Progetto di
"Teorie e Tecniche dei Nuovi Media"

Analisi letteraria di due testi inglesi del periodo vittoriano:

Alice nel paese delle meraviglie
&
Attraverso lo specchio

di Lewis Carroll

Docente
dott. P. Ceravolo

Studente
Mattia Cavenaghi
736856

A.A. 2009/10

Indice

1 Introduzione

Con questo elaborato si vogliono approfondire le tematiche dell'Information Retrieval (IR) presentate nel corso, introducendo l'Analisi Lessicale (AL) su testi narrative realizzata mediante un semplice software di analisi realizzato sfruttando le potenzialità di Microsoft Access.

Sebbene il fine ultimo delle due tipologie di analisi sia simile, condividendo alcune delle fasi di elaborazione dei documenti, è stata scelta l'AL poiché applicabile a campi differenti dall'ambito web (es. Neurologia, studi sull'apprendimento, studi linguistici, etc) e più attinenti agli studi umanistici e medici. Con l'AL si è voluto produrre non una serie di vettori di termini, ma un thesauro ossia una lista di parole senza definizione, che identifica il contenuto dei due testi, associandovi un grado di leggibilità, lavoro inseribile in un ipotetico contesto relativo agli studi sull'apprendimento.

Dopo aver dato una breve panoramica biografica sull'autore dei documenti in analisi, si sono riportati due brevi riassunti degli stessi (sezione 2); successivamente si sono riportate le descrizioni delle fasi di analisi e del software impiegato (sezione 3) per poi concludere con la discussione dei risultati ottenuti e delle relative conclusioni (sezioni 4 e 5).

2 I testi letterari di Lewis Carroll: Alice in Wonderland e Through the Looking-Glass

2.1 Perché questi libri?

La scelta è ricaduta su questi libri poiché è una lettura effettuata in concomitanza del corso, inoltre non essendoci una continuità ben definita ma simile nella trama degli stessi, come ci si aspettava si sono osservati interessanti sviluppi nell'analisi (sezione 3).

I testi acquisiti di *Alice in Wonderland* (AIW) e *Through the Looking-Glass* (TTLG), sono in formato "txt" liberamente scaricabili dal sito del progetto Gutenberg (`http://www.gutenberg.org/wiki/Main_Page`), si è scelto inoltre di utilizzare una trascrizione della versione originale ottocentesca in lingua inglese poiché le varie versioni in italiano, nelle prime fasi di analisi hanno dato luogo ad ambiguità del linguaggio e difficoltà di comprensione delle vicende narrative, cosa già di per se abbastanza impegnativa in due testi del *genere nonsense*[1].

2.2 L'autore

Lewis Carroll, pseudonimo di *Charles Lutwidge Dogdson*, nacque a Daresbury nel Cheshire nel 1832. Studiò a Rugby e dal 1851 a Oxford, come allievo del Christ Church College, dove rimase fino al 1881 come lettore di Matematica pura.

Nel 1861 venne ordinato diacono ma non prenderà mai gli ordini superiori. Di carattere timido e sensibile, fu amico e fotografo di alcune bambine: si ispirò a una di esse, *Alice Liddell*, figlia del decano del Christ Church e coautore del celebre dizionario greco-inglese Liddell-Scott, per scrivere ALICE NEL PAESE DELLE MERAVIGLIE (1865), opera molto amata nell'ambito della letteratura infantile inglese ma apprezzata anche dal pubblico adulto per i giochi logici e verbali. Il libro ebbe un seguito, *Attraverso lo specchio* (1871), che riprende i temi di

[1] La *letteratura nonsense*, sia poesia che prosa, si basa sull'equilibrio tra ordine e caos, tra senso compiuto e nonsense. Spesso presenta un mondo capovolto o alterato, ma è distinto dal fantasy. Presenta frequentemente, ma non sempre, una matrice umoristica, che nasce però da uno spunto diverso rispetto ad uno scherzo: il nonsense suscita l'ilarità perché non ha senso, mentre lo scherzo perché ha un senso particolare. Il nonsense è un genere parassita, che appare all'interno degli altri generi o tipi letterari, come i versi, le poesie, i romanzi, i racconti brevi, le canzoni, il giornalismo e le ricette. La correttezza formale è spesso bilanciata da un caos semantico o dai doppi significati (fonte: Wikipedia, 2010).

Alice con la variante che i personaggi, che nel primo libro sono carte da gioco, diventano pezzi degli scacchi.

Notevole è anche La caccia allo Snark (1876), apparentemente una buffa poesia nonsense ma che nasconde affascinanti chiavi di interpretazione simbolica. Nel 1874 fa uscire con il suo vero nome alcune opere di Matematica; da allora si immergerà sempre di più in studi di Logica e Matematica, di cui sono testimonianza opere come Euclide e i suoi rivali moderni (1879), Il gioco della logica (1887), Che cosa disse la tartaruga ad Achille (1894) e Logica simbolica (1896); scrive anche numerosi articoli sulla rappresentanza proporzionale.

