10.12871/CLICIT201415

Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado

Alessia Barbagli•, Pietro Lucisano•,
Felice Dell'Orletta◇, Simonetta Montemagni◇, Giulia Venturi◇
•Dipartimento di Psicologia dei processi di Sviluppo e socializzazione, Università di Roma "La Sapienza"
alessia.barbagli@gmail.com,pietro.lucisano@uniroma1.it
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{nome.cognome}@ilc.cnr.it

Abstract

Italiano. L'ultimo decennio ha visto l'affermarsi a livello internazionale dell'uso di tecnologie del linguaggio per lo studio dei processi di apprendimento. Questo contributo, che si colloca all'interno di una ricerca più ampia di pedagogia sperimentale, riporta i primi e promettenti risultati di uno studio finalizzato al monitoraggio dell'evoluzione del processo di apprendimento della lingua italiana condotto a partire dalle produzione scritte degli studenti con strumenti di annotazione linguistica automatica e di estrazione di conoscenza.

English. *Over the last ten years, the use of language technologies was successfully extended to the study of learning processes. The paper reports the first results of a study, which is part of a broader experimental pedagogy project, aimed at monitoring the evolution of the learning process of the Italian language based on a corpus of written productions by students and exploiting automatic linguistic annotation and knowledge extraction tools.*

1 Introduzione

L'uso di tecnologie del linguaggio per lo studio dei processi di apprendimento e, in termini più applicativi, di costruzione dei cosiddetti *Intelligent Computer–Assisted Language Learning systems (ICALL)* è sempre più al centro di ricerche interdisciplinari che mirano a mettere in luce come metodi e strumenti di annotazione linguistica automatica e di estrazione della conoscenza siano oggi maturi per essere usati anche nel contesto educativo e scolastico. A livello internazionale, ciò è dimostrato dal successo del *Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, arrivato nel 2014 alla sua nona edizione[1].

Il presente contributo si pone in questo contesto di ricerca, riportando i primi risultati di uno studio tuttora in corso, finalizzato a descrivere, con strumenti di carattere quantitativo e qualitativo, l'evoluzione delle abilità di scrittura, sia a livello del contenuto testuale sia delle competenze linguistiche, dalla prima alla seconda classe della scuola secondaria di primo grado. Si tratta di un lavoro esplorativo finalizzato a costruire un modello di analisi empirica in grado di consentire l'osservazione dei processi e dei prodotti della didattica della produzione scritta. Il carattere innovativo di questa ricerca nel panorama nazionale e internazionale si colloca a vari livelli.

Sul versante metodologico, la ricerca qui delineata rappresenta il primo studio finalizzato al monitoraggio dell'evoluzione del processo di apprendimento linguistico della lingua italiana condotto a partire dalle produzione scritte degli studenti e con strumenti di annotazione linguistica automatica e di estrazione di conoscenza. L'utilizzo di tecnologie del linguaggio per il monitoraggio dell'evoluzione della competenza linguistica di apprendenti affonda le radici in un filone di studi avviato a livello internazionale nell'ultimo decennio e all'interno del quale analisi linguistiche generate da strumenti di trattamento automatico del linguaggio sono usate, ad esempio, per: monitorare lo sviluppo della sintassi nel linguaggio infantile (Sagae et al., 2005; Lu, 2007); identificare deficit cognitivi attraverso misure di complessità sintattica (Roark et al., 2007) o di associazione semantica (Rouhizadeh et al., 2013); monitorare la capacità di lettura come componente centrale della competenza linguistica (Schwarm e Ostendorf, 2005; Petersen e Ostendorf, 2009). Prendendo le mosse da questo filone di ricerca, Dell'Orletta e Montemagni (2012) e Dell'Orletta et al. (2011) hanno di-

[1] http://www.cs.rochester.edu/∼tetreaul/acl-bea9.html

mostrato all'interno di due studi di fattibilità che le tecnologie linguistico–computazionali possono giocare un ruolo centrale nella valutazione della competenza linguistica italiana di studenti in ambito scolastico e nel tracciarne l'evoluzione attraverso il tempo. Questo contributo rappresenta uno sviluppo originale e innovativo di questa linea di ricerca, in quanto la metodologia di monitoraggio linguistico proposta è utilizzata all'interno di uno studio più ampio di pedagogia sperimentale, basato su un corpus significativo di produzioni scritte di studenti e finalizzato a rintracciare l'evoluzione delle competenze in una prospettiva diacronica e/o socio–culturale.

