Il ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: primi risultati

Alessia Barbagli
Dipartimento di Psicologia dei processi di Sviluppo e socializzazione, Università di Roma "La Sapienza"

Felice Dell'Orletta
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR), ItaliaNLP Lab

Giulia Venturi
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR), ItaliaNLP Lab

Pietro Lucisano
Dipartimento di Psicologia dei processi di Sviluppo e socializzazione, Università di Roma "La Sapienza"

Simonetta Montemagni
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR), ItaliaNLP Lab

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 riporta i primi e promettenti risultati di uno studio interdisciplinare che si è avvalso di metodi e tecniche di analisi propri della linguistica computazionale, della linguistica e della pedagogia sperimentale. Lo studio, finalizzato al monitoraggio dell'evoluzione del processo di apprendimento della lingua italiana, è stato condotto a partire dalle produzione scritte di studenti della scuola secondaria di primo grado con strumenti di annotazione linguistica automatica e di estrazione di conoscenza e ha portato all'identificazione di un insieme di tratti qualificanti il processo di apprendimento linguistico.*

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 and promising results of an interdisciplinary study 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

Gli ultimi dieci anni hanno visto un crescente interesse verso le tecnologie del linguaggio come punto di partenza per ricerche interdisciplinari finalizzate allo studio delle competenze linguistiche di apprendenti la propria lingua madre (L1) o una lingua straniera (L2). Sebbene con obiettivi diversi, le ricerche condotte a livello internazionale sono accomunate da una medesima metodologia basata sull'uso di strumenti di annotazione linguistica automatica e condividono il medesimo obiettivo di indagare la 'forma linguistica' di corpora di produzioni spontanee. In questo senso il testo linguisticamente annotato costituisce il punto di partenza all'interno del quale rintracciare una serie di caratteristiche linguistiche (lessicali, grammaticali, sintattiche, ecc.) che

possano essere considerate indicatori affidabili per ricostruire il profilo linguistico delle produzioni. Lo scopo è ad esempio quello di studiare in che modo tali caratteristiche sono rivelatrici della qualità di scrittura di apprendenti una lingua straniera (Deane and Quinlan 2010) o quello di monitorare la capacità di lettura come componente centrale della competenza linguistica (Schwarm e Ostendorf 2005; Petersen e Ostendorf 2009). La medesima metodologia è stata utilizzata per monitorare lo sviluppo nel tempo della sintassi nel linguaggio infantile a partire da trascrizioni del parlato (Sagae et al. 2005; Lu 2007; Lubetich and Sagae 2014). L'analisi automatica della 'forma linguistica' di produzioni di apprendenti rappresenta il punto di partenza anche per identificare eventuali deficit cognitivi attraverso misure di complessità sintattica (Roark et al. 2007) o di associazione semantica (Rouhizadeh et al. 2013).

Da un punto di vista più applicativo, tecnologie basate sul trattamento automatico del linguaggio sono oggi impiegate nella costruzione di *Intelligent Computer–Assisted Language Learning systems (ICALL)* (Granger 2003), per sviluppare strumenti di valutazione automatica delle produzioni scritte per lo più di apprendenti una lingua straniera (Attali and Burstein 2006) o per mettere a punto programmi di correzione automatica degli errori commessi da apprendenti una L2 (Ng et al. 2013, 2014). A livello internazionale, ciò è dimostrato dall'organizzazione di numerose conferenze sull'argomento come ad esempio il *Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, arrivato nel 2015 alla sua decima edizione[1].

A questa panoramica va aggiunto il fatto che strumenti di estrazione della conoscenza sono oggi utilizzati per analizzare il 'contenuto' di produzioni per lo più scritte. A livello internazionale, i metodi tradizionalmente impiegati a questo scopo (*Knowledge Tracing systems*) fanno riferimento a un comune paradigma che permette di modellare il processo di apprendimento delle conoscenze attraverso l'analisi di una serie di compiti svolti nel tempo dagli studenti e valutati dagli insegnanti (Corbett and Anderson 1994). Tali metodi non sono basati su strumenti di trattamento automatico del linguaggio, ma stanno diventando sempre più d'interesse all'interno della comunità di Machine Learning[2] in contesti di apprendimento personalizzato a distanza (*Adaptive E–learning*) (Piech et al. 2015; Ekanadham and Karklin 2015).