Ammalatosi di bronchite, morì a Guildford nel Surrey il 14 gennaio del 1898.

2.3 Riassunto dei libri

2.3.1 Alice nel paese delle meraviglie

4 Maggio, Alice seduta all'aperto con sua sorella maggiore si sta quasi addormentando dalla noia, quando vede un Bianconiglio che guarda l'orologio e parlotta fra sé dicendo "E' tardi! E' tardi!". Fatto così curioso che decide di seguirlo giù per una tana di coniglio molto profonda, finendo sottoterra in una saletta con una porticina chiusa a chiave che conduce ad un bel giardino.

La bambina pur di entrare nel giardino mangia e beve cose che le fanno cambiare di dimensione, però non riesce ad attraversare la porticina, dopo l'ultima trasformazione che la porta a diventare enorme, piange di frustrazione e quando si rimpicciolisce, si ritrova a nuotare in un mare fatto delle sue stesse lacrime.

Nel mare, incontra molte creature, fra cui un Topo, le creature ed Alice riescono ad uscire dal mare e ad asciugarsi, ma Alice viene subito lasciata sola, decidendo così di incamminarsi verso una radura che la porta alla casa del Bianconiglio, dove trovando uno strano liquido ridiventa grande. Spaventa il coniglio ed i suoi vicini sebbene non riescano a scacciarla, riescono a farla ridivenire molto piccola, riuscendo a scappare.

Successivamente Alice incontra un Bruco che sta fumando la hookah, ma che riesce a far irritare la bambina chiedendole di recitare una poesia, cosa che al momento non le riesce. Il Bruco informa Alice che mangiare da una parte del fungo su cui è seduto la farà crescere, ma mangiare dall'altra parte la farà rimpicciolire: cosa interessante poiché essa sta ancora provando a diventare della misura giusta per raggiungere il giardino.

Arriva ad una casetta nel bosco, dove risiedono una Duchessa, il suo bambino bruttino, la sua Cuoca ribelle, e il suo Gatto del Cheshire. La cucina è piena di pepe e di piatti vengono tirati dalla cuoca alla Duchessa per la rabbia. Alice prova a salvare il bambino da tutto quel pandemonio, ma il bambino si trasforma in un porcellino, così deve lasciarlo andare.

A questo punto riappare il Gatto del Cheshire, sogghigna ad Alice, e le raccomanda di visitare il Cappellaio Matto o la Lepre Marzolina. Il Gatto del Cheshire appare e scompare all'improvviso. Finalmente, scompare gradualmente e rimane solo il suo sogghigno.

La protagonista va a casa della Lepre Marzolina, dove è in corso una festa, si siede a tavola con la Lepre, il Cappellaio, e il Ghiro ma li trova maleducati e se ne stanca presto, così se ne va.

Decide di attraversare una porta in un albero e si ritrova ancora una volta nella saletta con la porticina che porta al giardino, questa volta riesce ad entrare.

Nel giardino trova tre giardinieri che stanno dipingendo di rosso delle rose bianche poiché hanno paura che la Regina di Cuori tagli loro la testa, per aver trapiantato rose del colore sbagliato. Ad un tratto appare la Regina di Cuori con il suo seguito di carte da gioco, che invita Alice a giocare a croquet un gioco molto singolare.

Qui scopre che la Duchessa deve essere decapitata e nel frattempo la testa del Gatto del Cheshire appare sul campo da gioco causando un parapiglia. La Duchessa in persona viene chiamata dalla prigione a risolvere l'assunto ed inizia a parlare con Alice della morale di ogni cosa.

La Regina decide allora che Alice deve andare a trovare la Finta Tartaruga, scortata dal Grifone dove ne assiste alla storia ed assiste ad un ballo chiamato Quadriglia dei Gamberi. Alice riprova a recitare una poesia con ben poco successo ma viene richiamata in tribunale, dove sta iniziando il processo.

Il Fante di Cuori è accusato di aver rubato le torte della Regina. Per Alice è molto eccitante essere in tribunale ed ascoltare la testimonianza del Cappellaio e della Cuoca, essa stessa è chiamata a testimoniare dopo essere di nuovo inesplicabilmente cresciuta.

La bambina si rivela impertinente ed il Re le ordina di lasciare il tribunale, ma lei rifiuta. É scandalizzata dall'ingiustizia del procedere della corte e provoca la Regina, che ordina anche la sua esecuzione. Alice dice alla corte ed ai giurati che non sono nient'altro che un mazzo di carte, ed essi si alzano per attaccarla.

