Molto rumore per nulla ? Gli effetti della lemmatizzazione sull'analisi di un corpus di interviste con Alceste

Carlo Tomasetto[1], Alberto Cattaneo[2], Patrizia Selleri[1]

[1]Università di Bologna – Dip. di Scienze dell'Educazione – Facoltà di Psicologia – I-47023 Cesena (FC) – Italia

[2] Istituto Svizzero di Pedagogia per la Formazione Professionale – Dip. Ricerca e Sviluppo – CH-6900 Lugano – Svizzera

Abstract

Before proceeding to statistical analysis of textual data, the whole corpus usually undergoes several preliminary processing, in order to improve the quality of data and in turn to improve the validity of the analysis. Treatments such as normalization, lemmatization and/or lexicalization allow to reduce ambiguities and to remove not significant semantic variations ; at the same time, they involve the risk of modifying the integrity of the text, with unpredictable outcomes. The aim of this paper is to illustrate *if* and *to what extent* preliminary processing on a *corpus*, drawn from semi-structured interviews, affect the results of the analysis produced by ALCESTE. Such processing have been carried out in our *corpus* in three different phases : the first level consist on the automatic lemmatization operated by the software (based on a small dictionary and on an algorithm for deriving headwords), plus the text normalization. At a second level we corrected the undue inclusions operated by ALCESTE and added a lexicalization limited to the most relevant words in the semantic field of our research ; at a third level, all the words recurring more than a minimum threshold were lemmatized. Then, the three different Hierarchical Descending Classifications provided by ALCESTE – corresponding with the three mentioned versions of the same *corpus* – have been compared. Results show that the efforts to increase the linguistic quality and accuracy of the data do not produce relevant effects on the analysis, although they contribute to increase the stability of the classes and the content care of their specific vocabularies. We note that only the most systematic processing induces appreciable advantages as regards as to the automatic lemmatization provided by ALCESTE.

Riassunto

Nelle fasi che precedono un'analisi statistica di dati testuali, il *corpus* da analizzare viene solitamente sottoposto a diversi trattamenti preliminari, nel tentativo di migliorare la qualità del dato analizzato e di conseguenza la validità dei risultati delle elaborazioni statistiche successive. Operazioni di normalizzazione, lemmatizzazione e/o lessicalizzazione permettono di ridurre le ambiguità ed eliminare variazioni non significative sul piano semantico (il "rumore di fondo"), ma allo stesso tempo rischiano di modificare l'integrità del testo raccolto, con incognite ed insidie non trascurabili. Scopo di questo contributo è indagare se e in che misura i trattamenti preliminari su di un *corpus* costituito da interviste semi-strutturate incidono sui risultati ottenuti quando il *corpus* stesso viene analizzato con il *software* ALCESTE. Tali trattamenti sono stati introdotti per tre livelli progressivi : il primo livello è quello realizzato automaticamente dal software (lemmatizzazione basata su un dizionario ristretto e un algoritmo di derivazione dei lemmi), accompagnato dalla normalizzazione del testo. Al secondo livello sono state corrette le lemmatizzazioni indebite prodotte da ALCESTE e sono state lessicalizzate le forme più pertinenti rispetto al campo semantico della ricerca ; al terzo livello sono state sistematicamente lemmatizzate tutte le forme al di sopra di una soglia predefinita di occorrenze. Sono stati quindi confrontati i risultati della Classificazione Discendente Gerarchica ottenuta da ALCESTE a partire dalle tre versioni dello stesso *corpus*. I risultati, che certo non possono dirsi conclusivi, evidenziano che gli interventi non producono effetti rilevanti sui risultati dell'analisi, per quanto facciano comunque registrare alcuni miglioramenti oggettivi nella stabilità delle classi e nel contenuto dei vocabolari specifici delle stesse. In particolare, soltanto l'intervento di lemmatizzazione più sistematico induce variazioni apprezzabili rispetto alla lemmatizzazione automatica effettuata da ALCESTE.

Mots-clés : lemmatizzazione, normalizzazione, lessicalizzazione, Alceste.

1. Introduzione

Sottoporre ad un'analisi statistica il dato testuale non è un'operazione delle più evidenti. La lingua non costituisce in effetti un universo statistico propriamente detto, sfuggendo a qualsiasi definizione operativa accettabile ed essendo composta da elementi – le parole – di per sé mai univoche e fortemente dipendenti dal contesto. Nel tentativo di migliorare la qualità linguistica dei *corpora* da studiare, e di conseguenza la validità dei risultati di qualsiasi successiva analisi (Bolasco, 1997 ; 2004), negli ultimi anni sono cresciuti i tentativi di integrare conoscenze e strumenti propri *e* della statistica, *e* della linguistica già nelle fasi che precedono l'analisi statistica vera e propria, ad esempio con l'impiego di dizionari e grammatiche ai fini della lemmatizzazione o della lessicalizzazione del testo grezzo.

Intervenire sul testo che verrà sottoposto ad analisi statistica è però tutt'altro che semplice e soprattutto tutt'altro che privo di insidie. Spesso il ricercatore è chiamato a prendere decisioni a priori, senza sapere esattamente quali effetti esse avranno ai fini dell'analisi ; senza sapere, cioè, se gli interventi miglioreranno effettivamente la base dati sottoposta ad analisi, oppure produrranno più confusione che vantaggi (Brunet, 2000). Se da un lato si dice che un intervento preliminare sul testo, per avere senso, deve essere completo e sistematico, "*à zéro fautes*" (Labbé, 2003 : 7), dall'altro è evidente che tale sistematicità ha dei costi rilevanti in termini di tempo e risorse ; per questo altri autori suggeriscono interventi parsimoniosi, mirati e commisurati allo scopo della ricerca (Bolasco, 1999). L'interrogativo sugli effetti *reali* che i trattamenti preliminari del testo avranno sulle analisi che intendiamo compiere, alla luce di quanto sopra, appare più che lecito.