Il presente contributo si pone in questo contesto di ricerca, riportando i primi risultati di uno studio più ampio, tuttora in corso, condotto a partire da un corpus di produzioni scritte di studenti italiani nel primo e nel secondo anno della scuola secondaria di primo grado. Si tratta di uno studio finalizzato a costruire un modello di analisi empirica in grado di monitorare l'evoluzione sia della 'forma linguistica' sia del 'contenuto' utilizzando strumenti di annotazione linguistica automatica uniti a tecnologie di estrazione automatica di conoscenza da testi. Come discusso in quanto segue, l'approccio messo a punto si ripropone di monitorare tale evoluzione sia nel tempo (nel passaggio cioè dal primo al secondo anno di scuola) sia rispetto ad una serie di variabili di sfondo (come ad esempio il background familiare, le abitudini personali, ecc.) rintracciate grazie ad un questionario studenti distribuito in classe.

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 (come lingua madre) condotto con strumenti di annotazione linguistica automatica e di estrazione della conoscenza. Come preceden-

1 http://www.cs.rochester.edu/∼tetreaul/naacl-bea10.html
2 http://dsp.rice.edu/ML4Ed_ICML2015

temente discusso, sino ad oggi le ricerche a livello internazionale che si sono basate sull'uso di tecnologie del linguaggio per monitorare l'evoluzione nel tempo di competenze linguistiche di apprendenti una lingua madre si sono per lo più concentrate sull'analisi di produzioni orali infantili. Al contrario, chi si è interessato allo studio dell'evoluzione delle abilità di scrittura lo ha fatto a partire da produzioni di apprendenti una lingua straniera. Minore attenzione è stata dunque dedicata all'uso di tali tecnologie per lo studio diacronico di come evolvono le abilità di scrittura di studenti madrelingua. Per quanto riguarda la lingua italiana, all'interno di due precedenti studi di fattibilità, (Dell'Orletta e Montemagni 2012) e (Dell'Orletta et al. 2011) hanno dimostrato che le tecnologie linguistico–computazionali possono giocare un ruolo centrale nella valutazione della competenza linguistica di studenti madrelingua in ambito scolastico e nel tracciarne l'evoluzione attraverso il tempo. Questo contributo rappresenta uno sviluppo originale e innovativo di questa linea di ricerca all'interno della quale l'uso congiunto di strumenti di annotazione linguistica automatica e di estrazione di conoscenza rappresenta un'ulteriore innovazione metodologica. Ciò è reso possibile dalla particolare conformazione interna del corpus di produzioni scritte utilizzato in questo lavoro e descritto nei paragrafi successivi.

La scelta del ciclo scolastico e dei tipi di produzioni scritte analizzate è un altro elemento di novità di questo studio, soprattutto sotto il profilo di ricerca in pedagogia sperimentale. Non solo infatti è stato scelto il primo biennio della scuola secondaria di primo grado come ambito scolastico da analizzare perché poco indagato dalle ricerche empiriche, ma sono stati anche analizzati i temi di studenti ai quali era stato richiesto di dare ad un coetaneo dei consigli per scrivere un buon tema. Questo ha permesso di indagare come cambia (a livello di 'contenuti') la percezione dell'insegnamento della scrittura nel passaggio dal primo al secondo anno di scuola attraverso la pratica di scrittura (analisi della 'forma linguistica'). Poche sono state infatti sino ad oggi le indagini che hanno verificato i risultati dell'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. Al contrario, gli studi si sono per lo più concentrati sull'educazione linguistica (Rigo 2005) e in modo specifico sulla competenza testuale anche in termini di produzione.

In quanto segue, nel Paragrafo 2 introdurremo l'approccio del più ampio contesto di ricerca in cui questo contributo si inserisce. Dopo aver illustrato la metodologia e gli strumenti di analisi linguistico–computazionale qui adottati (Paragrafo 3), nei Paragrafi 4 e 5 riporteremo i primi risultati ottenuti. Infine, nel Paragrafo 6 trarremo alcune conclusioni e tratteggeremo gli sviluppi futuri di questa ricerca.

2. Il contesto e i dati della ricerca

Il contesto a cui fa riferimento questo studio è quello della ricerca IEA IPS (*Association for the Evaluation of Educational Achievement, Indagine sulla Produzione Scritta*) (Purvues 1992), un'indagine sull'insegnamento e sull'apprendimento della produzione scritta nella scuola, che agli inizi degli anni '80 coinvolse quattordici paesi di tutto il mondo, tra cui l'Italia (Lucisano 1988; Lucisano e Benvenuto 1991). Prendendo le mosse dai risultati raggiunti, il presente contributo si basa sull'ipotesi che nei primi due anni della scuola media superiore di primo grado si realizzino dei cambiamenti rilevanti sia nel modo in cui gli studenti si approcciano alla scrittura sia nel modo stesso in cui essi scrivono. L'intuizione è che ciò sia dovuto al fatto che gli studenti sono sottoposti nel passaggio dal primo al secondo anno di scuola a un insegnamento più formale della scrittura.