A questo punto, Alice si rende conto di aver dormito per un bel po' di tempo in grembo a sua sorella, le racconta del suo sogno meraviglioso e poi rientra per

il the. La sorella è rapita dal sogno ed immagina Alice da grande, conservando intatto il suo senso infantile del meraviglioso.

2.3.2 Attraverso lo specchio

E' il 4 Luglio, Alice sta giocando con i suoi gattini, uno bianco chiamato Bucaneve ed uno nero chiamato Kitty, quando si chiede come sia il mondo dall'altra parte dello specchio. Sale quindi sul camino e si affaccia allo specchio appeso, scoprendo che vi è un altro mondo. In questa versione riflessa del mondo scopre un libro il "Jabberwocky" leggibile solo tramite lo specchio a causa della sua scrittura capovolta. La ragazzina scopre inoltre che i pezzi degli scacchi sono vivi finché rimangono piccoli abbastanza da essere presi in mano.

Alice lascia la casa in una notte fredda e nevosa, entrando in un assolato giardino primaverile, dove i fiori hanno la capacità di parlare con le persone ed intrattengono la bambina su alcune bizzarrie del giardino. Procedendo nel giardino Alice incontra la Regina Rossa in dimensioni umane, la quale ha l'abilità di correre molto velocemente, poiché negli scacchi il pezzo della regina si muove fino a sette caselle in qualsiasi direzione. La Regina Rossa rivela ad Alice che l'intera contea è un'enorme scacchiera, e le offre la possibilità di divenire a sua volta una regina se riesce ad arrivare nella ottava riga in una partita di scacchi; dopo essere stata schierata di fronte alla Regina bianca, la partita comincia ed Alice comincia a sua volta un viaggio in treno per tutta la scacchiera.

Durante il suo viaggio incontra per primi i fratelli Tweedledum e Tweedledee, che riconosce grazie ad una famosa filastrocca per bambini. Dopo aver recitato il poema "The Walrus and the Carpenter", i due gemelli fanno notare ad Alice la presenza del Re Rosso, addormentato sotto un albero, cosa che la porta ad essere coinvolta in una disputa semi-filosofica. Infine i Tweedle recitando una nuova filastrocca, si vestono per una battaglia ma vengono messi in fuga da un gigantesco corvo.

Alice proseguendo il suo cammino incontra la Regina Bianca, smemorata ma in grado di predire il futuro, entrambe avanzano poi lungo la scacchiera fino a che la Regina si trasforma in una pecora, la quale emettendo suoni senza senso comincia a dar noia ad Alice.

Attraversando un ruscello posto nella sesta fila della scacchiera, la ragazzina incontra Humpty Dumpty, che le da la sua interpretazione del termine "Jabberwocky" prima di cadere per terra. "Il re di tutti i cavalli e di tutti gli uomini" arriva ad aiutare Humpty Dumpty accompagnato dal Leone e l'Unicorno

e recitando una filastrocca, nel frattempo i due animali si danno battaglia. In questo capitolo il Leprotto Marzolino ed il Cappellaio Matto fanno una breve riapparizione nei panni dei "messaggeri Anglo-Sassoni" chiamati "Haigha" ed "Hatta".

Lasciando il Leone e l'Unicorno a combattersi, Alice raggiunge la settima linea ed attraversando l'ennesimo ruscello entra nel territorio del Cavaliere Rosso, intento a catturare il Pedone Bianco (Alice), ma viene salvata del Cavaliere Bianco. Il gentiluomo scorta la donzella attraverso la foresta e recitando un lungo poema, ma cadendo ripetutamente da cavallo (a causa del suo movimento ad L negli scacchi, simile ad un balzo).

Accomiatandosi dal Cavaliere Bianco, Alice attraversa l'ultimo ruscello e viene automaticamente incoronata Regina. Trovando la compagnia delle due Regine, le viene dedicata una nuova partita a scacchi, che si rivela un parapiglia in cui Alice afferra la Regina Rossa credendola responsabile del suo nonsenso, cosa che la porta a risvegliarsi in un armadio tenendo in braccio il suo gattino nero.

La storia si conclude ricordando le parole dei fratelli Tweedle, dove tutto è un sogno del Re Rosso, compreso Alice. Il poema finale è un omaggio dell'autore, il quale considera la vita anch'esso un sogno.

