UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Tecnologie linguistico-computazionali e analisi della leggibilità di testi per l'infanzia

Candidato: *Irene Valenzano*

Relatore: *Alessandro Lenci*

Anno Accademico 2011-2012

INDICE

INTRODUZIONE

"Riconoscere la difficoltà dello studente partendo da un livello di complessità di lettura più adatto a lui può permettere di incentivare la sua motivazione e il suo interesse, in modo da impostare un percorso di crescita che porti anche il più debole verso la lettura e relativa piena comprensione dei testi originali." (Lotti, Manuale utente FacilTesto, 2010)

Spesso, soprattutto in ambito scolastico, gli insegnanti notano una certa svogliatezza negli studenti che devono affrontare la lettura di testi, ma altrettanto spesso questo non è dovuto ad un indole pigra o demotivata quanto alla difficoltà di apprendimento dovuta ad un genere di scrittura non adatto.

Scrivere un documento leggibile risulta più complicato del previsto se lo scopo è, prima di tutto, quello di comunicare un determinato messaggio nella maniera corretta. Inoltre l'autore, dovendo far fronte all'attuale contesto socio-linguistico multietnico, ha difficoltà ulteriori nel proporre un testo che risulti comprensibile e chiaro. Per questo motivo la valutazione del grado di leggibilità di un documento ha riscosso sempre più interesse da parte di un vasto settore di studiosi.

Cominciamo l'approfondimento dell'argomento con una definizione: *l'indice di leggibilità* è generalmente una formula matematica usata per la predizione della facilità/difficoltà di lettura di un testo.

Come in tutti i campi sperimentali però, anche per quanto riguarda lo studio della leggibilità dei testi, si sono sviluppati diversi metodi di calcolo, tant'è che possiamo dire di essere di fronte a due generazioni di valutazione automatica della leggibilità basate su diverse caratteristiche.

Il progetto sulla valutazione della leggibilità, all'interno del quale si colloca questo lavoro, nasce dall'intuizione di partenza riguardante il "potere diagnostico" delle tecnologie linguistico-computazionali in compiti di monitoraggio linguistico (S.Montemagni & Dell'Orletta, 2011).

Grazie a questi sviluppi oggi è possibile valutare la leggibilità di testi a diversi livelli di descrizione linguistica: a livello lessicale, morfo-sinttattico e sintattico. Questi diversi livelli di analisi - superiori rispetto a quelli alla base del calcolo degli indici della generazione precedente - danno risultati sempre più affidabili e difficili da ottenere mediante l'analisi manuale, permettendo una valutazione più approfondita della complessità di un testo.

La mia relazione rappresenta, perciò, una raccolta di riflessioni e osservazioni riguardo le caratteristiche che rendono un testo un documento complesso da comprendere per un particolare tipo di pubblico, quale gli studenti di scuole elementari.

Ho focalizzato il lavoro in ambito scolastico, perché è caratterizzato dal graduale aumento di studenti immigrati, i quali dovrebbero ritrovare negli insegnamenti e nei

testi didattici un primo livello di integrazione con la società circostante. Non sono esclusi dal contesto generale, i molti studenti madrelingua, che mostrano difficoltà d'apprendimento a vari livelli nello studio delle stesse discipline scolastiche.

I nuovi metodi di valutazione della leggibilità potrebbero supportare l'insegnante nel difficile compito di scegliere il tipo di lettura adatto agli studenti, compito fino ad ora basato sulle sole capacità di intuizione. Allo stesso modo, si può pensare di facilitare chi scrive per questo particolare tipo di pubblico.

Nella valutazione della leggibilità, come vedremo, sono stati considerati anche altri generi di testi di lettura infantile, per cercare di inquadrare meglio il tema della ricerca.

CAPITOLO I : GLI INDICI DI LEGGIBILITÀ

Come detto in precedenza, si può parlare di due generazioni di calcolo della leggibilità, basate su caratteristiche diverse. Gli indici di prima generazione consistevano generalmente in una formula matematica che consentiva di calcolare automaticamente la difficoltà di un testo a livello lessicale, quindi andando a considerare la lunghezza di parole e frasi.

Gli indici di seconda generazione invece, grazie all'ausilio delle nuove tecnologie linguistiche permettono di approfondire il problema della complessità di un testo, dando la possibilità di analizzare il documento rispetto a diversi livelli linguistici.

1.1 INDICI DI PRIMA GENERAZIONE.

Il primo indice di leggibilità prendeva in considerazione come variabili:

- lunghezza delle parole in sillabe;
- lunghezza delle frasi in parole;

Questo è il caso dell' Indice di Flesch, messo a punto nel 1948 da Rudolf Flesch, tarato sulla lingua inglese. L'indice di Flesch consiste essenzialmente in un metodo di valutazione della leggibilità di un testo in lingua inglese calcolando il numero di sillabe contenute in 100 parole di un campione tratto dal brano in esame e la lunghezza media delle frasi del campione calcolata in numero di parole.

La formula di Flesch è la seguente:

$$F = 206 - (0{,}6 * S) - P$$

dove:

- F è la leggibilità misurata secondo questi parametri:
- S è il numero delle sillabe, calcolato su un campione di 100 parole;
- P il numero medio di parole per frase.

In seguito l'indice è stato rivisto per la lingua italiana da Roberto Vacca nel 1972 (indice Flesh-Vacca) mantenendo essenzialmente le stesse caratteristiche . Entrambi gli indici oscillano su una scala di valori compresa tra 0 e 100, dove 100 indica la leggibilità più alta e 0 quella più bassa.

Il secondo indice di leggibilità tarato sulla lingua italiana è Gulpease, definito nel 1988 nell'ambito delle ricerche del GULP (Gruppo Universitario Linguistico Pedagogico) presso il Seminario di Scienze dell'Educazione dell'Università degli studi di Roma "La Sapienza". L'indice nasce dal bisogno di aggirare il problema della sillabazione delle parole che poteva risultare ostica in alcuni casi, per questo ha il vantaggio di calcolare la lunghezza delle parole in lettere, anziché in sillabe.

Infatti l'indice Gulpease considera come variabili linguistiche:

- lunghezza delle parole in lettere (non in sillabe);
- il numero delle parole che compongono il testo;
- il numero delle frasi che compongono il testo;

La formula per il suo calcolo è la seguente:

$$G=89 - (Lp / 10) + (3 \times Fr)$$

dove:

- Lp = (100 × totale lettere) / totale parole
- Fr = (100 × totale frasi) / totale parole

I risultati dell'indice Gulpease oscillano sulla stessa scala di valori 0-100 di Flesch ma in più questo indice valuta la leggibilità di un testo anche rispetto al grado di scolarizzazione del lettore.

In particolare i risultati possono classificarsi come:

- risultato inferiore a 80 per testi difficili da leggere per chi possiede la licenza elementare;
- risultato inferiore a 60 per testi difficili da leggere per chi possiede la licenza media;
- risultato inferiore a 40 per testi difficili da leggere per chi possiede il diploma superiore;

Questi modelli di misurazione della leggibilità tendono ad esaurirsi in una formula matematica che stabilisce, attraverso un calcolo statistico, il grado di difficoltà di un testo.

Questi indici di prima generazione, calcolati sulla base di caratteristiche legate prettamente alla struttura superficiale del testo – le uniche calcolabili in modo automatico a quel tempo - presupponevano per esempio che a frasi lunghe corrispondesse una maggiore difficoltà grammaticale e viceversa, non dando informazioni aggiuntive su quali fossero gli specifici livelli di difficoltà del testo.

1.2 INDICI DI SECONDA GENERAZIONE

Il ricorso a tecnologie linguistico - computazionali avanzate per l'analisi e il monitoraggio linguistico rendono oggi possibili analisi sempre più affidabili e accurate, che coprono aspetti della struttura linguistica rimasti fino ad ora inesplorati in quanto difficilmente attingibili mediante un'analisi manuale del testo (S.Montemagni & Dell'Orletta, 2011)

Con la graduale maturazione di questi sistemi in grado di accedere al contenuto informativo dei testi cosi come alla loro struttura linguistica attraverso l'elaborazione automatica del linguaggio, si cerca di valutare la leggibilità di un testo basandosi su un ampio spettro di parametri che spaziano tra i diversi livelli di descrizione linguistica monitorati[1] in modo automatico.

Recentemente è stato sviluppato per l'italiano il primo software avanzato per in calcolo della leggibilità: READ-IT (F. Dell'Orletta, 2011). READ-IT valuta la leggibilità di un corpus attraverso l'analisi combinata di tratti tradizionali di un testo (tipici di Gulpease e Flesch) con caratteristiche lessicali, morfo-sintattiche e sintattiche.

READ-IT si applica a testi analizzati sintatticamente e assegna ad ogni testo considerato un punteggio che quantifica la sua leggibilità. Dati una serie di parametri e un corpus di riferimento, il software crea un modello statistico basato sulle caratteristiche statistiche estratte da tale corpus.

Questa generazione di indici tratta la misura della leggibilità come un compito di classificazione, basata sul tipo di corpora disponibili per la lingua italiana e sul particolare tipo di utente di riferimento. Tale *classificazione* oscilla su una scala di valori che stabilisce quanto un testo sia di " facile " o " difficile " lettura.

READ-IT effettua la sua valutazione di leggibilità rispetto a due corpus di riferimento come rappresentativi di varietà di lingua semplice e complessa.

La valutazione della leggibilità avviene in relazione al testo linguisticamente annotano, ovvero sottoposto a "tokenizzazione" - segmentazione del testo in parole ortografiche – "lemmatizzazione[2]" e "analisi morfo-sintattica" del testo tokenizzato, fino ad arrivare alla " analisi della struttura sintattica " della frase in termini di relazioni di dipendenza.

La tabella 1 mostra il risultato di una parte di testo annotato linguisticamente.

			Lemmatizzazione	**Annotazione Morfo-Sintattica**			**Annotazione a dipendenze**	
	Id	Forma	Lemma	Cat.gram.	Cat.gram 2	Tratti	Testa	Tipo
1	1	La	Il	R	RD	num=s|gen=f	2	Det
	2	Madre	Madre	S	S	num=s|gen=f	4	Subj

[1] Monitorare un testo vuol dire analizzare le caratteristiche linguistiche sottostanti all'indice sintetico di leggibilità ottenuto.

[2] La "lemmatizzazione" è un operazione automatica o manuale con cui si riconduce ciascuna occorrenza o parola di un testo al suo lemma fondamentale o entrata di dizionario, individuandone contemporaneamente una serie di informazioni (cat. Grammaticale,genere,numero, etc.)che variano secondo gli obiettivi prestabiliti.