Scopo della ricerca è inoltre quello di monitorare come tali cambiamenti si verifichino non solo nell'arco temporale preso in esame, ma anche rispetto ad alcune caratteristiche descrittive del campione di studenti esaminato. Per questo motivo è stato messo a punto un questionario somministrato in classe dai docenti agli studenti e composto da circa trenta domande corrispondenti ad altrettante variabili di sfondo considerate. Le domande contenute riguardano diversi aspetti che vanno dall'inquadramento anagrafico degli studenti, la caratterizzazione socio–culturale della famiglia (professione dei genitori, titolo di studio, libri in casa ecc...) e la rilevazione delle loro abitudini (ad esempio, tempo dedicato alla lettura e alla scrittura, tempo dedicato ad ascoltare musica, ecc...), per arrivare a domande che vanno a indagare le idee, le credenze e i convincimenti degli studenti a proposito della scrittura e il loro rapporto con la scrittura scolastica.

Allo scopo di monitorare i cambiamenti abbiamo preso in esame un corpus di 240 prove scritte da 156 studenti di sette diverse scuole secondarie di primo grado di Roma; la scelta delle scuole è avvenuta basandosi sul presupposto che esista una forte 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, selezionati come rappresentativi rispettivamente di un ambiente socio–culturale medio–alto e medio–basso. In ogni scuola è stata individuata una classe e, benché le scuole di periferia siano quattro mentre quelle del centro siano tre, il numero degli studenti è quasi equivalente (77 in centro e 79 in periferia) dal momento che le classi delle scuole del centro sono più numerose.

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. Alla fine del secondo anno è stata somministrata la traccia della 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], mentre per la prova dell'anno precedente ne è stata utilizzata una forma adattata alla classe e all'età[4].

In questo studio ci siamo focalizzati sull'analisi di una porzione dell'intero corpus raccolto. Si tratta della collezione di prove comuni di scrittura somministrate nel primo e secondo anno, composta da 109 testi. La scelta di prendere in esame questa sottoporzione ci ha permesso di mostrare come i cambiamenti che avevamo supposto esistere sia nel modo in cui gli studenti si approcciano alla scrittura sia nel modo stesso in cui essi scrivono possano essere verificati utilizzando sia strumenti di annotazione linguistica automatica del testo sia di estrazione automatica della conoscenza. Mentre i primi infatti, come vedremo, permettono di monitorare le variazioni di 'forma linguistica' nella pratica della scrittura, i secondi consentono di analizzare anche come cambi che cosa gli studenti scrivono a proposito della pratica della scrittura (come muti dunque il 'contenuto' dei temi).

3 La traccia somministrata al termine del secondo anno è la seguente "Un ragazzo più giovane di te ha deciso di iscriversi alla tua scuola. Ti ha scritto per chiederti come fare un tema che possa essere valutato bene dai tuoi insegnanti. Mandagli una lettera cordiale nella quale descrivi almeno cinque punti che tu pensi importanti per gli insegnanti quando valutano i temi"

4 La traccia somministrata al termine del primo anno "Un tuo amico sta per iniziare la quinta elementare con le tue maestre e ti ha confessato di aver paura soprattutto dei lavori di scrittura che gli saranno chiesti. Scrivigli una lettera raccontando la tua esperienza, gli aspetti positivi e anche le tue difficoltà nei compiti di scrittura che hai fatto in quinta elementare. Raccontagli dei compiti che ti sono piaciuti di più e di quelli che ti sono piaciuti di meno e anche dei suggerimenti che le maestre ti davano per insegnarti a scrivere bene e di come correggevano i compiti scritti. Dagli consigli utili per cavarsela."

3. Analisi linguistico–computazionale delle produzioni scritte degli studenti

Il corpus di 109 prove comuni oggetto di questo studio è stato analizzato impiegando strumenti e metodologie di analisi automatica del testo che hanno permesso di accedere sia alla 'forma linguistica' sia al 'contenuto' delle prove.

Il corpus di produzioni scritte, una volta digitalizzato, è stato prima di tutto arricchito automaticamente con annotazione morfo–sintattica e sintattica. A tal fine è stata utilizzata una piattaforma consolidata e ampiamente sperimentata di metodi e strumenti per il trattamento automatico dell'italiano sviluppati congiuntamente dall'ILC–CNR e dall'Università di Pisa[5]. Per quanto riguarda l'annotazione morfo–sintattica, lo strumento utilizzato è descritto in (Dell'Orletta 2009); sul versante dell'analisi sintattica a dipendenze, abbiamo utilizzato DeSR (Attardi et al. 2009). Entrambi sono in linea con lo "stato dell'arte" per il trattamento automatico della lingua italiana, considerata anche la loro qualificazione tra gli strumenti più precisi e affidabili per l'annotazione morfo–sintattica e sintattica a dipendenze nella campagna di valutazione di strumenti per l'analisi automatica dell'italiano, EVALITA[6]. Il testo linguisticamente annotato costituisce il punto di partenza per le analisi successive: *i)* l'identificazione dei contenuti più salienti e *ii)* la definizione del profilo linguistico sottostante al testo a partire dal quale è possibile ricostruire un quadro delle competenze linguistiche di chi lo ha prodotto.

