

Valutazione automatica della leggibilità nei testi per l'infanzia

Assiterm 2021 10 dicembre '21

Università degli Studi di Salerno – Italia

Dipartimento di Scienze Politiche e della Comunicazione

Lorenza Melillo e Alessandro Maisto.

Comprensione

- Il processo di comprensione di un testo parte dalla decodifica dello stesso, e si fonda sulle abilità linguistiche di base.



➤ Il processo di comprensione si basa su:

Conoscenze enciclopediche

Processi inferenziali

Abilità linguistiche

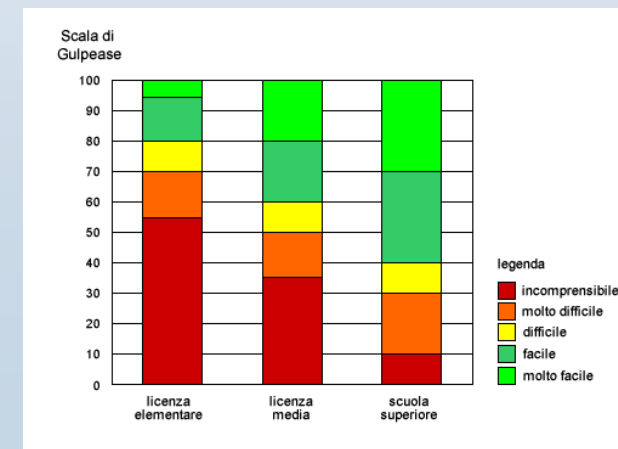
Leggibilità

Formula di Flesch

- L'indice di Flesch nato per l'inglese è stato rielaborato da Roberto Vacca, il quale, nel 1972, ha adattato i parametri della formula alla lingua italiana (vedi [Franchina-Vacca 1986]).

Indice Gulpease

- Creato da Lucisano e Piemontese, 1988.
- Si basa su: lunghezza delle frasi (in quantità di parole) e lunghezza delle parole (in quantità di lettere).





Comprensione e leggibilità

Comprensione	Leggibilità
Analisi qualitativa	Analisi quantitativa
Organizzazione logica e concettuale	Aspetti linguistici: Sintassi e lessico

Complessità lessicale e vocabolario di base

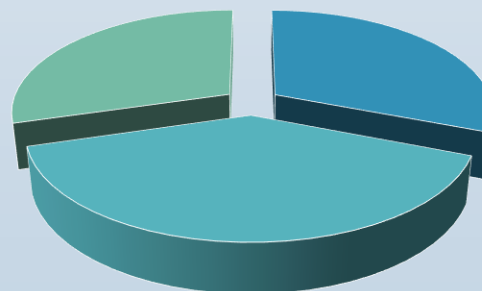
- Vocabolario di base della lingua Italiana (VdB)
- La lista dei vocaboli fu pubblicata per la prima volta nel 1980 in appendice al libro Guido all'uso delle parole ed è stata poi utilizzata in diverse opere lessicografiche come il Gradit.
- Il VdB include circa 7000 parole e si fonda sui dati di frequenza e sui dati di uso delle parole offerti dal Lif, il *Lessico di frequenza della lingua italiana contemporanea*, pubblicato dalla Ibm a Pisa nel 1970 (Zampolli et al.)



Table 1. Composizione del VDB (Gradit 1999-2007)

Marche d'uso	Vocaboli
FO – vocabolario fondamentale	2.077
AU – alto uso	2.663
AD – alta disponibilità	1.988

Lexemes



■ FO – fundamental vocabulary ■ AU – high usage ■ AD – high availability

Complessità lessicale e vocabolario di base

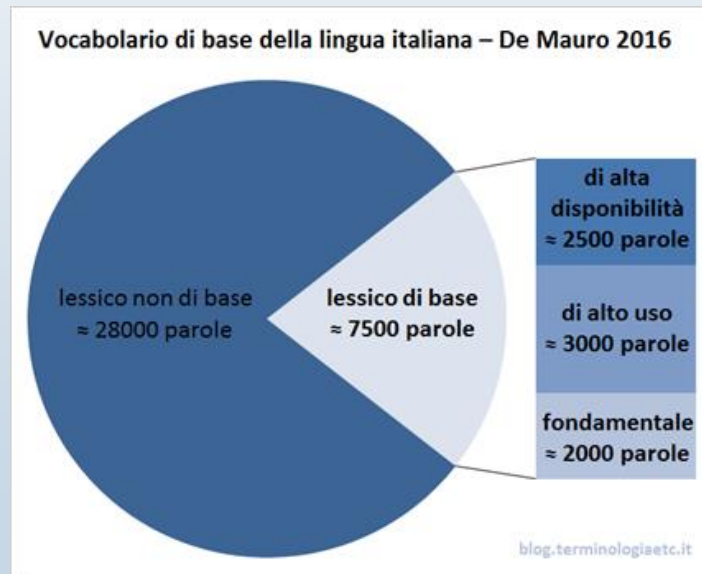
- Negli anni il VdB è stato utilizzato per marcare le parole, infatti i principali dizionari d'uso della lingua italiana hanno utilizzato le marche d'uso seppur con differenze di etichettatura.
 - a) marche d'uso del GRADIT (Grande dizionario dell'Italiano, dir. da T. De Mauro, 8 voll., Torino, UTET, 1999-2007)