Il presente contributo si propone dunque di indagare se e in che misura i trattamenti preliminari a cui viene sottoposto un *corpus* esteso, in particolare per quanto concerne la sua lemmatizzazione, incidono sui risultati dell'analisi statistica del testo, ottenuta attraverso il *software* ALCESTE. Dopo aver rilevato alcuni punti salienti dell'attuale dibattito sulla lemmatizzazione nell'ambito della statistica testuale, presenteremo un raffronto tra i risultati ottenuti con ALCESTE su un medesimo *corpus* tratto da interviste semi-strutturate e sottoposto a tre diversi e successivi livelli di lemmatizzazione, secondo criteri linguistici definiti a priori dagli autori. Lo studio ha finalità puramente descrittive, in quanto il *corpus* analizzato non ha valore rappresentativo rispetto a tutti i possibili casi in cui ALCESTE può trovare impiego nell'analisi dei testi su base statistica. Riteniamo tuttavia che i risultati emersi offrano spunti di interesse per chi, come noi, utilizza in ambito socio-psicologico strumenti di analisi lessicale e del contenuto su base statistica.

2. Qualche riflessione sulla lemmatizzazione

L'obiettivo generale degli interventi preliminari su un *corpus* testuale è quello di ridurre per quanto possibile le fonti di ambiguità inevitabilmente presenti nel dato linguistico, conservando «distinte […] le variazioni significative in termini semantici e fonde[ndo] le forme che costituiscono degli invarianti semantici» (Bolasco, 1999 : 213). Due sono quindi le direzioni di intervento possibili : la distinzione forzata degli omografi, e l'accorpamento di forme con uguale significato ma significante non coincidente (voci flesse riconducibili a una stessa radice ed *eventualmente* sinonimi appartenenti a radici diverse).

A dire il vero, esiste anche un primissimo livello di trattamento, definito di *normalizzazione* del testo, che di solito si limita a interventi di tipo ortografico e formale. In questa fase si

provvede all'eliminazione del cosiddetto "rumore di fondo" (Giuliano, 2004) : inesattezze nella trascrizione o nella raccolta del testo, errori ortografici o di pronuncia dei locutori, forme tronche o parole lasciate a metà (tipiche del parlato, come nel nostro caso, ma anche di scambi in *chat-line*, via sms, ecc.), e via dicendo (Labbé, 1990). Benché di questi interventi di base spesso non venga nemmeno reso conto nella presentazione dei risultati di una ricerca, il loro impatto potrebbe rivelarsi già decisivo per un'analisi di tipo socio-psicologico, ad esempio facendo perdere in partenza l'informazione sull'occorrenza sistematica di una dizione o grafia "sbagliata" della stessa parola in alcuni gruppi sociali e non in altri. La rilevanza dell'informazione persa varierebbe considerevolmente a seconda degli obiettivi che la ricerca si prefigura, ma è evidente che non sempre il ricercatore riesce a predire a priori ciò che sarà effettivamente rilevante in seguito.

Ad un livello di complessità maggiore troviamo invece gli interventi che spostano l'attenzione del ricercatore sugli aspetti morfologici (ma anche semantici) del testo. Si tratta di interventi di *lemmatizzazione*, ovvero «di riconoscimento della categoria grammaticale di una parola, che produce la riconduzione della forma grafica al lemma di appartenenza» (Bolasco, 1999 : 191). Questa operazione è assai meno intuitiva della precedente, dal momento che è spesso arduo distinguere le sfumature o variazioni semantiche significative che possono accompagnare lemmi di eguale grafia. Illuminante in proposito un esempio presente nel nostro *corpus*, nella frase : «…perché poi appunto volevo discutere ma non sono riuscito sui *termini*, non *termini* che cosa vuol dire questo, cosa vuol dire quest'altro,…» : troviamo qui utilizzate entrambe le accezioni del lemma "termine", e pare che sia del locutore stesso (prima ancora che del ricercatore) la necessità dialogica di disambiguarne il senso.

Benché già un intervento di lemmatizzazione esponga a incertezze non sempre risolvibili, l'intervento sul testo può – e in molti casi deve – spingersi oltre. In effetti, prendere il lemma come unità linguistica di riferimento comporta la perdita di buona parte della ricchezza semantica di un testo, dal momento che molte *lessie complesse* costituiscono unità minime dotate di significato del tutto distinte dai singoli lemmi che le compongono (Bolasco, 1999). La *lessicalizzazione di poliformi e poliremathiche* consente di identificare sequenze di vocaboli con significato autonomo, ad esempio frasi idiomatiche (come_dire, per_esempio, ecc.), ma anche locuzioni dotate di senso proprio (formazione_a_distanza, per restare nell'ambito del nostro *corpus*). Occorre però anche in questo caso riservare particolare attenzione al contesto in cui ricorrono le forme poliremathiche, dal momento che, ad esempio, l'uso della medesima sequenza nella frase "la *formazione a distanza* di tempo è ancora efficace" non dovrebbe dare luogo ad alcuna lessicalizzazione. È stato calcolato che, tra le 30 forme più frequenti di un *corpus*, circa il 30% sono generatrici di poliformi ; la loro frequenza reale nel lessico di quel *corpus* viene ad essere ridotta a ranghi più bassi dopo il processo di lessicalizzazione (id.)

