occurrences, sa fréquence relative est de deux millionièmes». www.diseur.unict.it/ddi/html/definizioni.html: «La frequenza percentuale (o frequenza relativa) è la frequenza assoluta sul totale di tutte le parole-occorrenze del testo. La percentuale è al millesimo e va arrotondata per difetto». L'arrotondamento che viene utilizzato in questo volume, in modo da avere dei dati più precisi e fruibili, è al centomillesimo. Inoltre ho trovato più proficuo utilizzare, al posto dell'indice di frequenza relativa, l'indice di «frequenza relativa normalizzata», il quale si riferisce alla percentuale di frequenza della parola nel testo e si ottiene moltiplicando la frequenza relativa per 100. Nelle liste di frequenza uso l'indicazione «frequenza relativa» per l'indice di frequenza relativa normalizzata.

35 I confronti lessicologici devono tener conto: 1) delle scelte editoriali dei curatori delle edizioni moderne delle opere dei tre autori; 2) delle diversità delle tipologie di fonte (una pubblicazione letteraria come una poesia o un'opera teatrale ha una fisionomia linguistica diversa rispetto ad un discorso pubblico di stampo politico). I confronti attraverso l'utilizzo degli indici di frequenza assoluta e di frequenza relativa devono essere limitati – per questi motivi – a considerazioni di massima sugli ordini di grandezza. Le conclusioni che se ne possono trarre devono essere necessariamente generiche ed indicare le linee guida. Solo l'analisi puntuale ed i confronti filologici possono concretizzare le indicazioni proposte dalla linguistica computazionale.

36 Cfr., a proposito delle questioni trattate in questo paragrafo, M. Bouloiseau, *Note des éditeurs*, in M. Robespierre, *Œuvres*, cit., t. IX, pp. 5-12 , in particolare p. 11; Idem, *Note des éditeurs*, in M. Robespierre, *Œuvres*, cit., t. X, pp. 5-7: «Ceci explique pourquoi nous avons parfois retenu, pour une même intervention, plusieurs extraits dont le sens général était identique, mais entre lesquels existaient des différences sensibles dans la forme».

37 Cfr. I. Chiari, *Informatica e lingue naturali*, cit., p. 131 ed il sito del programma http://www.concordancesoftware.co.uk/.

38 I. Chiari, *Informatica e lingue naturali*, cit., p. 59: «L'aggiunta di informazioni di tipo linguistico si dice *annotazione o etichettatura linguistica*. L'annotazione è una forma di codifica linguistica. Praticamente essa consiste nell'associazione di una *etichetta* (*tag* o *mark-up*) a una porzione specifica e ben delimitata di testo. L'etichettatura può riguardare qualunque aspetto del testo, indicazioni fonetiche, morfologiche, sintattiche, semantiche. L'annotazione di un corpus serve principalmente per poter estrarre successivamente in modo agile e veloce una gran quantità di dati linguistici e non linguistici sul testo».

39 Questa funzione mi ha permesso di creare – risparmiando tempo ed energie – dei file *txt* contenenti l'opera di ciascun autore analizzato. Questi file sono stati necessari per utilizzare il DBT.

40 Cfr. più avanti.

41 La stessa operazione preliminare deve essere compiuta sui file che si vogliono processare con il *DBT*. L'unica differenza è che il *DBT* non funziona affatto se le linee di testo superano la dimensione di 255 caratteri.

42 Il *DBT* (*Data Base Testuale*) è stato creato e sviluppato da Eugenio Picchi presso l'Istituto di Linguistica Computazionale (*ILC*) del Consiglio Nazionale delle Ricerche di Pisa. L'*ILC* è un centro d'eccellenza in ambito nazionale.

43 È analogo a ciò che permette di svolgere il sito dell'*ATILF*, mediante il programma *STELLA*.

44 Cfr. sopra, nota 38.

45 Rimando al manuale del *DBT* per le altre questioni tecniche e ulteriori precisazioni anche sui *tag* specifici. Cfr. www.aracnoidea.it.

46 La ricerca delle occorrenze – dopo l'operazione preliminare di creazione del database – è immediata, come quella del *Concordance* e del *DBT*.

