Trattamento automatico della lingua a supporto dell'editoria: primi esperimenti con il Devoto-Oli Junior

Irene Dini[1], **Felice Dell'Orletta**[1], **Fabio Ferri**[2], **Biancamaria Gismondi**[2], **Simonetta Montemagni**[1]
1. Istituto di Linguistica Computazionale "A. Zampolli" – CNR
2. Mondadori Education
{irene.dini,felice.dellorletta,simonetta.montemagni}@ilc.cnr.it
{fabio.ferri,bianca.gismondi}@mondadori.it

Abstract

English. The paper illustrates the results of a first experiment in which Natural Language Processing was used to support the revision of a children's dictionary, in particular for what concerns style and wording of definitions and the enrichment of the list of lemmas. The results achieved are promising and demonstrate the potential of a synergy to be strengthened in the publishing sector.

Italiano. *L'articolo illustra i risultati di un esperimento all'interno del quale tecnologie di TAL sono state utilizzate a supporto della redazione di un dizionario per bambini, in particolare per quanto riguarda la formulazione delle definizioni e l'aggiornamento del lemmario. I risultati raggiunti sono promettenti e mostrano il potenziale di una sinergia da rafforzare nel settore dell'Editoria.*

1 Introduzione

La consapevolezza delle potenzialità di metodi e tecniche di Intelligenza Artificiale (IA) nel settore dell'Editoria sta diffondendosi rapidamente. Il libro bianco su *The Future Impact of Artificial Intelligence on the Publishing Industry* (2019) riporta i risultati di un'indagine internazionale dalla quale emerge che il 25% delle case editrici intervistate ha già investito in applicazioni di tecniche di IA all'interno di diversi settori, che spaziano dal marketing e la distribuzione alla produzione editoriale.

All'interno dello scenario appena delineato, un ruolo centrale è svolto da metodi e tecniche per il Trattamento Automatico della Lingua (TAL), che sono oggi mature per poter contribuire in modo significativo alle diverse fasi del processo editoriale, permettendo - ad esempio - di indicizzare su base semantica il contenuto informativo di un testo, di monitorarne la complessità e l'efficacia comunicativa in relazione alla tipologia dei destinatari, di guidare la sua eventuale riformulazione, di verificare l'eventuale presenza di plagi, oppure di fornire supporto alle fasi di controllo linguistico e tipografico.

In questo contributo, riportiamo i risultati di un primo e promettente esperimento condotto congiuntamente dalla Casa editrice Mondadori Education e dall'Istituto di Linguistica Computazionale del CNR, all'interno del quale tecnologie di TAL sono state utilizzate a supporto della progettazione della nuova edizione di un dizionario per bambini: il *Devoto-Oli Junior* (DJ). In particolare, sono stati affrontati i temi del controllo, della valutazione e della specializzazione del dizionario rispetto alla platea dei destinatari, cercando di conciliare due prospettive apparentemente in contrasto, l'accessibilità dei contenuti da un lato e la loro informatività dall'altro.

2 Il prodotto Dizionario

Una Casa editrice ha con il proprio dizionario un rapporto complesso: opera di notevole impegno redazionale ed economico; pubblicazione di prestigio e, come usa dire, di *brand positioning*; prodotto con diffusione e profitti calanti.

Sperimentare — modi, tempi, target — sbagliando è un lusso che appartiene al passato; da qui l'esigenza di un approccio più certo, più rapido, senza sprechi: dunque, scientifico-tecnologico. E, come sarà descritto meglio sotto, il TAL avvantaggia una Redazione lessicografica

nella costruzione del lemmario, anche in relazione al target di mercato.

Quale che sia l'impostazione lessicografica — positivista, storico-linguistica o dal sapore valenziale —, la scelta del lemmario è, prima di ogni cosa, una faccenda di marketing: non vi è dizionario oggi sul mercato che non sbandieri numero di voci o lemmi, di significati, di neologismi.

È facile comprendere che, al momento dell'acquisto, un lemmario sterminato sia garanzia della capacità del dizionario stesso di risolvere i nostri problemi (almeno quelli lessicali e ortografici, s'intende). Eppure la seduzione di un *universo per ordine alfabetico* si scontra con due ineluttabili problemi industriali: il numero di pagine e il costo. Un libro, inteso come oggetto fisico, ha una sua ergonomicità e ci sono limiti fisici oltre i quali le operazioni di rilegatura divengono insensate e la consultazione sgradita. Vi è poi un rapporto matematico diretto — come ricordano incessantemente i Direttori commerciali — tra numero di pagine e costo: nel mondo di Google e Wiktionary, il prezzo è un affare assai delicato, se non per gli acquirenti istituzionali, certo per le famiglie.

