LED ON LINE
STUDI E RICERCHE

Luca Giuliano

L'ANALISI AUTOMATICA DEI DATI TESTUALI

SOFTWARE E ISTRUZIONI PER L'USO

LED
Edizioni Universitarie di Lettere Economia Diritto

ISBN 88-7916-247-0
Published in *Led on Line* - Electronic Archive by

LED Edizioni Universitarie di Lettere Economia Diritto

http: //www.ledonline.it - http://www.lededizioni.it

http:// www.ledonline.it/ledonline/giulianoanalisi.html
Ottobre 2004

ledonline@lededizioni.it
luca.giuliano@uniroma1.it

In copertina:

Romberch, *Congestorium artificiose memori*
1533

Videoimpaginazione e redazione grafica: Studio Venturini
Stampa: Globalprint

INDICE

1.
INTRODUZIONE ALL'ANALISI AUTOMATICA DEI TESTI

I giochi linguistici sono affascinanti proprio perché giocano con la vertigine che viene suscitata dalla combinatoria delle lettere dell'alfabeto cui si associa un senso che presenta sempre delle ambiguità. Questa poesia di Arrigo Boito ne è un'espressione sublime.

> Noi siamo tre Romei
> Madonna, fa che si diventi sei.
> Scesi dall'Alpi algenti
> Ove dan morte turbinando i venti.
> Qui ne venimmo dove
> Preghiam dal viso tuo dolcezze nove.
> Fa che tu ne promette
> Sul bel colle, lontan dall'empie sette.
> Tanto coll'occhio bruno
> Che sembri dire: intorno a me vi aduno [ad-uno].
> E ne farai felici
> se "l'assenso richiesto a voi do", dici [do-dici].
> Ché se rivolgi ad altre
> Estranie cose le pupille scaltre [scal-tre]
> Noi sentiremo il fiotto
> Stagnar nel cor e piangerem dirotto [dir-otto].
> Esaudi i tre Romei,
> Se buona, se gentil se santa sei [se(s)-santa-sei].

L'ultima cifra dell'ultimo verso è la somma delle cifre degli otto versi precedenti. Se i testi da sottoporre ad analisi testuale avessero tutti questo grado di ambiguità la nostra sarebbe un'impresa impari. Al gioco linguistico dei numeri

e delle lettere (segnalato da Stefano Bartezzaghi, "Lessico & Nuvole", *La Repubblica – Il Venerdì*, 22 aprile 2004) si aggiunge una difficile decodificazione del contesto. La poesia è stata scritta da Arrigo Boito nel 1884 per Eleonora Duse; Boito era in vacanza sulle Alpi piemontesi con Giuseppe Giacosa e Giovanni Verga; il secondo verso si riferisce al fatto che a loro tre si sarebbero aggiunti – se l'invito fosse stato accolto - la Duse, il marito Tebaldo Checchi e un attore loro ospite che si chiamava Zolis.

1. 1. TESTO, SIGNIFICATO E INTERPRETAZIONE

Textus deriva dal latino, come participio passato di *texĕre*, una parola antichissima dalla radice TEKY che indica il lavoro del taglialegna e del carpentiere. In questo senso è stata rilevata nelle aree indoiranica, greca (*tékton*, "carpentiere"), slava, germanica e celtica (Devoto, 1979).

'Tèssere', pertanto, va inteso come ordire una trama di fili, una tela, così come il carpentiere disponeva i blocchi di legno. Il verbo viene usato anche in senso figurato, "ordire una macchinazione o un inganno". Da qui 'tessuto', 'intreccio' e quindi "complesso linguistico del discorso" (Segre, 1981) così come appare nella *Institutio oratoria* (IX,4,13) di Quintiliano. E' proprio dall'affermasi del *textus* nel latino dell'era cristiana che Cesare Segre vede il trionfo della scrittura, delle religioni del Libro a fronte della diffidenza che i Greci avevano per la parola scritta intesa come semplice trascrizione del discorso orale.

D'altra parte la società antica è caratterizzata dalla comunicazione orale, una comunicazione in cui i messaggi vengono recepiti nella **situazione** in cui vengono emessi. "Io parlo, tu ascolti." La comunicazione viene prodotta nel mondo della vita e il discorso prende senso solo dai rapporti **particolaristici** tra i parlanti. E' una comunicazione sincronica, delimitata nel tempo e nello spazio. La conoscenza nella società della comunicazione orale è esperienza diretta e trasmissione di esperienze tra le generazioni. L'esperienza ricondotta alla forma orale si trasmette in modo ciclico: è sempre uguale a se stessa e ritorna all'inizio di ogni ciclo generazionale per trasmettersi identica nel ciclo successivo. La società si racconta attraverso il mito e il racconto è dotato di una forza di conservazione eccezionale. Nessun racconto è eterno quanto il mito. Le società prive di scrittura conservano paradossalmente la propria cultura in modo più rigido e stabile di quanto non accada nelle società della scrittura. Il messaggio nella trasmissione orale è attuale nel momento stesso in cui viene

emesso. Il testo del discorso non esiste. Il discorso è "in atto".

Nelle società della scrittura il discorso viene trascritto per essere conservato. Il testo della trascrizione orale permette la comunicazione asincrona, la comunicazione diacronica. Il testo si separa dal mondo della vita, dalla situazione in cui viene creato il discorso. "Io scrivo, tu leggi." Un messaggio scritto su una tavoletta o su un papiro varca i limiti del tempo e dello spazio per essere letto a grande distanza e anche molti secoli dopo. La situazione in cui viene emesso il messaggio e quindi la trascrizione (**contesto**) diventa molto importante per il suo significato, per la sua comprensibilità, per la sua corretta interpretazione e ricezione. La situazione in cui viene emesso il messaggio non è più una necessità. La trasformazione della situazione in contesto diventa una scelta: un modo per concepire messaggi che nascono con l'intenzione di essere **universali** (Stele di Rosetta). La scrittura nasce nelle società antiche perché i discorsi non sono più limitati alla immediatezza della situazione in cui vengono enunciati. I discorsi si trascrivono perché sono destinati a sopravvivere alla dimenticanza. La scrittura al suo esordio è sacra.

Nelle società della scrittura l'esperienza è lineare. Non è più legata al rapporto inter-generazionale. Si formano identità collettive forti che organizzano la trasmissione delle esperienze. Sono società centralizzate, spesso dotate di una casta sacerdotale che è depositaria dell'interpretazione dei testi. Sono società che si raccontano attraverso il libro. E' così che nascono le "religioni del Libro".

La scrittura virtualizza la memoria trasferendola in un testo scritto. Il senso del testo non è indipendente dal lettore e dalla sue scelte. Il testo assume un significato solo in presenza di un lettore. Il lettore collabora all'attualizzazione del testo mettendo in collegamento il testo, anche inconsapevolmente, con un mondo di significati che gli appartengono in quanto lettore. Vi è un extra-testo che contribuisce al testo. Lo stesso testo letto da persone diverse produce diversi "significati" del testo.