	3	Però	Però	B	B	_	4	Mod
	4	Preferiva	Preferire	V	V	num=s|per=3|mod=i|ten=i	0	ROOT
	5	La	Il	R	RD	num=s|gen=f	6	Det
	6	Figlia	Figlio	S	S	num=s|gen=f	4	Obj
	7	Maggiore	Maggiore	A	A	num=s|gen=n	6	Mod
	8	,	,	F	FF	_	10	Punc
	9	La	Il	R	RD	num=s|gen=f	10	Det
	10	Viziava	Viziare	V	V	num=s|per=3|mod=i|ten=i	4	Mod
	11	E	E	C	CC	_	10	Con
	12	Costringeva	Costringere	V	V	num=s|per=3|mod=i|ten=i	10	Conj
	13	Rosetta	Rosetta	S	SP	_	12	Subj
	14	A	A	E	E	_	10	Arg
	15	Fare	Fare	V	V	mod=f	14	Prep
	16	Tutti	Tutto	T	T	num=p|gen=m	18	Mod
	17	I	Il	R	RD	num=p|gen=m	18	Det
	18	Lavori	Lavoro	S	S	num=p|gen=m	15	Obj
	19	Faticosi	Faticoso	A	A	num=p|gen=m	18	Mod
	20	.	.	F	FS	_	4	Punc

Tabella 1: Esempio di rappresentazione tabellare del testo annotato linguisticamente

Nella tabella 1 ad ogni riga corrisponde un'occorrenza[3] di forma di parola ("token") e ogni colonna specifica la proprietà di questa forma a diversi livelli di analisi. La prima colonna "Id" associa ad ogni forma un numero progressivo identificativo; la colonna Lemma esprime il relativo esponente lessicale della forma; nelle colonne di annotazione morfo-sintattica (Cat.gram/Cat.gram.2) si associa ad ogni forma l'informazione relativa alla categoria grammaticale di appartenenza e l'eventuale sottocategoria (V= verbo; R= articolo; S= sostantivo; RD= articolo definito). In più la colonna "Tratti" integra ulteriori informazioni morfologiche, quali le categorie flessionali come persona, genere, numero; le colonne dell'annotazione a dipendenze forniscono una descrizione della frase in termini di relazioni tra parole (es. "soggetto", "oggetto" ecc.). Per ogni parola la colonna "Testa" riporta l'identificatore univoco della forma ("Id") che costituisce la testa da cui dipende (0 per il verbo della proposizione principale, assunto come radice dell'albero sintattico), mentre la colonna "Tipo di relazione" esplicita il tipo di dipendenza.

[3] Il termine "occorrenza" indica ogni parola di un testo, computata ogni volta che compare all'interno del testo stesso.

Partire da un testo linguisticamente annotato, offre l'opportunità di nuove elaborazioni automatiche "…per l'identificazione di una vasta tipologia di parametri che possono essere ulteriormente sfruttati nel compito di monitoraggio linguistico." (Montemagni, 2011).

L'indice di leggibilità diventa, perciò, uno strumento di studio molto più articolato e motivato linguisticamente.

Per la mia ricerca quindi, ho utilizzato il software READ-IT messo a punto presso l'Istituto di Linguistica Computazionale " Antonio Zampolli"del CNR di Pisa, che basa la sua valutazione della leggibilità su una vasta tipologia di parametri:

- *parametri lessicali* quali: indice di ricchezza lessicale (rapporto tipo/unità); tipologia del vocabolario usato (lemmi appartenenti al Vocabolario di base,VdB); appartenenza ai repertori del VdB (Fondamentale, Alto Uso, Alta Disponibilità) etc.

- *parametri morfosintattici* quali: calcolo della densità lessicale; calcolo della percentuale di appartenenza a determinate categorie morfosintattiche; la distribuzione dei verbi in accordo con il modo, il tempo, il numero e la persona; etc.

- *parametri sintattici* quali: livello di incassamento gerarchico; rapporto tra clausole principali e subordinate; ordine relativo tra principale e subordinata; etc.

  Il risultato di READ-IT è costituito dall'indice Dylan che oscilla su una scala di valori che va da 0 a 1, registrando così la probabilità di un testo di appartenere alla classe dei testi selezionati come particolarmente complessi (più vicino allo 0=facile, più vicino a 1=complesso).

CAPITOLO II: DESCRIZIONE E FINALITÀ DEL PROGETTO

L'obbiettivo del mio progetto è stato verificare il grado di leggibilità di testi per l'infanzia mediante analisi condotte dal software READ-IT.

Il lavoro è stato organizzato in cinque fasi:

1. Calcolo dell'indice di leggibilità globale. (sezione 2.3)

2. Analisi dei parametri sottostanti la valutazione della leggibilità globale fornita dall'indice Dylan, in particolare parametri lessicali, morfo-sintattiti e sintattici. (sezione 2.4)

3. Analisi contrastiva dei risultati ottenuti (cap.3). I confronti utili all'analisi sono stati effettuati:
   - internamente al corpus di testi per l'infanzia selezionato.
   (sezione 3.1)
   - rispetto a corpora di riferimento della lingua italiana.
   (sezione 3.2)

4. Analisi specifiche di leggibilità su testi singoli. (cap. 4)

5. Confronto fra le due generazioni di indici per la valutazione della leggibilità: Gulpease vs Dylan. (cap.5)

2.1 I CORPORA USATI

Prima di illustrare le analisi fatte e i risultati ottenuti, approfondiamo le caratteristiche del corpus di lettura infantile usato per il mio progetto e dei corpora di riferimento esterni considerati per il confronto.

2.1.1 CORPUS DI LETTURA INFANTILE

La ricerca sul grado di leggibilità dei testi è stata effettuata sul corpus di testi alla base dell'opera "Lessico Elementare" (ed. Zanichelli, Bologna, 1994) messo a disposizione in formato elettronico dall'unità staccata di Genova dell'istituto di Linguistica Computazionale (ILC-CNR).

Il progetto del "Lessico Elementare" ha dato vita ad un vero e proprio dizionario contenente la lista dei lemmi più frequenti tratti dai testi scritti con lo scopo "…di creare un dizionario di frequenza aggiornato e completo che costituisse, da un lato, un saldo punto di partenza per ricerche linguistiche e dall'altro, uno strumento operativo di lavoro per chi si occupa di educazione di bambini e di lingua scritta."

(L. Marconi, 2010). Il "Lessico Elementare" è quindi un lessico di frequenza della lingua (italiana) scritta *dai* bambini e *per* i bambini.

La mia ricerca si è limitata allo studio del corpus di parole scritte *per* i bambini.

Vediamo i criteri di selezione dei testi:

- I libri di Lettura e Fiabe più diffusi sono stati identificati sulla base di una statistica nazionale non pubblicata[4] riferita al mese di Aprile 1988.

- Il panorama dei Giornalini e Fumetti è più variegato ma il campione è quello delle testate più diffuse nel periodo di riferimento 1987-1989. Fanno parte del campione solo quelle testate che hanno resistito nel corso degli anni, mantenendo una certa consistenza nella loro quota di mercato. Tale corpus di riferimento[5] è stato determinato in base alle statistiche ufficiali rilevate dall'ADS (Accertamenti Diffusione Stampa, 1988).

- I ricercatori si sono basati ancora su una statistica non ufficiale, per stilare la lista dei libri di Testo più adatti al progetto. Tale statistica fa riferimento all'anno 1987/88. Sono stati scelti i venti testi più venduti in Italia per le classi I e II elementare, mentre per le classi III, VI e V sono stati scelti i 10 testi scolastici di lettura e i dieci sussidiari più venduti.

La tabella 2 fornisce la caratterizzazione numerica delle tre sottocategorie del corpus di lettura infantile.

	Testo	**Lettura**	**Fumetti**
Numero di Frasi:	15.562	12.992	30.661
Numero di Token:	210.173	190.898	227.831

Tabella 2: Caratterizzazione numerica delle tre sottocategorie del corpus di lettura infantile

Il corpus sul quale ho lavorato è formato da circa 500.000 occorrenze ed è suddiviso in tre sottosettori: libri di Lettura e Fiabe (Lettura), Giornalini e Fumetti (Fumetti) e libri di Testo (Testo).

2.1.2 CORPORA DI RIFERIMENTO: 2PAROLE E REPUBBLICA

I corpora di riferimento della lingua italiana utilizzati per il confronto sono:

- la prosa giornalistica di *Repubblica* (Rep) caratterizzata da un alto livello di complessità linguistica.

[4] Fonte Demoskopea

[5] L'insieme dei testi sottoposti ad un analisi automatica viene comunemente definito col termine "corpus di riferimento"

- il periodico di “facile lettura” *2Parole* (2Par).

Al momento 2Parole risulta essere il solo corpus di testi semplificati disponibile, specificatamente indirizzati ad un pubblico particolare, caratterizzato da un basso livello di alfabetizzazione o con disabilità intellettuali.

I risultati dell’analisi di ognuna delle tre categorie di testi di lettura infantile sono stati messi a confronto con quelli dei corpora Repubblica e 2Parole per verificare in che modo e di quanto si avvicina il tipo di scrittura dei primi a uno o l’altro dei corpora giornalistici.

2.2 LEGGIBILITÀ DELLE DIVERSE SOTTOCATEGORIE

L’indice di leggibilità Dylan è stato calcolato automaticamente per ogni testo del corpus e complessivamente per ciascuna sottocategoria (Testo, Lettura, Fumetti).
Nella Tabella 3 troviamo i risultati ottenuti per ogni sottocategoria, ricordando che la scala di valori sui quali oscilla l’indice va da 0 a 1 (più vicino allo 0=facile, più vicino a 1=complesso).

Sottocategoria	Indice Dylan
Giornalini e Fumetti	0. 93604
Lettura e Fiabe	0. 98537
Libri di Testo	0. 02513

Tabella 3 : indice Dylan per ogni sottocategoria.

È interessante notare la bassa leggibilità del corpus di Giornalini e Fumetti. Per quanto il tipo di scrittura di tali testi possa essere breve e d’impatto, probabilmente agiscono sulla valutazione della leggibilità diversi ordini di fattori:

- La segmentazione delle frasi può risultare difficile da decifrare automaticamente.

- Testo e immagine - in questo caso più che in altri - sono strettamente collegati.

- Infine, nei testi fumettistici l’analisi diventa a tratti inattendibile per la frequente occorrenza di parole onomatopeiche, frasi ellittiche prive di verbo o di soggetto etc., complesse da annotare automaticamente a livello sintattico.

Per quel che riguarda la categoria della Lettura e Fiabe, il discorso non cambia in apparenza, ma cambia nella sostanza.
Il grado di leggibilità ottenuto è molto basso, quindi con un risultato prossimo a “1” come per la sottocategoria precedente, ma si suppone siano diversi – forse diametralmente opposti - i motivi di tale risultato.

Prima di tutto c'è un notevole cambiamento a livello sintattico avendo a che fare con frasi più lunghe e articolate, tipiche della narrazione. Di conseguenza abbiamo una maggiore complessità a livello sintattico, caratterizzata dalla presenza di più frasi subordinate, etc..