Una sfida particolare è poi un dizionario con un target scolastico di riferimento: se infatti un vocabolario dell'uso ha ambizioni totalizzanti, un vocabolario per la scuola è un'operazione ontologicamente editoriale in quanto si fonda sulla capacità di scegliere e ritagliare un mondo linguistico plausibile e utile.

Operazione non così banale qualora si consideri l'ambivalenza della lingua a cui gli studenti sono esposti: da un lato, lessico di base che impiegano con maggior o minor *proficiency*; dall'altro, lessico disciplinare tecnico e tecnico-scientifico di cui sono comprensibilmente ricchi i testi scolastici (*onnisciente*, *antagonista*, *esarcato*, *tettonica*, *fosfolipidico* ecc.). E, come è facile immaginare, questa ambivalenza investirà sia la scelta delle voci sia la costruzione della singola definizione.

3 La nuova edizione di un dizionario

La progettazione della nuova edizione del DJ si è concentrata su due questioni principali:

i. il linguaggio utilizzato nelle definizioni, la sua complessità ed effettiva accessibilità per l'utenza a cui l'opera è destinata, ovvero bambini in età compresa tra gli 8 e i 13 anni;
ii. il lemmario, la sua verifica e il suo aggiornamento a distanza di quasi dieci anni dalla prima edizione, data alle stampe nel 2012.

3.1 La complessità del linguaggio

La scrittura delle definizioni è un punto cruciale e, in genere, molto caratterizzante di questo tipo di opere. Fin dalla prima edizione quindi ci si è molto concentrati su questo aspetto. Definire le parole, sia quelle comuni e "di base" sia quelle meno comuni, più specialistiche o elevate, con altre parole semplici e accessibili a un'utenza con competenze linguistiche in fase evolutiva richiede molte scelte e un piano di scrittura ben definito.

Dal punto di vista lessicale, in prima battuta, è sembrato naturale cercare di definire le parole selezionate utilizzando soltanto le ca. 7.000 voci del *Vocabolario di Base* (VdB) di Tullio De Mauro. Tuttavia, questo metodo ha mostrato presto i suoi limiti, soprattutto quando si è trattato di definire voci o significati tecnico-scientifici. Inoltre, come è emerso nelle interviste effettuate su campioni significativi di insegnanti, i docenti cercano in un dizionario uno strumento didattico che in primo luogo consenta loro di aumentare le competenze lessicali degli alunni, oltre che potenziare quelle già possedute.

Da qui la scelta di utilizzare nelle definizioni qualche parola in più rispetto a quelle del VdB. Coerentemente con questa decisione, ad esempio, nelle definizioni esclusivamente sinonimiche, tipiche degli aggettivi dove l'uso delle perifrasi spesso complica e appesantisce la spiegazione del significato, sono state impiegate triplette di parole, organizzate in un climax che procede dalla parola semanticamente più vicina al lemma a quella più lontana, ma anche da quella più comune a quella più elevata. Purtroppo non sempre, però, i due criteri coincidono, per cui talvolta la parola a più alta complessità lessicale è anche la prima, essendo quella più vicina di significato.

Un altro esempio ci viene fornito dai demotici, una classe chiusa di lemmi per le cui definizioni in genere si approntano delle formule fisse. Proprio a causa della loro ripetitività, queste voci sono sembrate quelle giuste per azzardare l'uso di una parola non comune come *nativo*, inserita nella breve definizione formulare "Abitante, nativo di…", contando anche sulla trasparenza del termine *nativo*, facilmente collegabile a *nato*. Così, in lemmi come *napoletano* troviamo definizioni brevi, come appunto "Abitante, nativo di Napoli", che introducono l'utente a una parola nuova.

C'è poi il problema della complessità sintattica delle definizioni, che merita una riflessione

preliminare. Le definizioni dei lemmi di un dizionario obbediscono a regole precise (verbi definiti con verbi, sostantivi con sostantivi, aggettivi con aggettivi o perifrasi attributive, ecc.): Inoltre, per ragioni di spazio, le frasi definitorie sono spesso ellittiche; nel DJ i due casi più frequenti di definizioni ellittiche sono: i) "Abitante di Napoli", dove il determinante è privo di determinato; ii) nei verbi intransitivi, è spesso indispensabile specificare chi è il potenziale soggetto, utilizzando formule tipo "Di mezzo di trasporto, procedere".