FO	fondamentale	}	VOCABOLARIO di BASE
AU	alto uso		
AD	alta disponibilità		
CO	comune		
TS	legato ad un uso tecnico specialistico*		
LE	uso letterario		
RE	regionale		
DI	dialettale		
ES	esotismo		
BU	Basso uso		
OB	Obsoleto		

*Marca seguita dalla specifica di settore (stor., med., mus., arte etc)

Complessità lessicale e vocabolario di base

Dopo anni di ricerche nel 2016 esce la versione rinnovata del VdB il Nuovo Vocabolario di Base della lingua italiana (NVdB) (Chiari, De Mauro 2016) con 7.500 parole. L'NVdb è recentemente stato informatizzato (Elia et al. (2021)) per lo sviluppo di progetti di linguistica computazionale.



- Il VdB si basa su un corpus di 500.000 occorrenze di parole raccolte in cinque marche d'uso.
- Il NVdB si basa su un corpus di 18.000.000 occorrenze di parole che derivano dall'analisi di un corpus appositamente costruito di italiano contemporaneo e comprende multiwords.

Complessità lessicale e vocabolario di base

Internazionale

Ultimi articoli I più letti Sezioni ▼ Il

principe

s.m., agg.

in. XIII sec.; dal lat. princīpe(m) propr. "chi occupa il primo posto", comp. di primus "primo" e del tema di capĕre "prendere".

FO

1. s.m., chi ha una posizione preminente per autorità e potere in uno stato o vi esercita una sovranità di tipo monarchico: *il P. di Machiavelli* | *principe della Repubblica di Venezia*, il doge

2a. s.m., chi è investito del titolo nobiliare, originariamente proprio dei signori feudali appartenenti al primo ordine dopo l'imperatore, precedente quello di duca, attribuito spec. ai membri delle famiglie regnanti: *i principi di casa Savoia*; *vivere come un principe*, *stare da principe*, in modo molto agiato, lussuoso; *comportarsi come un principe*, in modo molto garbato, con maniere raffinate

2b. s.m., presunto successore legittimo in una monarchia

3a. s.m., estens., persona di grande autorità e prestigio, illustre esponente di un gruppo, di una comunità: *il principe dei romanzieri*; anche spreg.: *il principe dei ladri*

3b. s.m., persona ricca e agiata; anche, persona che si distingue per i modi raffinati: *essere un principe*

4. s.m. **TS** stor. in Roma antica, soldato di fanteria pesante

5. agg. **CO** principale, più importante: *l'argomento principe dell'accusa*

6. agg. **OB** il primo, il più antico

Polirematiche

edizione principe

loc.s.f.

TS filol.

→ editio princeps

principe azzurro

loc.s.m.

CO

1. nelle fiabe, il figlio del re, giovane e bello, che salva e sposa la protagonista

2. estens., lo sposo ideale sognato dalle ragazze

principe consorte

loc.s.m.

CO

il marito della regina quando non sia re

principe degli apostoli

loc.s.m.

CO

san Pietro

principe del foro

loc.s.m.