La categoria dei Libri di Testo ha presentato invece un grado di leggibilità globale molto alto, con indice Dylan prossimo a "0". A livello individuale, la situazione mostra esempi contrastanti. Per il momento resta interessante notare che globalmente i testi dai quali i bambini ricavano l'informazione necessaria al loro sviluppo linguistico e cognitivo, risultano essere consoni alle loro capacità di apprendimento.

2.3 RISULTATI DEL MONITORAGGIO DELLE CARATTERISTICHE LINGUISTICHE DEI DIVERSI SOTTOCATEGORIE

Cerchiamo ora di capire le motivazioni sottostanti ai risultati descritti sopra.
In particolare, passiamo ora a considerare singolarmente i parametri sui quali si basa il calcolo dell'indice Dylan.

2.3.1 PARAMETRI GENERALI DEL TESTO

Prima di tutto abbiamo una serie di parametri legati alla struttura del testo (trattati anche dagli indici di leggibilità di prima generazione), quali il Numero di token per frase e Numero di caratteri per token (Tabella 4).

	Testo	**Lettura**	**Fumetti**
Numero di token per frase	13,5	15,24	7,43
numero di caratteri per token	4,75	4,82	4,72

Tabella 4: Parametri generali del testo

Questi parametri sono importanti nella valutazione della leggibilità di un testo, perché sia la lunghezza di una parola sia la lunghezza di una frase possono dare un importante indizio iniziale riguardo la complessità generale del documento. Questi sono i parametri usati dagli indici di prima generazione (es. Gulpease) utili, come già detto, a caratterizzare genericamente la complessità lessicale del testo, ma insufficienti per la valutazione delle caratteristiche sintattiche legate a tale complessità.

2.3.2 PARAMETRI LESSICALI

La tabella 5 riporta l'elenco dei *parametri lessicali* presi in considerazione per ogni sottocategoria di testi:

		Testo	Lettura	Fumetti
Type(lemmi)/token		0,37	0,48	0,48
Perc. di parole trovate nel dizionario:		37,1	34,78	28,69
	Aut	39,32	40,65	39,53
	Adt	21,11	19,75	19,03
	Fot	39,56	39,58	41,43

Tabella 5: Parametri lessicali

Il rapporto "Type(lemmi)/token" viene calcolato rispetto alle prime 1000 parole del testo in relazione ai *lemmi* . Il lemma è la forma rappresentativa di tutte le altre forme flesse che una classe di parole può avere. Questo rapporto rappresenta uno dei metodi più diffusi per misurare la *varietà lessicale* di un corpus. È un indice che mette in rapporto il numero delle occorrenze delle parole unità del vocabolario di un testo (al denominatore) con il numero di parole tipo (al numeratore).
I valori di tale parametro oscillano su una scala che va da 0 e 1 dove, più elevato è il rapporto (Type/Token), maggiore è la varietà lessicale del testo.
Diventa necessario avere una lunghezza fissa di parole perché la misurazione della varietà lessicale è sensibile alla lunghezza del testo. Più lungo è il testo, più il rapporto è destinato a calare, visto che il numero di parole nuove cresce con un ritmo minore rispetto all'aumento della frequenza delle parole che si ripetono.

Passiamo ora a considerare la percentuale di parole trovate nel dizionario.
L'uso di parole chiare e familiari facilita la comprensione del messaggio che si vuole inviare. Ricordiamo che i bambini hanno spesso difficoltà nella lettura proprio per colpa della non conoscenza di alcuni vocaboli o dal mancato riconoscimento di elementi morfologici. Per questo motivo è preferibile scegliere le parole che appartengono al vocabolario di base della lingua italiana. Nell'analisi si andrà a considerare la percentuale di parole (forme) del testo che appartengono al Vocabolario Di Base di De Mauro (VdB). Tale vocabolario comprende circa 7000 parole, quelle che hanno la maggiore frequenza statistica nella nostra lingua, cioè quelle che più usiamo, che più ci sono familiari.

A tale percentuale segue la ripartizione nei repertori lessicali di :

- Parole Fondamentali (FO), cioè vocaboli ad altissima frequenza che fanno parte dei lemmi principali e le cui ricorrenze costituiscono circa l'80% delle occorrenze lessicali nell'insieme di tutti i testi scritti o dei discorsi parlati.

- Parole ad Alto Uso (AU), cioè vocaboli ad alta frequenza le cui occorrenze costituiscono il 6% circa delle occorrenze lessicali nell'insieme di tutti i testi scritti o dei discorsi parlati.

- Parole ad Alta Disponibilità (AD), vocaboli utilizzati raramente nel parlare o nello scrivere ma che sono conosciuti poiché relativi ad atti o oggetti di grande importanza nella vita quotidiana. Questo gruppo di parole richiede un aggiornamento frequente, essendo maggiormente esposto al mutare della cultura materiale (Il Giornale dell' E- Learning).

2.3.3 PARAMETRI MORFOSINTATTICI

Ora passiamo a considerare i *parametri morfosintattici* considerati, elencati nella Tabella 6.

	Testo	**Lettura**	**Fumetti**
Densita Lessicale:	0,57	0,56	0,59
Categorie grammaticali:			
A – Aggettivi	6,15	6,44	5,38
C – Congiunzioni	4,75	5,23	3,87
B – Avverbi	4,87	6,19	6,97
E – Preposizioni	12,17	12,38	9,3
F - Segni di punteggiatura	14,23	14,17	19,63
I – Interiezioni	0,08	0,13	1,12
N - Numeri Cardinali	1,31	1,05	1,3
P – Pronomi	5,85	6,75	6,46
S – Sostantivi	24,11	21,69	21,75
R – Articoli	9,77	8,35	6,46
V - Verbi principali	14,99	15,95	15,32
X - Classi residuali	0,02	0	0,04
Verbi+Modo:			
VA+c -Perc. di verbi ausiliari di modo congiuntivo	0,18	0,53	0,26
VA+f -Perc. di verbi ausiliari di modo infinito	0,37	0,46	0,44
VA+g -Perc. di verbi ausiliari al gerundio	0,007	0	
VA+d -Perc. di verbi ausiliari di modo condizionale	0,2	0,61	0,27
V+p - Perc. di participi	6,69	7,7	6,33
VA+i -Perc.di verbi ausiliari di modo indicativo	7,53	8,15	12,01
V+i - Perc. di verbi principali di modo indicativo	60,45	55,99	52,47
VA+p -Perc. di forme participiali di verbi ausiliari	0,29	0,38	0,4
V+m - Perc. di verbi principali di modo imperativo	1,8	1,23	2,65
V+c - Perc. di verbi principali di modo congiuntivo	1,56	2,46	2,26
V+d - Perc. di verbi principali di modo condizionale	0,37	0,56	0,87
V+f -Perc. di verbi principali di modo infinito	17,78	18,75	20,62
V+g -Perc. di verbi principali al gerundio	2,65	3,09	1,37

Tabella 6 : Parametri morfosintattici

Fra questi parametri abbiamo la Densità Lessicale, data dal rapporto tra parole dotate di significato – nomi, verbi, aggettivi e avverbi - e il totale delle occorrenze di parole nel testo, il cui ruolo principale è quello grammaticale – articoli, pronomi, preposizioni e congiunzioni. È un dato importante, perché più numerosi sono gli elementi lessicali rispetto a quelli grammaticali, maggiore è il carico informativo trasmesso dal testo e quindi la difficoltà di lettura, che si presume minore nei testi divulgativi.
Seguono la distribuzione delle categorie morfosintattiche.
Per ciascuna categoria morfosintattica viene riportata la percentuale di occorrenze rilevate nel testo oggetto di analisi, ad esempio avremo una percentuale del 6,15 di Aggettivi (A) nei Testi Scolastici verso quella del 6,44 dei libri di Lettura etc.
Inoltre è stata considerata la distribuzione dei modi dei verbi principali e ausiliari presentati nel corpus. Come visto in precedenza, avremo anche qui una serie di percentuali ad indicare la presenza di un verbo al modo infinito, al modo condizionale etc.

2.3.4 PARAMETRI SINTATTICI

Nella Tabella 7 infine sono elencati i *parametri sintattici* considerati nella composizione dell'indice Dylan.

		Testo	**Lettura**	**Fumetti**
Caratteristiche dell'albero sintattico				
-Media delle profondità massime degli alberi sintattici		3,87	4,25	2,45
-Lunghezza media delle catene preposizionali		1,24	1,25	1,22
-Distribuzione delle catene preposizionali		Percent.	Percent.	Percent.
	1	80,07	79,2	81,73
	2	16,36	16,88	14,64
	3	2,81	3,16	2,78
	4	0,52	0,6	0,61
	5	0,1	0,1	0,14
-Distribuzione delle catene di subordinate		Percent.	Percent.	Percent.
	1	85,6	82,11	87,39
	2	12,14	15,02	11,27
	3	1,85	2,49	1,14
	4	0,27	0,32	0,09
Caratteristiche dei predicati verbali				
-Percentuale delle radici verbali		0,61	0,65	0,48
-n° medio di dipendenti per teste verbali		1,81	1,85	1,75

-Distribuzione delle teste verbali per n° di dipendenti		Percent.	Percent.	Percent.
	0	8,54	8,32	9,73
	1	33,03	33,42	35,23
	2	35,64	34,16	33,43
	3	16,48	16,54	15,58
	4	4,8	5,52	4,57
	5	1,06	1,33	1,05
Subordinazione				
-Divisione frasi:		Percent.	Percent.	Percent.
frasi principali		72,03	66,71	78,94
frasi subordinate		27,96	33,28	21,05
-Distribuzione posizionale delle Subordinate				
prima della frase principale		13,71	11,49	8,1
dopo la frase principale		86,28	88,5	91,89
Caratterstiche delle ralazioni di dipendenza				
-Media delle lunghezze delle relazioni di dipendenza		2,34	2,31	1,95
-Medie delle lunghezze delle relazioni di dipendenza massime		5,41	6	2,73

Tabella 7: Parametri sintattici

La struttura delle tabella 7 mostra in che modo READ-IT divida i parametri sintattici in quattro insiemi di categorie fondamentali.

Il primo insieme è quello che riguarda le "Caratteristiche dell'albero sintattico[6]"; questo insieme di caratteristiche è destinato a valutare i diversi aspetti delle profondità dell'albero sintattico attraverso le seguenti misurazioni:

1) La media delle profondità massime degli alberi sintattici, ovvero la distanza massima che intercorre tra una parola del testo senza dipendenti (foglia) e la radice dell'albero, espressa come il numero di relazioni di dipendenza attraversati nel cammino foglia-radice.
2) La profondità media delle catene di dipendenza a testa nominale, misura l'indice di incidenza di strutture nominali complesse, contraddistinte dalla presenza di modificatori aggettivali, nominali e preposizionali.
3) La distribuzione delle catene preposizionali, misura la distanza delle catene di dipendenza a testa nominale per profondità. Il dato che segue riporta la percentuale di occorrenze di catene di dipendenza a testa nominale con profondità uguale a 1, 2, 3…

[6] Un **parse tree** o **albero sintattico** (**concreto**) è un albero che rappresenta la struttura sintattica di una stringa in accordo a determinate forme grammaticali. Un programma che produce quest'albero viene chiamato parser. I *parse tree* possono essere generati per frasi dei linguaggi naturali, così come durante l'elaborazione di linguaggi formali e linguaggi di programmazione.