La lettura non è una attualizzazione del testo virtualmente rappresentato da segni o ideogrammi. La lettura è una **attualizzazione dei significati** del testo. Alla lettura del testo, alla sua memoria e alla cura posta contro la soggettività e le sue distorsioni viene data grande importanza, per esempio, nella religione musulmana che pone il Libro (il Corano) nel suo centro. L'assoluta superiorità del testo rappresentato dal Corano è un vero e proprio dogma di fede nell'Islam. Il Corano è la Parola di Dio. Il Corano è Dio che si fa testo. Anche la religione cristiana è una religione del Libro, ma nel cristianesimo Dio si fa uomo. Il Corano viene memorizzato nelle scuole coraniche attraverso la recitazione salmodiante. Così come viene fatto per recitazione della Torah nella

sinagoga. La sacralità del Sefer Torah (il Pentateuco) si esprime anche nella sua trascrizione che avviene secondo regole molto rigide sulla preparazione dell'inchiostro, su come effettuare le eventuali correzioni, ecc. Il Sefer Torah è sacro non solo nella enunciazione ma anche nella sua materialità: quando è danneggiato ed inutilizzabile viene chiuso in un contenitore di terracotta e sepolto nel cimitero ebraico.

Il testo subisce una ulteriore trasformazione nella società della comunicazione digitale. La comunicazione digitale non conosce tempo e non conosce confini. E' contemporaneamente sincronica e diacronica. Il testo, nella comunicazione digitale, appare là dove qualcuno lo richiede. La sua **universalità** dipende dalla compresenza di altri testi, dalla interconnessione dei messaggi tra loro. Il contesto non è più il mondo della vita cristallizzato in un sistema di conoscenze che ne permettono la ricostruzione. Il contesto è rappresentato da tutti gli altri testi che sono collegati al testo in una **rete**. Nelle società della comunicazione digitale l'**esperienza** è **reticolare**. Priva di centro, molteplice, diversificata.

La società digitale si racconta attraverso l'**ipertesto**. L'ipertesto è una virtualizzazione del testo, un suo potenziamento. Il testo reticolare è l'insieme delle memorie virtualizzate, l'intelligenza collettiva. Il testo in rete si attualizza in una forma particolare di lettura che è la navigazione ipertestuale, intertestuale e intratestuale.

> Il testo è dunque il tessuto linguistico del discorso (Segre, 1981, p. 269).

Il tessuto linguistico è realizzato attraverso una successione di lettere e accenti interrotti da uno spazio o da segni d'interpunzione che costituiscono le parole, disposte in righe parallele. La lingua si presenta all'osservazione attraverso i testi. Un testo scritto è composto di segni che sono ordinati in una certa sequenza per formare una catena (Hjelmslev, 1970, p. 37). L'ordinamento è determinato: da sinistra a destra per l'alfabeto latino, da destra a sinistra per l'alfabeto ebraico, dall'alto al basso per l'alfabeto mongolo, ecc.).

> Chiamiamo *testo* la totalità di una catena linguistica così sottoposta ad analisi (Hjelmslev, 1970, p. 111).

La parola *testo* viene utilizzata sia per indicare il contenuto di un discorso (il **significato**) quanto per indicare i segni da leggere, il *veicolo* materiale della trascrizione (il **significante**). Quando vi trovate di fronte a un testo scritto in una lingua straniera di cui non sapete nulla, nemmeno i segni dell'alfabeto, allora vi trovate di fronte a un significante allo stato puro. Naturalmente c'è un

rapporto tra significante e significato, perché senza significante mancano le condizioni in cui si manifesta il significato. Senza significante non c'è possibilità di esprimersi e dunque non c'è più significato. Il significante, secondo De Saussure, è un'immagine acustica che, associata a un significato, forma un segno (Marchese, 1978, p 248).

Il rapporto tra significante e significato è rappresentato dalla **significazione**. La significazione è indipendente dalla natura del significante sulla base del quale si manifesta.

Alcuni autori hanno criticato questa impostazione (Abrham Moles, Marshall McLuhan), sostenendo che il modo di presentare qualche cosa (per esempio un prodotto pubblicitario) ha un impatto maggiore di ciò che viene presentato realmente. E chiamano questo (per esempio l'immagine a colori) come "contenente" che affascina il pubblico e agisce su di esso al di là del suo contenuto letterale. In realtà essi sembrano confondere due piani diversi. L'immagine non è significante, ma è già significato. Il vero significante è la carta, il contrasto cromatico dei colori, ecc.

Il testo ha un'oggettività? In un certo senso sì, la sua oggettività è rappresentata dalla materialità del significante. Un discorso orale non può che essere soggettivo, singolare, irripetibile. Padre Roberto Busa, il gesuita che con il suo studio lessicografico sull'opera di S. Tommaso d'Aquino (*Index Tomisticus*, 1974-1980) è stato tra i pionieri dell'analisi testuale, ha descritto il suo immenso lavoro come quello di un ricercatore che percorre il greto di un fiume ormai inaridito e raccoglie i sassi sui quali l'acqua ha lasciato i segni del suo passaggio. Lo scorrere dell'acqua rappresenta la voce melodiosa della discorsività di Tommaso; i sassi rappresentano quanto resta nei testi delle sue parole (testimonianza raccolta personalmente durante la presentazione a Roma, all'Università Gregoriana, di un'analisi testuale delle encicliche papali, 27 marzo 2001).

Solo in tempi recentissimi, in un tratto brevissimo della storia della parola, sì è resa possibile la registrazione e la riproduzione della voce, ma non per questo l'oralità conquista una oggettività. L'oggettività del testo attraverso la scrittura è potenzialmente possibile nel momento in cui la sua attualizzazione è differita, ma l'intervento del lettore, e quindi della significazione, porta con sé l'irrompere della soggettività (Segre, 1981). Infatti è all'interno di questa intersoggettività, attraverso la competenza dei codici e dei riferimenti al contesto e all'extra-testo, che si organizza l'**interpretazione**: il confronto, il dibattito, la critica, la problematicità continua che è alla base del lavoro di una comunità scientifica.

L'interpretazione del testo è una approssimazione alla verità nella consa-

pevolezza che la verità è una meta che potrebbe non essere mai raggiunta. La verità non si può trovare da qualche parte, non si può "scoprire" ma si "produce" attraverso i meccanismi dell'interpretazione (Rorty, 1979). Ogni testo ospita contemporaneamente più testi. Ogni testo è soggetto di più interpretazioni. Ma non solo per la intersoggettività del testo. Un testo autografo di solito contiene delle aggiunte, delle correzioni, delle varianti sulle quali il critico esercita le sue scelte per stabilire il "testo autentico". Il problema della autenticità del testo non viene eliminato nella comunicazione digitale. Un errore di ortografia, una punteggiatura messa nel posto sbagliato, una decodifica errata compiuta nel passaggio da un sistema operativo all'altro, possono far convivere più testi in un unico testo digitale.