CO

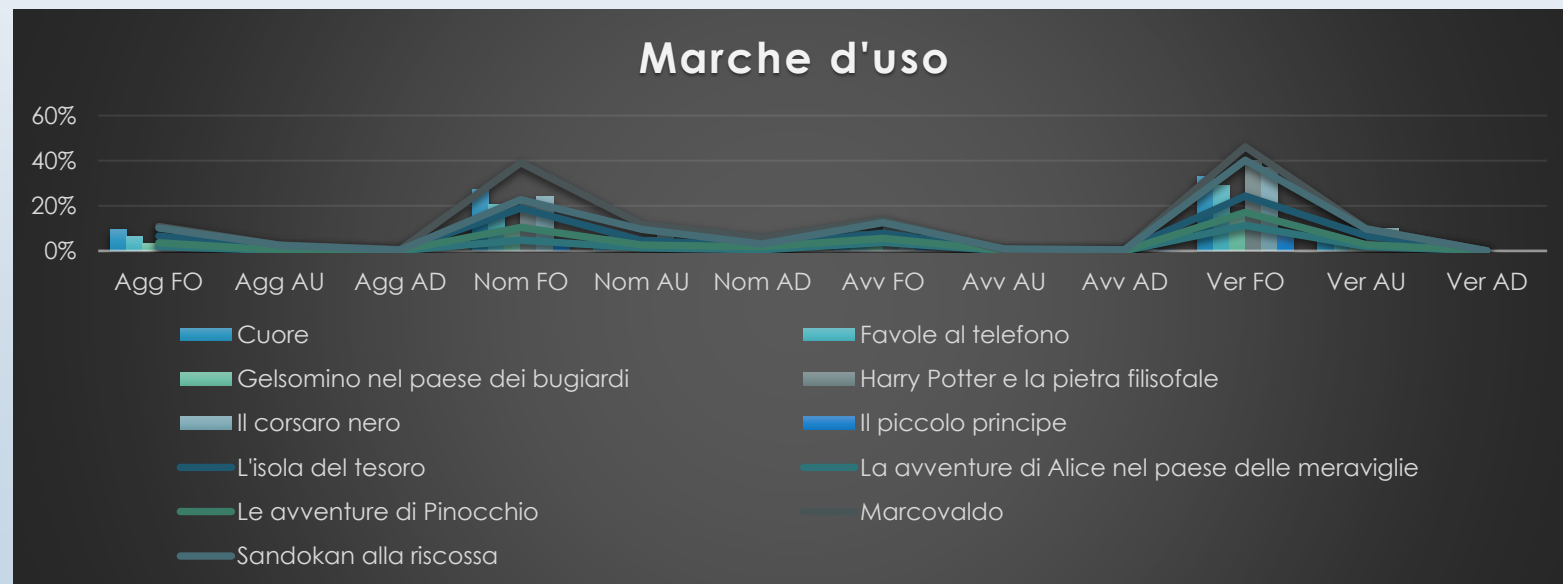
avvocato di grande notorietà e abilità

principe del sangue

loc.s.m.

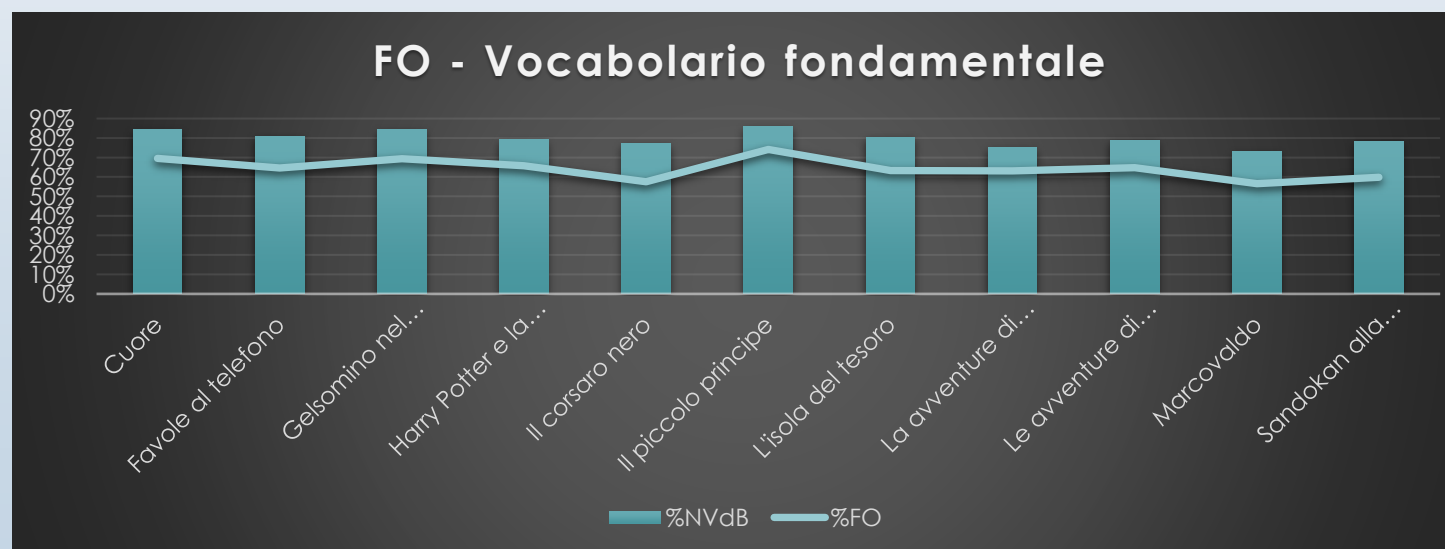
Valutazione automatica della leggibilità

- In questo lavoro presentiamo una ricerca che ha l'obiettivo di valutare, grazie a strumenti computazionali, il livello di leggibilità del lessico dei testi della letteratura per l'infanzia, proponendo un nuovo **Indice di Leggibilità** calcolato automaticamente sulla base di *feature* lessicali e sintattiche.