4) La distribuzione delle catene di subordinate, riporta percentuale e numero di occorrenze di catene di subordinazione con profondità uguale a 1, 2, 3…

Il secondo insieme riguarda le "Caratteristiche dei predicati verbali"; questo insieme di caratteristiche valuta i differenti aspetti del comportamento del predicato verbale attraverso le seguenti misurazioni:

1) Percentuale delle radici verbali, calcolata rispetto a tutte le radici degli alberi sintattici costruiti per il testo.
2) Il numero medio di dipendenti per testa verbale, siano essi argomenti del verbo (es. oggetto, soggetto, etc.) oppure modificatori (es. modificatori temporali, locativi, etc.).
3) La distribuzione delle teste verbali per numero di dipendenti istanziati. Il dato riporta la percentuale di occorrenze di verbi con un numero di valenze istanziate uguale a 0, 1, 2, 3…

Il terzo insieme è quello dedicato alla "Subordinazione" notoriamente riconosciuta come indice di complessità strutturale in una lingua. In particolare sono state considerate:

1) La ripartizione tra frasi principali e frasi subordinate, calcolata a partire dal rapporto tra le radici verbali (corrispondenti alle frasi principali) e le clausole argomentali, temporali, causali, locative, etc.
2) La distribuzione posizionale delle frasi subordinate, che considera l'ordine relativo delle subordinate rispetto alla principale.

Ultimo insieme è quello riguardante le " Caratteristiche delle relazioni di dipendenza" misurate in termini di parole (token) che occorrono tra la testa e il dipendente, con l'esclusione delle relazioni che riguardano la punteggiatura. Anche questo è un parametro importante per la valutazione della leggibilità in quanto, il numero di token tra la testa e il dipendente regola il grado di complessità di un'intera frase. Maggiore è il numero di token nel cammino Testa - Dipendente maggiore sarà la difficoltà di lettura.

CAPITOLO III: ANALISI CONTRASTIVA DEI RISULTATI DEL MONITORAGGIO LINGUISTICO.

Cosa fa di un testo un documento difficile da comprendere? Su quali basi possiamo affermare che un testo è adatto ad un determinato tipo di pubblico?
Trovandomi di fronte ad una vasta lista di valori - ricavata dal monitoraggio linguistico automatico dei testi (cap. II sez.2.4) - ho selezionato quelli che erano i dati più rilevanti al raggiungimento del mio scopo, cioè quelli che meglio evidenziano le caratteristiche che fanno di un testo un esempio di facile o difficile lettura.
Ho provato ad effettuare una serie di confronti, partendo dal dato generale per arrivare a quello specifico.

3.1 CONFRONTO TRA LE SOTTOCATEGORIE DEL CORPUS DI LETTURA INFANTILE

Diamo nuovamente uno sguardo alle Tabelle 5, 6 e 7. Queste elencano per ciascuna sottocategoria di testi i parametri monitorati, utili alla comprensione del risultato dell'indice di leggibilità; sappiamo che i "Libri di Testo" hanno leggibilità pari a 0,02513; i testi di "Lettura e Fiabe" leggibilità pari a 0,98537; "Giornalini e Fumetti" leggibilità pari a 0,93604 (Tabella3).
Proviamo a motivare tali risultati in modo da spiegare perché i Libri di Testo risultano essere più leggibili rispetto ai testi delle altre due sottocategorie e viceversa.
Partendo dai parametri lessicali (Tab.5) ci accorgiamo che il numero medio di token per frase, in altre parole, la media della lunghezza delle frasi, è notevolmente differente.

	Testo	**Lettura**	**Fumetti**
Numero di Token per frase	13,51	15,25	7,44

Non è un dato inaspettato. Sappiamo che i fumetti sono composti da frasi brevi e concise soprattutto rispetto alle frasi che invece compongono un testo di lettura o una favola (vedi es. A e B).

Esempio A: **"**Nessun record è stato ancora battuto !
Sei la nostra unica speranza !
E cerca di non farti male !
Battere il record di salto dei camion sarà una formalità !
Sto volando verso la gloria !
Ha urtato contro l' ultimo centimetro dell' ultimo camion !
Anche quest' anno il record è sfumato !" (tratto da "Topolino 1653")

Esempio B:	“ Il geografo oggi , ponendo una in rapporto all' altra notizie provenienti da settori diversi , cerca di stabilire analogie e differenze ; in tal modo ci aiuta a porre ordine nel mare di informazioni in cui siamo immersi .”(tratto da “Il grande libro della geografia”)

Ad ogni modo, come abbiamo visto, questo non fa dei testi fumettistici dei documenti particolarmente leggibili.
Proseguendo con l’analisi degli altri risultati arriviamo ai dati relativi alla Varietà Lessicale (Type/token) che mostra per i Fumetti e per i libri di Lettura e Fiabe un dato maggiore del 0,10% rispetto ai Libri di Testo.

	Testo	**Lettura**	**Fumetti**
Type(lemmi)/token:	0,37	0,48	0,48

Questo vuol dire che i primi richiedono uno sforzo di lettura superiore, in quanto maggiore è il numero di *parole tipo* presenti nei testi. Ricordiamo che il rapporto Type/Token oscilla su valori da 0 a 1, dove 0 indica un vocabolario meno vario e 1 indica la presenza di un vocabolario più complesso.
Anche la “Percentuale di parole trovate nel dizionario” testimonia la minore difficoltà di lettura dei Libri di Testo rispetto agli altri, presentando una percentuale del 37% contro quella del 34% dei libri di Lettura e del 28% dei Fumetti. È vero che la percentuale maggiore indica la presenza di un lessico più chiaro, ma è anche vero che per ciò che riguarda i fumetti, siamo in presenza di testi ricchi di parole onomatopeiche, esclamazioni etc., che falsano il risultato ottenuto (vedi es. C).

Esempio C:	“Ffrrsshaa Corpo di mille paguri ! Glu .... che diav .... glu ....”

Proviamo ad osservare i risultati dell’annotazione di una frase di questo tipo (tab.8):

Frase	ID	forma	lemma	Cat. Gram.	Cat. Gram.2	Tratti	Testa	Tipo di dipendenza
1	1	**Ffrrsshaa**	**Ffrrsshaa**	**S**	**SP**	**_**	**2**	**Mod**
	2	Corpo	Corpo	S	SP	_	0	ROOT
	3	di	di	E	E	_	2	Comp
	4	mille	mille	N	N	_	5	Mod
	5	paguri	paguro	S	S	num=p|gen=m	3	Prep

	6	!	!	F	FS	_	2	Punc
2	1	**Glu**	**Glu**	**S**	**SP**	**_**	**0**	**ROOT**
	2	.....	.....	F	FF	_	0	ROOT
	3	che	che	P	PR	num=n|gen=n	0	ROOT
	4	diav	diav	S	S	num=n|gen=f	3	Mod
	5	.....	.....	F	FF	_	6	Punc
	6	**glu**	**glu**	**S**	**S**	**num=s|gen=f**	**4**	**Mod**
	7	.....	.....	F	FS	_	4	Punc

Tabella 8: Risultati dell'annotazione di una frase tratta da un fumetto

Notiamo che le espressioni onomatopeiche analizzate automaticamente sono considerate in generale (Cat. Gram.) sostantivi (S) e nello specifico (Cat. Gram.2) entrano a far parte dell'insieme dei nomi propri (SP).
Possiamo immaginare però, che non riusciremo a trovare sul dizionario sostantivi/nomi propri di questo tipo, scoprendo in questo modo una sorta di conflitto nell'annotazione linguistica .

Vedremo in seguito che nessuna delle tre sottocategorie ha una percentuale realmente alta di parole ritrovate sul dizionario se messe a confronto con i risultati ottenuti dall' analisi dei testi di una tipologia di corpus di "facile lettura" come il periodico 2Parole.

Passiamo ora ai parametri morfo-sintattici elencati nella Tabella 6. I dati relativi alla Densità lessicale mostrano una sottile differenza, legata in particolar modo alla categoria dei Fumetti. La densità lessicale maggiore in questo caso indica la maggior presenza di parole "semanticamente piene". Ricordando che in questo insieme di parole rientrano i sostantivi (S) potremmo associare tale risultato alla situazione riscontrata in precedenza riguardo le parole onomatopeiche considerate nomi (S).

	Testo	**Lettura**	**Fumetti**
Densità Lessicale:	0,57	0,56	0,59

I dati oscillano in maniera più evidente se andiamo a considerare una per una le categorie grammaticali.
Analizziamo quelle che mostrano valori maggiormente differenti.
Le "congiunzioni" (C) mostrano un valore più alto nei testi di Lettura. Nelle fiabe infatti, abbiamo la presenza di frasi più lunghe ed articolate, per questo è logico aspettarsi un maggior uso di congiunzioni, rispetto ai testi più brevi dei fumetti.

	Testo	**Lettura**	**Fumetti**
C – Congiunzioni	4,75	5,23	3,87

I dati sugli "avverbi" (B) *(bene, fortemente, malissimo, domani,* etc) mostrano una percentuale del quasi 7% per i Fumetti e del quasi 5% per i Libri di Testo. A cosa è dovuto tale risultato?

	Testo	Lettura	Fumetti
B – Avverbi	4,87	6,19	6,97

Pensando ai Giornalini e ai Fumetti, immaginiamo vignette con stralci di dialoghi. Proprio nei dialoghi l'uso degli avverbi, come quelli di tempo (*ieri, oggi, ora, subito, tardi*, etc) o di luogo (*qui, vicino, lontano*,etc) ha un riscontro importante (vedi es. D e E).

Esempio D: "A questa velocità saremo al castello orbitante fra due minuti !
Ci vorrà molto **meno**, ci vorrà due secondi capi !
Ci vorranno 2 minuti ! " minuti , non 2 secondi !
Ehi bambini , gli facciamo vedere ?
Date una spinta proprio **qui** .
E io cosa avevo detto ?
Del resto so **perfettamente** che in questa atmosfera si viaggia veloci .
Proprio come nella galassia maggiore ."

Esempio E: Proprio **vicino** alla grande valle del' eco avanzammo con circospezione e . . . e **po**i io gridai .
Eco **vicino** , eco **lontano** il più coraggioso è .
. . e **poi** io gridai .
. . e **poi** gridai **ancora** .
Eco **lontano** eco **vicino** il più pauroso è .
Ma **adesso** basta cosi , entriamo nel castello !