Per quanto si sia cercato di evitare le formule ellittiche più pesanti, è chiaro che la complessità sintattica di queste frasi costituisce una delle questioni più spinose da affrontare.

3.2 Il lemmario

I dizionari pensati per questo target sono in genere costituiti da un numero di voci compreso tra un minimo di ca. 15.000, come il *Dizionario Italiano di Base* di Tullio De Mauro (DIB), e un massimo di ca. 23/25.000, come il DJ. Si tratta quindi di repertori lessicografici estremamente selettivi, risultato di scelte molto meditate.

Nel caso del DJ, si è partiti dalle ca. 7.000 voci del VdB, che includono 1.991 parole fondamentali, ca. 2.750 di alto uso e ulteriori 2.337 appartenenti al vocabolario ad alta disponibilità. Grazie a questo primo nucleo, fin dalla prima edizione del DJ sono stati poi lemmatizzati:

i. i derivati più comuni delle 7.000 parole non compresi nel VdB, in modo da fornire agli studenti famiglie di voci il più possibile complete;
ii. molti sinonimi o contrari, utili per collocare ciascun lemma all'interno di una rete cognitiva di collegamenti che ne favorisca la reciproca comprensione e memorizzazione;
iii. i termini non inclusi tra i lemmi del VdB, ma necessari per definirli senza dover ricorrere a complicati giri di parole. Com'è noto, infatti, un dizionario è un sistema chiuso, per cui ogni parola utilizzata per definire deve essere a sua volta definita all'interno dell'opera.

Tuttavia, parole come *sostantivo*, *transitivo* o *coordinata* e *sottrazione*, non incluse nel VdB, rischiavano di non rientrare nel corpus del DJ anche seguendo gli altri criteri individuati. Termini specialistici e disciplinari "di base" come questi non potevano non essere presenti in un dizionario progettato per essere impiegato da insegnanti della scuola primaria e della secondaria di primo grado. L'individuazione dei termini settoriali adatti a questa utenza per numero e livello di specializzazione è dunque il vero nodo da sciogliere. In occasione della prima edizione la soluzione è stata trovata facendo lo spoglio dei manuali delle varie materie della scuola secondaria di primo grado corredati da glossari, un metodo che richiede un considerevole dispendio di risorse e non garantisce risultati soddisfacenti.

4 Il ruolo del TAL nella revisione del DJ

Nella progettazione della nuova versione del DJ, sono state utilizzate tecniche avanzate di TAL a supporto i) del controllo e possibile riformulazione delle definizioni, e ii) della revisione ed eventuale integrazione del lemmario. Le analisi sono state condotte sull'intero corpus dei dati del dizionario in formato XML, per un totale di più di 23.000 lemmi a cui sono associate più di 41.000 definizioni. Come passo preliminare, il corpus delle definizioni è stato linguisticamente annotato con LinguA (Dell'Orletta, 2009; Attardi e Dell'Orletta, 2009; Attardi et al., 2009). I livelli di annotazione alla base delle elaborazioni che seguono sono quello morfo-sintattico e lemmatizzazione, e sintattico a dipendenze.

4.1 Analisi delle definizioni

L'analisi delle definizioni ha riguardato due facce delle complessità linguistica, quella lessicale e quella sintattica. Attraverso questo tipo di analisi è stato possibile identificare quali definizioni contenessero termini e/o strutture sintattiche di difficile comprensione.

La complessità lessicale della definizione è stata calcolata in funzione della complessità lessicale delle parole semanticamente piene che vi ricorrono, sia nella forma in cui effettivamente compaiono, sia in relazione al lemma associato. Numerosi sono i fattori che contribuiscono a rendere un termine complesso, che spaziano dalla frequenza, al grado di ambiguità o di astrattezza, alla lunghezza, per menzionarne solo alcuni (cfr. Shardlow et al. (2021) per una rassegna delle caratteristiche connesse alla complessità lessicale). Seguendo Rayner e Duffy (1986), in questo esperimento ci siamo focalizzati sul fattore frequenza.