1. 2. CLASSIFICAZIONE DEI TESTI E FORMAZIONE DEL CORPUS

Nel parlare di "testi" che cosa intendiamo? Ci sono sicuramente i testi poetici e letterari, gli articoli dei giornali, ma anche gli atti notarili, come il contratto d'acquisto di una casa, gli atti pubblici come i bandi di concorso, oppure i documenti legislativi. Ci sono le lettere che si scambiano i privati, oggi ci sono anche le lettere in formato elettronico (e-mail), oppure i verbali delle assemblee societarie. Ci sono testi che sono trascrizioni di interrogazioni verbali (gli interrogatori giudiziari oppure le interviste non direttive) e testi che sono semplicemente schede di rilevazione di dati, come il modello di rilevazione del censimento.

Per classificare i testi prendiamo come riferimento una ricerca internazionale promossa dell'OCSE (Organizzazione per lo Sviluppo e la Programmazione Economica) nel 2000 che aveva come scopo la rilevazione delle competenze linguistiche, matematiche e scientifiche dei ragazzi di 15 anni (*PISA – Programme for International Students Assessments*). In questa ricerca sono stati assunti tre criteri principali di classificazione:

- *genere*: letteratura di fantasia e letteratura empirica;
- *categoria*: testi descrittivi, narrativi, informativi, argomentativi, conativi, documenti, ipertesti;
- *struttura*: testi continui e testi discontinui.

La prima è una distinzione di *genere* e riguarda i testi che si riferiscono alla *fiction* e alla *non fiction*. La distinzione è tutt'altro che semplice. I **testi finzionali** sono prodotti di attività estetiche sottoposti a convenzioni letterarie. I testi non finzionali, definiti anche come **testi empirici**, fanno riferimento a dati di

esperienza. *Le ultime lettere di Jacopo Ortis* (1806) di Ugo Foscolo è un testo finzionale, mentre le lettere degli immigrati polacchi analizzate da W. Thomas e F. Znaniecki in *Il contadino Polacco in Europa e in America* (1918) sono testi empirici.

Il secondo criterio è riferito alle *categorie di organizzazione del contenuto* e cioè lo scopo per cui i testi sono stati scritti.

I **testi descrittivi** rispondono alla domanda "che cosa?". Possono contenere descrizioni "soggettive" che esprimono il punto di vista di chi scrive (o di chi parla, se il testo è una trascrizione) oppure possono essere descrizioni tecniche e scientifiche che, almeno nelle intenzioni, puntano ad essere "oggettive".

I **testi narrativi** rispondo alle domande "quando?" e "in che ordine?". Possono essere "racconti" (in cui il punto di vista è quello del narratore), "rapporti" (le informazioni contenute nel testo possono essere verificate/falsificate da altri che non sono il narratore), "testi di attualità" (la narrazione viene effettuata da un giornalista e permette al lettore di farsi una sua opinione della realtà in cui vive).

I **testi informativi** sono prevalentemente orientati a rispondere alla domanda "come?". Possono prendere la forma del "saggio esplicativo" (presentazione di concetti più o meno complessi), "riassunti" (sintesi di informazioni contenute in un testo originario), "verbali" (trascrizioni di quanto è stato detto durante una riunione), "interpretazione di testi" (commento a un testo al fine di dare una spiegazione di quanto vi è contenuto).

I **testi argomentativi** rispondono soprattutto alla domanda "perché?". Si tratta di "argomentazioni scientifiche" (interpretazione di idee o sistemi di pensiero e spiegazione di eventi su cui è possibile effettuare controlli di validità) oppure "commenti" (interpretazioni e spiegazioni che hanno un carattere del tutto personale e soggettivo).

I **testi conativi** forniscono indicazioni su come si devono svolgere determinati compiti. Appartengono a questa categoria le "istruzioni" (come i libretti che accompagnano le strumentazioni tecnologiche) e i "regolamenti" (dagli statuti associativi fino alle leggi istituzionali).

I **documenti** sono testi che servono a conservare le informazioni in una forma predefinita (certificati anagrafici, contratti di compra-vendita, fatture, ecc.).

Gli **ipertesti** costituiscono la forma più recente di testi e sono costituiti da parti di testo collegati tra loro in modo che il lettore possa costruire un suo percorso personale durante la lettura.

Il terzo criterio è riferito alla *struttura fisica* del testo.

I **testi continui** sono quelli più consueti, come quello che state leggendo in questo momento. Vi sono delle frasi organizzate in capoversi e in paragrafi. Vi sono (a volte) dei titoli dei paragrafi e i paragrafi sono raggruppati in capitoli o in sezioni dotate a loro volta di titoli che ne riassumono il contenuto o suggeriscono qualche idea. Alcune parole sono evidenziate in qualche modo (tra virgolette, in corsivo, i grassetto, ecc.) e alcune parti del testo sono organizzate in modo da facilitare la lettura (elenchi) o da renderne riconoscibile la funzione speciale (note, indicazioni bibliografiche, commenti tra parentesi, ecc.).

I **testi discontinui** sono quelli che normalmente non vengono considerati testi ma che nel programma PISA sono stati comunque classificati (elenchi, moduli, questionari, tagliandi, tabelle, ecc.).

Per noi è ancora degno di nota un altro criterio di classificazione che non è stato considerato rilevante nel programma PISA ma che oggi è di assoluta preminenza: **testo digitale** o **testo a stampa.** Probabilmente la distinzione non è del tutto ortodossa perché oggi gran parte dei testi assumono o tendono ad assumere la forma digitale. Tuttavia vi è una differenza sostanziale tra i testi che sono già in origine digitali (e-mail, pagine web, blog, messaggi in forum e newsgroup, conversazioni in chat, MUD, ecc.) e testi che sono destinati alla stampa, provengono dalla digitalizzazione del testo a stampa oppure sono solo occasionalmente redatti in forma digitale ma sono orientati alla lettura stampata (come questi appunti che sto scrivendo sul monitor utilizzando la tastiera del computer). In alcuni casi vi sono testi che non sono pubblicabili in una forma diversa da quella elettronica (http://www.bibletimp.it).

I testi, di per sé, non sono immediatamente oggetti di analisi scientifica. Noi ci avviciniamo ad essi con l'intento di analizzarli mediante strumenti quantitativi, di dare una interpretazione e classificazione logica del loro contenuto a partire dalla spiegazione dei valori semantici delle parole. Questo è un approccio **logico-semantico** che va tenuto distinto da un approccio più propriamente **linguistico** rivolto all'analisi delle tipologie discorsive, dello stile, del "come" viene prodotto un discorso anziché del "che cosa" contiene.