Anche i testi di Lettura e Fiabe presentano un tipo di scrittura ricca di avverbi, dimostrata dal dato di poco inferiore a quello dei Fumetti.

Altro risultato interessante, è quello che riguarda i "segni di punteggiatura " (F). Come sappiamo la punteggiatura è usata nella forma scritta e serve a conferire tonalità ed espressione al testo. È indispensabile per la corretta lettura dei testi e ne facilita la comprensione. Per la stessa motivazione vista in precedenza - cioè la grande ricchezza di discordi diretti - non deve stupirci la percentuale ottenuta per la sottocategoria dei Fumetti, superiore del 5% rispetto alle altre.

	Testo	Lettura	Fumetti
F - Segni di punteggiatura	14,23	14,17	19,63

Le percentuali per i "Sostantivi" (S) e "Articoli" (R) risultano maggiori nella categoria dei Libri di Testo. Questa distribuzione, legata in particolare all'uso dei

sostantivi, riflette il maggior grado di informatività dei materiali didattici (M.Voghera, 2004). Riprenderò la questione in seguito (cap. III sez 3.2.2).

	Testo	Lettura	Fumetti
S – Sostantivi	24,11	21,69	21,75
R – Articoli	9,77	8,35	6,46

Per quanto riguarda la sezione dedicata allo studio sulla distribuzione del "Modo del Verbo" , la situazione è abbastanza omogenea e le tre categorie mostrano percentuali paritarie; l'eccezione è costituita dai casi del "Verbo Ausiliare di Modo Indicativo" (Va+i) che mostra una percentuale per la categoria dei Fumetti maggiore del quasi il 5% rispetto alle altre due.

	Testo	Lettura	Fumetti
VA+i -Perc. di verbi ausiliari di modo indicativo	7,53	8,15	12,01

Non stupisce la percentuale del 60% dei "Verbi Principali al Modo Indicativo" dei Libri di Testo, se consideriamo il tipo di pubblico a cui i testi fanno riferimento.
Dando uno sguardo ad alcune percentuali ottenute per la sola categoria dei libri di Lettura, ritroviamo dati maggiori rispetto alle altre categorie nel caso dei "Verbi Ausiliari al Modo Congiuntivo" (0,53%) "Verbi Ausiliari al Modo Condizionale" (0,61%), e soprattutto "Verbi principali al Gerundio" (3,09%), a testimonianza di un uso più complesso del lessico che giustifica anche il minor grado di leggibilità.
Passiamo alla Tabella 7 e analizziamo i parametri sintattici. Partiamo dalla percentuale di Radici Verbali all'interno di un testo rispetto a tutte le radici degli alberi sintattici costruiti. Il dato presenta un 6% per i Libri di Testo e i libri di Lettura contro il 4% dei Fumetti. Probabilmente la differenza di percentuale sta nel fatto che nei Fumetti spesso ci troviamo davanti a frasi con radice nominale o a frasi che riproducono suoni onomatopeici o esclamazioni (vedi es. F).

Esempio F:	Zap zap Oh , no ! Zap Cras Ecco Rockerduck e il suo pilota ! Brrr .... che acque gelide !

Differente è anche il risultato della "Media delle lunghezze delle relazioni di dipendenza massime ";

	Testo	Lettura	Fumetti
Media delle lunghezze delle relazioni di dipendenza massime	5,41	6	2,73

Ricordando che tale media considera la distanza in Token tra la radice e il suo dipendente, riconduciamo tale disparità di risultati alla maggiore lunghezza delle frasi dei libri di Lettura e dei Libri di Testo rispetto a quelle dei Fumetti.

La “Divisione in frasi” per ciascuna categoria, dà ancora ragione alle nostre intuizioni riguardo le tre tipologie ti testi.

	Testo	Lettura	Fumetti
Divisioni Frasi:			
Frasi Principali	72,03	66,71	78,94
Frasi Subordinate	27,96	33,28	21,05
Distribuzione posizionale delle Subordinate:			
Prima della frase principale	13,71	11,49	8,1
Dopo la frase principale	86,28	88,50	91,89

Infatti il 79% di frasi principali presenti nei Fumetti, contro il 66% di quelle dei testi di Lettura, ci riconferma che questi ultimi sono caratterizzati da periodi più lunghi e articolati, dimostrato inoltre dalla percentuale del 33% di frasi subordinate, rispetto a quella del 21% risultata per i Fumetti e del 28% per i libri di Testo.
Sorprende, invece, il dato ottenuto per la “Distribuzione Posizionale delle Subordinate” riguardo i Libri di Testo; il 13% di subordinate pre-principale indica la sostanziale presenza di costruzioni “indirette”, dove la proposizione principale segue la proposizione secondaria, notoriamente più complesse rispetto alle costruzioni “dirette”, riconosciute come di più facile comprensione, in cui la proposizione principale è seguita dalla subordinata. Infatti la percentuale di subordinate post-principale dei Libri di Testo è inferiore di più del 2% rispetto alle percentuali delle altre due sottocategorie.

Il confronto con altre tipologie di testo (2Par e Rep), potrebbe rivelare ulteriori dati interessanti, soprattutto se ricordiamo il tema di fondo, ovvero, capire se i testi “letti” dai bambini sono effettivamente alla portata delle loro capacità di comprensione.

3.2 CONFRONTO TRA IL CORPUS DI LETTURA INFANTILE E ALTRE TIPOLOGIE DI TESTI

L’indice di leggibilità Dylan ha stabilito il livello di comprensione delle diverse sottocategorie di testi del corpus lettura infantile.
A questo punto possiamo provare a fare confronti con altri generi di testi per approfondire la ricerca. Ho pensato fosse interessante confrontare in parallelo due tipologie di testi definiti di alto livello di leggibilità e altri due considerati di basso livello di leggibilità.
Nello specifico ho messo a paragone:

- la categoria dei Libri di Testo (indice di Dylan più basso = maggiore leggibilità) con il periodico di “facile lettura” 2Parole (sez. 3.2.1)
- la categoria dei testi di Lettura e Fiabe (indice di Dylan più alto = minore leggibilità) con la prosa giornalistica di Repubblica, caratterizzata da un discreto grado di complessità linguistica (sez. 3.2.2)

Tali comparazioni potrebbero offrirci la possibilità di monitorare differenze e similarità linguistiche a diversi livelli e capire perché una categoria viene definita più complessa da elaborare rispetto ad un'altra.

3.2.1 SOTTOCATEGORIA DI LIBRI DI TESTO A CONFRONTO CON IL CORPUS 2PAROLE

Nell'analisi dei risultati di tutti i parametri presi in considerazione per il monitoraggio dei corpora, ho selezionato i dati che presentavano per ognuna un oscillazione maggiore al 2% (Tabella 9)

	Testo	**2Par**
Numero di Token per Frase	13,5	18,66
Percentuale di parole trovate nel dizionario:	37,1	47,92
Aut	39,32	31,13
ADt	21,11	13,79
Fot	39,56	55,06
B- Avverbi	4,87	3,44
E –Preposizioni	12,17	15,44
F -Segni di punteggiatura	14,23	11,01
P- Pronomi	5,85	2,24
V – verbi	14,99	13,5
S –Sostantivi	24,11	29,7
VA+i	7,53	19,8
V+p	6,69	4,86
V+i	60,45	48,98
V+f	17,78	24,87
Media delle lunghezze delle relazioni di dipendenza massime:	5,41	7,71
Media delle profondità Massime degli alberi:	3,87	5,25
Frasi Principali:	72,03	72,67
Frasi Subordinate:	27,96	27,32
Distribuzione posizionale delle Subordinate		
Prima della principale:	13,71	11,31
Dopo la principale:	86,28	88,68

Tabella 9:Risultati del monitoraggio del corpus di Libri di testo e di 2Parole

Il primo dato, in cui ho riscontrato un notevole scarto è quello riferito alla "Percentuale di parole trovate nel dizionario", che mostra per i Libri di Testo un risultato inferiore del 10% rispetto al corpus di 2Parole. La differenza rimane rilevante anche nella ripartizione di tali parole (AU, AD, FO), arrivando quasi al 20%. Questo non era stato colto nel confronto tra le singole sottocategorie del corpus di lettura, che al contrario aveva visto il risultato dei Libri di Testo decisamente superiore agli altri; certo, è logico pensare che cambiando i termini di paragone cambino anche i risultati, ma stupisce la grande differenza di tali riscontri. Ci si

sarebbe aspettati una situazione diversa, una maggiore somiglianza tra le due tipologie, soprattutto perché si stanno analizzando testi scolastici di scuole primarie.
Anche le percentuali, rispettivamente dei "Pronomi" e dei "Sostantivi", fanno luce sulla diversità dei testi oggetto d'analisi. La prima maggiore del 3% nei Libri di Testo rispetto a quella di 2Par e la seconda inferiore del 5%. L'uso dei pronomi, in sostituzione del nome è indice di una scrittura più difficile da elaborare. Spesso un bambino (e non solo) può trovare notevoli difficoltà ad orientarsi in rimandi testuali, a volte espressi anche in forme a lui poco familiari, per questo è sempre buona norma preferire l'uso di sostantivi.
Inoltre, la percentuale di "Pronomi" riscontrata nei Libri di Testo è addirittura superiore a quella del corpus di Repubblica, rispettivamente del 5,85% e del 4,09%.

A livello morfo-sintattico inoltre, sarebbe meglio preferire i Verbi principali di modo Indicativo; per questo la percentuale del 60% (V+i) della categoria dei Libri di Testo è un dato molto positivo. Lo è meno riscontrare una percentuale del quasi 7% di Verbi principali al Participio (V+p); di certo non è un dato "pesante" dal punto di vista statistico ma l'uso del participio ricorre principalmente in testi particolarmente articolati, prodotti in contesti spesso formali, per questo da evitare in testi che devono mantenere una certa facilità di elaborazione dei contenuti.

La media delle lunghezze massime (per frase) delle relazioni di dipendenza risulta essere maggiore in 2Parole, che mostra anche un numero maggiore di token per frase.
Per quanto riguarda la "Distribuzione posizionale delle frasi subordinate rispetto alla principali" vale lo stesso discorso fatto in precedenza per il confronto tra le categorie del corpus di lettura (sez. 3.1);
Rimando alla fine di questo paragrafo una discussione dell'uso del verbo e del sostantivo, allargando lo sguardo anche ai testi di Lettura e Fiabe e di Repubblica.

3.2.2. SOTTOCATEGORIA DI TESTI DI LETTURA E FIABE A CONFRONTO CON IL CORPUS DI REPUBBLICA

Anche in questo secondo caso ho selezionato i parametri che presentavano dati oscillanti in maniera evidente (Tabella 10). Mettendo a confronto la categoria del testi di Lettura e Fiabe con quella di Repubblica ho voluto capire perché la prima ha registrato un indice di leggibilità cosi alto , essendo la prosa giornalistica un corpus complessivamente molto elaborato.