L'oggetto delle analisi rappresenta un altro elemento di novità: è stato scelto il primo biennio della scuola secondaria di primo grado come ambito scolastico da analizzare perché poco indagato dalle ricerche empiriche e poiché poche sono state sino ad oggi le indagini che hanno verificato l'effettiva pratica didattica derivata dalle indicazioni previste dai programmi ministeriali relativi a questo ciclo scolastico, a partire dal 1979 fino alle Indicazioni Nazionali del 2012.

2 Il contesto e i dati della ricerca

Il contesto di riferimento è rappresentato dalla ricerca IEA IPS (*Association for the Evaluation of Educational Achievement, Indagine sulla Produzione Scritta*) che agli inizi degli anni '80 ha coinvolto quattordici paesi di tutto il mondo (tra cui l'Italia) in un'indagine sull'insegnamento e sull'apprendimento della produzione scritta nella scuola. I risultati dell'indagine sono riportati in Purves (1992), e per l'Italia in Lucisano (1988) e Lucisano e Benvenuto (1991).

Lo studio più ampio, tuttora in corso, in cui il presente contributo si colloca si basa sull'ipotesi che nei due anni presi in esame si realizzino dei cambiamenti rilevanti nelle modalità di approccio alla scrittura degli studenti e nella loro produzione scritta, e che tali cambiamenti siano dovuti allo stimolo di un insegnamento più formale. Si ritiene che tali cambiamenti possono essere verificati osservando le variazioni che risultano dai prodotti dell'attività di scrittura scolastica.

La ricerca è stata organizzata individuando tre tipi di variabili: di sfondo (es. background familiare, territoriale, personale), di processo (es. misura di abilità linguistiche degli studenti) e di prodotto (es. misure sui testi degli studenti).

Abbiamo preso come riferimento un campione di giudizio composto da studenti di sette diverse scuole secondarie di primo grado di Roma; la scelta delle scuole è avvenuta basandosi sul presupposto che esista una relazione tra l'area territoriale in cui è collocata la scuola e l'ambiente socio–culturale di riferimento. Sono state individuate due aree territoriali: il centro storico e la periferia rappresentativi rispettivamente di un ambiente socio–culturale medio–alto e medio–basso. Per ogni scuola è stata individuata una classe, per un totale di 77 studenti in centro e 79 in periferia. Per ogni studente, sono state raccolte due tipologie di produzioni scritte: le tracce assegnate dai docenti nei due anni scolastici e due prove comuni relative alla percezione dell'insegnamento della scrittura, svolte dalle classi al termine del primo e del secondo anno[2]. È stato così possibile raccogliere un corpus di 1.352 testi che sono stati digitalizzati per le successive fasi di analisi. Per entrambe le tipologie di produzioni, l'analisi ha riguardato sia il contenuto sia la struttura linguistica sottostante. In quanto segue, ci focalizzeremo sull'analisi delle prove comuni sull'insegnamento della scrittura.

3 Analisi delle produzioni scritte

Il corpus di produzioni scritte, una volta digitalizzato, è stato arricchito automaticamente con annotazione morfo–sintattica e sintattica. A tal fine è stata utilizzata una piattaforma consolidata di metodi e strumenti per il trattamento automatico dell'italiano sviluppati congiuntamente dall'ILC–CNR e dall'Università di Pisa[3]. Per quanto riguarda l'annotazione morfo–sintattica, lo strumento utilizzato è descritto in Dell'Orletta (2009); sul versante dell'analisi a dipendenze, abbiamo utilizzato DeSR (Attardi et al., 2009).

Il testo arricchito con informazione linguistica ("linguisticamente annotato") costituisce il punto di partenza per le analisi successive, riconducibili a due filoni principali finalizzati rispettivamente all'identificazione dei contenuti più salienti e alla definizione del profilo delle competenze linguistiche di chi lo ha prodotto. L'accuratezza dell'annotazione, seppur decrescente attraverso i diversi livelli, è sempre più che accettabile da permettere la tracciabilità nel testo di una vasta tipologia di

[2]Le tracce somministrate derivano dalla Prova 9 della Ricerca IEA–IPS (Lucisano, 1984; Corda Costa e Visalberghi, 1995) che consiste in una lettera di consigli indirizzata a un coetaneo su come scrivere un tema.