Da un punto di vista operativo, per la costituzione di un oggetto di ricerca l'osservazione non è indipendente dalla teoria. Non possiamo delimitare il campo di interesse delle nostre osservazioni senza formulare delle ipotesi che saranno la guida per la raccolta e l'analisi dei dati. Le ipotesi sono risposte provvisorie a problemi. Il processo di ricerca inizia con una domanda che è in attesa di risposta.

I testi, pertanto, hanno un interesse e sono analizzabili solo se costituiscono un **corpus** di testi:

> Per corpus s'intende un qualsiasi insieme di testi, fra loro confrontabili sotto un qualche punto di interesse (Bolasco, 1999, p. 182).

Il corpus è l'insieme dei testi sui quali si deve effettuare l'analisi. In molti casi il corpus si costituisce facilmente: per esempio, la totalità delle risposte ad una domanda aperta in un questionario sottoposto a un numero finito di persone; ciascuna delle risposte costituisce un testo. I testi possono essere raggruppati secondo le caratteristiche dei soggetti intervistati (maschi e femmine, classe di età, ecc.). In un'analisi comparata degli articoli dei giornali quotidiani su un certo avvenimento, il corpus sarà rappresentato dalla totalità degli articoli nel corso di un certo periodo. Anche in questo caso gli articoli (o semplicemente i titoli) sono testi che possono essere classificati, oltre che secondo la testata, secondo la posizione, secondo il rilievo che hanno sulla pagina, ecc.

Ci sono casi in cui il corpus è più difficile da determinare. Il corpus è sempre una conseguenza di decisioni operative in un certo contesto di ricerca: discorsi politici, testi di interviste o storie di vita, messaggi pubblicitari, trascrizione di *focus groups*, raccolte di e-mail su un certo argomento, ecc. Il corpus è un insieme ragionato di testi che corrispondono ad un obiettivo e quindi a precise ipotesi di ricerca. E' impossibile dire a priori se un corpus è costituito adeguatamente senza assumere come riferimento lo scopo per cui verrà analizzato (Habert – Fabre - Issac, 1998, p. 35).

1. 3. GLI ELEMENTI COSTITUTIVI DEL TESTO: LE PAROLE

Il corpus quindi è costituito di testi e ogni testo rappresenta una delle possibili partizioni di un corpus. Ogni testo si può suddividere a sua volta in **frammenti**. Il tessuto linguistico del discorso è suddiviso o costituito di **frasi**, cioè di partizioni del testo in cui si esprimono pensieri come enunciati completi. Un testo può essere costituito di una sola frase.

Senza alcuna pretesa di voler affrontare materie complesse come la semiologia, la linguistica, la retorica e la stilistica, ci sono però delle conoscenze fondamentali nello studio della lingua che devono essere ben comprese per affrontare lo studio quantitativo dei testi e delle parole.

I principali elementi costitutivi di un testo sono le **parole**. Di ogni parola dobbiamo conoscere la pronunzia, la grafia e il significato. Gli elementi fondamentali di questa conoscenza costituiscono il **lessico**. Gli oggetti del lessico sono raccolti in opere che descrivono le parole all'interno dei settori in cui es-

se vengono utilizzate: i **vocabolari**. E' inappropriato il termine dizionario perché potrebbe indicare opere che non hanno nulla a che fare con la linguistica come, per esempio, il *Dizionario biografico degli italiani* (Lepschy, 1979, p. 129).

Le regole che descrivono e spiegano le strutture della lingua costituiscono la **grammatica**. La **morfologia** è quella parte della grammatica che si occupa delle forme (flessione, declinazione e coniugazione) che le parole assumono. La **sintassi** detta invece le regole per una disposizione ordinata e corretta delle parole all'interno del **discorso**.

Senza entrare nel dettaglio della nomenclatura grammaticale, ci sono alcune definizioni che è bene rammentare perché sono rilevanti ai fini di una corretta applicazione dell'analisi testuale.

Di fatto la distinzione tra grammatica e lessico oggi passa attraverso la nozione di regolarità della lingua. La grammatica è il regno della sistematicità, mentre il lessico è il regno dell'irregolarità, si sottrae alle generalizzazioni (Lepschy, 1979, p. 134).

Le parti del discorso nella sintassi:
- *soggetto*: ciò di cui si parla e a cui si riferisce l'azione (compiuta o subita);
- *verbo* o *predicato verbale*: l'azione riferita al soggetto, il modo di essere, ecc.;
- *oggetto*: lo scopo dell'azione;
- *complementi*: parti della proposizione che ne completano il significato esprimendo relazioni e circostanze in cui l'azione viene svolta.

I termini grammaticali:
- *aggettivo*, *attributo*: si aggiunge al sostantivo per determinarlo e qualificarlo;
- *articolo* (determinativo o indeterminativo): particella premessa al sostantivo che precisa il genere e il numero del nome che la segue;
- *avverbio*: si pone vicino al verbo o all'aggettivo per modificarlo;
- *congiunzione*: serve a legare parti di una frase tra loro;
- *interiezione*: esprime un'esclamazione di tipo emotivo (dolore, gioia, sorpresa, ecc.);
- *numero*: definisce la singolarità o pluralità di sostantivi e verbi;
- *preposizione*: serve a legare una parola (nome, aggettivi o verbi) all'altra;
- *pronome*: sostituisce il nome senza specificarlo;
- *sostantivo*, *nome*: indica esseri animati, inanimati, oggetti concreti o astratti;
- *verbo*: indica l'azione o il modo di essere di una entità, la relazione tra entità, l'appartenenza ad una categoria esistenziale.

Le parole costituiscono le proposizioni (le frasi). Le frasi messe in sequenza costituiscono un discorso (o un testo, quando il discorso è trascritto). Il testo

però non è una somma di frasi: è qualcosa di più che contribuisce a costruire il senso. Il testo permette di restituire il significato delle parole anche quando queste, prese isolatamente, sono **ambigue**. La parola con grafia <rosa> può essere riferita al nome del fiore, ad un colore, al participio passato del verbo *rodere* o all'aggettivo corrispondente, espresso nel genere femminile. La pronunzia è solo in parte sufficiente a disambiguare la parola (se si parla in modo corretto): *róso*, nel senso di *corróso* ha l'accento acuto come *fióre*; mentre *ròsa* (il fiore) ha l'accento grave. Tuttavia *ròsa* (il colore, il gruppo, la serie) è indistinguibile da *ròsa*. Solo il contesto della frase permette di rendere univoco il significato della parola. Così intere frasi possono essere ambigue e solo il testo permette di comprenderne il senso. "Quel cane del tenore ulula da un'ora" assume un senso completamente diverso secondo il testo in cui la frase è inserita: potrebbe riferirsi ad una povera bestia (il cane) che aspetta il ritorno del suo padrone (il tenore); oppure essere riferito al tenore (un cane) che canta in modo sgradevole per il pubblico (l'esempio è tratto da Marchese, 1978, p. 67).