47 Bisogna indicare che il *DBT* non prevede la correzione del testo archiviato dal programma. Il *Concordance* prevede che si effettuino correzioni ma – mancando una lista dei riferimenti attendibile – ne rende difficile l'attuazione. Il *Bruco* – diversamente – prevede la possibilità che un *corpus* non sia ancora del tutto pulito e ne facilita la correzione (tramite la lista dei riferimenti ed il suggeritore implementato in *Microsoft Word*).

48 Esempio: ricercando (*licité*) nelle *Œuvres* di Robespierre, la lista delle parole è: «catholicité», «complicité», «duplicité», «explicitée», «félicité», «multiplicité», «publicité», «simplicité», «sollicité», «sollicitée», «sollicitées», «sollicités».

49 Soprattutto nella ricerca delle lessie – come *sans-culotte* – che possono presentarsi sia con il trattino che senza il trattino.

50 Cfr. G. Laurent, *Introduction*, in M. Robespierre, *Œuvres*, cit., t. IV (1939), pp. I-XXXVIII, in particolare p. XXXV; M. Bouloiseau, *Introduction*, in M. Robespierre, *Œuvres*, cit., t. VI (1950), pp. XI-XXX, in partico-

lare pp. XIV, XXVII e ss.; G. Laurent, *Complément d'introduction*, in M. Robespierre, *Œuvres*, cit., t. V (1961), pp. 7-11, in particolare pp. 10-11.

51 Cfr. L.-A. Saint-Just, *Œuvres Complètes*, cit., p. 5; J.-P. Marat, *Œuvres politiques*, cit., t. I, pp. 208-210. A tal proposito sottolineo che nelle liste di frequenza di Saint-Just e di Marat non indico le varianti grafiche che non compaiono. In fase di ricerca, sono state, comunque, inserite nella lista delle parole da cercare (*pick list*). Esempio: «puissan(/t)s» compare in questo modo nella lista di frequenza delle parole nelle *Œuvres* di Robespierre e così «puissants» nella lista di frequenza delle parole delle *Œuvres politiques* di Marat e delle *Œuvres complètes* di Saint-Just.

52 Cfr. C. Vetter, *Il dispotismo della libertà. Dittatura e rivoluzione dall'Illuminismo al 1848*, Milano, 1993, p. 206, nota 14.

53 Cfr., per esempio, Étienne Brunet, http://www.cavi.univ-paris3.fr/lexicometrica/article/numero1/hypertexte.htm: «Les dictionnaires électroniques du français classique ou préclassique posent des problèmes spécifiques, dont beaucoup sont liés au traitement complexe des graphies anciennes. La tâche est plus facile quand il s'agit d'une orthographe normalisée et d'un état de langue plus proche de nous».

54 Mediante l'opzione «trova e sostituisci», presente in un qualsiasi programma di video scrittura.

55 La parola «être» – che può essere sia verbo che sostantivo maschile – si riscontra – nelle *Œuvres* di Robespierre – nelle seguenti forme grafiche (tra parentesi il numero delle occorrenze): «être» (4287), «etre» (82), «étre» (54), «ètre» (46); la parola «suprême» nelle seguenti forme: «suprême» (159); «supreme» (1). Si tratta di fare, quindi, otto ricerche per una sola espressione. Per svolgere questo lavoro il *DBT* risulta d'aiuto. Nella ricerca della cooccorrenza «lois» («loix») – «foibles» («faibles»), però, neppure il *DBT* può venirci incontro e dovremo necessariamente cercare quattro cooccorrenze per un'unica espressione.

56 Cfr. in merito a queste problematiche Maurizio Lana: http://www.cisi.unito.it/arachne/num2/lana2.html.

57 Lessie semplici e lessie composte.

58 Per questo motivo ho scelto di affiancare alle liste di lessie scelte di ogni autore, l'indice delle cento parole (vuote e piene) più frequenti e l'indice delle cento parole con contenuto semantico significativo più frequenti. In questi indici sono stati omessi i nomi «Robespierre», «Marat» e «Saint-Just».