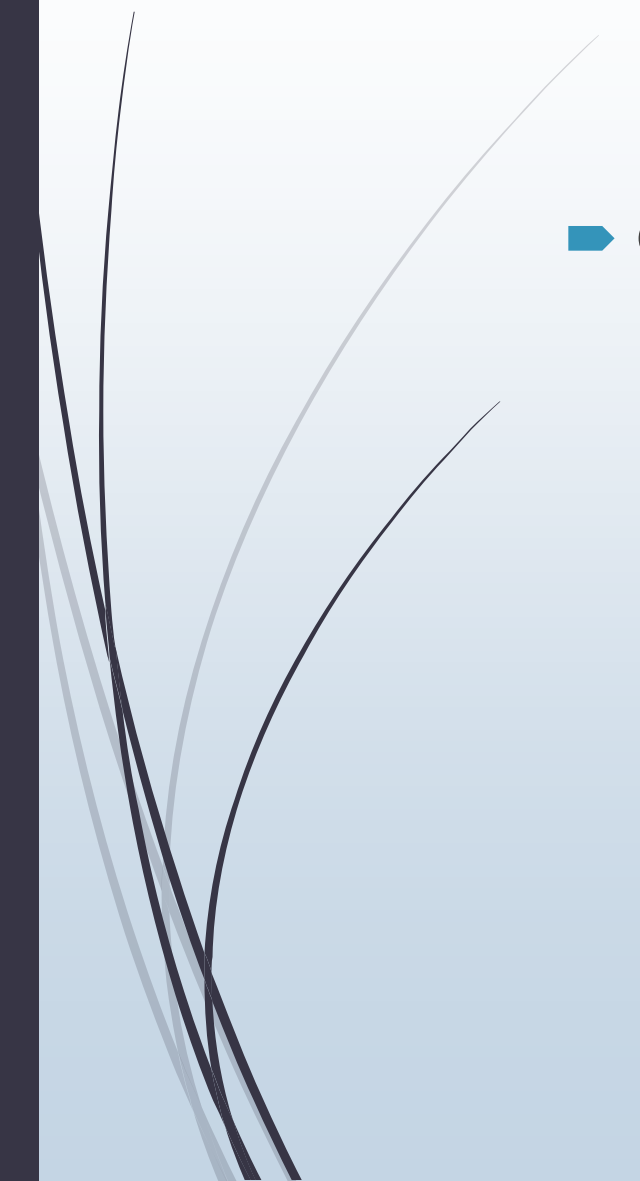
Valutazione automatica della leggibilità

- La letteratura per l'infanzia è caratterizzata, generalmente, da una scelta di vocaboli che non ostacoli la comprensione semantica.
- Come evidenziato da Tullio De Mauro nel VdB, questo tipo di letteratura dovrebbe utilizzare un gran numero di termini appartenenti al *lessico fondamentale*.



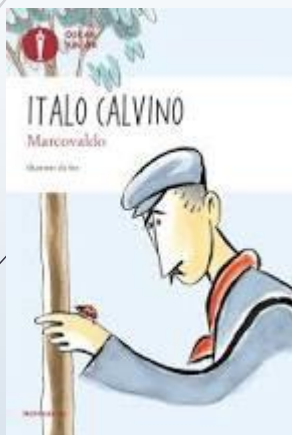


Obiettivi



- Gli obiettivi sono tre:
 - verificare la scelta dei vocaboli in relazione all'ipotesi De Mauro;
 - sperimentare un sistema di valutazione della complessità sintattica;
 - elaborare un sistema integrato automatico per produrre un Indice di Leggibilità lessicale e sintattico.

Struttura dell'analisi



- L'analisi che noi intendiamo portare avanti riguarda la comprensione dei testi da parte di bambini con una età compresa tra sette e dieci anni frequentanti la scuola primaria.
- La ricerca si è focalizzata sulla scelta di un corpus di undici testi integrali di letteratura per l'infanzia,
- I testi sono stati scelti tra i libri pubblicati e tradotti tra il 1865 e il 1997, anno di pubblicazione del primo libro della serie Harry Potter di Joanne Kathleen Rowling
- Inoltre sono stati selezionati dieci libri ad alta leggibilità tratti dalle collane *Il Mulino a Vento* e *Attacca Parole*, edite da Raffaello.