La parola scritta ha quindi tante repliche diverse che possono derivare da significati diversi (**polisemia**) con pronunzia identica (parole **omofone**) e da significati diversi con grafia identica (parole **omografe**). D'altra parte ci possono essere significati che hanno una stessa forma grafica ma che appartengono a due parole diverse (**omonimia**); oppure due forme grafiche diverse che esprimono lo stesso significato (**sinonimia**). Le omonimie sono frequenti nell'inglese, in cui la pronunzia e la grafia sono in un rapporto complesso.

Le parole possono anche essere raccolte in lemmi, cioè essere classificate in forme che hanno fra di loro qualche attributo comune, per esempio il fatto di essere coniugazioni dello stesso verbo (*venire* come lemma di *vengo*, *vieni*, *viene*, ecc.). Le parole così come sono classificate nei vocabolari costituiscono dei lemmi. Tuttavia la lemmatizzazione non è un'operazione univoca né semplice da definire. Si possono classificare in un unico lemma parole che hanno una funzione grammaticale diversa (*termine*, *termini*, *terminare*, *terminante*) oppure parole che appartengono ad uno stesso ambito semantico perché legate ad un'origine comune (che hanno lo stesso etimo) come il lemma *fare* che comprende il verbo *fare* con tutte le sue coniugazioni, ma anche parole come *faccenda*, *faccendiere*, *facile*, *facinoroso*, *facoltà*, *fazione* e *fazioso* (Gianni, 1988).

Le parole sono costituite, come si è detto, da un significante e da un significato. Il significante è la rappresentazione grafica della fonetica della parola e consiste di una stringa di lettere ordinate linearmente. L'inventario dei significanti è limitato, si presta a vari tipi di ordinamento (l'ordine alfabetico è quello solitamente utilizzato nei vocabolari) ed è analizzabile con i sistemi di trattamento automatico dell'informazione. I significati fanno riferimento a un

campo dai confini sfumati, non sono ordinabili e non sono mai stati inventariati in modo esaustivo in nessuna lingua. Noi cerchiamo nei vocabolari il significato di una parola a partire da un significante noto. Non possiamo fare l'inverso e non c'è nessun vocabolario che permetta di farlo, salvo alcuni prodotti culturali interessanti come il *Dizionario analogico della Lingua italiana* (UTET, Torino, 1991) o il *Dizionario alla rovescia* (Selezione del Reader's Digest, Milano, 1992).

Tra le varie categorie di vocabolari utilizzati nella linguistica, per noi sono particolarmente interessanti i **vocabolari (o lessici) di frequenza**, in cui le parole sono ordinate per ordine di frequenza. Il problema naturalmente è quello della rappresentatività dei testi sui quali è stato effettuato lo spoglio. Fino a che punto possiamo dire che un determinato numero di frasi, tratte da campi diversi del discorso, possono essere considerati un campione rappresentativo della lingua italiana?

I lessici di frequenza più significativi per la lingua italiana sono:

- U. Bortolini, G. Tagliavini e A. Zampolli. *Lessico di frequenza della lingua italiana contemporanea* (LIF), Garzanti, Milano, 1972.
- T. De Mauro, F. Mancini, M. Vedovelli, M. Voghera. *Lessico di frequenza dell'italiano parlato.* ETASLIBRI, Milano, 1993.
- A.M Thornton, C. Iacobini, e C. Burani. BDVDB. *Una base di dati per il vocabolario di base della lingua italiana.* Bulzoni Ed., Roma, 1997.
- T. De Mauro. *Vocabolario di Base della lingua italiana* (VdB), in Guida all'uso delle parole. Editori Riuniti, Roma, 1997 (12a edizione).

La domanda che si pone il lessicografo è: quante parole ha una lingua? Di solito un vocabolario contiene da cinquantamila a cinquecentomila voci per i più estesi. Un individuo di media cultura conosce normalmente da 3.000 a un massimo di 20.000 parole. Gli scrittori utilizzano mediamente da 5.000 a 15.000 parole (Lepschy, 1979, p. 146).

La frammentazione minima di un testo è la parola, una sequenza di caratteri alfabetici delimitata da due separatori. L'insieme dei separatori deve essere convenzionalmente definito come un insieme di caratteri che non appartengono all'alfabeto. Questo non è un problema banale perché in un alfabeto possono esserci dei caratteri che in alcune circostanze possono essere considerati dei separatori (per esempio i numeri). Sono quasi sempre da considerare come separatori: lo spazio bianco (*blank*), la punteggiatura (, . : ; ? !), le virgolette, i "trattini" (- / |) e le parentesi ({ } ( ) [ ] . Per i caratteri speciali (# @ $ £ § ° % & ^ * < >) e per i numeri si dovrà decidere caso per caso secondo gli obiettivi dell'analisi.

Una sequenza di caratteri delimitata da due separatori definisce una **for-**

ma grafica. Le forme grafiche intese come unità di conto vengono definite **occorrenze** (*words token*). L'analisi automatica di un testo fornisce come primo risultato un conteggio delle forme grafiche (delle parole). Ad ogni parola diversa viene associato un codice numerico, pertanto è possibile costruire un indice delle forme grafiche che sarà rappresentato dalle **parole distinte** (*words type*).

Non tutte le parole di un testo possono essere considerate come equivalenti dal punto di vista semantico. Vi sono delle difficoltà nel considerare la parola una unità di base elementare della semantica. Pensiamo a parole come *gatto*, *finestra*, *mela*, *amare*, *odiare*. Ognuna di queste parole presente in una frase rinvia a un contenuto e noi possiamo scegliere se utilizzare l'una o l'altra secondo il concetto che vogliamo esprimere. Parole come *il*, *e*, *che*, *di* sono presenti in una frase in rapporto con altre parole e non possono essere sostituite con parole equivalenti sebbene con un altro significato. "Amare il gatto" può diventare "odiare il gatto", ma l'articolo *il* è insostituibile senza generare una frase grammaticalmente diversa. Uno studioso inglese del XIX secolo, Henry Sweet, propose di distinguere tra **parole piene** (*full words*) e **parole forma** (*form words*). Le parole forma hanno un significato esclusivamente grammaticale che può essere stabilito solo in relazione con le altre parole.