3 L'analisi lessicale

3.1 Le fasi

Avvalendoci dell'articolo tratto da `http://www.funsci.com/fun3_en/lexicon/handbook.htm` distinguiamo le seguenti fasi tramite cui si sono analizzati i documenti, fasi che si sono rivelate ricorsive e senza un ordine preciso:

- *normalizzazione*: consiste nell'inserimento del testo del documento in una tabella, eliminandone tutti i caratteri non alfabetici e convertendo le lettere maiuscole in minuscole. Questa operazione è successiva alla normalizzazione manuale dei testi recuperati, in particolare per il testo di AIW si è reso necessario inserire il breve componimento poetico introduttivo;
- *calcolo delle frequenze e delle ricorrenze:* dopo la normalizzazione del documento si calcola il numero di volte che un determinato termine t ricorre nel testo x e la sua frequenza all'interno dello stesso, osservando che:

$$\text{Ricorrenza (R)} = \#t(x)$$

$$\text{Frequenza (F)} = \frac{\#t(x)}{\#p(x)}$$

$$\text{Ricchezza lessicale (RL)} = \frac{\sum t(x)}{\#p(x)}$$

- *operazioni logiche (tra due documenti):* le seguenti operazioni consentono di operare su due documenti normalizzati, quindi in formato tabellare, producendo una tabella contenente i dati risultanti:
  - A - B: operazione di sottrazione dal documento normalizzato A di tutti i termini contenuti nel documento normalizzato B:
    - dati due testi (es. testo moderno A ed uno ottocentesco B) possiamo capire quali sono i termini antichi e moderni, quali sono i termini decaduti nella letteratura contemporanea, etc...;
    - dati due testi (es. un romanzo A ed un testo scientifico B) possiamo evidenziare i termini specialistici;
    - dati due testi (es. un componimento poetico dialettale A ed un componimento poetico in italiano B) possiamo estrarre i termini propri e caratteristici del documento A.

  - $A \times B$: operazione di prodotto cartesiano tra i due documenti normalizzati A e B ottenendo una tabella contenente tutti i termini comuni:
    * dati due testi di autori differenti (es. Alessandro Manzoni e Luigi Pirandello) otteniamo i termini di impiego comune rispetto ai due stili di scrittura;
    * dati due testi viene calcolato il *rapporto delle frequenze*, se tale valore si avvicina ad 1 significa che i termini comuni si presentano con uguale frequenza in entrambi i documenti:

$$\text{Rapporto delle frequenze (RF)} = \frac{F(t(A))}{F(t(B))}$$

  - $A + B$: operazione di somma di tutti i termini contenuti nel documento normalizzato A e B, producendo un thesauro ossia una lista di parole senza definizioni.

- *individuazione delle locuzioni:* questa operazione viene eseguita sui documenti normalizzati ed a seconda nel numero di parole minime che compone una locuzione, consente di ottenere la frequenza ed il numero di ricorrenze con cui queste si presentano nel documento. Le locuzioni individuate permettono di analizzare lo stile di scrittura di un autore, nel nostro caso sarà utile mantenere i termini grammaticali e ci consentirà di definire il thesauro di identificazione dei due testi;

- *calcolo dell'Indice di Leggibilità (IL):* la lingua inglese, a differenza di quella italiana, dispone di parole diverse per distinguere la leggibilità della calligrafia o del carattere tipografico (*legibility*) dalla scorrevolezza della lettura in funzione della struttura linguistica (*readability*). Lo stesso testo può essere *legible* ma non *readable*. Il software adottato utilizza una formula IL non documentata nella letteratura, se adottassimo questo strumento di analisi rischieremmo di compromettere i risultati del lavoro, motivo per cui in questa fase si farà ricorso a software disponibili on-line che utilizzano il *Flesh Reading Ease* il quale indica il grado di difficoltà che si riscontra nella lettura di un testo anglosassone in scala 0-100, più il valore numerico calcolato è alto e più il documento è di facile comprensione. La formula applicata è:

$$206.876 - 1.015\left(\frac{\#p(x)}{\#s(x)}\right) - 84.6\left(\frac{\#sill(x)}{\#p(x)}\right)$$

dove:

- *#p(x):* è il numero di parole che compone il testo;
- *#s(x):* è il numero di frasi che compone il testo;
- *#sill(x):* è il numero di sillabe che compone il testo.

3.2 Il software di analisi

Il programma adottato nel presente elaborato, fa parte di una serie di progetti di ambito amatoriale, incentrati sullo studio scientifico. Realizzato mediante Microsoft Access è quindi costituito tabelle e maschere che sfruttano il modulo *Lexicon*, con cui l'utente può interagire e le cui funzionalità sono descritte di seguito.

Maschere: sono interfacce grafiche tramite cui l'utente può operare sui dati contenuti nelle tabelle:

- *Normalize:* normalizza il testo di un documento e lo inserisce in una tabella;
- *Frequencies:* calcola le ricorrenze e le frequenze dei termini contenuti in una tabella normalizzata;
- *A - B:* ricava i termini presenti nella tabella A e che non sono presenti nella tabella B;
- *A X B:* ricava i termini comuni alle tabelle A e B, calcolando il rapporto delle frequenze;
- *A + B:* somma i termini delle tabelle A e B ricalcolandone le frequenze;
- *A => Thesaurus:* aggiunge i termini della tabella A al thesauro, ricalcolandone le frequenze;
- *A <= Restore:* rimuove i termini presenti nel documento A dal thesauro;
- *Text - GrammEn:* rimuove dalla tabella normalizzata associata al testo da analizzare, tutti i termini grammaticali presenti nella tabella GrammEn;