Struttura dell'analisi

- Il lavoro si è concentrato sull'attribuzione automatica di un **Indice di Leggibilità** originale basato su due diversi valori: la comprensibilità lessicale e la complessità sintattica.
- Ad una fase di pre-processing, comprendente POS Tagging, Lemmatizzazione e Parsing dei testi, è seguita una fase di *matching* dei termini appartenenti al NVdB basata sulla sua informatizzazione.
 - Elia A., Maisto A., Melillo L., Pelosi S. (2021) Lexical complexity and basic vocabulary of the Italian language In Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities 14th International Conference, NooJ 2020, Zagreb, Croatia, June 5–7, 2020, Revised Selected Papers
- Per il pre-processing è stato utilizzato **Tint** (The Italian NLP Tool), un software per l'elaborazione del linguaggio naturale (NLP) in italiano. Tint è un software open-source.



Struttura dell'analisi

- Comprensibilità lessicale:
 - Sono stati selezionati i termini Fondamentali, di Alto Uso e di Alta Disponibilità del NVdB, suddivisi per categoria grammaticale (Aggettivi, Avverbi, Nomi e Verbi) andando a formare 12 classi di termini.
 - Questi termini sono stati individuati all'interno dei testi ed è stato attribuito un peso differente ai termini delle diverse classi.
- Complessità Sintattica:
 - tramite un'analisi del risultato del *parsing* sono stati prese in esame diverse *feature* sintattiche come la presenza di subordinate, la linearità dei periodi, la presenza di modificatori ed incisi.

Indice di complessità sintattica

► Gli indicatori sintattici presi in esame sono i seguenti:

1. V_{mod} : include frasi subordinate (soggettive, preposizionali, infinitive, relative, avverbiali).
2. V_{punt} : presenza di coppie di segni di punteggiatura interne alla frase
3. V_{ordine} : ordine di frase diverso da Soggetto-Verbo-Complemento e presenza di soggetti non riconducibili ad un Sostantivo

Indice di complessità sintattica

- Il punteggio complessivo è calcolato con questa formula:

$$S_t = \frac{\sum S_{ft}}{N_{ft}}$$

- La complessità sintattica del testo è data dalla sommatoria delle complessità sintattiche di ogni frase del testo diviso per il numero di frasi del testo.



Indice di complessità sintattica

- ▶ Quando Gelsomino nacque la gente del paese si alzò nel cuor della notte, credendo di aver udito le sirene delle fabbriche che chiamavano al lavoro: era soltanto Gelsomino che piangeva per provare la voce, come fanno tutti i bambini appena nati. (*Gelsomino nel paese dei bugiardi*, Rodari)
- ▶ Alle otto e mezzo, il signor Dursley prese la sua valigetta ventiquattr'ore, sfiorò con le labbra la guancia della moglie, e tentò di dare un bacio a Dudley, ma lo mancò perché, in quel momento, in preda a un furioso capriccio, il pupo stava scagliando i suoi fiocchi d'avena contro il muro. (*Harry Potter e la pietra filosofale*, Rowling)

Indice di complessità sintattica

- ▶ Quando Gelsomino **nacque (advcl)** la gente del paese si alzò nel cuor della notte, **credendo (advcl)** di **aver udito (xcomp)** le sirene delle fabbriche che **chiamavano (acl:relcl)** al lavoro: era soltanto Gelsomino che **piangeva (acl:relcl)** per **provare (advcl)** la voce, **come fanno tutti (advcl)** i bambini appena nati.
 - ▶ Lunghezza: 45 elementi
 - ▶ V_mod: 7
 - ▶ V_punt : 2
 - ▶ V_ordine: 2
 - ▶ SL: 0.837, CS: 2.398: ILA: -1.56
- ▶ Alle otto e mezzo, il signor Dursley prese la sua valigetta ventiquattr'ore, **sfiorò (advcl)** con le labbra la guancia della moglie, e tentò **di dare un bacio (xcomp)** a Dudley, ma lo mancò perché, in quel momento, in preda a un furioso capriccio, il pupo stava **scagliando (advcl)** i suoi fiocchi d'avena contro il muro.
 - ▶ Lunghezza: 62 elementi
 - ▶ V_mod: 3
 - ▶ V_punt : 7
 - ▶ V_ordine: 0
 - ▶ SL: 0.931, CS: 1.872: ILA: -0.941

Indice di complessità sintattica

- La mia vita è monotona. Vado a caccia di polli, gli uomini cacciano me. Tutti i polli si somigliano, e tutti gli uomini si somigliano. Dunque mi annoio un po'. Ma se tu mi addomestichi, nella mia vita ci sarà un sole. Riconoscerò un rumore di passi che sarà differente da qualsiasi altro. Gli altri passi mi faranno nascondere sotto terra, il tuo mi chiamerà fuori dalla tana, come una musica.
 - Semplicità Lessicale: 0.642
 - Complessità Sintattica: 0.651
 - Totale: -0,008
- La mia vita è monotona perché vado a caccia di polli e gli uomini cacciano me, ma tutti i polli si somigliano, e tutti gli uomini si somigliano, dunque mi annoio un po'. Ma se tu mi addomestichi, nella mia vita ci sarà un sole e riconoscerò un rumore di passi che sarà differente da qualsiasi altro perché gli altri passi mi faranno nascondere sotto terra, mentre il tuo mi chiamerà fuori dalla tana, come una musica.
 - Semplicità Lessicale: 0.642
 - Complessità Sintattica: 2.014
 - Totale: -1.373