	Lettura	**Rep**
Numero di Token per Frase	15,24	24,93
Type(lemmi)/token:	0,48	0,39
Percentuale di parole trovate nel dizionario:	34,78	25,13
Aut	40,65	41,75
ADt	19,75	17,58
Fot	39,58	40,65

B- Avverbi	6,19	5,01
E –Preposizioni	12,38	15,91
F -Segni di punteggiatura	14,17	13,31
P- Pronomi	6,75	4,09
V-Verbi	15,95	13,05
S –Sostantivi	21,69	26,5
VA+i	8,15	17,1
V+p	7,7	13,14
V+i	55,99	40,72
V+f	18,75	19.47
Media delle lunghezze delle relazioni di dipendenza massime:	6	9,44
Media delle profondità massime degli alberi:	4,25	6,09
Frasi Principali	66,71	63,84
Frasi Subordinate	33,28	36,15
Distribuzione posizionale delle Subordinate:		
Prima della principale:	11,49	11,93
Dopo la principale:	88,5	88,06

Tabella 10: Risultati del monitoraggio tra il corpus Lettura & Fiabe e Repubblica

In primo luogo il numero di Token per frase è molto diverso. Repubblica mostra un risultato maggiore, quindi caratteristica del corpus è essere formato da frasi più lunghe; procedendo, notiamo però che la varietà lessicale (type/token) è più alta per i testi di Lettura, diversamente da quanto ci si potrebbe aspettare. In definitiva avere un risultato per questo rapporto più vicino a 1 sottolinea che il vocabolario usato è più vario.

Mentre nel confronto precedente tra i libri di Testo e 2Par la differenza di percentuale era minima (ca 0,4%) in questo caso è sia più alta (ca 0,10%) ed è soprattutto a favore della categoria che in linea di principio avrebbe dovuto mostrare un uso del vocabolario meno articolato, ovvero i testi di Lettura e Fiabe.

Nonostante questo, il numero delle parole ritrovate nel vocabolario di base per quest'ultima categoria è maggiore di ca il 10% rispetto alla prosa giornalistica; perciò potremmo dedurre che l'uso di un linguaggio più vario, non implica necessariamente l'uso di parole non comuni. In questo caso come nel precedente, con una percentuale del 6,75% contro quella del 4,09% di Repubblica, i testi di Lettura presentano un maggiore uso di Pronomi ed un sostanziale minor uso dei Sostantivi.

Proseguendo con l'analisi dei dati ci accorgiamo che caratteristica della prosa giornalistica (Rep) è il maggior uso di Verbi Ausiliari al modo indicativo (VA+i) e dei Verbi Principali al participio (V+p), mentre i testi di Lettura mostrano una percentuale maggiore di Verbi Principali al modo indicativo (V+i). La Media delle lunghezze delle relazioni di dipendenza massime è maggiore nei testi di Repubblica, cosi come la percentuale di frasi Subordinate, invece risulta inferiore rispetto al corpus di Lettura la percentuale di frasi Principali.

Vorrei a questo punto aprire una parentesi sull'uso dei Verbi e dei Sostantivi, allargando l'analisi alle quattro categorie considerate nei due confronti (Tabella 11).

	Testo	**Lettura**	**2Par**	**Rep**
V- Verbi	14,99	15,95	13,5	13,05
S –Sostantivi	24,11	21,69	29,7	26,5
Rapporto Nomi/Verbi	1,6	1,35	2,2	2,02

Tabella 11: Analisi dell'uso dei verbi e dei sostantivi per le diverse categorie

I dati mostrano che tutte le tipologie di testi in questione hanno una maggior frequenza di Nomi rispetto ai Verbi, ma nello specifico abbiamo risultati alquanto diversi per le due fasce di corpora considerati. Infatti i libri di Testo e di Lettura presentano nell'uso dei Sostantivi una differenza rilevante rispetto ai corpora giornalistici.

Tale differenza è meglio sottolineata dal rapporto Nomi/verbi per ciascun corpus. Quello che emerge dall'analisi automatica dei corpora coincide con quanto riportato in Voghera (2004); per quel che riguarda la lingua scritta si nota una maggiore frequenza di nomi nei testi caratterizzati da un' alta densità informativa, verso una minore frequenza degli stessi nei testi più vicini alla lingua parlata, appunto quelli di lettura e scrittura creativa; queste tipologie di testi- quelli di lettura appunto- sono caratterizzati da una frequenza maggiore di verbi. Voghera (2005) registra lo stesso tipo di tendenza negli schemi di distribuzione dei nomi e dei verbi in corpora di testi informativi italiani. Anche Biber (1995) rivela una correlazione positiva tra la frequenza dei nomi e testi scritti ad alta densità informativa come articoli di giornale e testi accademici, cosi come tra frequenza dei verbi e testi di scrittura creativa.

CAPITOLO IV: TESTI FACILI E TESTI DIFFICILI, PERCHÉ?

Abbiamo fino ad ora considerato le intere sottocategorie di corpora di lettura infantile e quindi l'indice di leggibilità generale.
Ora potremmo entrare nello specifico e capire cosa rende un singolo documento appartenente al corpus difficile o facile da comprendere e quali sono le caratteristiche che influiscono più di altre nella valutazione del grado di leggibilità.
A questo proposito i confronti da fare potrebbero essere tanti; ho deciso di capire perché un testo può avere leggibilità pari a 1(leggibilità molto bassa) in una categoria con indice generale pari a 0 (leggibilità molto alta) e viceversa.
Per fare questo ho analizzato tre testi, uno per ogni sottocategoria.
Cominciamo dal corpus di Libri di Testo. Ricordiamo che questa categoria ha ottenuto indice Dylan pari a 0,02513 – ovvero, grado di leggibilità molto alto - per questo analizzeremo un singolo testo appartenente alla stessa categoria che ha presentato indice Dylan molto alto, segno di una bassa leggibilità. Nella Tabella 12 ho raccolto alcuni dati ricavati dall'analisi della sezione di testo tratto da "Scuola nuova due" che ha ottenuto indice Dylan pari a 0,9699. Precisamente la selezione dei parametri riportati in tabella nasce dal confronto del suddetto testo oggetto d'analisi con uno dei testi risultati più difficile da leggere. I parametri riportati sono quelli che hanno mostrato oscillazioni maggiori. Per ulteriore chiarezza ho sintetizzato nella terza colonna della tabella i dati riferiti all'intero corpus di Libri di Testo.

CARATTERISTICHE	**TESTO SINGOLO**	**CORPUS UNICO DI LIBRI DI TESTO**
Type(lemmi)Token:	0,67	0,37
Media Lunghezza Frasi:	23	13,5
V- verbi	0,16	14,99
S- sostantivi	0,22	24,11
Profondità relazioni di dipendenza:	2,55	2,34
Profondità relazioni di dipendenza Massime:	10,42	5,41
Profondità media Alberi sintattici:	7	3,87
Profondità media delle catene di subordinazione	2,28	1,17
Frasi Principali	0,46	72,03
Frasi Subordinate	0,53	27,09
Leggibilità:	**0,9699**	**0,0251**

Tabella 12: Analisi del testo tratto da "Scuola nuova due"

Prima di tutto notiamo che i dati riguardanti la Varietà Lessicale (Type/Token) sono prossimi a 1 il che, come già detto nei paragrafi precedenti, indica un vocabolario

piuttosto vario. Un dato importante è quello della Media della lunghezza di una frase pari a 23,27 Token per frase; considerando che altri singoli testi della stessa categoria (con indice più basso) mostrano per lo stesso parametro risultati che oscillano tra i 10 e i 15 Token per frase (dimostrato anche dalla media generale ottenuta dall'analisi dell'intero sottocorpora), si comprende la "pesantezza" di tale risultato ai fini della valutazione della leggibilità. I dati legati alla Media delle profondità degli alberi sintattici e alla Lunghezza delle relazioni di dipendenza massime sono altrettanto indizi importanti di bassa leggibilità. Il primo risultato ci dice che la profondità massima dell'albero -calcolata come la distanza massima che intercorre tra le parole del testo senza dipendenti (foglie) e la radice dell'albero, espressa come il numero di relazioni di dipendenza attraversati nel cammino foglia-radice- è pari a 7, verso un dato generale di 3,87; il secondo risultato riguardante la media delle lunghezze massime delle relazioni di dipendenza è pari a 10,42 contro il dato generale del 5,41. In più osserviamo una presenza maggiore di frasi subordinate rispetto alle principali, il che giustifica in qualche modo i dati precedenti. Vedremo in seguito l'importanza di questi risultati al fine della mia ricerca.

Analizziamo nella Tabella 13 alcuni risultati dell'analisi di un testo preso dal corpus di Lettura e Fiabe (leggibilità molto bassa, pari a 0,98537). Il testo in esame (tratto da " 366 e più storie della Bibbia ") ha un indice Dylan pari a 0, 0538 (leggibilità molto alta). I criteri di selezione dei parametri mostrati nella tabella che segue sono gli stessi descritti per il caso precedente.

CARATTERISTICHE	**TESTO SINGOLO**	**CORPUS LETTURA E FIABE**
Type(lemmi)Token	0,52	0,48
Media Lunghezza Frasi	17	15,48
V – verbi	0,16	15,95
S – sostantivi	0,22	21,69
Profondità relazioni di dipendenza	2	2,3
Profondità relazioni di dipendenza Massime:	7	6
Profondità degli Alberi sintattici	5	4,25
Profondità media delle catene di subordinazione	1	1,21
frasi Principali	0,59	66,71
frasi subordinate	0,4	33,28
Leggibilità	**0,0538**	**0,9853**

Tabella 13: Analisi del testo tratto da "366 e più storie della Bibbia".

I parametri mostrati sono gli stessi della tabella 12, il che lascerebbe supporre che le caratteristiche selezionate sono appunto, quelle che più di altre fanno oscillare il valore sulla scala dell'indice di leggibilità. Globalmente notiamo che i dati riferiti al testo singolo sono tutti inferiori ai dati ottenuti per il testo precedente e in più

notiamo una inversione di percentuale per quel che riguarda le frasi principali (0,59) e le subordinate (0,4), indice di un tipo di scrittura meno articolata.