Quanto sin qui descritto si presenta quindi molto complesso e ricco di insidie. Senza un riferimento al contesto in cui si presenta un lemma "ambiguo", è in molti casi impossibile decidere sul suo trattamento, e se è vero che frequentemente il contesto di cui parliamo è interno al testo – si basa cioè sull'esame delle concordanze, ovvero sul co-testo – altrettanto spesso non può prescindere da conoscenze extra-testuali. Saranno dunque – rispettivamente – le parole che accompagnano un determinato termine, o la conoscenza della situazione entro il quale esso viene usato, a fornirci le indicazioni necessarie per comprendere appieno di cosa si sta parlando, risolvendo le ambiguità semantiche. Già a questo livello puramente "tecnico", il cui obiettivo è apparentemente limitato alla necessità di migliorare la qualità del testo, ci accorgiamo quindi che è necessario uno spostamento di accento dalla *lingua* in quanto dispositivo e congegno lessicale, al *discorso* in quanto artefatto socio-culturale.

Oltre a ciò, il trattamento preliminare del testo è un passaggio estremamente dispendioso in termini di tempo e di risorse (Mellet, 2003 ; Brunet, 2000) ; se teniamo conto che un *software* come ALCESTE prende in considerazione per l'analisi anche parole con frequenze molto basse (di *default* la soglia è a 3 occorrenze), si comprende che la quantità di testo ambigua da sottoporre a trattamento può apparire ingovernabile. A questo proposito, diversi autori suggeriscono criteri di parsimonia ai quali attenersi nel definire le unità di testo sulle quali è vantaggioso intervenire (Morrone, 1993 ; Bolasco, 1999). Resta però da chiedersi : quali sono effettivamente le ricadute dei trattamenti preliminari sulla qualità dell'analisi finale ? In altre parole : il santo vale la candela ?

Stranamente, se pure da un punto di vista teorico appaiano convincenti gli argomenti addotti a favore degli interventi di lemmatizzazione e lessicalizzazione, molti autori concordano però sul fatto che le ricadute effettive ai fini dell'analisi statistica dei testi sono tutto sommato limitate (Lebart e Salem, 1994 ; Brunet, 2000), tanto che alcuni insinuano il dubbio, sia pure non condividendolo, che tutto ciò costituisca un lusso inutile, almeno in alcune circostanze (Mellet, 2003). Sembra cioè che il trattamento preliminare del testo sia sì un passaggio indispensabile per chi ha scopi di ricerca di base e mira a incrementare il più possibile la validità delle sue analisi, ma risulti un costo ingiustificato per chi si pone scopi di immediato utilizzo applicativo dei risultati ottenuti (Brunet, 2000).

Gli studi empirici che hanno cercato di verificare l'impatto della lemmatizzazione preliminare sui risultati di analisi statistiche di *corpora* testuali reali sono al momento poco numerosi. Brunet (2000) ha mostrato che le distanze calcolate tra 9 sub-testi raccolti in un grande *corpus* letterario rimangono praticamente invariate, sia analizzando le forme grezze, che quelle lemmatizzate. Kastberg Sjöblom (2002) ha invece confrontato tra loro i risultati dell'analisi di un *corpus* non trattato (forme grafiche), con quelli prodotti da due lemmatizzazioni condotte secondo criteri diversi. L'analisi sui *corpora* così trattati è stata effettuata attraverso due diversi *software*, entrambi largamente diffusi (Hyperbase e Lexicométrie) : i risultati delle analisi sono estremamente simili tra loro, quale che sia il metodo di lemmatizzazione impiegato (o non impiegato affatto). Il vantaggio relativo dei trattamenti preliminari appare soltanto quando si vogliano condurre specifiche analisi di tipo morfologico, grammaticale o sintattico, impossibili senza un etichettamento preliminare dei lemmi.

3.1. ALCESTE e il problema della lemmatizzazione

A nostra conoscenza, nessuno studio ha finora esplorato gli effetti della lemmatizzazione – o meglio, di diversi livelli di trattamento preliminare del *corpus* – sui risultati dell'analisi condotta con ALCESTE, un *software* di analisi dei dati testuali basato sull'assunto distribuzionale per cui il ricorrere costante di determinate parole o gruppi di parole all'interno dei medesimi contesti discorsivi *non* può essere considerato semplicemente un fatto casuale. Segmentando il testo in enunciati, è possibile secondo Reinert (2003) accedere al senso sottostante ad un testo basandosi su una logica di *ripetizione* delle parole, evidenziando cioè la probabilità che parole diverse ricorrano "sistematicamente insieme" in molti enunciati diversi. L'analisi principale di ALCESTE è la Classificazione Discendente Gerarchica (C.D.G.), il cui scopo è quello di raggruppare gli enunciati con un profilo lessicale simile – ovvero caratterizzati dalla co-occorrenza sistematica delle stesse parole – in classi il più possibile omogenee al loro interno, e il più possibile dissimili le une dalle altre (Reinert, 1993). Solo a questo punto ALCESTE procede a calcolare il vocabolario specifico di ciascuna classe di enunciati, ovvero l'insieme di quelle parole che ricorrono più frequentemente negli enunciati di una classe rispetto al resto del *corpus*. L'esame del vocabolario così estratto

permette al ricercatore, questa volta grazie ad un lavoro prettamente interpretativo, di avere accesso ai mondi lessicali sottostanti a ciascuna classe, ovvero ai diversi nuclei di significato che il discorso prodotto dai locutori implica e produce.

Il principio statistico distribuzionale di cui ALCESTE si serve per accedere ai mondi di significato sottostanti ad un testo – a partire dalla co-occorrenza di più parole all'interno degli stessi enunciati – rende evidente l'importanza della lemmatizzazione preliminare del testo. In effetti, la frammentazione di uno stesso lemma nelle sue diverse forme flesse renderebbe quanto mai arduo individuare le co-occorrenze di due lemmi nel testo. Un certo grado di lemmatizzazione del testo è realizzato di *default* da ALCESTE, che attraverso un semplice algoritmo provvede a ridurre alla radice lessematica molte forme flesse regolari di verbi, sostantivi, aggettivi, avverbi e preposizioni composte (Reinert, 1997) [1].

