Emozioni e testo: costruzione di risorse per il *tagging* automatico

Nadia Battisti[1], Francesca Dolcetti[2],

[1] Sapienza - Università di Roma – nadia.battisti@uniroma1.it

[2] Studio RisorseObiettiviStrumenti – francescadolcetti@studio-ros.it

Abstract

This contribution gives an account of the first results of a broader project aimed at working up automated resources for 'tagging' emotionally dense words, simultaneously trying to put their dependability to the test. The need began within the Emotional Analysis of the Text (AET), a method originating from psychoanalytic and psycho-sociologic work, which, starting from the '90s, has focussed on the affective-symbolic plane of words. To work with AET, words are classified on the basis of two references: 1) a linguistic one, from an etymological matrix, with the reduction of types into lexemes; 2) the other one, psychological, through the identification of words, with or without emotional density, only the former are being analysed. The construct of a dense word is close to the concept of the full word used in semiometry, but also close to some models emerging from studies on the affective meaning of the words of C. E. Osgood.

In order to create basic lists of dense and non-dense words, the research project moved in two intertwined directions: a) analyse the dictionaries with active and non-active words in the analysis, used by two groups of researchers, and construct a database of dictionaries of 79 AETs, created between 1998 and 2009; b) explore various aspects and the variability of the choices of the researchers themselves, using different statistical methods, and provide some initial results of these analyses. The study discusses the hypothesis that affective invariants exist in words and incline towards the idea that the relationship between word and emotion can be represented along a continuum, in which the context takes on a discriminating function. In other words, the construct of the density would be presented with fuzzy borders, more than being dichotomously polarisable. The results of the research account for some criteria that organise these borders.

Keywords: Emotional Analysis of the Text (AET), dense word, *sentiment*, *tagging*, reliability, TaLTaC, SPAD.

Riassunto

Il presente contributo vuole dar conto dei primi risultati di un progetto volto ad elaborare risorse automatizzate per il *tagging* di parole emozionalmente dense, cercando al contempo di metterne alla prova l'affidabilità. Tale esigenza è nata nell'ambito dell'Analisi Emozionale del Testo (AET), una metodologia nata dal lavoro psicoanalitico e psico-sociologico, che fin dagli anni '90 ha posto il *focus* sul piano affettivo-simbolico delle parole. Per realizzare un'AET, le parole sono classificate sulla base di due riferimenti: 1) uno linguistico di matrice etimologica, con riduzione a lessema delle forme grafiche; 2) l'altro psicologico, con identificazione delle parole aventi o meno densità emozionale, solo le prime sono poste in analisi. Il costrutto di parola densa è prossimo al concetto di parola piena usato dalla semiometria, ma anche vicino ad alcuni modelli emersi dagli studi sul significato affettivo delle parole di C. E. Osgood.

Per realizzare liste di base di parole dense e non-dense, il progetto di ricerca si è mosso in due direzioni tra loro intrecciate: a) analizzare i dizionari con le parole attive e non nell'analisi, usati da due gruppi di ricercatori, costruendo un *database* dei dizionari di 79 AET realizzate dal 1998 al 2009; b) esplorare diversi aspetti e con diversi metodi statistici la variabilità delle scelte dei ricercatori stessi e dare alcuni primi risultati di queste analisi. Lo studio discute sull'ipotesi che esistano delle invarianti affettive nelle parole e propende per l'idea che il rapporto fra parola ed emozione sia rappresentabile lungo un *continuum* in cui il contesto assume una funzione discriminante. In altre parole il costrutto della densità si presenterebbe con dei confini sfumati (*fuzzy*) più che

polarizzabile in modo dicotomico. Il risultati della ricerca danno conto di alcuni criteri che organizzano tali confini.

Parole chiave: Analisi Emozionale del Testo (AET), parola densa, *sentiment*, *tagging*, affidabilità, TaLTaC, SPAD.

1. Introduzione all'Analisi Emozionale del Testo

La classificazione delle forme grafiche di un testo, in nuove categorie da usare come variabili attive, è un'area di lavoro storica nell'analisi del testo. Il presente contributo vuole dar conto dei primi risultati di un progetto volto ad elaborare risorse automatizzate per il *tagging* di parole emozionalmente dense, cercando al contempo di metterne alla prova l'affidabilità e rientra nel più generale interesse per il *sentiment* di un testo[1]. Si tratta di un costrutto messo a punto dall'Analisi Emozionale del Testo (AET) che cerca di rintracciare emozioni entro le produzioni discorsive e testuali per conoscere e intervenire nelle relazioni sociali (Carli e Paniccia, 1999, 2000, 2002; Carli, Dolcetti e Battisti, 2004). L'AET si situa all'incrocio fra alcuni modelli della psicoanalisi, con la sociologia e l'antropologia, orientati dalla linguistica e la semiologia.

Proviamo a delineare il costrutto di emozione utilizzato in AET. Innanzitutto viene superata la distinzione fra emozioni "primarie" la cui espressione è innata (Ekman 1973; Izard, 1977) e "secondarie" apprese e influenzate dalle culture (Damasio, 1994; Izard, 1977)[2], non solo per una mancanza di unanimità tra i diversi Autori su quali siano queste emozioni, ma poiché entrambi i gruppi concorrono al più ampio processo di significazione affettiva che investe il linguaggio. Le parole sono dunque tracce del processo affettivo e simbolico condiviso, espressione soggettiva e al tempo stesso sociale, in accordo con Reinert (1993) "Lieux d'énonciation entre les représentations personnelles et les pre-construits sociaux donc partagés" (p. 12).

In questo quadro teorico e operativo l'emozionalità espressa attraverso le parole costituisce una proposta collusiva, articolata per e con altri, e viene valorizzata l'idea del linguaggio come atto (Austin, 1975). Di nuovo è un fenomeno individuale ma a valenza sociale, il modo emozionale infinito, si riversa e si esprime tanto attraverso ognuno di quegli angusti canali che sono dati dalle singole parole (Matte Blanco 1975), quanto attraverso la trama che più parole tra loro co-occorrenti compongono. Atto rilevato dalla collusione, quel costrutto creato da Carli che integra modelli delle rappresentazioni sociali con l'uso di meccanismi psichici individuali ben noti in psicologia, come ad esempio l'identificazione proiettiva (Carli, 1987, 1990, 1995, 2006).