3.1 L'identificazione dei contenuti

Il corpus di produzioni scritte è stato sottoposto ad un processo di estrazione terminologica finalizzato all'identificazione e all'estrazione delle unità lessicali monorematiche e polirematiche rappresentative del contenuto. L'ipotesi di partenza è che i termini costituiscono l'istanza linguistica dei concetti più salienti di una collezione documentale e che per questo motivo il compito di estrazione terminologica costituisce il primo e fondamentale passo verso l'accesso al suo contenuto. A tal fine è stato utilizzato $T2K^2$ (Text–to–Knowledge)[7], una piattaforma web che trasforma le conoscenze implicitamente codificate all'interno di un corpus di testi in conoscenza esplicitamente strutturata (Dell'Orletta et al. 2014). Il componente di estrazione terminologica all'interno di $T2K^2$ opera in due fasi: la prima volta all'identificazione all'interno del corpus di acquisizione di unità terminologiche rilevanti per il contesto indagato, la seconda finalizzata alla validazione della salienza dei termini estratti nella fase precedente.

Per quanto riguarda la prima fase, il processo estrattivo opera sul testo annotato a livello morfo–sintattico e lemmatizzato. Mentre l'acquisizione di unità monorematiche avviene sulla base della loro frequenza, l'acquisizione delle unità polirematiche si articola in due passaggi: il primo finalizzato all'identificazione dei potenziali termini sulla base di una mini–grammatica operante sul testo annotato morfo–sintatticamente e deputata al riconoscimento di sequenze di categorie grammaticali corrispondenti a potenziali unità polirematiche; il secondo basato sul metodo denominato C/NC–value (Frantzi et al. 2000), che appartiene alla classe delle misure di rilevanza rispetto al dominio (o "termhood") e che rappresenta ancora oggi uno standard *de facto* nel settore dell'estrazione terminologica (Vu et al. 2008).

5 http://linguistic-annotation-tool.italianlp.it/
6 http://www.evalita.it/
7 http://www.italianlp.it/demo/t2k-text-to-knowledge/

Le unità monorematiche e polirematiche estratte durante la prima fase vengono successivamente filtrate sulla base di una funzione, chiamata "funzione di contrasto", che valuta dal punto di vista quantitativo quanto un termine della lista estratta al passo precedente sia specifico della collezione di documenti analizzati. Per calcolare la salienza del termine, viene confrontata la sua distribuzione sia nel corpus di acquisizione sia in un corpus differente, detto "corpus di contrasto". La funzione utilizzata, chiamata "Contrastive Selection of multi–word terms" (CSmw), si è rivelata particolarmente adatta per l'analisi di variazioni distribuzionali di eventi a bassa frequenza (come appunto l'occorrenza di un termine polirematico). Se per una descrizione dettagliata del metodo si rinvia a (Bonin et al. 2010), vale la pena qui sottolineare come questa fase di filtraggio contrastivo si sia rivelata particolarmente efficace per identificare i concetti caratterizzanti le prove comuni del primo anno *per contrasto* rispetto ai concetti caratterizzanti le prove del secondo anno, e viceversa.

3.2 La ricostruzione del profilo linguistico

Il secondo tipo di analisi a cui sono state sottoposte le produzioni scritte degli studenti riguarda la struttura linguistica sottostante al testo. L'ipotesi di partenza è che l'informazione che è possibile estrarre dall'analisi automatica della 'forma linguistica' del testo rappresenti un indicatore affidabile per monitorare l'evoluzione delle competenze e abilità linguistiche degli apprendenti.

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. 2013a). I parametri sui quali si sono concentrate le analisi spaziano tra i diversi livelli di descrizione linguistica e mirano a catturare diversi aspetti della competenza linguistica di un apprendente, aspetti che spaziano dalla competenza semantico–lessicale a quella sintattica. Nella tipologia di parametri indagati, l'aspetto di maggiore novità riguarda quelli rintracciati a partire dal testo annotato al livello sintattico. Come discusso in quanto segue, questo livello di analisi, per quanto includa un inevitabile margine di errore, se appropriatamente esplorato rende possibile l'indagine di aspetti della struttura linguistica altrimenti difficilmente investigabili e quantificabili su larga scala.