[3]http://linguistic-annotation-tool.italianlp.it/

tratti riguardanti diversi livelli di descrizione linguistica, che possono essere sfruttati nei compiti di monitoraggio linguistico (Montemagni, 2013).

Partendo dall'assunto di base che i termini costituiscono la rappresentazione linguistica dei concetti più salienti di una collezione documentale e per questo motivo il compito di estrazione terminologica costituisce il primo e fondamentale passo verso l'accesso al suo contenuto, il corpus delle prove comuni morfo–sintatticamente annotato è stato sottoposto ad un processo di estrazione terminologica finalizzato all'identificazione e all'estrazione delle unità lessicali monorematiche e polirematiche rappresentative del contenuto. A tal fine è stato utilizzato $T2K^2$ (Text–to–Knowledge)[4], una piattaforma web finalizzata all'acquisizione di informazione semantico–lessicale da corpora di dominio (Dell'Orletta et al., 2014). Il monitoraggio delle competenze e abilità linguistiche degli apprendenti, che rappresenta un ambito inesplorato di applicazione di tali tecnologie, ha riguardato sia il livello lessicale sia aspetti della struttura linguistica (in particolare, morfo–sintattica e sintattica). A questo scopo è stato usato MONITOR–IT, lo strumento che, implementando la strategia di monitoraggio descritta in Montemagni (2013), analizza la distribuzione di un'ampia gamma di caratteristiche linguistiche (di base, lessicali, morfo–sintattiche e sintattiche) rintracciate automaticamente in un corpus a partire dall'output dei diversi livelli di annotazione linguistica (Dell'Orletta et al., 2013): la Tabella 1 riporta una selezione dei tratti più significativi utilizzati.

4 Primi risultati

I risultati riguardano il corpus delle prove comuni di scrittura somministrate nel primo e secondo anno per un totale di 240 testi. L'analisi di ciascuna collezione è stata condotta sia in rapporto al contenuto che alla struttura linguistica in relazione ad un'ampia gamma di tratti linguistici.

4.1 Analisi del contenuto

La Tabella 2 riporta i primi 15 termini estratti da $T2K^2$ a partire dalle prove comuni del primo e del secondo anno, ordinati per rilevanza. Tra i termini più salienti emersi dall'analisi delle prove del primo anno si segnalano *paura dei compiti, paura dei lavori di scrittura* così come *difficoltà nei compiti, esperienza in quinta*, che rivelano una tipologia di consigli appartenente alla sfera psico–emotiva. Nel secondo anno, i termini più significativi estratti dal testo fanno riferimento a consigli che riguardano aspetti più 'tecnici' come *uso di parole, pertinenza alla traccia, uso dei verbi*.

È interessante qui far notare come tali risultati siano in linea con la codifica manuale del contenuto condotta sulla base della griglia predisposta dalla ricerca IEA (Fabi e Pavan De Gregorio, 1988; Asquini, 1993; Asquini et al., 1993), che divide i consigli contenuti nei testi in sei macro aree (Contenuto, Organizzazione, Stile e registro, Presentazione, Procedimento, Tattica). Analizzando i risultati ottenuti sono evidenti i cambiamenti avvenuti tra il primo e il secondo anno. Nel primo anno la maggior parte dei consigli dati riflettono la didattica della scuola primaria e pertengono all'area della tattica; anche i consigli relativi al procedimento, sono focalizzati sulla sfera del comportamento e della realtà psico–emotiva. Nel secondo anno l'attenzione si sposta verso gli aspetti più prettamente linguistici come il contenuto e la presentazione (che comprende ortografia, calligrafia e grammatica), riflettendo la didattica della scuola secondaria di primo grado. La differenza appare ancora più significativa nel confronto tra i consigli più frequenti delle prove dei due anni. I consigli che hanno registrato le maggiori frequenze nelle prove del primo anno riguardavano esclusivamente l'aspetto psico–emotivo e il comportamento (es. *Aspetta un po', rifletti prima di scrivere*, *Leggi/scrivi molto*, *Non avere paura*) mentre nelle prove del secondo anno tra i dieci consigli più frequenti (es. *Scrivi con calligrafia ordinata*, *Usa una corretta ortografia*, *Attieniti all'argomento; solo i punti pertinenti*) non compare nessun consiglio di tattica.