- *Locutions:* crea una tabella contenente frasi costituite da n parole, il valore di default di n è 2;
- *Readability:* determina l'indice di leggibilità di un testo (non utilizzato nella nostra analisi);
- *Sentences:* produce una tabella di periodi contenuti nel testo (non utilizzato nella nostra analisi);
- *Paragraphs:* produce una tabella di paragrafi contenuti nel testo (non utilizzato nella nostra analisi);
- *Punctuation:* produce una tabella contenente i segni di punteggiatura utilizzati (non utilizzato nella nostra analisi).

Tabelle: contengono i dati prodotti dal modulo Lexicon, le tabelle con suffisso "ZZ_" sono di sistema, non direttamente modificabili dall'utente ma necessarie al funzionamento del programma di analisi:

- *Name:* tabella di tutte le parole;
- *Name_freq:* tabella delle frequenze;
- *NameA-NameB_freq:* tabella delle frequenze di tutti i termini presenti in A ma non in B (risultato della maschera *A - B*);
- *NameAxNameB_rats:* tabella dei termini comuni ad A e B (rapporto di frequenze, risultato della maschera *A X B*);
- *NameA+NameB_freq:* tabella delle frequenze dei termini somma tra A e B (risultato della maschera *A + B*);
- *GrammEn:* tabella contenente i termini grammaticali della lingua inglese;
- *Name-G:* tabella di tutte le parole del documento in analisi, a cui sono sottratti i termini grammaticali (risultato della maschera *Text - GrammEn*);
- *Name_n:* tabella contenente le parole prese in sequenze di tre elementi (risultato della maschera *Locutions*);
- *Th_Name-freq:* thesauro dei termini e loro frequenze;
- *Th_Name_list:* lista dei documenti contenuti nel thesauro;

- *Statistics:* tabella prodotta dalla maschera Readability (non utilizzato nella nostra analisi).

3.3 Gli indici di Readability

Per l'analisi della Readability si utilizza un software web di analisi automatica, presente all'indirizzo `http://www.read-able.com/`. Questo strumento consente di analizzare pagine web e testi in formato "txt" utili al nostro scopo, ritornando quattro indici:

- *Flesch Kincaid Reading Ease (FKRE);*
- *Flesch Kincaid Grade Level (FKGE o FKRA Flesch Kincaid Reading Age)*: questo indice viene utilizzato nel campo degli studi educativi. L'FKGL converte l'FKRE in un punteggio riferito al grado di istruzione conseguibile negli Stati Uniti, consentendo ad insegnati, genitori ed istituzioni una più facile l'identificazione della readability di un testo;
- *Gunning Fog Score (GFS):* è un indice di readability di testi in lingua anglo-sassone. Il risultato è una stima del numero di anni di educazione scolastica necessari affinché un individuo capisca ad una prima lettura, il contenuto di un documento;
- *SMOG Index (SI):* questo indice di readability è simile al precedente indice (GFS), consente inoltre di verificare la validità di un particolare messaggio;
- *Coleman Liau Index (CLI):* analogamente al KFGL, al GFS ed al ARI, consente di misurare il grado di readability di un testo scritto in lingua anglo-sassone: l'unica eccezione consiste nel valutare il numero di caratteri piuttosto che il numero di sillabe che compone il testo in esame;
- *Automated Readability Index:* simile al CLI, consente anch'esso di misurare il grado di readability di un testo anglo-sassone.

La nostra analisi terrà conto solamente del FKRE, poiché esprime in una scala che va da 0 (difficile) a 100 (facile) il livello di difficoltà che una persona può incontrare nella lettura dei un testo anglo-sassone.

4 L'analisi dei risultati

Di seguito vengono riportate le tabelle contenti i dati statistici relativi ai due testi analizzati, per praticità di lettura e scrittura con la lettera A ci si riferisce ad *Alice in Wonderland*, mentre con la lettera B ci si riferisce a *Through the Looking-Glass*.

Testo	*n. parole*	*n. termini*	*RL*
A	27487	2608	0.095
B	30549	2731	0.089

Tabella 1: calcolo della Ricchezza Lessicale dei due testi completi.

Grazie alla tabella 1 possiamo in prima analisi dedurre che il testo A è qualitativamente più ricco e contiene più informazioni rispetto B, l'RL ha in entrambi i casi un valore troppo basso per essere significativo nella nostra analisi, proviamo quindi a diminuire il numero di parole prese in esame, calcolare nuovamente l'RL ed aumentare gradatamente il numero di parole prese in esame.