L'utilizzo dell'annotazione linguistica prodotta in modo automatico come punto di partenza del monitoraggio delle abilità di scrittura pone con forza la questione della sua accuratezza. Si noti che l'accuratezza dell'annotazione automatica, inevitabilmente decrescente attraverso i diversi livelli, è sempre più che accettabile da permettere la tracciabilità nel testo di una vasta tipologia di tratti riguardanti diversi livelli di descrizione linguistica, che possono essere sfruttati in compiti di monitoraggio linguistico. Come dimostrato in (Montemagni 2013) per la lingua italiana e in (Dell'Orletta et al. 2013b) per testi in lingua inglese di dominio bio–medico, il profilo linguistico ricostruito a partire da corpora annotati automaticamente è in linea con quello definito a partire da corpora la cui annotazione è stata validata manualmente. Questo risultato rende legittima la scelta di operare all'interno di questo studio sul testo arricchito con annotazione linguistica automatica, nonostante esso includa inevitabilmente un margine di errore che varia a seconda del livello e del tipo di informazione linguistica considerata.

La tipologia di parametri che abbiamo monitorato in questo studio è varia: la Tabella 1 riporta una selezione di quelli più significativi. A partire dall'annotazione morfo–sintattica del testo è stato possibile calcolare come varia ad esempio la distribuzione

delle categorie morfo–sintattiche o di sequenze di categorie grammaticali e/o lemmi. Mentre la struttura sintattica a dipendenze sottostante il testo rappresenta il punto di partenza per arrivare a caratteristiche strutturali dell'albero sintattico, quali ad esempio l'altezza massima dell'albero calcolata come la massima distanza (espressa come numero di relazioni attraversate) che intercorre tra una foglia (rappresentata da parole del testo senza dipendenti) e la radice dell'albero, oppure la lunghezza delle relazioni di dipendenza (calcolata come la distanza in parole tra la testa e il dipendente), oppure la "valenza" media per testa verbale (calcolata come numero medio di dipendenti effettivamente istanziati – sia argomenti che modificatori – governati dallo stesso verbo).

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 calcolata come la proporzione delle parole semanticamente "piene" (nomi, aggettivi, verbi e avverbi) rispetto al totale dei tokens
– Distribuzione dei verbi rispetto al modo, tempo e persona
Catteristiche sintattiche
– Distribuzione delle relazioni di dipendenza
– "Valenza" media per testa verbale
– Caratteristiche della struttura dell'albero sintattico (es. altezza media dell'albero sintattico, lunghezza media delle relazioni di dipendenza)
– Uso della subordinazione (es. distribuzione di proposizioni principali vs. subordinate, livelli di incassamento gerarchico di subordinate)
– Modificazione nominale (es. profondità media dei livelli di incassamento in strutture nominali complesse)

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

4. Analisi del contenuto: primi risultati

La Tabella 2 riporta i primi 20 termini estratti in modo automatico da $T2K^2$ a partire dalle prove comuni del primo e del secondo anno, ordinati per rilevanza decrescente sulla base della funzione statistica *contrastiva* che consente di definire un ordinamento di rilevanza dei termini estratti da una collezione di documenti *per contrasto* rispetto ad un corpus di riferimento ("corpus di contrasto").

Rispetto a questa funzione, la rilevanza dei termini estratti dal corpus di prove del primo anno è stata dunque definita sulla base del contrasto con il corpus di prove del secondo anno e viceversa le prove del primo anno sono state utilizzate come "corpus di contrasto" per calcolare la rilevanza di termini estratti dalle prove del secondo anno. Come mostra la Tabella 2, tra i termini più salienti emersi dall'analisi delle prove del primo anno si segnalano 'paura dei compiti, paura dei lavori di scrittura' o anche 'difficoltà nei compiti, esperienza in quinta'. Sono tutti termini che rivelano una tipologia di consigli appartenente alla sfera psico–emotiva. Nel secondo anno, invece, i termini

Prova I anno	**Prova II anno**
compiti di scrittura	errori di ortografia
maestra di italiano	professoressa di italiano
lavori di scrittura	uso di parole
compiti in classe	tema in classe
errori di ortografia	compiti in classe
paura dei compiti	pertinenza alla traccia
compiti in classe d'italiano	professoressa di lettere
anno di elementari	tema
classe d'italiano	voti al tema
compiti di italiano	temi a piacere
maestra	contenuto del tema
compiti per casa	errori di distrazione
esperienze in quinta	professoressa
maestra delle elementari	frasi
maestra di matematica	traccia
compiti a casa	uso dei verbi
paura dei lavori	consiglio
compiti	parte destra del cervello
paura dei lavori di scrittura	bava alla bocca
difficoltà nei compiti	conoscenza dell'argomento

Tabella 2
I primi 20 termini estratti in modo automatico dal corpus delle prove comuni del I e II anno e ordinati per salienza decrescente.

più significativi estratti dal testo fanno riferimento a consigli che riguardano aspetti più "tecnici" come ad esempio 'uso di parole, pertinenza alla traccia, uso dei verbi', ecc.