L'AET assume quella tradizione che stabilisce delle differenze fra parole vuote, strumentali, ambigue e parole che portano con sé un senso pieno, marcando, nel loro occorrere, il significato del discorso anche affettivo (Reinert, 1993; Lebart, Piron e Steiner 2003).

[1] Si tratta di studi con una lunga tradizione, nei quali sono stati adottati diversi approcci e metodologie di ricerca, a partire da quelli pioneristici di C. H. Osgood ed altri avviati sin dagli anni '50 (Osgood, May e Miron, 1975).

[2] Infatti, si passa dall'identificarne alcune al rilevarne centinaia. Quelle primarie classicamente sono: gioia, tristezza, rabbia, sorpresa, interesse e disgusto; a queste ne vengono aggiunte, di volta in volta, molte altre quali: vergogna, colpa, imbarazzo, disprezzo, timidezza e via dicendo.

Con l'AET si mira ad attuare un processo di decostruzione dei nessi linguistici tipici del modo dividente ed asimmetrizzante della mente (funzione operativa del linguaggio), per addivenire, alla ricostruzione delle più ricorrenti catene associative tra parole dense con il supporto statistico del calcolo delle co-occorrenze. E' un processo isomorfo, come si può comprendere, a quello delle libere associazioni sviluppato nella psicoanalisi, che Sigmund Freud aveva elaborato per i suoi pazienti e che da tempo è utilizzato anche nell'intervento psico-sociologico.

Per l'AET, attualmente, vengono utilizzati il *software* Alceste, sviluppato com'è noto da Reinert (1993, 1996) e T-Lab di Lancia (2004). Sul piano operazionale, con l'AET non sono prese in considerazione le singole parole intese come *types*. Il criterio per lo *stemming*[3] è anch'esso di tipo emozionale e corrisponde in gran parte a un'operazione di lessematizzazione, dove il riferimento è alla comune radice etimologica di un gruppo di *types*. Etimologia che riesce a dar conto di quel precipitato storico-culturale che la parola è e di quel suo essere un modo di simbolizzare la relazione, che immesso nella relazione stessa ha lo "scopo" di tesserla entro qualche direzione. Questa direzione in seguito verrà raccolta, confermata o rimessa in gioco, dalla parola successiva, e così via dicendo, sia entro il testo di un medesimo soggetto, sia nel dialogo tra più soggetti. La combinazione quasi infinita delle parole tra loro e la possibilità di inventare nuove parole ci consente un lavoro interminabile nel costruire nuovi modi per affrontare la convivenza, di cui i testi possono essere testimonianza, dove possiamo trovare somiglianze con modi precedenti ma mai identità. Tornando, quindi, su questioni operazionali, oltre il riferimento al lessema, per operare lo *stemming* delle parole nell'AET resta il forte ancoraggio alla valutazione psicologica su cui è impegnato il ricercatore.

2. La ricerca

2.1. Obiettivi, metodologia e costruzione dei dati

Come si è già anticipato, il presente contributo vuole dar conto dei primi risultati di un progetto volto a elaborare risorse automatizzate per il *tagging* di parole emozionalmente dense, cercando al contempo di metterne alla prova l'affidabilità.

Per realizzare liste di base di parole dense e non-dense, il progetto di ricerca si è mosso in due direzioni tra loro intrecciate: a) raccogliere assieme i dizionari di 79 analisi realizzate con AET, comprendenti sia le parole attive che quelle escluse dall'analisi; b) esplorare la variabilità delle scelte dei ricercatori sotto diversi profili e con diversi metodi statistici, riportandone i primi risultati.

Vorremmo soffermarci sulla caratteristica di questo lavoro: si tratta di una meta-analisi dove la concordanza tra i ricercatori è stata studiata sulla base di scelte già effettuate, nel corso di

[3] Abbiamo utilizzato il termine *stemming* per segnalare in tal modo di operare nell'AET un'etichettatura svincolata da puntuali modelli della linguistica o della semantica, e per certi versi della semiometria. Se è vero, come si è detto, che il primo lavoro *tagging* che si compie nell'AET è in sostanza una riduzione della parola alla sua radice etimologica, è vero anche che il ricercatore ha come scopo quello di selezionare le parole in base ad una loro densità emozionale. Infine, come si è detto anche in altra sede (Carli, Dolcetti e Battisti, 2004), lavorando con l'AET si può sentire l'esigenza di fare dei *tagging* anche di tipo diverso, per esempio per formare delle polirematiche o usare solo alcune delle forme grafiche di un medesimo lessema, in base ad una valenza emozionale che ad altre sue forme non viene riconosciuta.

interventi, ognuno con delle sue specificità. Diversamente sarebbe stato mettere i ricercatori nel compito di scegliere la densità emozionale o meno delle parole entro liste, al di fuori d'interventi e lontano dal loro contesto d'uso.

I 79 dizionari raccolti provengono da AET realizzate tra gli anni 1998 e 2009, da due gruppi di ricercatori[4], che abbiamo scelto di denominare *Senior* e *Junior*. Questi si differenziano, in primo luogo, per il livello di esperienza: il gruppo *Senior* avendo elaborato la metodologia l'ha anche usata per primo; quello *Junior* vi si è formato in un secondo momento. Più esattamente presso il gruppo *Senior* sono stati raccolti 38 dizionari di AET realizzate dal 1998 al 2002, mentre presso quello gruppo *Junior* 41 dizionari di AET realizzate dal 2002 al 2009. Il gruppo *Senior* è anche formato da ricercatori con un rapporto di lavoro tra loro più stabile rispetto a quello *Junior*: a quest'ultimo alcuni ricercatori appartengono più continuativamente, mentre altri di volta in volta vi partecipano nell'obiettivo di portare a compimento specifiche AET, quest'ultime realizzate, oltre che per committenti esterni, anche con lo scopo di formare al metodo ricercatori meno esperti.