Passiamo ora ad un testo tratto dal corpus dei Fumetti e Giornalini . Anche in questo caso, per inverso, scelgo un testo tratto dal "Topolino" con leggibilità alta pari a 0.1817, essendo compreso in una categoria del corpus di lettura infantile con Indice Dylan pari a 0,93604. (Tabella 14)

CARATTERISTICHE	**TESTO SINGOLO**	**CORPUS UNICO DI GIORNALI-NI E FUMETTI**
TypeTokenL	0,61	0,48
Media Lunghezza Frasi	7,4	7,43
S –sostantivi	0,17	21,75
V – verbi	0,20	15,32
Profondità relazioni di dipendenza:	1,61	1,95
Profondità relazioni di dipendenza Massime:	3,22	2,73
Profondità media Alberi sintattici:	3,26	2,45
Profondità media delle catene di subordinazione	1,33	1,14
Frasi Principali	0,72	78,94
Frasi Subordinate	0,27	21,94
Leggibilità	**0,1817**	**0,93604**

Tabella 14: Analisi del testo tratto dal "Topolino"

In questo caso, prendendo in considerazione gli stessi parametri, ci accorgiamo che il maggior grado di leggibilità viene a dipendere, non dalla Varietà Lessicale, alta quanto si era osservato per il testo tratto dal corpus dei Libri di Testo, ma dalla sostanziale inferiorità dei dati riferiti alla media delle lunghezza delle frasi, alla profondità degli alberi sintattici e alla lunghezza delle relazioni di dipendenza massime. Inoltre abbiamo una maggiore percentuale di frasi principali, lo 0,72% contro lo 0,27% delle frasi subordinate, anche questo è un dato molto importante per quelli che sono gli standard studiati a proposito di un livello di scrittura più semplice da elaborare.

I dati che per i tre testi delle diverse sottocategorie e con diversi gradi di leggibilità-restano costanti sono quelli legati alla percentuale di Verbi e Nomi.

Soffermiamoci sul risultato ottenuto dalla Media della lunghezza delle frasi; nel primo documento appartenente alla sottocategoria dei Libri di Testo il dato è alto confronto ai dati degli altri due documenti; rispettivamente per ogni sottocategoria abbiamo una media di 23,27, 16,17 e 7,4 Token per frase. Questo è un dato

importante, considerando che nella maggior parte delle occasioni, quanto più è lunga una frase, tanto è più complicata da capire. Studi in questo campo hanno rilevato che una frase non dovrebbe contenere più di 25 parole ca. Scrivere frasi brevi e che contengano una sola informazione fondamentale ha tra le sue conseguenze quella di limitare la subordinazione, cioè di ridurre al minimo il numero di proposizioni presenti in un periodo e i dati in nostro possesso danno ragione a questa affermazione. In definitiva la media della lunghezza delle frasi non rende da sola un documento illeggibile, ma se ad essa si accompagna anche una certa complessità sintattica, con tante subordinate, incisi, parentesi etc., la questione diventa rilevante.

CAPITOLO V: DUE GENERAZIONI DI INDICI DI LEGGIBILITÀ A CONFRONTO*: Gulpease verso DyLan*

Sembra giusto a questo punto mettere a confronto le due generazioni di indici per la valutazione della leggibilità. Paragoniamo allora, l'indice di Gulpease – più recente e più semplice rispetto all'indice Flesch- con l'indice Dylan per capirne le differenze.

Riprendiamo alcune nozioni su Gulpease, già enunciate precedentemente. La formula dell'indice Gulpease considera come variabili linguistiche: la lunghezza delle parole in lettere, il numero di parole che compongono il testo e il numero delle frasi che compongono il testo.

La misurazione della leggibilità con l'indice Dylan si sviluppa tramite un classificatore statistico basato su metodi di apprendimento automatico. Come abbiamo visto si basa su caratteristiche linguistiche lessicali, morfosintattiche e sintattiche.

In questo primo caso ho sottoposto nuovamente ad analisi il documento della categoria Libri di Testo, tratto da "Scuola nuova due" (tab. 15).

CARATTERISTICHE	**SCUOLA NUOVA DUE**
Type(lemmi)Token:	0,67
Media Lunghezza Frasi:	23
Parole trovate sul Dizionario:	0,77
V- verbi	0,16
S- sostantivi	0,22
Profondità relazioni di dipendenza:	2,55
Profondità relazioni di dipendenza Massime:	10,42
Profondità media Alberi sintattici:	7
Profondità media delle catene di suborbinazione	2,28
Frasi Principali	0,46
Frasi Subordinate	0,53
Leggibilità:	**0,9699**

Tabella 15: Analisi de testo tratto da "Scuola nuova due"

Le analisi di Dylan restituiscono una situazione abbastanza chiara in particolar modo per ciò che riguarda la parte sintattica dell'analisi. Infatti le percentuali ottenute per il calcolo della "profondità degli alberi sintattici", "la lunghezza delle relazioni di dipendenza massime" e la "divisione delle frasi" sottolineano la difficoltà di elaborazione del testo. Tutto questo è la traduzione dell'Indice Dylan pari a 0,9699 (

ovvero quasi il massimo nella scala di valori che va da 0= facile lettura a 1= difficile lettura) .

L'indice Gulpease presenta un risultato pari a 54,6 (tab. 16), anch'esso non consono a quello che ci si aspetterebbe da un documento pensato e scritto per bambini . Nella scala di valori sulla quale oscillano i risultati dell'indice - che va da 0 a 100 (0= difficile lettura, 100= facile lettura) - un simile risultato corrisponde al livello di "difficile di lettura per coloro che possiedono la sola licenza media" (sez. 1.1.1.). Infatti l'età minima richiesta per la comprensione del testo è di 15 anni, con un livello di complessità totale del 45,4%. Possiamo dire che, per quanto i risultati degli indici siano sostanzialmente diversi per poter essere paragonati, il dato generale non cambia; il testo in esame non è adatto alla lettura da parte di utenti con un'età compresa tra i 6 e gli 11 anni.

Indice di leggibilità	Lingua	Valore (0-100)	Età richiesta	Livello di difficoltà
Gulpease	**IT**	**54,6**	**14,9 years**	45,4%

Tabella 16: Indice Gulpease per il testo tratto da "Scuola nuova due"

Ad ogni modo è necessario notare che, l'indice Gulpease si discosta di poco dal risultato reputato corretto, al contrario dell'indice Dylan. Questo perché l'indice Gulpease è ottenuto sulla sola base dei risultati riguardanti le "Statistiche del testo" (tab.17) e in particolar modo in base al:

- numero delle parole, pari a 1687;
- numero delle frasi, pari a 73;
- la media di lettere per 100 parole che è di 474,1 traducibile in lunghezza media delle parole pari a 4,74 token per parola.

Statistiche del testo	Valore
Letters	7998,0
Syllables	3311,0
Words	1687,0
Sentences	73,0
Average letters per 100 words	474,1
Average syllables per 100 words	196,3
Average sentences per 100 words	4,3

Tabella 17: Statistiche del testo per il testo tratto da "Scuola nuova due"

A questo punto analizziamo un altro testo, per capire come varia l'indice Gulpease in base a tali parametri.

Per il secondo esempio ho preso in considerazione il testo “Biancaneve e i sette nani” della categoria Lettura e Fiabe. Il corpus è stato analizzato e per tutte la caratteristiche linguistiche monitorate abbiamo ottenuto una percentuale di difficoltà di lettura molto bassa , tradotta dall’ indice Dylan pari praticamente a 0. I parametri maggiormente significativi sono elencati nella tabella 18.

CARATTERISTICHE	TESTO SINGOLO
Type(lemmi)/Token	0,56
Media Lunghezza Frasi	9
Parole trovate sul Dizionario:	0,80
S – sostantivi	0,18
V – verbi	0,19
Profondità relazioni di dipendenza:	1,77
Profondità relazioni di dipendenza massime:	4
Profondità media Alberi sintattici	3
Profondità media delle catene di subordinazione	1,3
frasi Principali	0,74
frasi Subordinate	0,25
Leggibilità	**0,06**

Tabella 18: Analisi del testo “Biancaneve e i sette nani”

Possiamo ben notare che alla base di un indice di leggibilità cosi basso, ci sono dati che sottolineano la semplicità di scrittura del testo, come la media della lunghezza delle frasi pari a 9 token per frase, la profondità delle relazioni di dipendenza massime pari a 4 token, e la profondità media degli alberi sintattici pari a 3 token. Inoltre, risulta molto ridotto l’uso della subordinazione.

Passiamo al risultato dell’indice Gulpease pari a 85,9 (tab.19). Anche in questo caso il livello di difficoltà del testo è basso, pari al 14,1%.

Indice di leggibilità	Lingua	Valore (0-100)	Età richiesta	Livello di difficoltà
Gulpease	**IT**	**85,9**	**9 years**	14,1%

Tabella 19: Indice Gulpease per il testo “Biancaneve e i sette nani”

La traduzione di questo risultato non va ricercata nelle caratteristiche linguistiche del testo - utili a capire il risultato dell’indice Dylan - ma, come abbiamo già visto, nei dati ottenuti dalle “Statistiche del testo”(tab. 20).

Statistiche del testo	Valore
Letters	4358,0
Syllables	1852,0
Words	891,0
Sentences	136,0
Average letters per 100 words	489,1
Average syllables per 100 words	207,9
Average sentences per 100 words	15,3

Tabella 20: Statistiche del testo "Biancaneve e i sette nani"

In questo secondo caso il numero delle parole è mediamente basso rispetto al caso precedente (le 891 parole attuali contro 1687 del primo testo) , ma il numero delle frasi e la media delle lettere che compongono una parola aumentano. Abbiamo 136 frasi contro le 73 precedenti ed una lunghezza media di parola di 4,89 lettere contro la precedente di 4,74 lettere.

Concludiamo questo paragrafo con un ultimo caso, prendendo in esame un testo della categoria Giornalini e Fumetti. Il testo preso in esame è il fumetto comparso sul giornalino "Topolino, n° 1653" . L'indice Dylan a riguardo è pari a 0,77 (scala di valori 0-1 ,dove 0 indica la leggibilità più alta e 1 quella più bassa) e nella tabella 21 ho riportato la caratteristiche del testo che hanno mostrato i risultati più interessanti.

CARATTERISTICHE	TESTO UNICO
Type(lemmi)/Token	0,77
Media Lunghezza Frasi	7,2
Parole trovate sul Dizionario	0,73
S - sostantivi	0,17
V – verbi	0,18
Profondità relazioni di dipendenza:	1,77
Profondità relazioni di dipendenza massime:	3,5
Profondità media Alberi sintattici	3,15
Profondità media delle catene di suborbinazione	1,56
frasi Principali	0,76
frasi Subordinate	0,23
Leggibilità	**0,77**

Tabella 21: Analisi del testo tratto da "Topolino, n° 1653"

L'indice Dylan pari a 0,77 in questo caso, non traduce le difficoltà di elaborazione dovute alla sintassi del testo. Infatti si notano grandi differenze tra i dati ottenuti per i parametri sintattici del testo precedente (vedi tab.18) caratterizzato da una leggibilità

molto alta e il testo attualmente in esame, caratterizzato da una leggibilità piuttosto bassa. L'indice Dylan traduce le difficoltà derivanti dalla maggiore varietà lessicale (Type/Token) e dalla minore percentuale di parole ritrovate nel Vocabolario di base (VdB). Riassumo per chiarezza i dati delle diverse categorie di testi nella tabella seguente:

CARATTERISTICHE	Scuola nuova due	Biancaneve	Topolino
Type(lemmi)/Token	0,67	0,56	0,77
Parole trovate sul Dizionario	0,77	0,80	0,73

Ma ad un indice Dylan che sottolinea la difficoltà di elaborazione di un testo fumettistico soprattutto su basi lessicali, si contrappone l'indice Gulpease pari a 86, che indica un testo assolutamente adeguato ad un pubblico di bambini. Nello specifico, la tabella 22 mostra che l'età richiesta per la corretta comprensione di questo testo è di 9 anni.