Funzioni/n. parole	*2000*	*4000*	*6000*	*8000*
t(A)	636	912	1184	1353
t(B)	535	874	1110	1293
RL(A)	0.318	0.228	0.197	0.169
RL(B)	0.267	0.218	0.185	0.161
Funzioni/n. parole	*10000*	*12000*	*14000*	*16000*
t(A)	1533	1698	1850	1964
t(B)	1464	1683	1799	1929
RL(A)	0.153	0.141	0.132	0.122
RL(B)	0.146	0.140	0.128	0.120
Funzioni/n. parole	*18000*	*20000*	*22000*	*24000*
t(A)	2107	2200	2320	2433
t(B)	2060	2173	2289	2364
RL(A)	0.117	0.110	0.105	0.101
RL(B)	0.114	0.108	0.104	0.098
Funzioni/n. parole	*26000*	*28000*	*30000*	*32000*
t(A)	2527	2608	-	-
t(B)	2510	2593	2699	2731
RL(A)	0.097	0.095	-	-
RL(B)	0.096	0.092	0.089	0.089

Tabella 2: calcolo della Ricchezza Lessicale per numero di parole crescente.

Dalla precedente tabella 2 ricaviamo il seguente grafico 1; possiamo osservare che i due testi risultano essere simili per quanto riguarda la RL, va osservato però che il testo A risulta inizialmente più ricco rispetto al testo B, poiché contiene un maggior numero di termini.

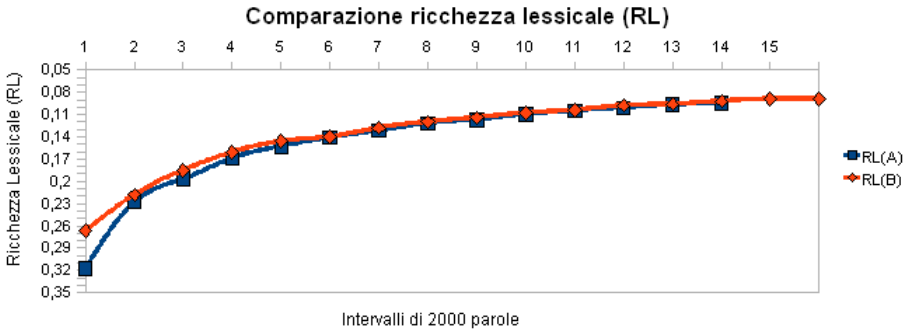

Figura 1: grafico comparativo della RL dei due testi analizzati.

Grazie alle operazioni logiche osserveremo diversi aspetti relativi ai due documenti consentendoci di fare alcune considerazioni:

- costruendo il thesauro dei due testi (A + B) e successivamente sottraendovi il thesauro dei termini grammaticali (Gr) otteniamo i termini ricorrenti in entrambi i racconti, con le relative frequenze (tabella 3, righe 1-3). Da questi dati possiamo dedurre che i termini i quali identificano i personaggi principali dei due racconti sono Alice, la Regina ed il Re.

#	*Termine*	*R(A + B - Gr)*	*F(A + B - Gr)*
1	alice	851	$1.47 \cdot 10^{-2}$
2	queen	258	$4.45 \cdot 10^{-3}$
3	king	129	$2.22 \cdot 10^{-3}$
#	*Termine*	*R(A - Gr)*	*F(A - Gr)*
4	alice	397	$5.56 \cdot 10^{-2}$
5	queen	74	$1.04 \cdot 10^{-2}$
6	king	63	$8.83 \cdot 10^{-3}$
#	*Termine*	*R(B - Gr)*	*F(B - Gr)*
7	alice	454	$5.57 \cdot 10^{-2}$
8	queen	184	$2.26 \cdot 10^{-2}$
9	king	66	$8.10 \cdot 10^{-3}$

Tabella 3: tabella contenente le ricorrenze e le frequenze con cui i termini descritti si presentano nei due testi letterari.

- nel testo narrativo *B* il termine *queen* si riferisce indiscriminatamente sia alla Regina Bianca che Rossa (tabella 3, riga 8), mediante un software di analisi automatica come possiamo identificare le due entità? Inoltre come possiamo identificare i personaggi coinvolti nelle due vicende?

  - Togliendo dai testi normalizzati tutti i termini grammaticali, possiamo ottenere tutti i termini chiave in essi contenuti: studiandone la frequenza otteniamo le parole di maggior rilievo, se da questa tabella eliminiamo i termini comuni ai due testi otteniamo indicazioni sugli altri personaggi o comunque sugli elementi caratteristici dei singoli testi aventi ricorrenza maggiore od uguale a dieci occorrenze (appendice A, tabelle 6 e 7).
  - abbiamo identificato i personaggi principali dei due libri ed i personaggi secondari, ma non siamo ancora riusciti ad identificare i personaggi "universalmente unici" come ad esempio la già citata Regina Bianca (White Queen) od il Vitello Tartaruga (Mock Turtle). Proviamo ad esaminare le locuzioni, ossia prendendo dal testo normalizzato tutti i gruppi di 3-parole (*L3(x)*) con le loro frequenze e vediamo cosa otteniamo:

(L3(A) - L3(B))	*R(L3(A) - L3(B))*	*F(L3(A) - L3(B))*
the_mock_turtle	52	$1.89 \cdot 10^{-3}$
the_march_hare	30	$1.09 \cdot 10^{-3}$
said_the_hatter	21	$7.64 \cdot 10^{-4}$
the_white_rabbit	21	$7.64 \cdot 10^{-4}$
said_the_mock	19	$6.91 \cdot 10^{-4}$
said_the_caterpillar	18	$6.55 \cdot 10^{-4}$
said_the_gryphon	17	$6.19 \cdot 10^{-4}$
said_the_duchess	15	$5.46 \cdot 10^{-4}$
said_the_cat	14	$5.09 \cdot 10^{-4}$

(a) locuzioni uniche riscontrate nel testo *A*.

(L3(B) - L3(A))	*R(L3(B) - L3(A))*	*F(L3(B) - L3(A))*
the_red_queen	54	$1.77 \cdot 10^{-3}$
the_white_queen	33	$1.08 \cdot 10^{-3}$
said_the_red	17	$5.57 \cdot 10^{-4}$
said_humpty_dumpty	14	$4.58 \cdot 10^{-4}$
the_knight_said	14	$4.58 \cdot 10^{-4}$
the_tiger_lily	11	$3.60 \cdot 10^{-4}$
alice_couldn_t	10	$3.27 \cdot 10^{-4}$

(b) locuzioni riscontrate nel testo *B*.

Tabella 4: Locuzioni di 3-parole estratte dai due testi narrativi.

... otteniamo proprio ALCUNI dei personaggi caratteristici dei testi.

Giunti a questo punto è possibile definire un thesauro dei nostri due documenti, costituito dai protagonisti e dai personaggi principali.

Come ultima fase del lavoro finora svolto si è voluto verificare l'Indice di Leggibilità dei testi, questa analisi ci permette di classificare i due documenti secondo indici standardizzati, in particolare il già citato indice Flesch Readaing Ease.

	Grade	
Readability Formula	*A*	*B*
Flesch Kincaid Reading Ease	88.7	91.7
Flesch Kincaid Grade Level	5.1	4.2
Gunning Fog Score	7.8	6.9
SMOG Index	4.3	4
Coleman Liau Index	6.9	6.8
Automated Readability Index	5.3	4.2

Tabella 5: tabella riassuntiva contenente gli indici di leggibilità.

5 Conclusioni

L'analisi effettuata ha consentito di individuare in modo automatico le entità o meglio i personaggi oggetto delle avventure narrate nei due documenti, sfruttando un'analisi statistica, molto rudimentale, sulle parole contenute nei due testi.

Tale attività ha portato alla creazione di un thesauro ossia una lista di termini senza definizioni, che identifica gli argomenti, o meglio le parole chiave dei racconti. Come ultima fase abbiamo calcolato l'indice di leggibilità dei documenti, ed avvalendoci dell'FKRE ne abbiamo scoperto la complessità: tale indice ci rivela infatti che sebbene siano di genere nonsense, sono molto semplici da comprendere anche per uno studente del 4° o 5° grado, di 9-10 anni.

Unendo quindi il thesauro con i termini chiave e l'indice calcolato, possiamo quindi classificare i nostri documenti secondo :

- in termini economici: a seconda della qualità di scrittura, adattamento o traduzione, i testi elaborati col sistema adottato permettono di di osservare la fascia di possibili acquirenti di una certa risorsa;
- in termini educativi: l'analisi svolta fornisce un aiuto ad educatori, genitori e tutte quelle persone che devono selezionare il materiale di studio per gli studenti;
- in termini tecnologici: il programma utilizzato, pur essendo molto semplice e presentando alcune deficienze (es. il dizionario dei termini grammaticali), non è di difficile implementazione, inoltre per l'analisi degli indici si è adottato uno dei tanti software on-line disponibili. L'utente finale che volesse replicare il lavoro svolto non deve necessariamente accedere a strumenti software e nozioni di complessità superiore (es. Formal Concept Analysis).