I suddetti dizionari originano ciascuno da un *corpus* composto con testi omogenei per tipologia: 41 *corpus* formati da interviste, 23 da temi scritti, 6 da resoconti, 4 da *focus group*, 2 da articoli di giornale, 1 da colloqui clinici, 1 da diari, in 1 caso dal testo di un libro.

In una prima fase del progetto sono stati costruiti i dati necessari per individuare e confrontare le scelte sulla densità emozionale delle parole, effettuate nei diversi dizionari da ciascun ricercatore. E' stato applicato un semplice scarto relativo normalizzato e abbiamo prodotto un'analisi sulla concordanza, su quanto *Senior* e *Junior* fossero d'accordo circa la densità o meno delle parole. Da questi dati abbiamo sviluppato un'analisi delle corrispondenze e alcune analisi delle specificità.

Per l'analisi della concordanza tra ricercatori è stato necessario compiere alcuni passi preliminari, non sempre di facile soluzione. Un primo scoglio sono state le *label* usate per lo *stemming* dei *word types* in ciascun dizionario, diverse da un caso all'altro, avendole i ricercatori eseguite manualmente analisi per analisi. Per renderle omogenee, di modo che fosse possibile un confronto tra i vari dizionari, un lavoro sui *file* di partenza è stato eseguito *ad hoc* [5].

Ottenuti, per tutti e 79 i dizionari, i soli dati relativi alle forme grafiche dei dizionari, i loro *stems* e l'informazione - "a" = accepté o "r" = rejeté - indicante l'uso di quella parola quindi

[4] Poiché un'AET viene realizzata entro ricerche-intervento è frutto del lavoro tra più ricercatori e la scelta delle parole su cui si basa l'analisi è solitamente elaborata nel confronto tra più ricercatori. In tutti i casi da noi esaminati le scelte erano frutto di una terna di "giudici".

[5] In partenza avevamo, infatti, dei *files* con estensione .dat e nome A2_DICO, così sono denominati i dizionari dei *corpus* dal *software* Alceste, utilizzato da entrambi i gruppi di ricercatori per le 79 AET oggetto di questo contributo. Questo tipo di *files* contiene diverse informazioni in riga: gli elenchi dei *types* del *corpus* e delle eventuali parole scritte in maiuscolo, entrambi in ordine alfabetico; l'elenco di eventuali numeri presenti nel *corpus*, le modalità delle variabili illustrative utilizzate in quell'analisi e la punteggiatura di quel *corpus*. Mentre in colonna: i codici numerici identificativi di ciascun elemento grafico formante il *corpus*; l'occorrenza delle forme grafiche presenti nel *corpus*, unitamente ad un codice ("a", "r" o "s") per indicare quell'elemento come variabile attiva (accepté), non attiva (rejeté) o supplementare; infine, la colonna per l'eventuale riunificazione di differenti *types* sotto uno *stem* comune. Ai fini della presente ricerca sono state utilizzate queste ultime tre colonne e le sole righe riguardanti le parole. Tra queste ultime, più esattamente abbiamo importato solo gli *stems* sopra la soglia di occorrenza usata per ciascuna analisi (questa soglia di norma pari a 3, nei 79 dizionari presentava una certa variabilità, fino ad un massimo di 6).

come emozionalmente densa o no, questi sono stati inseriti in un unico *file* e modificati in un modo tale che il *software* TaLTaC (Bolasco, 2010), grazie alla sua integrazione con TreeTagger[6], potesse importarli come una tabella parole (dense o non-dense) x dizionari.

In Excel, attraverso un meticoloso lavoro manuale, orientato a tenere conto tanto di somiglianze quanto di differenziazioni nel comportamento dei ricercatori in fatto di *stemming*, tutte le differenti *label* usate per uno stesso *stem* sono state ridotte a una solamente[7]. Di nuovo grazie a TaLTaC, sul risultato di questa tabella, dopo aver eliminato la colonna delle sole forme grafiche, è stata richiesta la fusione del campo dei nuovi *stems*.

La tabella, esito di quest'ultimo passo del lavoro, e contenente in due diverse righe lo stesso *stem* considerato sia denso che non, è stata elaborata da un esperto informatico consentendoci di arrivare ad una tabella della quale, a titolo di esempio, ne riportiamo qui sotto uno stralcio.

Stems	01_11settembre	02_islam	03_islam2	04_islam2005	05_terrorismo	06_terrorismo2005	07_analisidomanda	08_epg	…	74_comment	75_commentverifica	76_sviluppolocale	77_toscanasud	78_comport. econ.	79_aco	80_acorischio	81_pranzonatale
essere	2	2	2	2	2	2	2	2	…	2	2	2	2	2	2	2	2
sempre	2	2	2	2	2	2	2	2	…	2	2	2	2	2	2	2	2
fatica	1	0	0	1	1	0	1	1	…	1	0	1	1	1	1	1	1
crisi	1	0	1	1	1	0	1	1	…	1	1	1	1	1	0	1	1
rischiare	1	0	1	1	1	1	1	1	…	1	1	1	1	1	1	1	0
paura	1	1	1	1	1	1	1	1	…	1	1	1	1	1	1	1	1
diritto	1	1	2	1	2	2	1	0	…	0	1	1	2	2	1	1	0
dimostrare	2	0	2	2	2	2	0	2	…	1	1	1	2	2	0	1	1

Tabella 1. Database nuovi stems x dizionari (leggenda: 0=stem asssente; 1=stem denso; 2=stem non-denso).

Per quanto riguarda la modalità "0", che indica l'assenza della parola in quello specifico dizionario, sottolineiamo che questo accade quando il ricercatore in quell'analisi non ha incontrato quella parola e sulla stessa non ha potuto esprimere alcuna categorizzazione.

2.2. Alcune prime considerazioni sui dati raccolti

La tabella appena presentata, di tipo booleano in forma disgiuntiva completa, raccoglie complessivi 12.571 *stems*, individuati nei 79 dizionari. Si tratta di un dizionario complessivamente interessante in termini di variabilità linguistica, che originando in gran

[6] Uno strumento sviluppato da Helmut Schmid presso l'Istituto di Linguistica Computazionale, dell'Università di Stoccarda (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger), utile, tra l'altro, per annotare il testo con delle informazioni sui lemmi.