Si pone però anche il problema opposto, che è quello di considerare erroneamente ripetuto in più enunciati lo stesso lemma quando in realtà il locutore ha utilizzato flessioni o omografi con significato assolutamente diverso ; data la loro pertinenza sul piano semantico e pragmatico, tali fonti di ambiguità potrebbero da un lato compromettere per ALCESTE la possibilità di estrarre dalle classi di enunciati un vocabolario specifico non banale, e dall'altro lato potrebbero fuorviare l'interpretazione di tale vocabolario da parte del ricercatore. Vediamo alcuni esempi :

- *Fusione in un unico lemma di aggettivi, sostantivi, verbi e avverbi con radice comune*. Di per sé l'operazione, molto frequente in ALCESTE, è spesso utile e legittima. Molte volte capita però che, pur in assenza di una reale differenza sul piano semantico, l'uso delle diverse forme grammaticali abbia un senso molto forte a livello pragmatico, come nel caso degli avverbi utilizzati come interiezioni (non necessariamente l'avverbio "praticamente" è utilizzato in relazione alla "praticità" di uno oggetto)[2].
- *Fusione indebita di sostantivi della stessa radice, ma con campi semantici differenti*. Si pensi ad esempio alle coppie di sostantivi "motivo" e "motivazione", o a "metodo" e "metodologie".
- *Fusione di verbi in forma attiva e riflessiva.* Alcuni esempi presenti nel nostro *corpus* sono chiarificatori in proposito : «capire» una lezione è altra cosa dal «capirsi» in una relazione umana ; «programmare» in senso informatico non è la stessa cosa del «programmarsi» la giornata ; «trovare» un oggetto non evoca le medesime associazioni che «trovarsi» bene con qualcuno, come pure «sentire» un rumore rispetto a «sentirsi» a proprio agio ; infine, «riferire» attorno ad un tema particolare non è la stessa cosa che «riferirsi» a quello stesso tema[3].

Occorre insomma riuscire a disambiguare gli isoformi ma anche a evitare fusioni illecite di lemmi diversi. Anche questo aspetto è fondamentale, perché se è vero che ALCESTE si basa su

[1] Nel lessico di Alceste le parole sottoposte ad analisi sono indicate come "forme semplici" e comprendono sia le radici flemmatiche individuate dal software, sia le forme grafiche lasciate inalterate perché non riconosciute e/o ambigue. Per semplicità d'ora innanzi utilizzeremo "parola" anche nell'accezione di "forma semplice", mentre continueremo ad indicare con "forme grafiche" le flessioni eventualmente presenti nel corpus prima di qualsiasi trattamento.

[2] Inoltre, l'uso di un aggettivo – Tizio è aggressivo con Caio – piuttosto che di un verbo – Tizio aggredisce Caio – implica un livello di astrazione diverso e veicola significati specifici nel contesto dell'interazione sociale: nel caso specifico, l'aggettivo veicola una informazione disposizionale sulla personalità dell'attore, mentre l'uso del verbo di azione accentua il carattere episodico e situazionale del comportamento di Tizio (Semin e Fiedler, 1992).

[3] La forma riflessiva ci offre anche lo spunto per interrogarci sulle modalità di operare una lemmatizzazione che riesca a rendere conto, senza dunque perderli, anche dei pronomi personali oggetto e complemento che spesso divengono enclitici alla voce verbale.

una logica di *ripetizione* – ritrovare le stesse parole insieme in più enunciati – esso si basa anche su una logica di *rottura* (Reinert, 2003), poiché per accedere a universi di significato distinti occorre individuare le specificità di uso del lessico nei singoli enunciati, preservandone quindi il più possibile le peculiarità semantiche.

In conclusione, possiamo dire che uno dei motivi per cui ALCESTE viene utilizzato per scopi di analisi del contenuto è quello di ridurre al minimo l'intervento interpretativo del ricercatore nel processo di individuazione dei nuclei di significato (Klein e Licata, 2003). Il problema però potrebbe presentarsi a monte : senza un intervento consapevole, e spesso anche "interpretativo", del ricercatore, il dato testuale sottoposto ad analisi potrebbe rivelarsi poco affidabile, compromettendo così a priori la possibilità di accedere ai significati sottostanti[4].

3. Un caso di ricerca : la preparazione di un corpus tratto da interviste

Nell'esempio che illustreremo di seguito abbiamo voluto verificare se, intervenendo in maniera più o meno incisiva sul testo, oppure limitandosi ad utilizzare la lemmatizzazione eseguita di *default* dal *software*, i risultati dell'analisi peculiare di ALCESTE – la Classificazione Discendente Gerarchica – appaiono effettivamente diversi. Per fare questo non ci siamo serviti di *software* specifici per il trattamento preliminare del testo, che pure esistono e si prestano bene allo scopo – vedi ad esempio TALTAC (Bolasco, 2002) – ma ci siamo limitati ad un intervento manuale, basato sul controllo sistematico dei contesti locali dei vocaboli candidati al trattamento preliminare. Il nostro scopo era infatti quello di mantenere il controllo totale dell'intervento sul testo, subordinando i diversi livelli di trattamento a criteri definiti a priori, tendendo conto del fatto che i criteri di intervento devono comunque essere ponderati dal ricercatore anche quando si utilizzino *software* specifici.