Indice di Semplicità Lessicale

- Si parte dal matching di Nomi, Aggettivi, Avverbi e Verbi inclusi nel NVdB all'interno di un testo.
- Se una parola è ambigua o appartiene a più categorie, viene applicata la seguente scala di priorità per determinare a quale marca appartenga:

- FO - AU - AD

- Per ogni testo, la **semplicità lessicale** è determinata dalla seguente formula:

$$L_t = 2 * \log \frac{(2 * A_{FO} + N_{FO} + AVV_{FO} + V_{FO}) + A_{AU} + N_{AU} + AVV_{AU} + V_{AU} + A_{AD} + N_{AD} + AVV_{AD} + V_{AD}}{NVdB_t}$$

- L'indice di semplicità lessicale è uguale al doppio del logaritmo delle parole FO del testo per due, più le parole AU e AD del testo, fratto il totale delle parole del NVdB trovate nel testo.
- Punteggio massimo: $2 * \log(2) = 1.386295294$

Indice di Leggibilità Automatico

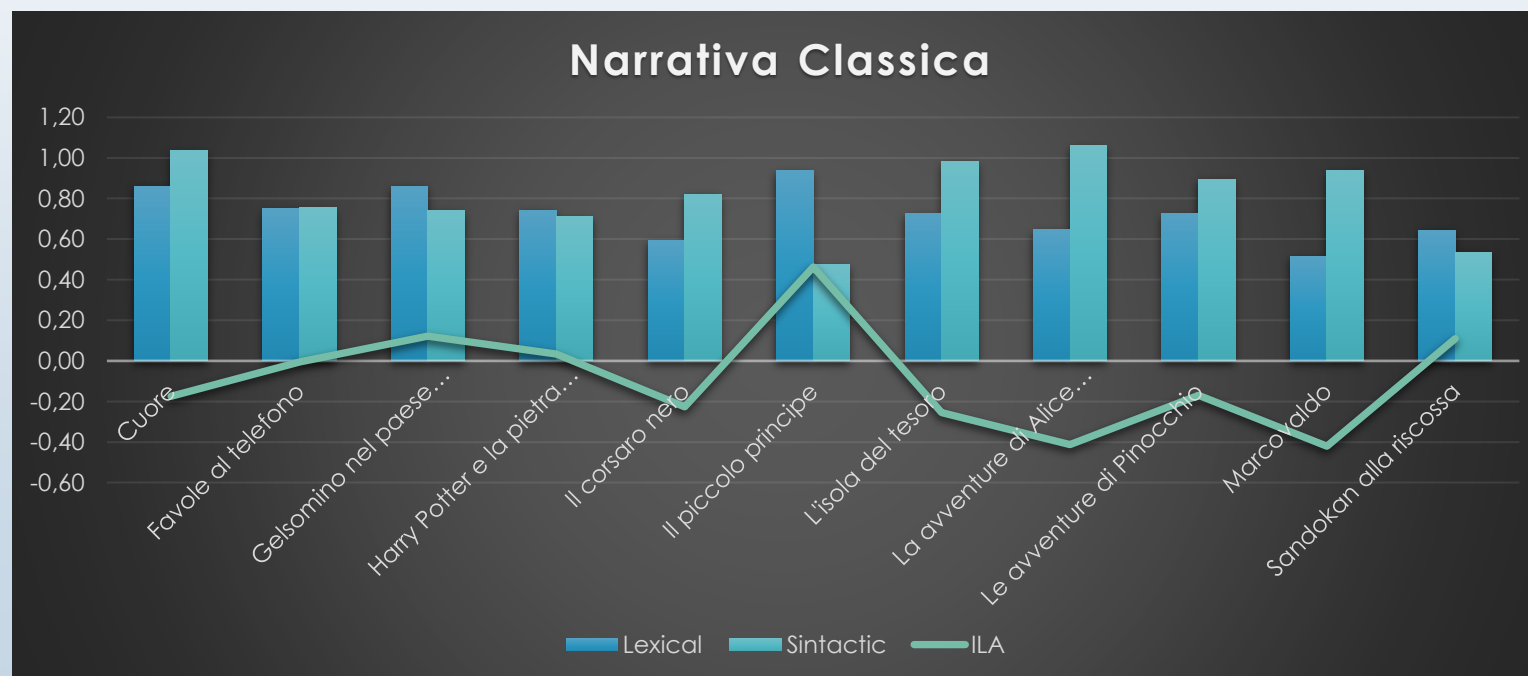
- I valori così ottenuti sono stati composti in un unico **Indice di Leggibilità** del testo, sperimentato sugli undici testi di letteratura per l'infanzia e sui dieci testi ad alta leggibilità.
- I risultati della sperimentazione hanno evidenziato valori coerenti con i risultati attesi. Alcuni libri sono risultati particolarmente complessi sia a livello lessicale che sintattico (ad esempio, *Marcovaldo*) ed in generale ai testi considerati ad alta leggibilità è stato attribuito un punteggio di leggibilità particolarmente alto.