4.2 Analisi della struttura linguistica

Il monitoraggio comparativo tra le caratteristiche linguistiche rintracciate nel corpus di prove comuni realizzate nel primo e nel secondo anno è stato condotto con l'obiettivo di tracciare l'evoluzione delle abilità linguistiche degli studenti nei due anni. Dall'ANOVA delle prove comuni risulta che esistono differenze significative tra primo e secondo anno a tutti i livelli di analisi linguistica considerati. Ad esempio, rispetto alle catteristiche 'di base' risulta che la variazione del *numero medio di token per frase* nelle due prove dei due anni

[4] http://www.italianlp.it/demo/t2k-text-to-knowledge/

Catteristiche di base
Lunghezza media dei periodi e delle parole
Catteristiche lessicali
Percentuale di lemmi appartenenti al *Vocabolario di Base (VdB)* del *Grande dizionario italiano dell'uso* (De Mauro, 2000)
Distribuzione dei lemmi rispetto ai repertori di uso (Fondamentale, Alto uso, Alta disponibilità)
Type/Token Ratio (TTR) rispetto ai primi 100 e 200 tokens
Catteristiche morfo–sintattiche
Distribuzione delle categorie morfo–sintattiche
Densità lessicale
Distribuzione dei verbi rispetto al modo, tempo e persona
Catteristiche sintattiche
Distribuzione dei vari tipi di relazioni di dipendenza
Arità verbale
Struttura dell'albero sintattico analizzato (es. altezza media dell'intero albero, lunghezza media delle relazioni di dipendenza)
Uso della subordinazione (es. distribuzione di frasi principali vs. subordinate, lunghezza media di sequenze di subordinate)
Modificazione nominale (es. lunghezza media dei complementi preposizionali dipendenti in sequenza da un nome)

Tabella 1: Selezione delle caratteristiche linguistiche più salienti oggetto di monitoraggio linguistico.

Prova I anno			**Prova II anno**		
DomainRel.	Termine	Freq.	DomainRel.	Termine	Freq.
1	compiti di scrittura	26	1	errori di ortografia	15
2	maestra di italiano	21	2	professoressa di italiano	10
3	lavori di scrittura	17	3	uso di parole	9
4	compiti in classe	30	4	tema in classe	7
5	errori di ortografia	11	5	compiti in classe	9
6	paura dei compiti	9	6	pertinenza alla traccia	4
7	compiti in classe d'italiano	7	7	professoressa di lettere	4
8	anno di elementari	7	8	tema	369
9	classe d'italiano	7	9	voti a tema	3
10	compiti di italiano	7	10	temi a piacere	3
11	maestra	405	11	contenuto del tema	3
12	compiti per casa	6	12	errori di distrazione	3
13	esperienze in quinta	4	13	professoressa	131
14	maestra delle elementari	4	14	frasi	80
15	maestra di matematica	4	15	traccia	81

Tabella 2: Un estratto dell'estrazione terminologica automatica dalle prove comuni del I e II anno.

è significativa. Mentre le prove scritte nel primo anno contengono frasi lunghe in media 23,82 token, la lunghezza media delle frasi delle prove del secondo anno è pari a 20,71 token. Significativa è anche la variazione nell'uso di voci riconducibili al *VdB*, che diminuisce dall'83% del vocabolario nelle prove del primo anno al 79% nel secondo anno, così come i valori di *TTR* (rispetto ai primi 100 tokens), che aumentano passando dallo 0,66 allo 0,69. In entrambi i casi, tali mutamenti possono essere visti come conseguenza di un arricchimento lessicale. Per quanto riguarda il livello morfo–sintattico, sono soprattutto le caratteristiche che catturano l'uso dei tempi e dei modi verbali a essere particolarmente significative. A livello del monitoraggio sintattico, è ad esempio l'uso del *complemento oggetto in posizione pre– o post–verbale* a variare significativamente. Se nelle prove del primo anno il 19% dei complementi oggetto è in posizione pre–verbale, nel secondo anno la percentuale diminuisce passando al 13%; mentre nel primo anno i complementi oggetti post–verbali sono l'81% e aumentano passando all'87% nel secondo anno. Nelle prove del secondo anno si osserva dunque un maggiore rispetto dell'ordinamento canonico soggetto–verbo–oggetto, più vicino alle norme dello scritto che del parlato.