3.1. Il corpus

Il *corpus* sul quale è stata effettuata l'analisi con ALCESTE è costituito da 33 interviste raccolte durante un'esperienza di *blended learning* svoltasi in Svizzera nel quadro del Progetto ICT[5], un'iniziativa coordinata dall'Istituto Svizzero di Pedagogia per la Formazione Professionale[6] (ISPFP) volta a sviluppare progetti di implementazione delle tecnologie nella didattica della formazione professionale. In particolare, si tratta di interviste semi-strutturate condotte con 9 apprendisti (intervistati 2 volte ciascuno), 2 docenti (intervistati 5 volte) e 2 assistenti di pratica (AP, una particolare figura di *tutoring*, anch'essi intervistati 5 volte) di una classe di apprendisti muratori del primo anno di corso, frequentanti la Scuola Professionale dell'Artigianato e dell'Industria (SPAI) di Mendrisio, nel Cantone Ticino[7]. Il *corpus* raccolto soddisfa i requisiti necessari per essere analizzato attraverso il *software* ALCESTE (Matteucci e Tomasetto, 2002 : 314) dal momento che consta di oltre duecentomila occorrenze ; anche il rapporto tra dimensione del *corpus* e ampiezza del vocabolario appare soddisfacente, situandosi molto al di sotto della soglia critica del 20%[8] (Bolasco, 1999 : 203).

[4] Occorre tenere presente che le inesattezze prodotte dalla lemmatizzazione automatica di Alceste non sono da addebitarsi a limiti del programma, ma alle dimensioni molto ridotte del dizionario italiano in dotazione.
[5] http://www.ict.sibp-ispfp.ch. Si veda anche Cattaneo et al. (2005).
[6] Vedi http://www.ispfp.ch.
[7] Una descrizione più dettagliata del corpus, del piano della ricerca e del contesto entro il quale era inserita è a disposizione in Cattaneo (2005), da cui sono tratti i dati che seguono.
[8] Ossia: V/N < 20%. Nel nostro caso, tale rapporto si assesta attorno al 5%.

3.1. L'intervento sul testo : tre tappe

La preparazione del testo è stata scomposta in tre tappe successive, in modo che fosse possibile anche verificare, valutare e quantificare gli effetti prodotti di fatto sull'analisi da trattamenti più o meno marcati del testo.

La **prima tappa** (che nelle analisi che seguono indicheremo con "I") ha coinciso con la *normalizzazione* del testo. A partire dalle trascrizioni iniziali, si è proceduto a poche operazioni, corrispondenti ad interventi sul testo il meno possibile "invasivi" : *eliminazione degli errori di battitura* (ortografia, lettere "intruse",…), *completamento di parole* (dove non ci fosse ambiguità), *traduzione di espressioni dialettali in lingua italiana* («l'è scià ammò» sostituito con «è ancora qui»), *snellimento di ripetizioni tipiche del linguaggio orale* (il passaggio «anche questo, questo, diciamo questo disagio tecnico» è diventato «anche questo disagio tecnico). La lemmatizzazione del testo è stata affidata in questa prima fase soltanto alle risorse interne di ALCESTE.

Quanto fatto non poteva però rivelarsi soddisfacente alla luce delle riflessioni sulla lemmatizzazione già affrontate. Ecco perché la **seconda tappa** (cui ci riferiremo con "II") ha visto un intervento di *lessicalizzazione* che ha operato in maggiore profondità, e ha potuto essere realizzata dopo l'esame del vocabolario lemmatizzato generato da ALCESTE nella prima analisi sul *corpus*. Tale intervento è riconducibile alle seguenti tipologie :

- *correzione delle lemmatizzazioni "indebite"* operate automaticamente da ALCESTE : casi come «Bellinzona» riconosciuta come forma flessa di «bell<» ; «spar+» interpretata come radice sia di «sparare» che di «sparire».
- *Disambiguazione* di vocaboli particolarmente rilevanti nel campo semantico della nostra ricerca («natel» per «cellulare»[9] ; le diverse forme legate alla radice «apprend-», che - data la rilevanza per i nostri interessi - sono state lemmatizzate separando sostantivo da verbo).
- *Lessicalizzazione :* questo tipo di intervento sul testo – per le ragioni riportate anche nella riflessione iniziale – non è stato operato con sistematicità assoluta, e di conseguenza sono poche le poliremathiche individuate e lessicalizzate. Tra queste possiamo citare «in_pratica», perché non fosse confuso con la locuzione «assistente/i_di_pratica» ; «d_accordo», così da distinguerlo dal sostantivo «accordo» ; «faccia_a_faccia», che altrimenti sarebbe stato riconosciuto da ALCESTE come congiuntivo del verbo «fare».

La scelta di non lessicalizzare in maniera troppo sistematica è dovuta ad una considerazione di ordine statistico : un'eccessiva disambiguazione – raggiunta con operazioni come la lessicalizzazione e l'individuazione delle poliremathiche – rischia di ridurre in maniera molto (troppo ?) significativa le frequenze delle parole, con effetti non necessariamente felici sull'analisi.

La **terza tappa** (identificata nelle tabelle con "III") ha avuto come fine quello di una *lemmatizzazione* più sistematica, e non più limitata all'ambito semantico più pertinente per i nostri obiettivi di ricerca. Quest' ultimo passaggio ha quindi permesso :

[9] Nella Svizzera italiana il telefono cellulare viene spesso identificato con il nome del primo modello commercializzato (Natel, appunto).

- la riconduzione alla loro radice corretta di lemmi individuati in automatico come afferenti a radici diverse, indipendentemente dal campo semantico di riferimento[10] ;
- la riconduzione delle flessioni dei verbi alle radici verbali corrette : molte lessie della medesima voce verbale venivano considerate dalla lemmatizzazione automatica come afferenti a radici diverse. Come è facile intuire, questo è accaduto quasi sistematicamente con i verbi irregolari.

Per un principio di parsimonia negli interventi abbiamo stabilito a priori di *non intervenire* quando la radice lemmatica correttamente individuata non avesse raggiunto le 5 occorrenze. Tale provvedimento ha potuto prendere forma sulla base della lemmatizzazione prodotta da ALCESTE nel corso delle analisi precedenti, ossia a partire dalle computazioni fornite dal *software* stesso. Laddove l'accorpamento delle diverse forme grafiche non avesse superato, a seguito dell'eventuale lemmatizzazione, il limite delle 5 occorrenze, non siamo dunque intervenuti sul testo[11].