Come precedentemente introdotto, i contenuti delle prove comuni del primo e del secondo sono stati analizzati allo scopo di monitorare il modo in cui evolve nei due anni la percezione dell'insegnamento della scrittura attraverso appunto i consigli che gli studenti stessi danno ai loro coetanei su come scrivere un buon tema. Per verificare l'affidabilità della metodologia di estrazione dei contenuti abbiamo messo a confronto i risultati di questo processo automatico con le valutazioni manuali delle prove. Tali valutazioni sono state condotte da uno degli autori, esperto in pedagogia sperimentale, che ha utilizzato la griglia predisposta dalla ricerca IEA (Fabi e Pavan De Gregorio 1988; Asquini 1993; Asquini et al. 1993). La griglia divide i consigli in sei macro–aree: Contenuto, Organizzazione, Stile e registro, Presentazione, Procedimento e Tattica (vedi Tabella 3)[8]. Inoltre, durante questa fase, sono stati individuati all'interno di ciascun tema i periodi che contenevano dei consigli e ad ogni consiglio è stato attribuito un codice identificativo a tre cifre con la rispettiva percentuale di occorrenza (vedi Tabella 4).

Analizzando i risultati della codifica manuale, possiamo notare come nel primo anno la maggior parte dei consigli dati riflettano la didattica della scuola primaria e pertengono alla macro–area della Tattica (41,8%) e del Procedimento (36,9%) focalizzandosi sulla sfera del comportamento e della realtà psico–emotiva. Come si può notare nella Tabella 4, a queste macro–aree corrispondono consigli che riguardano esclusiva-

8 Ogni area ha ulteriori articolazioni interne che identificano il consiglio in maniera sempre più puntuale: ad esempio l'area Contenuto comprende 'aspetti generali, informazione', ecc, l'area Organizzazione comprende 'introduzione, corpo del testo, conclusione', ecc., l'area Stile e registro comprende 'uniformità, chiarezza, scelte lessicali e sintattiche', ecc. e così via.

Area	I anno	II anno
Contenuto	5,3%	23,0%
Organizzazione	1,7%	5,2%
Stile e registro	5,3%	18,4%
Presentazione	9,0%	31,3%
Procedimento	36,9%	17,2%
Tattica	41,8%	5,0%

Tabella 3
Risultati della codifica manuale del contenuto delle prove comuni nel I e II anno rispetto alle sei macroaree IEA.

mente l'aspetto psico–emotivo e il comportamento (es. 'Aspetta un po', rifletti prima di scrivere', 'Leggi/scrivi molto', 'Non avere paura'). Si tratta appunto di consigli "più emotivi" che trovano un corrispettivo nei termini estratti automaticamente quali 'paura dei compiti, paura dei lavori di scrittura, difficoltà nei compiti, esperienza in quinta', ecc. Al contrario, nel secondo anno i consigli più frequenti sono quelli di Contenuto (23%) e Presentazione (31,3%): gli studenti tendono a mettere l'attenzione su aspetti più tecnico–linguistici, riflettendo il cambio della didattica della scuola secondaria di primo grado rispetto a quella della scuola primaria. Nelle prove del secondo anno infatti 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 riconducibile all'area della Tattica (vedi Tabella 4). Anche in questo caso i consigli corrispondono a termini estratti automaticamente quali ad esempio 'uso di parole, pertinenza alla traccia, uso dei verbi, conoscenza dell'argomento, contenuto del tema', ecc.

Questo confronto tra i risultati della fase di estrazione automatica e la fase di annotazione manuale dei consigli di scrittura dati apre nuovi scenari di ricerca. Le prime evidenze raccolte in questo esperimento preliminare suggeriscono infatti come le tecnologie di estrazione automatica del contenuto possano essere usate come supporto di studi finalizzati a definire metodologie di valutazione dell'effettiva pratica didattica, a indagare cioè come gli insegnanti insegnano a scrivere a partire dal modo in cui gli studenti percepiscono l'insegnamento della scrittura.

5. Analisi della struttura linguistica: primi risultati

L'analisi comparativa tra le caratteristiche linguistiche rintracciate nel corpus di prove comuni degli studenti del primo e secondo anno è stata condotta allo scopo *i)* di ricostruire le loro abilità di scrittura e *ii)* di monitorarne l'evoluzione rispetto alla variabile temporale e alle variabili di sfondo raccolte grazie al questionario somministrato nelle scuole.