Sebbene i risultati ottenuti siano ancora preliminari rispetto al più ampio contesto della ricerca, crediamo mostrino chiaramente le potenzialità dell'incontro tra linguistica computazionale ed educativa, aprendo nuove prospettive di ricerca. Le linee di attività in corso includono l'analisi della correlazione tra le evidenze acquisite attraverso il monitoraggio linguistico e le variabili di processo e di sfondo così come lo studio dell'evoluzione delle abilità linguistiche del singolo studente.

References

G. Asquini, G. De Martino, L. Menna. 1993. Analisi della prova 9. In AA.VV *La produzione scritta nel biennio superiore. Ricerca nelle scuole superiori del Molise*, IRRSAE MOLISE, Campobasso, Lampo, pp. 77–100.

G. Asquini. 1993. Prova 9 lettera di consigli. In AA.VV *La produzione scritta nel biennio superiore. Ricerca nelle scuole superiori del Molise*, IRRSAE MOLISE, Campobasso, Lampo, pp. 67–75.

G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, Reggio Emilia.

M. Corda Costa and A. Visalberghi. 1995. *Misurare e valutare le competenze linguistiche. Guida scientifico-pratica per gli insegnanti*. Firenze, ed. La Nuova Italia.

F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, Reggio Emilia.

F. Dell'Orletta, S. Montemagni, E.M. Vecchi, and G. Venturi. 2011. Tecnologie linguistico-computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria. In G.C. Bruno, I. Caruso, M. Sanna, I. Vellecco (a cura di) *Percorsi migranti: uomini, diritto, lavoro, linguaggi*, pp. 319–336, Milano, McGraw-Hill.

F. Dell'Orletta F. and S. Montemagni. 2012. Tecnologie linguistico–computazionali per la valutazione delle competenze linguistiche in ambito scolastico. In *Atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI 2010)*, 27-29 settembre, Viterbo.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2013. Linguistic Profiling of Texts Across Textual Genre and Readability Level. An Exploratory Study on Italian Fictional Prose. *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-2013)*, 7–11 September, Hissar, Bulgaria, pp. 189–197.

F. Dell'Orletta, G. Venturi, A. Cimino, S. Montemagni. 2014. T2K: a System for Automatically Extracting and Organizing Knowledge from Texts. In *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 2062–2070, 26-31 May, Reykjavik, Iceland.

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

A Fabi and G. Pavan De Gregorio. 1988. La prova 9: risultati di una ricerca sui contenuti in una prova di consigli sulla scrittura. In *Ricerca educativa*, 5, pp. 2–3.

X. Lu. 2007. Automatic measurement of syntactic complexity in child language acquisition. In *International Journal of Corpus Linguistics*, 14(1), pp. 3–28.

P. Lucisano. 1984. L'indagine IEA sulla produzione scritta. In *Ricerca educativa*, Numero 5.

P. Lucisano. 1988. La ricerca IEA sulla produzione scritta. In *Ricerca educativa*, 2–3, pp. 3–13.

P. Lucisano e G. Benvenuto. 1991. Insegnare a scrivere: dalla parte degli insegnanti. In *Scuola e Città*, 6, pp. 265–279.

S. Montemagni. 2013. Tecnologie linguistico–computazionali e monitoraggio della lingua italiana. In *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, Anno XLII, Numero 1, pp. 145–172.

S.E. Petersen e M. Ostendorf. 2009. A machine learning approach to reading level assessment. In *Computer Speech and Language (23)*, pp. 89–106.

A. Purvues. 1992. *The IEA Study of Written Composition II: Education and Performance in Fourteen Countries vol 6*. Oxford, Pergamon.

B. Roark, M. Mitchell, K. Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pp. 1–8.

M. Rouhizadeh, E. Prud'hommeaux, B. Roark, J. van Santen. 2013. Distributional semantic models for the evaluation of disordered language. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, pp. 709–714.

K. Sagae, A. Lavie, B. MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05)*, pp. 197–204.

S.E. Schwarm e M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05)*, pp. 523–530.