Le operazioni attuate nel terzo passaggio prefigurano interventi più sistematici ma allo stesso tempo più interpretativi e, dunque, potenzialmente più arbitrari da parte del ricercatore. Trattandosi però di un'operazione che si sposta sul piano semantico, pragmatico e di contenuto dell'enunciato, siamo propensi a credere che senza affidarsi alla sensibilità del ricercatore sia difficile ricostruire l'universo pragma-linguistico al quale il testo è ancorato.

3.3 Effetti della lemmatizzazione sui risultati

A questo punto occorre verificare se un intervento come quello qui descritto produca in definitiva dei riscontri oggettivi nei termini di differenze in sede di analisi, o piuttosto confermi gli esempi della letteratura già citati in generale a proposito della lemmatizzazione.

3.1.1. Stabilità dei dati strutturali

Cominciamo con il comparare qualche dato sulle caratteristiche generali del *corpus* : a seguito degli accorpamenti nella lemmatizzazione, possiamo notare (Cfr. Tabella 1.) che il numero di forme distinte – come pure il numero di *hapax legómenon* – diminuisce progressivamente passando dalla I alla III versione. Essendo gli accorpamenti della terza versione molto più sostanziosi della seconda, la differenza è molto più consistente tra la II e la III versione che tra la I e la II. Per quanto riguarda invece il numero di parole analizzate, notiamo come nella II versione esse siano maggiori rispetto alla I : il vocabolario della versione II è infatti stato aumentato a seguito delle distinzioni interne alle lemmatizzazioni indebite.

	III	**II**	**I**
Numero di forme distinte	10417	11665	11691
Numero di hapax	4782	5403	5424
Numero di parole analizzate	3285	4136	4070

Tabella 1. Consistenza del corpus nelle tre versioni

[10] Questo non significa considerare la lessicalizzazione un *minus* dal punto di vista assiologico rispetto alla lemmatizzazione: al contrario!

[11] Occorre inoltre precisare che non abbiamo operato una disambiguazione rispetto alle categorie grammaticali: quando sostantivi e verbi erano già accorpati, non siamo cioè intervenuti nel separarle, se non qualora ci fosse un'evidente differenza semantica; l'intervento non c'è stato nemmeno nel caso in cui le categorie grammaticali fossero invece già riconosciute e mantenute separate, al fine di evitare anche in questo caso un'eccessiva riduzione delle forme.

	III	II	I
Numero di occorrenze	203737	203798	204000
Frequenza media per forma grafica	20	17	17
Frequenza media per parola	27.41	24.44	24.66
Numero di occorrenze analizzabili (freq.>3)	38.46%	38.24%	38.26%

Tabella 2. Dettagli sulle occorrenze nelle tre versioni del corpus.

Analogamente, ed in senso opposto, il numero di occorrenze totali e di occorrenze considerate per l'analisi è più alto nella versione originale che nelle altre due (Tabella 2.). Essendo diminuito il numero di parole, la frequenza media generale di ogni forma è più alta nella versione III, in cui molti accorpamenti sono stati fatti, mentre nelle altre due è stabile. Lo stesso accade considerando la media per parola. Anche il rapporto tra V e N – come accennato in precedenza – migliora nel passaggio dalla I alla III versione : dal 5.73% scende infatti al 5.11%[12].

3.1.1. Stabilità delle classi

Dalla tabella che segue è anche possibile vedere come sono distribuite le u.c.e.[13] che sono state classificate nei tre passaggi. Il dato più significativo è la percentuale di u.c.e. che classificate in modo stabile dal programma : il fatto che tale percentuale aumenti di oltre due punti percentuali passando dalla I alla III versione del *corpus* potrebbe infatti indicare che il lavoro di lemmatizzazione migliora in modo apprezzabile il *corpus* stesso ai fini della conformazione delle classi.

	III	II	I
Numero di u.c.e.	5947	5955	5959
Percentuale di u.c.e. stabili nelle due C.D.G.	85.79%	82.42%	83.65%
C.D.G.1 (11 parole per u.c.) : n° di u.c. [14]	4059	4008	4020
C.D.G.1 (13 parole per u.c.) : n° di u.c.	3743	3657	3667

Tabella 3. Distribuzione delle u.c.e. nelle tre versioni del corpus.

Indicando con la stessa lettera le classi corrispondenti per vocabolario, ecco dunque i dendrogrammi rappresentanti il risultato della Classificazione Discendente Gerarchica :

III versione	II versione	I versione

[12] La differenza intermedia è minima: la II versione registra infatti un valore di 5.72%.

[13] Gli enunciati, o meglio, le "unità di contesto elementari" (u.c.e.), vengono definiti operazionalmente dal software a partire da due ordini di criteri: uno relativo alla punteggiatura, l'altro al numero di parole (forme semplici).

[14] Alceste realizza una doppia C.D.H. con unità di contesto (u.c.) di lunghezza diversa, al fine di ottenere classi dal contenuto stabile, quale che sia la lunghezza dei segmenti di testo utilizzati in fase di calcolo. In questo caso la prima C.D.H. ha utilizzato u.c. di lunghezza media di 11 parole, la seconda di 13.