L'analisi presenta però alcuni limiti:

- definizione dell'obiettivo: come tutti i tipi di analisi, l'utente finale deve mantenere bene a mente l'obiettivo prefisso;
- definizione dei thesauri: in conseguenza al primo punto, i dizionari grammaticali devono essere definiti il più precisamente possibile, in caso contrario si verifica un sovraccarico (es. locuzioni contenenti aggettivi) od una perdita (es. locuzioni senza aggettivi) di definizione dei termini ricercati;

- il genere letterario: l'analisi di un genere letterario (il nonsense ad esempio) può essere di non facile applicazione, basti pensare ai termini-concetti-personaggi come il Vitello simil-Tartaruga/Mock Turtle/Finto Vitello;
- limitato automatismo: essendo l'analisi statistica, utilizza il calcolo di ricorrenze e frequenze dei termini, fattore che incide molto nei casi in cui si abbiano parti di documento con elevato contenuto informativo, ma limitata "visibilità" (es. i poemetti);
- incapacità di dare una definizione semantica al contenuto del testo: ossia noi possiamo sapere cosa il documento contiene, ma non il suo significato, o per lo meno l'interpretazione che ne da l'autore.

A Tabelle dei dati

t(A) - t(B)	*R(t(A) - t(B))*	*F(t(A) - t(B))*
turtle	58	$8.13 \cdot 10^{-3}$
hatter	56	$7.85 \cdot 10^{-3}$
gryphon	55	$7.71 \cdot 10^{-3}$
mock	55	$7.71 \cdot 10^{-3}$
rabbit	49	$6.87 \cdot 10^{-3}$
duchess	42	$5.89 \cdot 10^{-3}$
dormouse	40	$5.61 \cdot 10^{-3}$
march	34	$4.76 \cdot 10^{-3}$
hare	31	$4.34 \cdot 10^{-3}$
caterpillar	27	$3.78 \cdot 10^{-}$
jury	22	$3.08 \cdot 10^{-3}$
court	18	$2.52 \cdot 10^{-3}$
bill	16	$2.24 \cdot 10^{-3}$
footman	14	$1.96 \cdot 10^{-3}$
mad	14	$1.96 \cdot 10^{-3}$
grow	13	$1.82 \cdot 10^{-3}$
dodo	13	$1.82 \cdot 10^{-3}$
gloves	11	$1.54 \cdot 10^{-3}$
pool	10	$1.40 \cdot 10^{-3}$
witness	10	$1.40 \cdot 10^{-3}$

Tabella 6: analisi delle ricorrenze e delle frequenze dei termini caratteristici presenti nel testo A.

t(B) - t(A)	*R(t(B) - t(A))*	*F(t(B) - t(A))*
knight	57	$6.99 \cdot 10^{-3}$
dumpty	52	$6.38 \cdot 10^{-3}$
humpty	52	$6.38 \cdot 10^{-3}$
tweedledum	33	$4.05 \cdot 10^{-3}$
kitty	25	$3.07 \cdot 10^{-3}$
tweedledee	25	$3.07 \cdot 10^{-3}$
kitten	24	$2.94 \cdot 10^{-3}$
unicorn	21	$2.58 \cdot 10^{-3}$
gnat	18	$2.21 \cdot 10^{-3}$
lion	17	$2.09 \cdot 10^{-3}$
lily	16	$1.96 \cdot 10^{-3}$
pudding	15	$1.84 \cdot 10^{-3}$
messenger	14	$1.72 \cdot 10^{-3}$
hill	14	$1.72 \cdot 10^{-3}$
carpenter	12	$1.47 \cdot 10^{-3}$
square	12	$1.47 \cdot 10^{-3}$
tiger	12	$1.47 \cdot 10^{-3}$
brook	12	$1.47 \cdot 10^{-3}$
road	12	$1.47 \cdot 10^{-3}$
rushes	11	$1.35 \cdot 10^{-3}$
boat	11	$1.35 \cdot 10^{-3}$
hatta	11	$1.35 \cdot 10^{-3}$
oysters	10	$1.23 \cdot 10^{-3}$
helmet	10	$1.23 \cdot 10^{-3}$
poetry	10	$1.23 \cdot 10^{-3}$
haigha	10	$1.23 \cdot 10^{-3}$

Tabella 7: analisi delle ricorrenze e delle frequenze dei termini caratteristici presenti nel testo *B*.

Riferimenti bibliografici

[1] Simona Balbi and Michelangelo Misuraca. Pesi e metriche nell'analisi dei dati testuali. 7, 2005.

[2] Lewis Carroll. *Alice's Adventures in Wonderland*. Project Gutenberg's, 2008.

[3] Lewis Carroll. *Through the Looking-Glass*. Project Gutenberg's, 2008.

[4] P. Ceravolo. Analisi del testo. In *Corso di Teorie e Tecniche dei Nuovi Media*, 2008.

[5] William B. Frakes and Ricardo Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall PTR, June 1992.

[6] C. Poli and G. Carboni. Lexical analysis of texts. *Fun Science Gallery*, 1998.

[7] AA. VV. Automated readability index. *Wikipedia*, 2010.

[8] AA. VV. Enciclopedia multimediale delle scienze filosofiche. 2010.

[9] AA. VV. Information retrieval. *Wikipedia*, 2010.

[10] AA. VV. Narratologia. *Wikipedia*, 2010.