[7] Ad esempio, lo *stem* "abbandonare" (che di norma raccoglie sia le forme verbali sia gli aggettivi), etichettato nei differenti dizionari, come abbandon<, abband<, abbandona<, e abbando<, è stato ricondotto alla sola *label* "abbandonare". Allo stesso tempo, se il lessema "competere" era stato rintracciato come talvolta distinto in "competere" e "competenza" e tal altra no, in questo secondo caso non ne è stata operata forzosamente l'unificazione.

parte dal parlato, possiamo paragonare al Lessico di frequenza dell'Italiano Parlato (LIP),[8] contenente circa 490.000 forme grafiche[9], classificate in 15.478 lemmi.

Da questa prima ricognizione è stato possibile osservare che gli *stems*, densi e non-densi, analizzati in ciascuna AET variano tra 314 e 4.188 (valore medio 1.399) mente quelli densi variano tra 93 e 609. Tradotto in percentuale, quelli emozionalmente densi variano tra il 16,5% e il 75%, del totale degli *stems*, nelle varie analisi. Da questi dati si comprende come una lista di base di parole dense potrebbe anche consentire di poter realizzare uno *screening* iniziale sulla quantità di emozionalità presente in un testo. Questo tipo di contributo può affiancarsi e/o trovare incroci con liste già esistenti e talvolta incorporate in *software*, come quella degli aggettivi "positivi" e "negativi" implementata in TaLTaC (Bolasco, Della Ratta 2004), ma anche diventare una risorsa in applicativi come quello per la *sentiment analysis* di SAS dove c'è una fase di implementazione alle risorse interne già fornite dal programma.

3.3. Concordanza sulla categorizzazione di densità e non-densità delle parole

Per mettere in relazione le osservazioni fatte sull'accordo fra *Senior* e *Junior*, entro i diversi dizionari, abbiamo elaborato un primo indice frutto di una semplice tecnica di combinazione di due valori (Nobile, 2008). Si tratta di uno scarto relativo normalizzato formulato come segue:

$$\frac{\text{Frequenza relativa } \textit{Junior} \text{ dense} - \text{Frequenza relativa } \textit{Senior} \text{ dense}}{\sqrt{\text{Frequenza relativa } \textit{Senior} \text{ dense}}}$$

Con questa formula si sono volute comparare le due condizioni, la scelta relativa da parte del gruppo *Senior* e la scelta relativa da parte del gruppo *Junior*, circa la densità di ciascuno *stem* del dizionario. Con questo indice si è scelta una prima prospettiva dalla quale esplorare l'accordo: si sono prese a "modello" le scelte del *Senior*. L'uso della radice al denominatore ha lo scopo di perequare gli effetti del confronto. L'indice varia intorno allo 0, massimo accordo tra i gruppi *Senior* e *Junior*, tra un minimo di – 1 e un massimo di 1 per il disaccordo. Il massimo accordo tra i gruppi si ha sia quando essi convergono sulla densità o sulla non-densità, sia quando convergono su frequenze relative di densità che ciascuno attribuisce allo *stem*. Il segno positivo o negativo dell'indice consente di sapere quale dei due gruppi scarti di più nello scegliere la parola come densa, o non, emozionalmente: se lo *Junior* o il *Senior*. L'indice è tanto più attendibile quanto più è uguale tra loro il numero di osservazioni fatte dai due gruppi di ricercatori. Esso invece non da un risultato in alcune condizioni: per gli *stems* che sono stati osservati solo da uno dei due gruppi (7.079 su 12.571); per gli *stems* che il *Senior* ha considerato sempre non-densi (ricordiamo, infatti, che la formula sopra presentata è impostata sulla frequenza relativa delle parole dense). Visto questo limite, per questi ultimi casi solamente (pari a 1.112 su 12.571) si è calcolato lo scarto relativo normalizzato, tra

[8] Nella ricerca linguistica sull'italiano si tratta della più importante raccolta di testi del parlato, formata, tra il 1990 e il 1992, da 469 testi (De Mauro, Mancini, Vedovelli e Voghera, 1993).

[9] Nel caso della presente ricerca le forme grafiche trattate sono state circa 97.000 derivanti dall'unione dei 79 dizionari, alla base dei quali avevamo circa 6.100.00 di occorrenze. I 12.571, *stems*, invece, originano da una lista ridotta di forme grafiche, pari a 49.170, essendo stare escluse le parole che nei 79 dizionari erano sotto soglia di analisi. La dimensione dei singoli dizionari presenta una certa variabilità, da un minimo di 1.182 *types* ad un massimo di 25.816 (con media 6.602). Il TTR (Type Token Ratio), il rapporto fra *types* e *token* (Tuzzi, 2003), è al di sotto del 20%, per il 92% dei dizionari del gruppo *Senior* e per l'81% del il gruppo *Junior*, indicando nel complesso una buona ricchezza lessicale dei *corpus*.

Senior e *Junior*, sulla base della frequenza relativa delle parole non-dense, individuando, tra l'altro, molti altri accordi sulla non densità delle parole.

In finale i casi sui quali abbiamo informazioni circa la concordanza tra gruppi di ricerca sono 5.492, pari al 43,69% della lista di tutti gli *stems* presenti nei 79 dizionari .

Prima di proseguire ad una valutazione di questi dati ci siamo posti il problema delle frequenze, molte sono infatti le parole osservate complessivamente in un numero esiguo di dizionari. Per le elaborazioni successive abbiamo così scelto di utilizzare solamente gli *stems* presenti in almeno 10 dei 79 dizionari, ritenendo questa una frequenza minima sulla quale poter fare con fiducia considerazioni sui risultati.

Tale questione ci ha fatto anche riflettere sull'utilità di costituire un osservatorio sui dizionari utilizzati mano a mano le AET da arricchire nel tempo con altri dizionari, messi a disposizione sia dai due gruppi di ricerca che hanno partecipato a questo primo lavoro sia da parte di altri ricercatori.