Calcolo dell'ILA

$$L_t - S_t$$

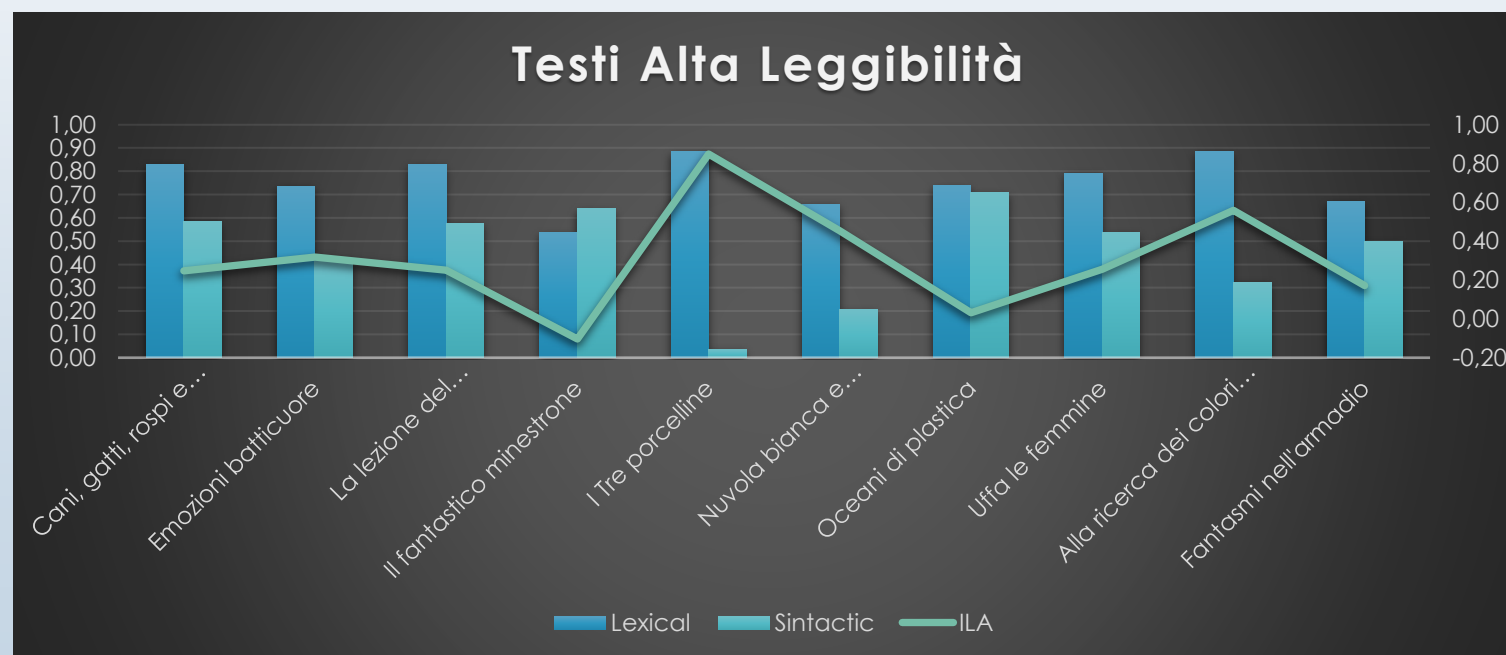
Indice di Leggibilità Automatico

- Nel nostro Indice di Leggibilità abbiamo integrato il modello di comprensione lessicale con altri strumenti di valutazione come la complessità sintattica.



Indice di Leggibilità Automatico

- Nel nostro Indice di Leggibilità abbiamo integrato il modello di comprensione lessicale con altri strumenti di valutazione come la complessità sintattica.