La complessità dei termini all'interno delle definizioni è stata calcolata in riferimento a un dizionario di frequenza organizzato in classi costruito a partire dal corpus itWaC (Baroni et al., 2009), ad oggi il corpus più esteso esistente per

l'italiano. La classe di frequenza di ciascun termine è stata calcolata in base al corpus utilizzando la seguente funzione:

$$C_{CT} = \lfloor log_2 \frac{freq(MFT)}{freq(CT)} \rfloor$$

dove MFT è il termine più frequente del corpus, CT è il termine considerato e *freq* è una funzione che associa ad un termine la sua frequenza assoluta nel corpus (Richter et al., 2015). Le classi di complessità sono state definite in relazione alle forme e ai lemmi: sono 27 per i lemmi (da 0 a 26) e 26 per le forme (da 0 a 25). Partendo dall'assunto che termini di uso comune vengono considerati semplici mentre termini utilizzati raramente vengono considerati difficili, alla classe 0 appartengono i termini (forme o lemmi) più frequenti e quindi più comprensibili, mentre alle classi 25 e 26 appartengono i termini (rispettivamente forme e lemmi) più rari e più difficili.

Oltre alla complessità lessicale, per ogni definizione è stato calcolato un punteggio di complessità sintattica, utilizzando READ-IT (Dell'Orletta et al., 2011), il primo strumento per la valutazione della leggibilità di testi in italiano basato su TAL. READ-IT si basa su un'analisi sofisticata delle strutture linguistiche sottostanti al testo e articolata su diversi livelli di descrizione linguistica. Per calcolare la complessità sintattica READ-IT si basa su un ampio spettro di tratti linguistici (in particolare morfo-sintattici e sintattici desunti a partire dall'annotazione linguistica condotta preliminarmente). La complessità è espressa con un valore compreso tra 0 (semplice) e 1 (difficile).

4.2 Revisione del lemmario

La revisione del lemmario del DJ è stata condotta attraverso una verifica interna volta a identificare se c'erano termini usati nelle definizioni il cui lemma non era definito nel dizionario, e una verifica rispetto a risorse esterne. Come risorse esterne sono stati usati:

- il lemmario del *Nuovo Vocabolario di Base* di Tullio De Mauro (NVdB), pubblicato nel 2016, oltre trent'anni dopo la prima versione (1980), con l'aggiunta di ca. 1.000 parole;
- il lemmario costruito automaticamente a partire dall'analisi di un corpus di testi per bambini selezionati all'interno della produzione scolastica Mondadori, che comprende l'intero curriculo della Scuola Primaria affiancato dalla cosiddetta parascolastica e da libri di narrativa.

Se l'aggiornamento rispetto al NVdB ha riguardato il lessico comune, l'integrazione rispetto al lemmario estratto dal corpus scolastico Mondadori ha invece comportato un aggiornamento terminologico settoriale, dal momento che il corpus, basato sulla produzione del II ciclo della Scuola Primaria, include libri di lettura e sussidiari antropologici e scientifici.

Nel caso della verifica interna (rispetto al corpus delle definizioni) e quella esterna (rispetto al corpus scolastico Mondadori) sono stati utilizzati lemmari costruiti in modo automatico a partire dall'annotazione morfo-sintattica e dalla lemmatizzazione. Confrontando la lista dei lemmi del dizionario e i lemmari di riferimento (VdB e quelli costruiti automaticamente) è stato possibile identificare i lemmi da valutare per l'eventuale inserimento nel nuovo DJ. Questo tipo di analisi ha portato a identificare più di 160 lemmi del NVdB che non facevano parte del lemmario del DJ, e circa 150 lemmi di parole che ricorrevano nel corpus delle definizioni ma non erano definiti. Più consistente è il numero di lemmi ricavati dall'analisi del corpus scolastico Mondadori, che ovviamente richiede un'analisi attenta mirata a discriminare la terminologia settoriale rilevante per un dizionario per bambini.

5 Elaborazioni: alcuni esempi

5.1 Complessità lessicale

Dopo aver associato le classi di complessità a tutte le parole piene, a ogni definizione sono stati assegnati 4 diversi indicatori di Complessità Lessicale (CL) riguardanti i) la CL dei termini più complessi che vi ricorrono, e ii) la media dei valori di CL di tutte le parole piene all'interno della definizione. In entrambi i casi, il valore di CL è stato calcolato in relazione sia alla forma che al lemma.