Tornando ad esaminare i risultati dell'analisi della concordanza tra di due gruppi di ricerca, il *database* utilizzato contiene 2.758 *stems* pari al 21, 94% dei 12.571 iniziali.

Vediamo ora alcuni risultati sull'andamento della concordanza nella categorizzazione delle parole come dense o non-dense tra Senior e Junior.

Intervalli di accordo	N. di stems	% di stems
da -1 a -0,76	54	1,96%
da -0,75 a -0,51	212	7,69%
da -0,50 a -0,26	526	19,07%
da -0,25 a -0,01	810	29,37%
0	699	25,34%
da 0,01 a 0,25	304	11,02%
da 0,26 a 0,50	91	3,30%
da 0,51 a 0,75	37	1,34%
da 0,76 a 1	25	0,91%
Totali	**2758**	**100,00%**

Tabella 2 – Dati sull'accordo per intervalli.

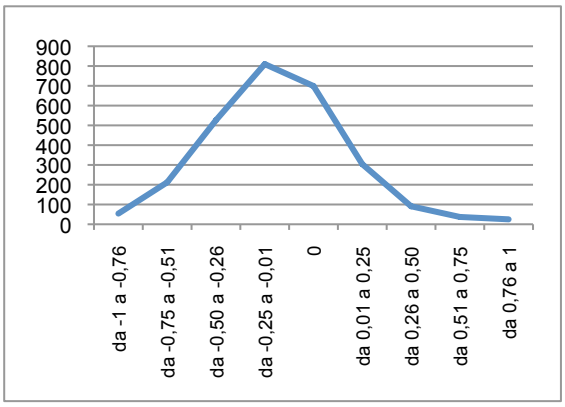

Grafico 1 – Dati sull'accordo per intervalli.

Possiamo anche notare, dal grafico sopra presentato, come gli accordi siano più numerosi nell'area dei punteggi con segno negativo. Questo dato è il risultato di una maggior fedeltà alle proprie delle scelte da parte del gruppo *Senior*, rispetto al gruppo *Junior*. Esso potrebbe essere in linea con quegli elementi con cui abbiamo già inizialmente inteso differenziare i due gruppi: attribuendo sia una possibile maggiore *expertise* al gruppo *Senior* sia rilevando una sua minore variabilità di gruppo, rispetto a quello *Junior*.

Intervalli di accordo	N. di *stems*
accordo	1813
disaccordo	945
Totale	**2758**

Tabella 3 – Dati sull'accordo.

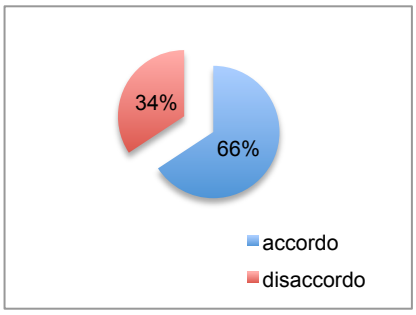

Grafico 2 – Ripartizione % dell'accordo e del disaccordo.

Si può notare da questi risultati come ci sia una quota significativa di accordo tra i due gruppi, *Senior* e *Junior*, si tratta del 65,74%, ottenuto utilizzando i valori di accordo che cadono nell'intervallo da -0,25 a +0,25.

Verificare una buona percentuale di concordanze tra i ricercatori ci ha fatto passare ad altre domande. Infatti, l'accordo ci dice ancora poco di come questo si distribuisca tra i diversi dizionari, o loro raggruppamenti, o come di come esso si distribuisca in base alla densità e non-densità; un'analisi delle corrispondenze può rispondere a queste domande.

3.4. Analisi delle corrispondenze delle parole dense e non-dense dei 79 dizionari

Per questa parte della ricerca abbiamo organizzato una nuova tabella dove ciascuno *stem* è stato annotato con la relativa fascia di accordo e sono stati pensati dei raggruppamenti con i dizionari i cui *corpus* sono vicini sotto il profilo tematico, all'incrocio di righe e colonne abbiamo inserito le relative occorrenze. Sono stati formati raggruppamenti di cui indichiamo il numero di dizionari contenuti: "crisi della convivenza sociale" (12 dizionari), "professione psicologica" (14), "vissuti" (6), "nuovi lavori" (9), "sanità" (9), "sistemi formativi" (12), *media* (3) e "sistemi produttivi" (14)[10].