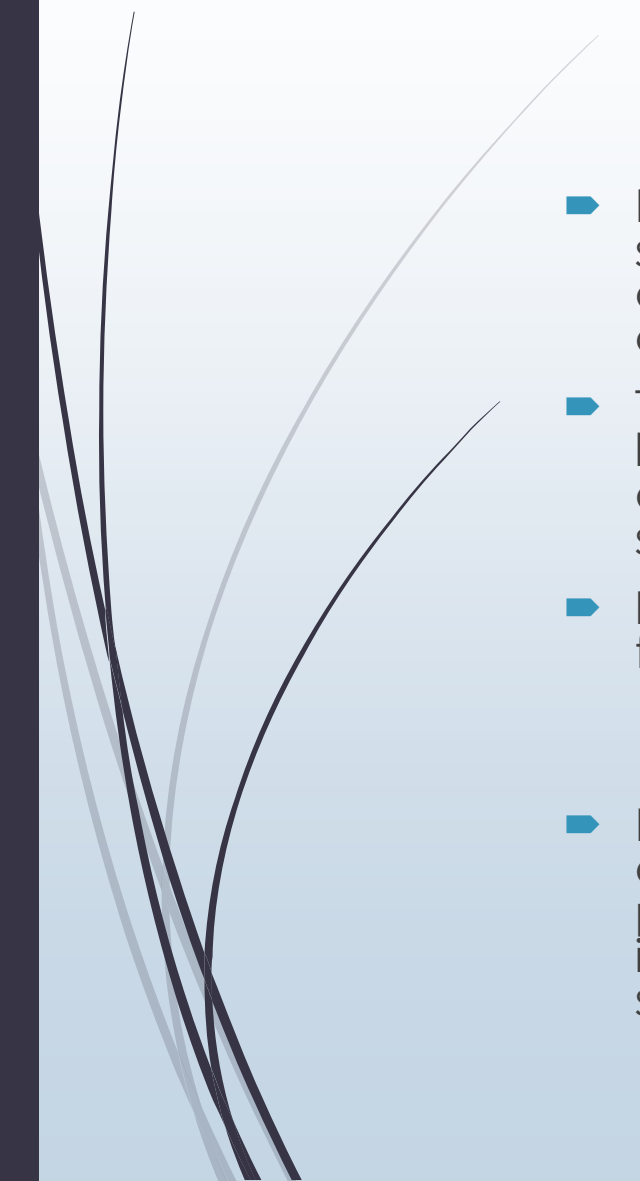
Conclusioni



- ▶ I dati che vengono fuori da questa analisi ci mostrano come il corpus ci dia un feedback sulla frequenza delle parole del vocabolario fondamentale (FO) completamente coerente con la nostra ipotesi di partenza
 - ▶ infatti i libri ad alta leggibilità dedicati esclusivamente ai bambini della scuola primaria hanno una elevatissima percentuale di termini appartenenti al *lessico fondamentale*.
- ▶ Inoltre i dati ci permettono di evidenziare che la complessità sintattica non sia associabile a mutamenti della lingua sull'asse diacronico, ma piuttosto a differenze funzionali.



Lavoro futuro



- Per validare i risultati ottenuti fino ad ora abbiamo cominciato una sperimentazione che si sta svolgendo presso una Scuola Primaria con il coinvolgimento di una III, IV e V classe. I testi per la comprensione sono tratti dal corpus sottoposto all'analisi linguistica.
- Tale necessità nasce dalla consapevolezza che la comprensione, non solo sia legata alla complessità di un testo ma sia legata al recupero e all' integrazione di tutte le conoscenze che il lettore possiede per la rielaborazione del testo stesso.
- Il processo di comprensione di un testo sia negli adulti che nei bambini è fortemente legato alle conoscenze enciclopediche.
 - Dunque l'imaturità di queste conoscenze nei bambini, la qualità inferiore e il difficile recupero rendono più complessa la loro comprensione.
- Nell'applicare le conoscenze enciclopediche i bambini tendono ad utilizzare quelle più familiari e immediatamente disponibili e solo nel corso della scuola primaria, grazie allo sviluppo di diverse abilità cognitive, riusciranno a realizzare inferenze. Pertanto si ritiene necessario effettuare parte del lavoro in ambito scolastico.



References

- Bertolini C. (2012) Senza Parole Promuovere la comprensione del testo fin dalla scuola dell'infanzia. Edizioni junior – Spiaggiari edizioni Srl
- Chiari I., De Mauro T. (2016) Nuovo vocabolario di base della lingua italiana. Casa Editrice Sapienza.
- De Mauro T. (1980), "Il vocabolario di base della lingua italiana", in Guida all'uso delle parole, Roma: Editori Riuniti, pp. 146-172.
- Elia A., Maisto A., Melillo L., Pelosi S. (2021) Lexical complexity and basic vocabulary of the Italian language In Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities 14th International Conference, NooJ 2020, Zagreb, Croatia, June 5–7, 2020, Revised Selected Papers
- Lumbelli L. (2009) La comprensione come problema. Il punto di vista cognitivo Laterza

Grazie per la vostra attenzione !

lmelillo@unisa.it, amaisto@unisa.it