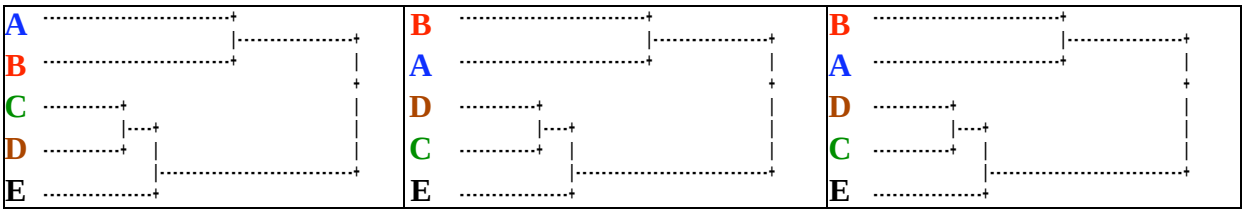

Tabella 4. Confronto tra i dendrogrammi corrispondenti alle C.D.G. sulle tre versioni del corpus.

Il dendrogramma si presenta pressoché uguale per conformazione in tutte e tre le versioni, pur non corrispondendo formalmente la successione delle classi. Varia come prevedibile il numero di u.c.e. di ciascuna classe, ma questo non incide sulla stabilità della classificazione. Nella Tabella 5. riportiamo invece alcuni dati che ci permettono di operare un confronto sulla distribuzione delle singole classi. Possiamo così verificare – dettagliando meglio rispetto al dendrogramma – come la distribuzione delle classi corrispondenti nelle tre versioni del *corpus* sia simile. La classe "A" e la "B" rappresentano infatti circa il 60% delle u.c.e. del testo, mentre il rimanente 40% è diviso tra la "C" e la "D", che insieme costituiscono circa il 25-30% delle u.c.e., e la "E" (10-15%). Ciò che possiamo ancora constatare a partire dalla Tabella 5. è l'influenza dell'intervento di lemmatizzazione sul *corpus* : a fronte dell'accorpamento dei termini evidente nel III passaggio, vediamo come in ciascuna classe, indipendentemente dalla variazione percentuale delle u.c.e. analizzate rispetto agli altri due passaggi, si verifica sempre un aumento del numero di parole analizzate in media in ciascuna u.c.e. Anche le parole stellate associate alle singole classi (dato non riportato in tabella) restano invariate nei tre passaggi.

	Numero di u.c.e.			Numero di parole analizzate per u.c.e.		
	III	II	I	III	II	I
Classe A	1224 *(23.99%)*	1257 *(25.61%)*	1240 *(24.87%)*	12.93	12.61	12.60
Classe B	1948 *(38.18%)*	1748 *(35.62%)*	1771 *(35.53%)*	8.04	7.86	8.00
Classe C	765 *(14.99%)*	632 *(12.88%)*	568 *(11.39%)*	12.44	11.85	12.00
Classe D	633 *(12.41%)*	560 *(11.41%)*	579 *(11.61%)*	12.94	12.58	12.59
Classe E	532 *(10.43%)*	711 *(14.49%)*	827 *(16.59%)*	13.81	13.29	13.19

Tabella 5 Confronto sulla distribuzione delle singole classi nelle tre versioni del corpus.

Per finire, abbiamo verificato i cambiamenti apportati dai trattamenti sul vocabolario specifico delle classi, calcolando un indice di somiglianza tra le parole piene contenute in ciascuna classe ai diversi passaggi (Tabella 6). La misura utilizzata è l'Indice di Jaccard, una misura di somiglianza tra due popolazioni che esprime la proporzione di individui in comune – in questo caso, le forme semplici presenti nella stessa classe in due passaggi successivi – rispetto al totale degli individui delle due popolazioni[15].

	Indice di somiglianza tra I e III	**Indice di somiglianza tra II e III**	**Indice di somiglianza tra I e II**
Classe "A"	.478	.439	.807

[15] La formula dell'indice di Jaccard è: s = c/(p+q-c), dove c rappresenta il numero di individui presenti simultaneamente nelle due popolazioni e p e q gli effettivi di ciascuna di esse. Se si preferisce è possibile trasformare la misura in una distanza d = 1-s. Se due popolazioni sono identiche, s è uguale a 1; se invece non vi sono individui in comune, s è uguale a 0 (Thioulouse, Chessel e Doufour, 2003).

Classe "B"	.562	.587	.867
Classe "C"	.552	.590	.593
Classe "D"	.554	.471	.675
Classe "E"	.482	.451	.667

Tabella 6. Misura della stabilità del vocabolario di ciascuna classe (indice s di Jaccard) .

I valori riportati mostrano che il vocabolario specifico delle classi rimane sostanzialmente stabile. In particolare i valori dell'indice di Jaccard indicano una somiglianza molto elevata nel vocabolario delle classi ottenute al passaggio I e al passaggio II, in particolare per le classi "A" e "B" - che da sole raggruppano più del 60% degli enunciati del corpus. Più bassa è invece la somiglianza tra le classi ottenute al passaggio II e al passaggio III, il che evidenzia che soltanto a questo livello si è avuta una reale perdita e/o acquisizione di forme semplici non presenti nel passaggio precedente. I valori dell'indice restano comunque moderatamente elevati sia tra il passaggio II e III, sia considerando la somiglianza complessiva tra i vocabolari ottenuti al I e al III passaggio, confermando quindi la sostanziale stabilità di buona parte del vocabolario specifico delle classi.

4. Conclusioni

I risultati ottenuti dal nostro raffronto non possono certo dirsi conclusivi, dal momento che il *corpus* analizzato non ha valore rappresentativo rispetto a tutti i possibili casi in cui ALCESTE può trovare impiego nell'analisi dei testi su base statistica. Ciò che abbiamo evidenziato è però che, intervenendo a correggere gli errori prodotti nella lemmatizzazione automatica dal *software*, e disambiguando con attenzione i termini pertinenti al nostro oggetto di indagine, non abbiamo ottenuto effetti apprezzabili sui risultati dell'analisi. Soltanto un intervento successivo molto più incisivo e sistematico – con l'esame di tutti lemmi con più di 5 occorrenze nel testo e la loro eventuale fusione/disambiguazione – ha apportato miglioramenti non banali a livello di stabilità delle classi e copertura del testo, oltre a cambiamenti apprezzabili nel contenuto del vocabolario specifico delle singole classi. In ogni caso, la struttura di queste ultime è restata sostanzialmente invariata attraverso i tre successivi passaggi.