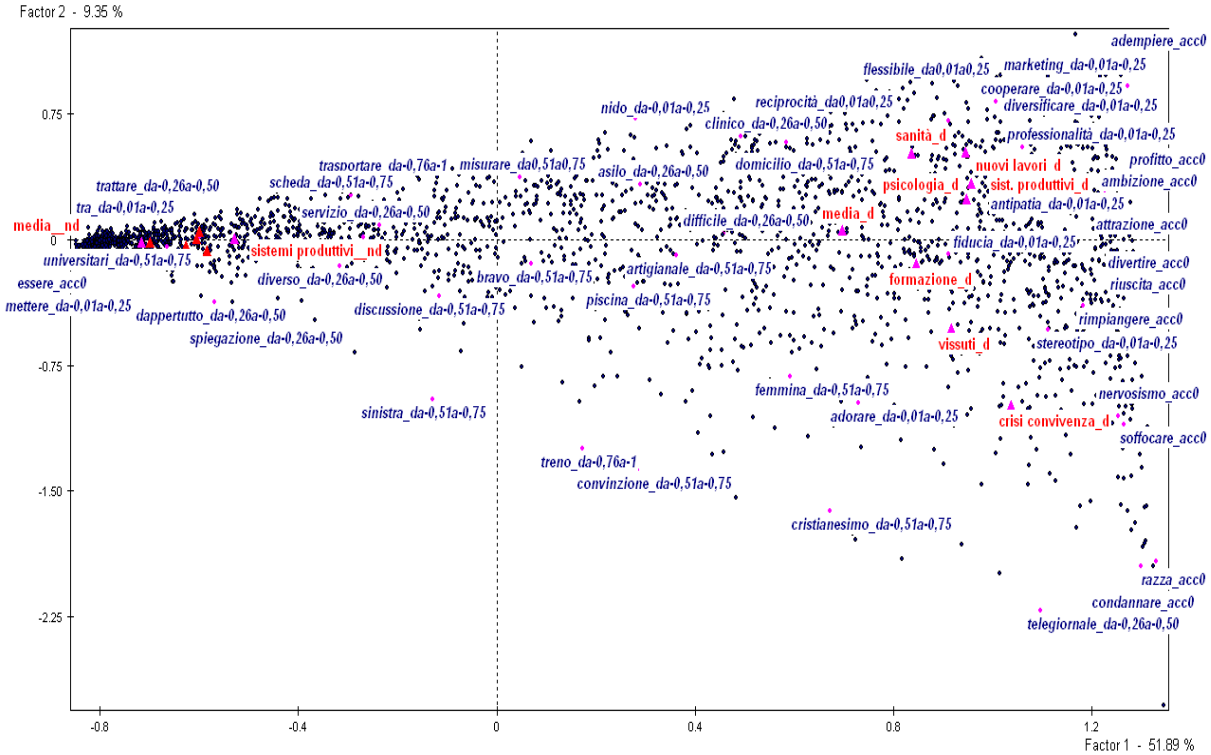

Grafico 4 – Analisi delle corrispondenze.

[10] Più in dettaglio: la tematica "crisi della convivenza sociale" comprende AET sull'11 settembre, sul terrorismo, sull'Islam, sulle rappresentazioni della sicurezza, del rischio sul lavoro e del traffico stradale; in "professione psicologica" vi sono analisi sul modo di usare determinati costrutti teorici, sulle aspettative e sugli obiettivi lavorativi; nella tematica "vissuti" ricadono analisi sulla sperimentazione di vissuti di sofferenza, di colpa, di vergogna, di esperienze di psicoterapia, infine, rispetto ad un evento come il pranzo di natale; nei "nuovi lavori" vi sono analisi sui contratti atipici e sull'avvio di lavori poco consueti; nell'area "sanità" analisi sono state realizzate per lo sviluppo organizzativo di contesti sanitari; la tematica "sistemi formativi" comprende analisi sulla domanda formativa e sulla rappresentazione del futuro dopo la formazione; quella definita "*media*" concerne analisi di articoli di quotidiani; infine, nei "sistemi produttivi" abbiamo analisi sul clima organizzativo, sulla reputazione delle imprese, sullo sviluppo locale.

Questi dati sono stati elaborati con SPAD (Lebart, Salem 2004) realizzando un'analisi delle corrispondenze i cui risultati sono riportati qui di seguito.

Nel grafico, poiché i casi sono numerosi, abbiamo limitato le etichette in modo che fossero rappresentative della variabilità con cui si distribuiscono nello spazio. I fattori successivi ai primi due non sembrano portare un ulteriore significativo contributo.

Vediamo come, sul primo fattore, le parole sostanzialmente si polarizzano in base alla loro densità e non-densità emozionale, a destra le prime a sinistra le seconde. Il secondo fattore sembra far emergere l'ampiezza della ricchezza lessicale dei diversi raggruppamenti di dizionari, se vogliamo le loro maggiori o minori specificità.

Se ci mettiamo ad osservare l'andamento dei dati lungo il primo fattore, ci troviamo di fronte ad una figura, che somiglia geometricamente ad un cono, che con una certa progressione si amplia da sinistra verso destra. Una prima caratteristica di questa polarizzazione è che nella zona delle parole non-dense c'è una minore variabilità sia tra le parole sia tra i raggruppamenti tematici. I dati tendenzialmente ammassati gli uni sugli altri fanno ipotizzare una forte comunanza secondo qualche criterio. Non sembra attribuibile al grado di accordo fra "giudici" in quanto i valori sono sostanzialmente gli stessi della polarità positiva sulla destra: tra - 0,25 e + 0,25, dove 0 è il massimo. Estraendo le parole in ordine di significatività sul fattore abbiamo: *sarebbe stato*, *a parte*, *ovunque*, *tanto che*, *infine*, *malgrado*, *un pò* – tutte con massimo accordo – parole verso le quali non solo nell'AET, ma anche nella letteratura c'è forte consenso sul loro appartenere a quell'universo definito come parole vuote, o parole forma, al contrario delle parole piene, o contenuto. Quest'area della non-densità ha a che fare con un uso di queste parole "in appoggio" ad altre, in base ai modelli grammaticali e sintattici della nostra lingua. Per riassumere, al lato sinistro del primo fattore troviamo parole non-dense, poca variabilità nelle parole e nelle categorie e livello di accordo alto.

Man mano che ci spostiamo verso la parte centrale del grafico, che corrisponde all'incrocio dei fattori, sempre mantenendo il riferimento al primo fattore, abbiamo una zona dove la convergenza di giudizio far *Senior* e *Junior* è meno decisa, come possiamo vedere dall'annotazione sull'accordo di queste parole, includendo diverse parole con basso accordo (valori tra -0,26 a 0,50 e tra 0,26 e 0,50 dell'indice qui utilizzato).

Se sul primo fattore la variabilità dei dati è soprattutto spiegabile nella distinzione fra densità e non-densità e con poche variazioni nell'accordo, nel secondo fattore troviamo sia una presenza maggiore del disaccordo fra i "giudici" sia una maggiore variabilità lessicale. Vediamo, in primo luogo, entro quali dimensioni specifiche si declina il disaccordo su questo secondo fattore. Abbiamo individuato alcuni criteri che possono raccogliere gruppi di parole con aspetti comuni. Fra questi troviamo nomi propri ad esempio di luoghi (Germania, Svizzera, Stati Uniti, Parigi, Napoli) mentre su Cina c'è il massimo accordo, segno dei tempi ci sembra di poter dire. Ci sono parole relative a professioni e/o ambiti produttivi con loro specificità (chimica, giornalismo, turismo, infermiere, architettura, sociologia, psicologia); e ancora, mezzi di trasporto (treno, autobus, automobile, aereo), luoghi caratterizzanti il quotidiano o lo svolgimento di un'attività (convegno, albergo, casa, corridoio, reparto, aula, bar, piscina, cinema, piazza). Altri gruppi individuati, sui quali contiamo di proseguire una riflessione, ci sembra riguardino ambiguità di due diversi livelli. Il primo riguarda parole come: ordine, contro, destra, sinistra, civile, statale, santo; il secondo è relativo in particolare ad alcuni verbi quali: abitare, ritirare, vestire, accendere, mirare, tracciare e ritirare.