Nell'analisi testuale si distingue più correntemente tra parole piene e **parole vuote**. La distinzione non è, come nel caso di Sweet, di carattere funzionale ma è strumentale ai fini della ricerca. Le parole vuote vengono definite di volta in volta come parole che non esprimono un contenuto interessante ai fini dell'analisi (e spesso sono parole grammaticali o di semplice legame nella frase), mentre le parole piene sono quelle che contribuiscono significativamente all'interpretazione del testo. In alcune analisi – come si vedrà in seguito – le parole grammaticali, di solito considerate come parole vuote, possono diventare molto importanti per l'interpretazione automatica del testo. Un testo che presenta un numero di occorrenze sopra la media nelle proposizioni *in* e *di* potrebbe essere un testo descrittivo; mentre la prevalenza di *ma* e *se* indicherebbe la presenza di elementi di incertezza (Bolasco, 1999, p. 193).

La frammentazione del testo è un problema di analisi che richiede delle decisioni meditate. Il problema non ha una soluzione univoca. Non si possono fornire indicazioni valide per tutti i casi. Una **frammentazione sintattica** (frasi delimitate da una punteggiatura che ha una rilevanza nella ricostruzione del senso), adeguata per un testo in prosa, può essere sostituita da una **frammentazione stilistica** (il verso di testo poetico), da una **frammentazione di senso** (un enunciato di senso compiuto ricavato in modo strumentale da una suddivisione del testo), oppure da una **frammentazione di comodo** (una riga

di testo). Ogni corpus necessita di una frammentazione del testo che sia adeguata alle ipotesi di lavoro sulla base delle quali è stato costruito.

2.
PREPARAZIONE E MODIFICA DEL CORPUS: TEXTPAD

Il modo più semplice e diretto di intraprendere un percorso didattico sull'analisi automatica dei testi è quello di esplorare un corpus con il sussidio di un software adeguato

Il corpus, che chiameremo convenzionalmente LEX è costituito dai seguenti testi:

- *Statuto del regno di Sardegna* (detto anche Statuto albertino), la costituzione concessa da Carlo Alberto nel 1848 e poi successivamente estesa al Regno d'Italia.
- *Costituzione della Repubblica romana del 1849*, promulgata il 1 luglio mentre le truppe francesi stavano per travolgere le difese dell'esercito di Giuseppe Garibaldi.
- *Costituzione della Repubblica italiana*, pubblicata sulla Gazzetta Ufficiale il 1° Gennaio 1948.
- *Dichiarazione Universale dei Diritti dell'Uomo*, adottata dall'Assemblea Generale delle Nazioni Unite il 10 Dicembre 1948.

Si tratta di un corpus (105 kb) di documenti politici fondamentali formato da quattro testi di genere empirico, conativi e continui. Sono stati scelti in quanto rappresentativi di un lessico politico di base. La loro appartenenza ad un corpus è dunque motivata dalla formulazione di una prima domanda: qual è il lessico caratteristico di questo corpus?

I testi sono stati scaricati dalla biblioteca elettronica del sito LIBER LIBER (http://www.liberliber.it), un'associazione culturale che cura il "progetto Manuzio" per la pubblicazione e la diffusione gratuita di opere letterarie

in formato elettronico.

TextPad è un eccellente editor testuale, particolarmente utile per testi di grandi dimensioni. Non è un word processing come Microsoft Word, piuttosto è un sostituto avanzato di Notepad per Windows. TextPad permette il trattamento del testo in formato ASCII (*American Standard Code for Infomation Interchange*). Il codice ASCII è stato introdotto nel 1963 negli Stati Uniti e successivamente, nel 1988, è stato adottato come norma internazionale sotto il nome di ISO 646 (*International Standards Organisation*). E' un codice binario di caratteri a 7 bit e quindi permette di codificare 128 caratteri diversi (27^8): alfabeto, punteggiatura, simboli grafici, caratteri di controllo non stampabili utilizzati per la trasmissione delle informazioni e per far funzionare determinate periferiche. Poiché i computer impiegano 1 byte (una sequenza di 8 bit) per inviare un carattere, alcuni produttori di computer e di software hanno sviluppato delle estensioni del codice aggiungendo un ottavo bit, portando così i caratteri codificabili a 256 (28^8). Così il codice ASCII, che permetteva di codificare solo i testi in inglese (senza accenti e senza lettere alfabetiche speciali), ha assunto diverse varianti. Tra queste varianti la più diffusa è la tabella ISO-8859-1 detta anche ISO-LATIN o LATIN-1 che permette di codificare le principali lingue dell'Europa occidentale (http://www.bbsinc.com/iso8859.html)

I codici Windows, MS-DOS e IBM sono estensioni diverse del codice ASCII. Senza addentrarci troppo in questo argomento di tipo informatico, occorre tenere presente che la conversione tra un codice e l'altro e da un sistema operativo all'altro (da Macintosh a Windows, per esempio) comporta dei cambiamenti di carattere, specialmente nella conversione delle lettere accentate. Per esempio, il codice 233 in Macintosh restituisce *é* mentre lo stesso codice in ANSI Windows restituisce *Ú* (i caratteri sono richiamabili direttamente dalla tastiera numerica – BlocNum – digitando Alt + il numero decimale corrispondente). Pertanto è necessario porre la massima attenzione nel salvataggio dei file quando si passa da un sistema operativo all'altro.

TextPad effettua il salvataggio sempre in formato ASCII e permette una gestione semplice ed efficace delle principali funzioni di modifica del testo. Con file di grandi dimensioni è più veloce di Word e permette di compiere agevolmente anche le operazioni più sofisticate. Il software, prodotto da *Helios Software Solutions*, è scaricabile per il consueto periodo di prova dal sito (http://www.textpad.com/).

2. 1. FUNZIONI DI MODIFICA DEL TESTO

Senza addentrarci troppo nella gestione del programma vediamo solo alcune delle funzioni principali, quelle di maggiore utilità per la preparazione dei testi ai fini dell'analisi testuale. Non mi soffermo sulle funzioni di routine come *Taglia*, *Copia* e *Incolla* o la funzione di *Cambia Maiuscole/minuscole* che sono del tutto ovvie per qualsiasi editor di testi. Oltre alle funzioni principali, TextPad permette di compiere molte operazioni utili come la gestione di più testi con modifica contemporanea di tutti i testi attivi e il controllo ortografico in quindici lingue. Per chi desidera approfondire c'è un'ottima guida in linea.

Il testo sul quale lavorare deve sempre essere importato in formati ASCII (*plain tex*). In WinWord c'è una procedura di salvataggio apposita nel menu **File**, voce *Salva con nome*, opzione "Tipo File: Testo normale". In Internet Explorer e in altri browser è spesso sufficiente fare un "copia e incolla" passando attraverso gli "Appunti" di Windows.

2.1.1. Modifica del margine destro a colonna 80

Di solito un file in formato *plain text* si presenta con righe molto lunghe che oltrepassano il margine destro della finestra di lavoro. Le righe vanno a capo solo dopo l'interruzione di riga o il segno di paragrafo. Per la maggior parte dei software di analisi testuale questo non comporta problemi, anzi è utile perché spesso ci viene richiesto di inserire all'inizio di ogni riga (o paragrafo) dei marcatori per la frammentazione del testo (v. *2.2. Funzioni di sostituzione*).