Sono state pertanto condotte una serie di esplorazioni statistiche rispetto alle distribuzioni nelle prove delle caratteristiche linguistiche estratte a partire dal testo linguisticamente annotato in modo automatico. A questo scopo, è stato utilizzato il test T per campioni accoppiati del programma SPSS v.22 che restituisce per ogni variabile media, dimensioni del campione, deviazione standard e errore standard della media e per ogni coppia di variabili correlazione, differenza media nelle medie, test T, e intervallo di confidenza per la differenza nella media, deviazione standard e deviazione standard della differenza media. Con il test T è possibile dunque verificare se le misure rilevate

Consigli con maggior frequenza					
Prova I anno			Prova II anno		
Cod.	Consiglio	%	Cod.	Consiglio	%
546	Aspetta un po', rifletti prima di scrivere	11,5	411	Scrivi con calligrafia ordinata	6,4
628	Leggi/scrivi molto	10,6	441	Usa una corretta ortografia	5,3
626	Lavora sodo, fai vedere che ti impegni	10,4	111	Attieniti all'argomento; solo i punti pertinenti	5,3
549	Non avere paura	7,1	443	Usa una corretta punteggiatura	3,3
548	Concentrati, resta concentrato	6,2	433	Usa correttamente i verbi (modi e tempi)	3,0
636	Segui sempre i consigli dell'insegnante	4,1	121	Cerca di essere originale/creativo/pieno di immaginazione	2,9
632	Non metterti a discutere con l'insegnante	3,2	351	Usa un vocabolario ricco ed espressivo	2,9
434	Usa correttamente pronomi, verbi, congiunzioni	3,0	355	Usa una terminologia/registro appropriata/o all'argomento	2,6
610	Abbigliamento e aspetto fisico in generale	2,0	440	Ortografia aspetti generali	2,6
622	Non bisbigliare e non fare chiasso	2,0	100	Aspetti di contenuto non specificati	2,2

Tabella 4
Alcuni dei consigli più frequenti nelle prove comuni del I e II anno sulla base della griglia IEA.

nelle prove del secondo anno presentino un miglioramento, un peggioramento o se le misure medie siano rimaste sostanzialmente uguali rispetto a quelle del primo anno. Mediante la correlazione verifichiamo se le variazioni rispettano o meno le differenze di partenza tra i soggetti esaminati e dunque se gli eventuali miglioramenti rappresentino uno sviluppo coerente con le condizioni di partenza degli studenti o se sia intervenuto qualche elemento di cambiamento che ha stimolato cambiamenti significativi.

5.1 Caratteristiche di base e morfo–sintattiche

Partendo dall'analisi delle variabili linguistiche di base riportate nella Tabella 5, possiamo notare che la lunghezza del testo, misurata in termini di numero totale di token, e la lunghezza media dei periodi, misurata in termini di token per periodo, variano in modo statisticamente significativo nel passaggio dal primo al secondo anno scolastico. Mentre nel primo anno gli studenti scrivono prove più lunghe e con periodi mediamente più lunghi, nel secondo anno le prove sono più brevi e contengono periodi mediamente più corti. Questi risultati potrebbero sembrare una prima spia di una inaspettata maggiore complessità delle prove del primo anno rispetto a quelle del secondo. La lunghezza del testo e dei periodi è infatti un elemento tipicamente associato ad una maggiore complessità linguistica. In questo caso tuttavia due sono i fattori che hanno influenza su questo e altri risultati discussi in quanto segue.

Da un lato la maggiore lunghezza del testo e dei periodi nelle prove del primo anno è sicuramente influenzata dal tipo di traccia assegnata: la traccia distribuita il primo anno prevedeva che gli studenti scrivessero di più, non soltanto dando dei consigli

Caratteristiche	I anno	II anno	Significatività
Lunghezza media delle prove (in token)	275,23	239,21	0,00
Lunghezza media dei periodi (in token)	24,02	20,97	0,01
Distribuzione di:			
– punteggiatura	9,70%	10,60%	0,00
– segni di punteggiatura "debole"	0,49%	1,11%	0,00
– congiunzioni	6,90%	5,92%	0,00
– congiunzioni subordinanti	2,78%	2,43%	0,01
– sostantivi	18,16%	19,73%	0,00
– preposizioni articolate	2,74%	3,47%	0,00
– determinanti dimostrativi	0,33%	0,47%	0,00
– pronomi	10,39%	7,72%	0,00
– pronomi personali	1,64%	0,76%	0,00
– pronomi clitici	5,78%	3,99%	0,00

Tabella 5
Caratteristiche di base e morfo–sintattiche e significatività della variazione tra I e II anno.