Nell'insieme dei gruppi suddetti, presenti sul secondo fattore e con alto disaccordo, la presenza di alcuni criteri ci sembra smentisca che il disaccordo sia solo frutto di idiosincrasie individuali dei "giudici", piuttosto alcune parole solo in alcuni contesti diverrebbero pregnanti dal punto di vista dell'evocare emozioni, sia in chi parla che chi ascolta.

Un altro aspetto del secondo fattore è il contributo di variabilità espressa nella parte destra del piano fattoriale dove le parole si distribuiscono in senso verticale di pari passo alla differenziazione dei diversi raggruppamenti tematici. Elemento questo confermato da una analisi delle specificità di cui qui non diamo conto per via dello spazio.

In finale, vogliamo valorizzare come questo secondo fattore, che per altro ad un esame più approfondito, nella porzione degli alti accordi, risulta formato solo di parole dense, esprima in generale la differenziazione contestuale del linguaggio. In particolare i raggruppamenti con significatività maggiore sono "crisi della convivenza" e "vissuti", nel quadrante in basso a destra, mentre su quello in alto a destra abbiamo "nuovi lavori" e "sistemi produttivi".

3.5. Definizione di una lista di base di parole dense e non-dense

Alla luce dei risultati dell'analisi della concordanza tra "giudici" e dell'analisi delle corrispondenze fin qui riportati e discussi, ci è sembrato di poter rintracciare, infine, alcuni criteri per definire una prima lista di parole dense e non-dense da poter mettere alla prova e monitorare nel tempo.

Utilizzando i risultati dell'analisi delle corrispondenze abbiamo raccolto in diverse liste le parole "appartenenti" ai quattro semiassi dei due fattori. Abbiamo così ottenuto quattro liste contenenti ciascuna 689 *stems* (tale è il numero di parole fornite dall'output dell'analisi). Da ciascuna di queste si è proceduto eliminando le parole con basso accordo, ottenendo così: 1) 618 *stems* con alto accordo, per il semiasse negativo del primo fattore, dove abbiamo detto che si concentrano le parole non-dense; 2) 638 *stems* con alto accordo, per il semiasse positivo del primo fattore, dove si concentra una parte della densità emozionale; 3) 840 *stems* con alto accordo, per il secondo fattore nel complesso, dove abbiamo visto concentrarsi un'ulteriore e sostanziale quota di parole emozionalmente dense, unitamente ad una quota davvero residua di parole non-dense (pari a 5).

In seguito ad ulteriori elaborazioni, attraverso l'unione e l'intersezione delle suddette liste, in conclusione, siamo giunti alla costruzione di una lista di base di 948 parole dense e ad una lista di 618 parole non-dense, per le quali nei 79 dizionari esaminati è stato riscontrato un alto accordo tra i due gruppi di ricerca qui definiti *Senior* e *Junior* (tra -0,25 e +0,25 dell'indice utilizzato). Questo risultato, lo ricordiamo, emerge da una lista iniziale di parole già selezionata: contenente *stems* presenti in almeno 10 dei 79 dizionari.

Diamo qui conto di ulteriori tipiche parole dense (oltre quelle che si potevano apprezzare nel piano fattoriale): urgente, affidare, guarire, aggredire, sperimentale, distruggere, scambio, accumulare, povero, contraddire, odiare, costringere, fortuna, spaventare, riformare, violenza, sfortuna, eccellenza, il dovere, rabbia, burocrazia, stimare, consolare, riuscirci, conflitto, colpa, accreditare, agevolare, benessere, supervisione, difendere, carriera, tradire, ecc.

In finale, con questo lavoro abbiamo una base di dati utile per compiere comparazioni con altre liste: con quella degli aggettivi "positivi" e "negativi", che si può trovare in TaLTaC o altre che si posono prelevare dal General Inquirer (Stone, 1962). Attualmente, ad esempio, possiamo dare già conto di un primo confronto realizzato con le parole 210 parole piene individuate negli studi di semiometria (Lebart, Piron e Steiner, 2003): di queste parole 202

sono state definite dense in almeno uno dei 79 dizionari delle AET qui considerate; le restanti 8 parole semplicemente sono assenti da tali dizionari.

4. Conclusioni

Ricordiamo le due coordinate entro cui questo lavoro si è mosso: a) l'operazionalizzazione della distinzione tra parole dense e non-dense, sotto il profilo delle emozioni; b) l'accordo fra "giudici" quale condizione per misurare quanto questa distinzione sia condivisa, non solo sotto il profilo teorico, ma entro una prassi di lavoro. Questo secondo criterio, di affidabilità del metodo, è stato posto come un possibile elemento in grado di falsificare l'ipotesi che le parole appartengano esclusivamente ad una classe, densa oppure non-densa. Facciamo qui una piccola digressione, perché se è vero che nella consapevolezza della storicità del linguaggio e del suo dinamismo può sembrare ingenuo pensare a una tale mutua esclusività, resta il fatto che nell'ambito del *sentiment* sembra affermata l'idea di cogliere nette distinzioni tra le parole, come quando viene diviso il significato affettivo in positivo e negativo (Stone, 1992).

Ma torniamo a come, attraverso questo primo lavoro, l'indice per osservare l'accordo tra i "giudici" permette di cogliere alcune specificità dei dati. Un primo aspetto è il sostanziale buon accordo fra i "giudici", nell'attribuire a una parola capacità di evocare o meno emozioni. Con l'analisi delle corrispondenze abbiamo, invece: in primo luogo potuto distingue l'accordo sulla densità da quello sulla non-densità; secondariamente, abbiamo osservato un'area che potremmo definire del cattivo accordo, che ha indicato l'esistenza di parole la cui non unanimità di giudizio sembra essere collegata in parte alla "tematicità" delle parole, in parte a dei mondi di ambiguità già conosciuti negli studi linguistici e della statistica testuale che ci inducono a pensare ad un'area di densità dai confini più sfumati.