Alcuni software di impostazione *old style* (ad esempio DTM – Data Text Mining) chiedono invece che il file di testo non oltrepassi la colonna 80. In questo caso:

1. Dal menu **Configura** scegliere la voce *Preferenze*.
2. Nel menu ad albero, fare clic sul "+" vicino a *Classi del documento*.
3. Selezionare la classe di documento "testo".
4. Spuntare *Vai a capo alla colonna numero...*
5. Inserire il numero di colonna 79 (è una misura prudenziale perché con la colonna 80 accade che un carattere alfabetico si ponga a colonna 81).
6. Fare clic su OK.
7. Dal menu **Modifica** fare clic sulla la voce *Seleziona tutto*.
8. Dal menu **Modifica** fare clic sulla la voce *Riformatta*.

Questa stessa operazione si può compiere – senza modificare il margine destro – con la seguente procedura:

1. Visualizzare la finestra di lavoro del testo in una dimensione adeguata (spostando con il cursore il margine sinistro della finestra).
2. Cliccare sul pulsante "a capo" della barra strumenti.
3. Selezionare il testo da dividere ("seleziona tutto")
4. Dal menu **Modifica**, scegliere la voce *Dividi righe disposte a capo*.

2.1.2. Unire le righe

L'unione delle righe è l'operazione opposta alla precedente che mira invece a suddividere le righe ponendo il margine a una colonna predefinita. L'unione delle righe può servire a modificare un testo in modo che abbia a colonna 1 tutti gli inizi di paragrafo o di qualsiasi altra partizione del testo. Per unire più righe alla riga corrente:
1. Selezionare tutte le righe che devono essere unite.
2. Scegliere *Unisci righe* dal menu **Modifica**.

2.1.3. Selezionare un blocco verticale di testo

In molti casi è necessario selezionare un blocco verticale di testo (anziché selezionarlo in modalità normale (per righe orizzontali).
1. Assicurarsi che nel menu **Configura** la voce *A capo automatico* sia disattivata.
2. Scegliere la voce *Selezione blocco* dal menu **Configura**.
3. Per ritornare alla selezione normale del testo deselezionare la voce *Selezione blocco* .

La funzione "blocco verticale" rimane attiva fino a quando non viene disattivata dal menu **Configura** deselezionando la voce corrispondente. La selezione del testo avviene normalmente tramite mouse oppure tramite tastiera (shift+freccia). Questa funzione è preziosa quando si devono effettuare operazioni di cerca/sostituisci su campi di testo in larghezza fissa, come nelle matrici di dati e nelle tabelle non formattate.

2.1.4. Visualizzare spazi e tabulazioni

Una funzione molto utile nelle operazioni di modifica del testo è la visualizzazione dei caratteri che generano spazi vuoti, tabulazioni ed interruzione di riga.

Nel menu **Visualizza** abilitare il comando *Spazi visibili*. Nel testo saranno

visualizzati i seguenti simboli grafici:
· per il carattere spazio;
» per il carattere tabulazione;
¶ per il carattere interruzione di riga.
La stessa operazione può essere attivata o disattivata cliccando il simbolo ¶ sulla barra degli strumenti.

2.1.5. Visualizzare i numeri di riga

La visualizzazione dei numeri di riga è molto utile durante la modifica del testo, le operazioni di cerca e sostituisci e, in generale, per l'esplorazione del testo. Il comando *Numeri di riga* nel menu **Visualizza** mostra o nasconde il numero di riga nel documento attivo.

Nella sezione successiva vedremo come inserire una numerazione progressiva all'inizio di ogni riga.

2. 2. FUNZIONI DI SOSTITUZIONE DEL TESTO

Le funzioni di sostituzione nel testo sono fondamentali, non solo per la correzione e pulitura del testo, ma anche per l'inserimento dei marcatori di partizione specificatamente richiesti da alcuni software. Non entro nel merito delle operazioni di sostituzione che si possono considerare già note ed acquisite per qualsiasi utente di word processing. Mi soffermo solo su qualche esempio di maggiore utilità. Anche in questo caso per sfruttare pienamente le potenzialità del software occorre consultare il manuale in linea.

2.2.1. La sostituzione di una stringa di testo utilizzando le espressioni regolari

Dal menu **Cerca**, selezionando *Sostituisci* si accede alla finestra con le funzioni principali del comando. Le opzioni di controllo del comando *Sostituisci* sono organizzate in "Condizioni" e "Applicazione".

Tra le opzioni "Condizioni":

- *Testo* indica che la stringa di ricerca è una stringa di testo.
- *Hex* specifica che la stringa di ricerca è in esadecimale (questa opzione è utile soprattutto per chi lavora in un linguaggio di programmazione).

- *Parola intera* cerca le corrispondenze del testo come una parola intera.
- *Maiuscole/minuscole* cerca il testo con le stesse lettere maiuscole o minuscole (*case sensitive*).
- *Espressione regolare* specifica che la stringa di ricerca è un'espressione regolare.

Tra le opzioni "Applicazione":

- *Documento attivo* indica che il comando viene eseguito solo sul documento corrente.
- *Testo selezionato* indica che il comando viene eseguito solo sul testo selezionato.
- *Tutti i documenti* indica che il comando viene eseguito su tutti i documenti aperti (attivi).

La sostituzione di una stringa di testo è una delle funzioni fondamentali di questo comando ed è abbastanza nota da non richiedere molte specificazioni. Invece è di grande utilità l'opzione Espressione regolare per la sua discreta facilità d'uso e la sua duttilità.

Un'espressione regolare è una stringa di testo che utilizza dei caratteri speciali per individuare una sequenza di testo. Di default, TextPad utilizza le espressioni regolari dello standard UNIX esteso (altre scelte possono essere compiute dal menu Preferenze).

Il sommario delle espressioni regolari più utilizzate si trova nella Guida in linea. Nelle sezioni che seguono possiamo solo fare alcuni esempi tra i più frequenti.

2.2.2. Eliminare le righe vuote

La riga vuota è una riga che contiene solo il carattere di interruzione di riga ¶ che nella sintassi delle espressioni regolari si indica con **\n**. Pertanto dopo aver marcato l'opzione *Espressione regolare* nella finestra del comando *Sostituisci*:

1. Nel campo "Trova" scrivere **\n\n**
2. Nel campo "Sostituisci con" scrivere **\n**
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.
4. Ripetere l'operazione fino a quando non viene più trovata l'espressione regolare indicata.

2.2.3. Sostituire un carattere di tabulazione con un altro carattere

1. Nel campo "Trova" scrivere **\t**
2. Nel campo "Sostituisci con" scrivere **\carattere da inserire**
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

E' possibile inserire qualsiasi carattere. Per inserire una spazio vuoto bisogna battere lo [spazio vuoto] nel campo "Sostituisci con".