59 Cfr. E. Soletti, *Stilistica*, in *Dizionario di linguistica*, diretto da G. L. Beccarla, Torino, 1994: «Per *hapax* si intenda ogni forma che ricorra una sola volta nel testo o *corpus* in esame».

60 Cfr. I. Chiari, *Informatica e lingue naturali*, cit., p. 39: «Molte parole grammaticali come le preposizioni, gli articoli, le congiunzioni sono spesso dette parole vuote, dato che hanno un contenuto semantico difficilmente definibile, e hanno soprattutto la funzione di mettere in relazione tra loro altre parole. Si chiamano parole piene in genere i sostantivi, i verbi, gli aggettivi che veicolano un contenuto semantico relativamente più autonomo. In una lista di frequenza troveremo quasi sempre ai primi posti le parole vuote, rispetto a quelle piene».

61 Ho accorpato singolare e plurale della stessa lessia, solamente nel caso in cui nessuna delle due parole presenti occorrenze nel testo in esame.

62 Cfr. I. Chiari, *Informatica e lingue naturali*, cit., p. 36.

63 Cfr. più avanti.

64 Non compaiono, nelle *Œuvres*, cooccorrenze «terreur» – «félicité».

65 Contiene una parte della corrispondenza di Maximilien e Augustin Robespierre, inviata e ricevuta. Per i lavori di compilazione degli indici delle concordanze è stata presa in considerazione solo la corrispondenza inviata da Robespierre. La datazione delle lettere rispetta quanto indicato nelle *Œuvres*. L'ordinale, nell'indicazione del riferimento, si riferisce alla numerazione delle *Œuvres*. Il tomo III presenta, inoltre, un'appendice. La numerazione delle pagine nell'appendice ricomincia da 1. La numerazione della corrispondenza inizia nuovamente da I.

66 Sono stati inseriti nella lista delle concordanze tutti i passi contenuti nel testo dei due giornali, anche dove si tratta di discorsi, lettere o scritti non composti da Robespierre ma solamente riportati da lui ne *Le défenseur* o nelle *Lettres*.

67 Per i lavori di compilazione degli indici delle concordanze non viene specificato, nel riferimento, se i discorsi sono stati pronunciati all'Assemblea Nazionale Costituente. Viene data indicazione dei discorsi pronunciati alla Società degli Amici della Costituzione (Club dei giacobini).

68 Vale lo stesso discorso fatto per il tomo VI: cfr. nota 67.

69 Per i lavori di compilazione degli indici delle concordanze non viene specificato, nel riferimento, se i discorsi sono stati pronunciati alla Società degli Amici della Costituzione (Club dei giacobini).

70 Per i lavori di compilazione degli indici delle concordanze non viene specificato, nel riferimento, se i discorsi sono stati pronunciati alla Convenzione.

71 Vale lo stesso discorso fatto per il tomo IX: cfr. nota 70.

72 Cfr. più avanti § 3.5 – Scelte metodologiche inerenti alle cooccorrenze.

73 In pochi casi, se il passo non presenta segni d'interpunzione forti, ho utilizzato le virgole come punto d'inizio o fine del contesto. In qualche occasione – per rendere più chiara la lettura – ho presentato un contesto formato da due brevi periodi contigui.

74 Avvertenza: I *Discours sur les peines infamantes* presentano numerose varianti fra la prima edizione del 1784 e la seconda del 1785. Rimando alle *Œuvres* per ulteriori chiarimenti. Nel caso non sia presente nessuna indicazione s'intende l'edizione del 1784.

75 Cfr. sopra le indicazioni alle note 67-71.

76 Cfr. sopra, nota 12.

77 Cfr. l'esempio «bon citoyen»/«citoyen bon» al § 3.3 – Guida alla lettura e scelte metodologiche inerenti alle liste di frequenza.

78 *Lettre de Robespierre à la Société des Amis de la Constitution de Versailles* (1er juin 1791), in *Œuvres*, cit., t. III, p. 107.

79 Per la definizione di «riferimento» cfr. sopra nota 7.

80 Cfr. sopra nota 23.