In conclusione, da questo primo lavoro di ricerca abbiamo ottenuto dei risultati che non confermano il costrutto di densità in chiave dicotomica ma ne propongono una struttura *fuzzy*. Sotto il profilo metodologico l'elemento che dovrebbe "falsificare" l'affidabilità dei "giudici" ci sembra invece ci dia un contributo in grado di complessificare tale costrutto.

Il lavoro fin qui realizzato ci porta ad un punto importante, rispetto agli obiettivi iniziali che ci eravamo dati; la definizione di una lista di base di 948 parole dense e 618 non-dense, da condividere con la comunità scientifica per farne oggetto di una comune riflessione.

Pensiamo di poter mettere a frutto l'esperienza realizzata nel processo di costruzione dei dati pensando di sviluppare uno strumento informatico *ad hoc*, che consenta di incrementare i dati già raccolti con dati di altre AET, fatte dai medesimi due gruppi oggetto di questa ricerca, come da altri che si vorranno aggiungere. Questo aumenterebbe, mano a mano, il numero di osservazioni e potrebbe offrirci maggiore fiducia circa le assunzioni di ipotesi su un certo andamento del fenomeno di categorizzazione da parte dei ricercatori, o se vogliamo della "natura" di densità delle parole.

Il progetto di ricerca, fin qui realizzato, incoraggia ad una prosecuzione verso la realizzazione di analisi che studino singolarmente l'andamento dei tre fenomeni osservati: il gruppo di parole dense con alto accordo, il gruppo di parole non-dense con alto accordo, e il gruppo di parole sulle quali c'è disaccordo. Ma si potrà anche approfondire un confronto tra liste usate nell'area del *sentiment* e mettere in relazione i diversi studi fattoriali che in quest'area insistono. Per quanto riguarda quest'ultimo punto, da un'analisi della letteratura abbiamo potuto riscontrare diverse interessanti analogie tra gli assi semiometrici, i fattori riportati nelle ricerche di Osgood e altri (Osgood, May e Miron, 1975) con i risultati che via via, nel corso degli anni, sono emersi con l'AET.

Bibliografia

Austin, J. L. (1975). *How to do things with words*, in J. O. Urmson e M. Sbisà, editors. Cambridge: Harvard University Press.

Bolasco S. (2010). *TaLTaC$^{2.10}$. Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi*. Roma: LED.

Bolasco S. e Della Ratta Rinaldi F. (2004). Experiments on semantic categorisation of texts: analysis of positive and negative dimension. in G. Purnelle, C. Fairon, A. Dister, editors, *JADT2004 Le Poids des mots. Actes des 7es journées internationales d'analyse statistique des données textuelle*. Louvain-la-Neuve : Presses Universitaires de Louvain, vol (1): 202-210.

Carli R. (1987). *Psicologia clinica. Introduzione alla teoria e alla tecnica*. Torino: UTET.

Carli R. (1990). Il processo di collusione nelle rappresentazioni sociali. *Rivista di Psicologia Clinica*, vol. (3): 282-296.

Carli R. (1995). Il rapporto individuo-contesto. *Psicologia Clinica,* vol. (2): 5-20.

Carli R. (2006). La collusione e le sue basi sperimentali. *Rivista di Psicologia Clinica*, vol (2/3): 179-189.

Carli R. e Paniccia R. M. (1999). *Psicologia della formazione*. Bologna: Il Mulino.

Carli R. and Paniccia R. M. (2000). Il colloquio come testo: l'analisi emozionale del testo. In G. Trentini (Ed.). *Oltre l'intervista. Il colloquio nei conteti sociali.* Torino: ISEDI, pp. 125-158.

Carli R., Paniccia R. M. (2002). *L'analisi emozionale del testo. Uno strumento psicologico per leggere testi e discorsi*. Milano: FrancoAngeli.

Carli R., Dolcetti F. e Battisti N. (2004). L'Analisi Emozionale del Testo (AET): un caso di verifica nella formazione professionale. In Purnelle G., Fairon C. and Dister A., editors, *JADT 2004. Le poids des pots. Actes des 7es journées internationales d'analyse statistique des données textuelles*. Louvain-la-Neuve : Presses Universitaires de Louvain, vol.1: 250-261.

Damasio A. R. (1994). *Descartes' error. Emotion, reason, and the human brain*. London: G.P. Putnam,.

De Mauro T., Mancini F., Vedovelli M. e Voghera M. (1993). *Lessico di frequenza dell'italiano parlato*. Milano: ETASlibri.

Ekman P., editor, (1973). *Darwin and facial expression. A century of research in review*. New York: Academic Press.

Izard C. E. (1977). *Human emotions*. New York: Springer-Verlag.

Lancia F. (2004). *Strumenti per l'analisi dei testi. Introduzione all'uso di T-LAB*. Milano: FrancoAngeli.

Lebart L., Piron M. e Steiner J.-F. (2003). *La sémiométrie. Essai de statistique structurale*. Paris: Dunod.

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

Matte Blanco I. (1975). *The unconscious as infinite sets. An essay in bi-logic*. London: Duckworth.

Nobile. S. (1999). *La credibilità dell'analisi del contenuto*. Milano: FrancoAngeli.

Osgood C. E., May, W. H. e Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. Urbana: University of Illinois Press.

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

Reinert M. (1996). Un logiciel d'analyse lexicale: ALCESTE. *Les Cahiers de l'analyse des données*, vol. (XI): 471-484.

Stone P. J. (1962). *The general inquirer. A computer system for content analysis and retrieval based on the sentence as a unit of information*. Harvard: Laboratory of Social Relations, Harvard University.

Tuzzi. A. (2003). *L'analisi del contenuto*. Roma: Carocci