La Tabella 1 esemplifica gli indicatori di CL associati ad alcune definizioni. I valori associati a Max $CL_{f/l}$ consentono di identificare definizioni in cui compaiono termini particolarmente difficili (CL_f riguarda le forme e CL_l i lemmi) di cui va valutata una possibile sostituzione con termini più semplici. D'altro canto, i valori associati a Media $CL_{f/l}$ forniscono una misura globale della complessità lessicale della singola definizione, calcolata come la media delle classi di complessità di tutte le parole piene della definizione. Le ultime due colonne della tabella esplicitano la forma/lemma corrispondente al

valore Max $CL_{f/l}$: è interessante notare come i valori di forma e lemma più difficili possano far riferimento a termini diversi (cfr. definizione del lemma *antipatia*).

Con questo tipo di analisi sono state identificate le definizioni con un alto grado di CL che richiedevano una revisione. Per esempio, nella definizione di *orda* la parola "scalmanate", con CL=19, è stata sostituita con la parola "agitate" (CL=14), rendendo così la definizione maggiormente comprensibile. Nel caso di una definizione sinonimica come quella di *adombrarsi*, è emerso che le classi associate a "offendersi" e "risentirsi" appartengono alla classe di complessità 14, mentre "indispettirsi" alla classe 20. Il sinonimo associato alla classe più alta di CL è stato quindi retrocesso in ultima posizione dopo quelli più usuali, rispettando il climax previsto.

Ci sono poi casi in cui il lessicografo ha ritenuto opportuno non intervenire per diversi ordini di motivi. Ad esempio, perché la definizione conteneva tecnicismi non sostituibili, nonostante ad alto grado di difficoltà di comprensione, come nel caso della definizione di *ovulazione* riportata in tabella.

Termine	**Definizione**	**Max CL_f**	**Media CL_f**	**Max CL_l**	**Media CL_l**	**Forma con max CL_f**	**Lemma con max CL_l**
adombrarsi	Offendersi, indispettirsi, risentirsi.	20	9	16	8,2	indispettirsi	indispettire
antipatia	Sentimento di avversione istintiva.	14	12,3	14	12,3	istintiva	avversione
orda	Insieme di persone rumorose e scalmanate.	19	11,7	17	12,5	scalmanate	scalmanato
ovulazione	Uscita dall'ovario dell'ovulo pronto per la fecondazione.	17	12	18	12,2	ovario	ovario

Tabella 1: Indicatori di complessità lessicale associati a ogni definizione

La Tabella 2 riporta, per ciascuna categoria grammaticale, le medie dei 4 punteggi di CL associati a ogni definizione. Congiunzioni e avverbi risultano essere le categorie grammaticali le cui definizioni sono complessivamente più semplici. Nomi, verbi, aggettivi, pronomi, articoli e interiezioni risultano invece caratterizzati da definizioni maggiormente complesse.

Classe grammaticale	**Max CL_f**	**Media CL_f**	**Max CL_l**	**Media CL_l**
Aggettivo	12	8,4	11,9	8,5
Articolo	12,8	9,8	13	9,4
Avverbio	10,3	8,2	10,5	8,5
Congiunzione	9,9	7,6	10,2	8
Interiezione	12,6	9,6	12,3	9,6
Nome	12,7	9,3	12,6	9,3
Preposizione	11	8,5	11,1	8,8
Pronome	13,1	8,1	13,2	8,3
Verbo	12,4	9,4	11,6	8,9

Tabella 2: CL media per categoria grammaticale

5.2 Complessità sintattica

Grazie ai punteggi di READ-IT assegnati per il livello sintattico, è stato possibile individuare costruzioni ricorrenti di difficile comprensione. In questo studio preliminare, READ-IT è stato usato nella sua versione corrente, addestrata su testi di tipo giornalistico, per cui i punteggi assegnati a definizioni vanno considerati come indicativi, ma non specializzati rispetto alle peculiarità del linguaggio delle definizioni. Nonostante ciò, è stato possibile identificare definizioni contenenti costruzioni complesse da valutare per un'eventuale riformulazione semplificata, ad es. quelle introdotte da sintagmi preposizionali che ne circoscrivono il dominio o il significato. Per esempio, la definizione di *andare* "Di mezzo di trasporto, procedere" ha associato un indice di complessità sintattica (CS) di 0,36 che si è ridotto significativamente trasformandola in "Detto di mezzo di trasporto, procedere" (CS=0,04). Un altro esempio è costituito dalle definizioni dei demotici come *canadese* la cui definizione è passata dalla forma ellittica "Del Canada" alla forma "Relativo al Canada".