Tutto ciò ci porta a concludere che, almeno in questo caso specifico, un intervento lungo e costoso di normalizzazione, lemmatizzazione e lessicalizzazione mirate non ha intaccato nella sostanza i risultati di un'analisi che si è rivelata già stabile e attendibile nella prima fase, ovvero in assenza di interventi peculiari dei ricercatori. In altre parole, ciò che ALCESTE ha prodotto partendo da dati testuali relativamente grezzi, e utilizzando soltanto le proprie e pur limitate risorse linguistiche interne, non è stato molto diverso da ciò che il software ha ottenuto partendo da dati di qualità linguistica decisamente più elevata. Ciò conferma quanto rilevato in precedenza da Brunet (2000), secondo il quale un'analisi a scopi applicativi immediati, orientata magari soltanto all'individuazione di tipologie testuali, non necessariamente si giova di lunghi e costosi interventi di trattamento preliminare del testo. Questo non toglie che una ricerca con obiettivi teorici e richieste di validità interna diverse, debba presentare risultati privi di ambiguità linguistiche grossolane quali quelle prodotte da lemmatizzazioni inadeguate.

Références

Bolasco, S. (1997). Meta-data and Strategies of Textual Data Analysis : Problems and Instruments. *Lexicometrica, 1*. http://www.cavi.univ-paris3.fr/lexicometrica/article/numero1/bolasco.htm

Bolasco, S. (1999). *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*. Roma, Carocci.

Bolasco, S. (2002). Integrazione statistico-linguistica nell'analisi del contenuto. In Mazzara, B. M. (Ed.), *Metodi qualitativi in psicologia sociale. Prospettive teoriche e strumenti operativi*. Milano, Carocci : 329-342.

Bolasco, S. (2004). L'analisi statistica dei dati testuali : intrecci problematici e prospettive. In Aureli Cutillo, E. e Bolasco, S. (Eds.), *Applicazioni di analisi statistica dei dati testuali*. Roma, Casa Editrice Università La Sapienza : 9-26.

Brunet, E. (2000). "Qui lemmatise dilemme attise". *Lexicometrica, 2* : 1-19. http://www.cavi.univ-paris3.fr/lexicometrica/article/numero2/brunet2000.PDF

Cattaneo, A. (2005). *Contesti senza spazio e spazi senza contesto ? Rilettura di un'esperienza di blended learning*. Unpublished Ph.D. Thesis, Alma Mater Studiorum, Bologna.

Cattaneo, A., Comi, G., Merlini, F., Sanz, M., e Arn, C. (2005). *ICT.SIBP-ISPFP. Un progetto d'innovazione. Un projet d'innovation. (Quaderno ISPFP n. 29)*. Zollikofen, Istituto Svizzero di Pedagogia per la Formazione Professionale.

Giuliano, L. (2004). L'analisi automatica dei testi ad alta componente di "rumore". In Aureli Cutillo, E. e Bolasco, S. (Eds.), *Applicazioni di analisi statistica dei dati testuali*. Roma, Casa Editrice Università La Sapienza : 41-54.

Kastberg Sjöblom, M. (2002). *Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus*. Proceedings of the JADT 2002. Saint-Malo, Irisa, Inria : 391-402. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/kastberg.pdf

Klein, O. e Licata, L. (2003). When group representations serve social change : The speeches of Patrice Lumumba during the Congolese decolonization. *British Journal of Social Psychology, 42* : 571-593.

Labbé, D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble, CERAT.

Labbé, D. (2003). "Analyse des données textuelles et statistique lexicale". *Lexicometria, Numéro thématique "Autour de la lemmatisation"*. http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema1/spec1-texte1.pdf

Lebart, L. e Salem, A. (1994). *Statistique textuelle*. Paris, Dunod.

Matteucci, M. C. e Tomasetto, C. (2002). Alceste : un software per l'analisi dei dati testuali. In Mazzara, B. M. (Ed.), *Metodi qualitativi in psicologia sociale. Prospettive teoriche e strumenti operativi*. Milano : Carocci : 305-327.

Mellet, S. (2003). "Lemmatisation et encodage grammatical : un luxe inutile ?" *Lexicometria, Numéro thématique "Autour de la lemmatisation"*. http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema1/spec1-texte2.pdf

Morrone, A. (1993). Alcuni criteri di valutazione della significatività dei segmenti ripetuti. In Anastex, S. J. (Ed.), *JADT 1993, Actes des Secondes Journées Internationales d'Analyse Statistique de Données Textuelles*. Paris, ENST-Telecom : 445-453.

Reinert, M. (1993). Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et Société, 66* : 5-39.

Reinert, M. (1997). Les "Mondes lexicaux" des six numéros de la revue "Le Surréalisme au Service de la Révolution". *Mélusine, Editions L'Age d'Homme, Lausanne, XVI* : 270-302.

Reinert, M. (2003). Le rôle de la répétition dans la représentation du sens et son approche statistique par la méthode «ALCESTE». *Semiotica, 147* (1/4) : 389-420.

Semin, G. R. e Fiedler, K. (1991). The linguistic category model, its bases, application and range. *European Review of Social Psychology, 2* : 1-30.

Thioulouse, J., Chessel, D. e Dufour, A. B. (2003). Classification automatique. In Chessel, D., Dufour, A. B., e Thioulouse, J. (Eds.), *Biométrie et Biologie Evolutive*. Lyon, Université Lyon1. http://pbil.univ-lyon1.fr/R/archives/stage7.pdf