2.2.4. Inserire l'interruzione di riga dopo un segno di punteggiatura

La necessità di suddividere il testo in frammenti dotati di senso compiuto rende particolarmente utile questa funzione. Occorre fare attenzione perché il segno di punteggiatura "punto" (**.**) nella sintassi delle espressioni regolari significa "ogni singolo carattere". Ad esempio l'espressione **d.l** trova le parole che contengono *dal*, *del*, *dil*, *dol*, *dul*. Per evitare questo occorre far precedere il carattere "punto" dal carattere "barra diagonale verso sinistra" (detto anche "slash retroverso"): ****.

1. Nel campo "Trova" scrivere **\.**
2. Nel campo "Sostituisci con" scrivere **\.\n**
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

Naturalmente è possibile ripetere l'operazione con gli altri caratteri di punteggiatura che identificano le proposizioni principali (? !) ricordando che anche il punto interrogativo deve essere preceduto da **\.**

2.2.5. Inserire un carattere all'inizio di ogni riga

Accade spesso di dover inserire un carattere (o più caratteri) come marcatore delle partizioni del testo all'inizio di una riga, assumendo che la riga sia una partizione del testo dotata di senso. Ad esempio, ogni riga potrebbe identificare una proposizione principale.

1. Nel campo "Trova" scrivere **^**
2. Nel campo "Sostituisci con" scrivere **§** seguito da [spazio vuoto] (o un altro carattere qualsiasi seguito da [spazio vuoto]).
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

L'inserimento dello [spazio vuoto] come separatore ovviamente ha la funzione di non aggiungere il carattere da inserire alla parola seguente creando una nuova forma grafica.

2.2.6. Inserire un carattere alla fine di ogni riga

La fine di una riga non è il carattere di interruzione di riga (¶), ma il carattere che precede il carattere di interruzione di una riga.
1. Nel campo "Trova" scrivere **$**
2. Nel campo "Sostituisci con" scrivere [spazio vuoto] seguito da **§#§** (o da un altro carattere qualsiasi o da una stringa di caratteri).
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

Lo [spazio vuoto] prima della stringa di caratteri ha la funzione di non aggiungere il carattere da inserire alla parola precedente.

2.2.7. Inserire una numerazione progressiva all'inizio di ogni riga

L'inserimento di una numerazione progressiva è utile soprattutto quando si vuole creare una corrispondenza tra i dati di due file (spesso si tratta di un file con dati numerici e di un file con dati testuali).
1. Nel campo "Trova" scrivere **^**
2. Nel campo "Sostituisci con" scrivere **\i** seguito da [spazio vuoto] se la riga deve cominciare da 1; oppure **\i(100)** se la riga deve partire da 100; ecc.
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

L'inserimento dello [spazio vuoto] come separatore ovviamente ha la funzione di non aggiungere il carattere da inserire alla parola seguente creando una nuova forma grafica.

2.2.8. Inserire una numerazione progressiva all'inizio di ogni riga seguita da una stringa di caratteri e da un'interruzione di riga

Questa è una modifica del testo che è di grande utilità per chi deve analizzare dati testuali con software come SPAD o DTM.
1. Nel campo "Trova" scrivere **^**
2. Nel campo "Sostituisci con" scrivere **----\i\n**
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

2.2.9. Cercare una qualsiasi parola alfabetica che inizia con una lettera maiuscola e inserire alla fine della parola una stringa di caratteri

L'espressione è piuttosto complessa ma dimostra molto bene le potenzialità di questa opzione.

1. Nel campo "Trova" scrivere **[A-Z][a-z]+**
2. Nel campo "Sostituisci con" scrivere **&_maiusc**
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

Dopo aver effettuato questa operazione tutte le parole con iniziale maiuscola avranno la forma: Re_maiusc, Deputati_maiusc, ecc. Va sottolineato il fatto che Art. diventerà Art_maiusc e non Art._maiusc.

2.2.10. Cercare una qualsiasi parola alfabetica che inizia con una lettera maiuscola e sostituirla con la stessa parola in maiuscolo

1. Nel campo "Trova" scrivere **[A-Z][a-z]+**
2. Nel campo "Sostituisci con" scrivere **\U&**
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

Dopo aver effettuato questa operazione tutte le parole con iniziale maiuscola avranno la forma: RE, DEPUTATI, ecc. La sintassi **\U** significa "carattere maiuscolo" (*upper case*), mentre **\L** significa "carattere minuscolo" (*lower case*).

2.2.11. Cercare un numero qualsiasi e sostituirlo con lo stesso numero seguito da un carattere di tabulazione

1. Nel campo "Trova" scrivere **[0-9]\>**
2. Nel campo "Sostituisci con" scrivere **&\t**
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

La sintassi **\>** indica la fine della forma grafica "numero". Una stringa **\<[0-9]** sarebbe interpretata dal comando come l'inizio della forma "numero".

2.2.12. Inserire gli "apici" all'inizio e alla fine di una parola

Nella gestione dei file numerici, specialmente quando si tratta di tabelle, accade di dover inserire gli "apici" (') all'inizio e alla fine delle parole che identificano le "etichette" delle modalità.

Per l'inserimento dell'apice prima delle parole quando queste si trovano all'inizio di ciascuna riga:

1. Nel campo "Trova" scrivere **^**
2. Nel campo "Sostituisci con" scrivere **'**
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

Per l'inserimento dell'apice alla fine delle parole:

1. Nel campo "Trova" scrivere **[a-z]\>**
2. Nel campo "Sostituisci con" scrivere **&'**
3. Cliccare su "Sostituisci tutto" per applicare il comando a tutto il documento.

Attenzione! la sintassi **[a-z]** indica un *range* di caratteri (in questo caso si tratta di caratteri dall'alfabeto in minuscolo). In molti casi le parole contengono caratteri speciali, come il trattino di congiunzione (-) e l'apostrofo ('), oppure caratteri alfabetici che non appartengono all'alfabeto (à, è, ì. ò, à, ù). E' sempre preferibile evitare la presenza di caratteri speciali tra gli apici anche perché l'apostrofo verrebbe riconosciuto come apice e manderebbe in errore il programma. La lineetta può essere sostituita dal segno di *underline* (_) che invece viene trattato come facente parte dei caratteri dell'alfabeto, mentre l'apostrofo può essere abolito. Ad esempio la forma *socio-economico* dovrà avere la grafia *socio_economico* e la forma *d'accordo* dovrà avere la grafia *d_accordo*. Per le parole che terminano con un carattere di vocale accentata sarà necessario procedere all'inserimento manuale dell'apice, se l'elenco delle parole non è troppo lungo, oppure effettuare una normale procedura di trova/sostituisci (trova *à* / sostituisci con *à'*).