6 Conclusioni

In un dizionario della lingua d'uso il parlante deve potersi rispecchiare, perché è contemporaneamente la fonte e il destinatario dell'opera. Se questo è vero per qualsiasi dizionario, a maggior ragione lo è per quelli rivolti al mercato della Scuola Primaria, nei quali tutto deve essere a misura di bambino: le dimensioni del volume e il prezzo, perché la Primaria è la scuola dell'obbligo per eccellenza, il livello di complessità della lingua, che deve essere proporzionato alle conoscenze e ai bisogni dei bambini e dei loro insegnanti. Dal momento che le esigenze sono tanto particolari, in un'opera come il DJ, dunque, è fondamentale l'impiego di tecniche di produzione che siano efficienti. Le tecnologie TAL hanno risposto perfettamente a questa richiesta di efficientamento. La verifica del lemmario esistente mediante lo spoglio di ampi corpora mirati sul target, la classificazione della complessità lessicale e sintattica delle definizioni individuata attraverso l'impiego di uno strumento come READ-IT, l'individuazione delle nuove voci da inserire grazie all'uso incrociato di tutte queste tecniche hanno prodotto in tempi brevi risultati certi e attendibili. Soprattutto hanno consentito al lessicografo di lavorare su obiettivi circoscritti e gerarchizzati, conciliando la prospettiva dell'accessibilità con quella dell'informatività dell'opera. Il lavoro fianco a fianco di redattori e ricercatori, inoltre, ha aperto nuovi ambiti di sperimentazione e di riflessione, come la ricerca di nuovi modelli definitori, più accessibili rispetto a quelli tradizionali.

Bruno Migliorini, ormai molti decenni fa, chiudeva la sua nitida prosa sul vocabolario con un'affermazione sfiduciata: «sull'avvenire della lessicografia italiana non è possibile far presagi». Oggi, grazie a esperimenti come questo, siamo in grado di dire qualcosa di più: il TAL non potrà non essere parte di questo avvenire.

Bibliografia

Giuseppe Attardi and Felice Dell'Orletta. *"Reverse Revision and Linear Tree Combination for Dependency Parsing"*. In: NAACL-HLT 2009 – North American Chapter of the Association for Computational Linguistics – In Proceedings of Human Language Technologies. Association for Computational Linguistics. June Boulder, Colorado, pp. 261 – 264 (2009)

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi and Joseph Turian. "*Accurate Dependency Parsing with a Stacked Multilayer Perceptron"*. In: EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009. Proceedings, vol. Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence article n. 14. EVALITA 2009. December, Reggio Emilia, Italy 2009)

Marco Baroni, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta. *"The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora"*. Journal of Language Resources and Evaluation 43(3), 209–226 (2009)

Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi. "*READ-IT: assessing readability of Italian texts with a view to text simplification*". In Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT '11), 30 July, Edimburgo, UK (2011)

Felice Dell'Orletta. "*Ensemble system for Part-of-Speech tagging*". In: Evaluation of NLP and Speech Tools for Italian, 2009. Proceedings Evalita 2009 Organizers, 2009. December, Reggio Emilia, Italy (2009)

Tullio De Mauro (a cura di). *Grande dizionario italiano dell'uso* (GRADIT). Torino: UTET (1999-2000)

Tullio De Mauro. *Il Nuovo vocabolario di base della lingua italiana*. Internazionale, disponibile all'indirizzo https://dizionario.internazionale.it/ (2016)

Giacomo Devoto, Gian Carlo Oli. *Il Devoto-Oli junior. Il mio primo vocabolario di italiano.* Mondadori Education, Le Monnier (2012)

Keith Rayner, Susan Duffy. *"Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity"*. Memory & Cognition, 14:191–201 (1986)

Stefan Richter, Andrea Cimino, Felice Dell'Orletta and Giulia Venturi. *"Tracking the Evolution of Written Language Competence: an NLP-based Approach"*. In Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it), 3-4 December, Trento, Italy, pp. 236-240 (2015)

Matthew Shardlow, Richard Evans, Marcos Zampieri. "*Predicting Lexical Complexity in English Texts*". Manuscript, arXiv:2102.08773v1 [cs.CL] (2021)

White Paper on the Future Impact of Artificial Intelligence on the Publishing Industry. Gould Finch and Frankfurt Book Fair, disponibile all'indirizzo https://www.buchmesse.de/files/media/pdf/WhitePaperAIPublishingGouldFinch2019EN.pdf (2019)