Indice di leggibilità	Lingua	Valore (0-100)	Età richiesta	Livello di difficoltà
Gulpease	**IT**	**86**	**9 years**	14,2%

Tabella 22: Indice Gulpease per il testo tratto da "Topolino, n° 1653"

Analizziamo le "Statistiche del testo" (tab. 23) per comprendere il dato ottenuto. Risultano presenti nel testo 173 frasi e 1209 parole, il che vuol dire avere la media di 7 token per frase (1209Par/173Fr). Sono le frasi molto brevi- per questo considerate semplici- a far oscillare il risultato dell'indice Gulpease verso un risultato cosi alto.

Statistiche del testo	Valore
Letters	5572,0
Syllables	2284,0
Words	1209,0
Sentences	173,0
Average letters per 100 words	460,9
Average syllables per 100 words	188,9
Average sentences per 100 words	14,3

Tabella 23: Statistiche del testo tratto da "Topolino, n° 1653"

Facendo qualche passo in dietro, confermiamo questa considerazione attraverso un semplice calcolo. Nel primo caso, indice Gulpease di 54,6 e la lunghezza media delle frasi era di 23 token per frase (1687 Par/73Fr) cioè a maggiore lunghezza media di frasi corrisponde un indice di leggibilità minore (maggiore difficoltà di lettura) . La stessa cosa non si può affermare per l'indice Dylan. Infatti un risultato di leggibilità pari a 0,77 - sostanzialmente alto - sottolinea la difficoltà di lettura del testo e

ribadisce quello che si è sostenuto inizialmente riguardo la lunghezza di una frase: non sempre a frase breve corrisponde semplicità di lettura. Per l'indice Dylan questo testo non è adatto ad un pubblico d'infanzia.

Non dobbiamo dimenticare che le generali "Statistiche del testo" fanno parte dei parametri di calcolo sui quali si basa l'Indice Dylan. A queste si aggiungono una serie di analisi di caratteristiche linguistiche che esulano invece dal calcolo dell'indice di Gulpease.

In linea di principio - anche se con risultati nettamente diversi- i due indici di valutazione della leggibilità hanno mostrano nei primi 2 casi, una certa concordanza di fondo. In quest'ultimo esempio, invece, i risultati ottenuti sono profondamente contrastanti.

Attraverso queste considerazioni si è dimostrato in maniera pratica quali sono le caratteristiche che fanno oscillare i risultati dei diversi modelli di misurazione della leggibilità e chiarito il ruolo dei parametri sui quali gli indici di prima generazione potevano basarsi. Gli indici attuali possono fornirci indicazioni sul luogo di complessità di una frase o di un intero testo, mentre quelli precedenti potevano solo (sulla base degli strumenti disponibili) indicarci difficoltà legate alla sua struttura generale.

CONCLUSIONE

La diffusione di una cultura più attenta alle difficoltà legate alla lingua scritta, come quella dello studio - e non solo - può diventare un veicolo vincente sia per aiutare gli allievi ad affrontare con successo il loro percorso scolastico sia per supportare l'insegnante nel faticoso compito di ricerca di testi adatti ad ogni fascia di età o situazione socio-culturale. Allo stesso tempo può migliorare il lavoro dell' autore che desidera essere attento a determinate necessità del lettore.

Per tutto questo è però necessario intervenire concretamente su queste difficoltà e a mio avviso gli strumenti linguistici attuali (come READ-IT) sono un valido supporto a questo problema.

Le ricerche effettuate sui dati che mi sono stati messi a disposizione, ricavati dall'analisi del corpus alla base del "Lessico Elementare", hanno sottolineato l'importanza dell'indice di leggibilità - come fonte sintetica di misurazione della complessità di un testo – ma hanno dimostrato anche l'importanza di " ciò che c'è dietro " questo indice sintetico: il monitoraggio linguistico basato su parametri lessicali, morfo-sintattici e sintattici, utili a motivare la complessità del documento.

Un esempio pratico ha riguardato le analisi interne al corpus di lettura infantile che hanno definito i Libri di Testo come i più semplici da comprendere. Lo studio dei risultati dei parametri monitorati ha chiarito a cosa fosse legata la minore complessità della sottocategoria in questione cosa non possibile da ottenere con strumenti di

prima generazione come l'indice Gulpease, legato ad analisi della sola struttura superficiale del testo e per questo non motivato linguisticamente.

L'indice Dylan, risultato sintetico del monitoraggio effettuato tramite il software READ-IT, considera aspetti che riguardano per esempio, le relazioni di dipendenza all'interno di una frase, la distribuzione dei verbi, la struttura dell'albero sintattico e cosi via, restituendoci informazioni che potrebbero aiutarci a comprendere meglio a che livello linguistico effettuare - se fosse questo lo scopo del nostro lavoro- delle semplificazioni.

Le potenzialità di uno strumento motivato linguisticamente come READ-IT sono state particolarmente interessanti per la situazione presentata dalla sottocategoria dei Giornalini e Fumetti. La brevità delle frasi che caratterizza questo genere di scrittura avrebbe potuto essere indizio di "facile lettura" – come per gli indici di prima generazione - ma come abbiamo visto, la realtà dei fatti è decisamente opposta. Infatti la sottocategoria oltre ad ottenere un indice globale di complessità molto alto, nell'analisi del singolo testo ha potuto meglio dimostrare il conflitto di base tra i diversi indici di leggibilità, Dylan e Gulpease. Se per Gulpease il testo risultava particolarmente semplice da leggere in base alle statistiche testuali, per Dylan il risultato è contrario e le motivazioni sono state riscontrate nei vari dati ottenuti dal monitoraggio del testo, quali la maggiore varietà lessicale (Type/Token), la minore percentuale di parole ritrovate nel Vocabolario di base, etc .

Dobbiamo anche dire che l'annotazione di testi complessi come quelli dei fumetti spesso ha presentato delle difficoltà tecniche, ma questo non ha impedito una analisi generale comunque più approfondita.

Abbiamo compreso che le tecnologie di analisi automatica del testo permetterebbero di valutare la complessità di un documento e soprattutto di identificarne i luoghi di complessità, in vista di una possibile semplificazione.

Scrivere senza riuscire a comunicare non serve a nulla, per questo motivo è necessario, oltre ad imparare a scrivere correttamente, imparare ad essere leggibili.

L'applicazione di un indice linguisticamente motivato per la misura della leggibilità di un testo può essere una possibile soluzione all'identificazione dei passi critici che necessitano di una riscrittura, supportando così il compito dell'autore interessato. In più, le tecnologie di questo tipo possono avvalorare il compito di chi, come un insegnate, deve scegliere il tipo di lettura adatto agli allievi che ha di fronte, non dovendo più fare affidamento sulle sole capacità di intuizione.

RINGRAZIAMENTI

Lo studio dell'indice Dylan è stato condotto all'interno del Dylan Lab "Laboratorio per lo studio delle dinamiche linguistico-cognitive". Ringrazio per questo i ricercatori dell'Istituto di Linguistica Computazionale " Antonio Zampolli" del CNR di Pisa in particolare la dott.ssa Simonetta Montemagni, Felice Dell'Orletta e Giulia Venturi che mi hanno fornito tutti gli strumenti utili all'analisi.

È doveroso ringraziare l'unità staccata di Genova dell'istituto di Linguistica Computazionale (ILC-CNR) che gentilmente ha messo a disposizione in formato elettronico il corpus di testi alla base dell'opera "Lessico Elementare" (di L Marconi et al. ,ed. Zanichelli, Bologna, 1994)

Riferimenti Bibliografici

Aluisio Sandra, Lucia Specia, Caroline Gasperin, Carolina Scarton (2010). *Readability assessment for text simplificatio.*

Biber, D. (1995) *Dimention of register variation: A cross - linguistic comparison.* Cambrige & New York, Cambrige University Press

De Mauro, T. *Grande dizionario italiano dell'uso*, Torino, UTET, 2000.

Dell'Orletta Felice , Simonetta Montemagni, Giulia Venturi. (2011). *READ-IT : Assessing Readability of Italian Text with a view to Text Semplification.*

Lotti, P. (2010). *Manuale utente FacilTesto.*

Marconi Lucia, Michela Ott, Elia Pesenti, Daniela Ratti, Mauro Tavella (1994). *"Lessico Elementare",* Zanichelli, Bologna.

Marconi Lucia, Michela Ott, Elia Pesenti, Daniela Ratti, Mauro Tavella (1994). *Introduzione a "Lessico Elementare",* Zanichelli, Bologna.

Montemagni, S. (2010). *Tecnologie linguistico-computazionali per il monitoraggio della lingua italiana.* Presentazione tenuta nell'ambito della Giornata di Studio " Lo stato della lingua. Il CNR e l'italiano nel terzo millenio" Roma, 8 marzo 2010, Consiglio Nazionale delle Ricerche - Dopartimento Identità Culturale.

Montemagni Simonetta, Felice Dell'Orletta (in stampa). *Tecnologie linguistico - computazionali nella valutazione delle competenze linguistiche in ambito scolastico,* in stampa negli *Atti del XLIV Congresso Internazionale di Studi della SLI. Linguistica educativa/Lessico e Lessicologia, Viterbo,* 27-29 settembre 2010.

Pitler, E., & Nenkova, A. (2010). Revisiting Readability: A unified framework for predicting text quality

Voghera M. (2004) *La distribuzione delle parti del discorso nel parlato e nello scritto,* in Van Deyck R. et Kabatèk J. (a cura di), *La variabilitè en langue, I. Langue et langue ècrite dans le present et dans le passè, II. Les quatre variations,* Grand, Communication & Cognition (Studies in Language, 8), pp. 261-284.

Voghera M. (2005) *La misura delle categorie sintattiche*, in Chiari Isabella/ De Mauro Tullio (a cura di) *Parole e numeri. Analisi quantitative dei fatti di lingua,* Aracne, Roma, pp 125-138.

Siti Web

Dell'Orletta Felice, Simonetta Montemagni , Giulia Venturi, Eva M. Vecchi, *DyLan Lab Laboratorio per lo studio delle dinamiche linguistico-cognitive.* http://www.ilc.cnr.it/dylanlab/

Nicola Mastidoro , Maurizio Amizzoni**,** *Linguistica applicata alla leggibilità: considerazioni teoriche e applicazioni.* http://www.swif.uniba.it/lei/sfi/bollettino/149_mastidoro_amizzoni (visitato il 13 Agosto 2011).