su come scrivere un buon tema (come richiesto anche dalla traccia del secondo anno), ma descrivendo anche le difficoltà di scrittura incontrate, i tipi di compiti che erano piaciuti di più, il modo in cui le maestre correggevano i temi, ecc... Ad influire è però d'altro canto il fatto che le prove del secondo anno sono scritte da studenti che hanno presumibilmente migliorato le proprie abilità di scrittura. Il prevedibile miglioramento nel passaggio dal primo al secondo anno di scuola implica che i temi del secondo anno siano scritti in modo più "canonico" a cominciare dall'ordinamento del testo in periodi delimitati da un segno di punteggiatura di fine periodo, elemento che permette agli strumenti di annotazione linguistica automatica di individuare l'unità di analisi di un testo scritto (il periodo appunto). Come si può infatti notare nella Tabella 5, nel passaggio dal primo al secondo anno i segni di punteggiatura in generale aumentano. Oltre ai punti di fine periodo sono i segni di punteggiatura "debole" che separano parole e/o proposizioni all'interno del periodo[9] ad aumentare in maniera statisticamente significativa, a testimonianza di una maggiore abilità di organizzazione interna del contenuto. Un testo più "canonico" è dunque un testo che gli strumenti di annotazione linguistica analizzano con una maggiore precisione di analisi perché caratterizzato da tratti linguistici più simili a quelli dei testi sui quali sono stati addestrati. Come discusso in quanto segue, tale precisione influisce anche sulle caratteristiche sintattiche monitorate.

Caratteristica legata alla variazione di lunghezza del periodo è anche la diminuzione nell'uso delle congiunzioni nel passaggio dal primo al secondo anno. Esiste infatti una correlazione statisticamente significativa tra la diminuzione della distribuzione percentuale delle congiunzioni e la lunghezza media dei periodi: a diminuire nelle prove del secondo anno sono soprattutto le congiunzioni subordinanti. Sebbene ciò possa essere considerato a prima vista spia di una diminuzione della complessità sintattica del testo, tradizionalmente associata ad un maggior andamento ipotattico (Beaman 1984; Givón 1991), tuttavia tale variazione può essere interpretata anche in questo caso come indice di un ordinamento più lineare del contenuto (Mortara Garavelli 2003). Ad aumentare in maniera statisticamente significativa sono invece i sostantivi, le

9 Sulla base dello schema di annotazione adottato in questo studio si tratta di punto e virgola e due punti.

preposizioni articolate e i determinanti dimostrativi a parziale testimonianza di come i temi diventino nel secondo anno più informativi e strutturati (Biber 1993).

Un'altra caratteristica morfo–sintattica che testimonia l'evoluzione verso una forma di scrittura più "canonica" è la diminuzione dei pronomi in generale e dei pronomi personali e clitici in particolare. Soprattutto nel caso dei pronomi personali ciò è spia di una maggiore abilità d'uso della possibilità propria della lingua italiana di omettere il pronome personale. Questo risultato, l'aumento della punteggiatura in funzione segmentatrice–sintattica e, vedremo, la diversa distribuzione di alcune caratteristiche sintattiche sono tutti elementi che possiamo ipotizzare siano spia del fatto che nei temi del secondo anno gli studenti abbandonano un modo di espressione che, pur scritta, ha più le caratteristiche del parlato e acquisiscono invece nuove abilità linguistiche di scrittura.

Anche rispetto alla variazione delle competenze d'uso dei verbi i risultati riportati nella Tabella 6 forniscono indicazioni degne di nota. Sebbene la semplice distribuzione percentuale dei verbi non sia statisticamente significativa, risulta invece discriminante nel passaggio dal primo al secondo anno l'uso maggiore dei verbi modali e dei verbi di modo condizionale e gerundio. Se da un lato gli studenti nelle prove del secondo anno usano modi verbali tipicamente inseriti in strutture verbali complesse (quali appunto il condizionale e il gerundio), dall'altro sembrano ridurre progressivamente un modo verbale più semplice come l'indicativo. Le variazioni d'uso dei tempi verbali sono invece da ricondurre più che altro al diverso tipo di traccia nei due anni considerati. La diminuzione di verbi all'imperfetto e al passato nel passaggio dal primo al secondo anno, da un lato, e l'aumento di verbi al tempo presente, dall'altro, sono senza dubbio riconducibili al fatto che la traccia del primo anno richiedeva di descrivere la propria passata esperienza scolastica in quinta elementare, mentre in base alla traccia del secondo anno gli studenti dovevano descrivere la loro attuale esperienza nella scuola secondaria di primo grado. Inoltre, la diminuzione dell'uso degli ausiliari potrebbe essere legata a questa variazione d'uso dei tempi, sebbene tale dato sia sovrastimato poiché lo schema di annotazione linguistica qui adottato non ci permette al momento di distinguere i tempi composti dalle forme passive. Va tuttavia fatto notare come alcune di queste variazioni d'uso dei tempi verbali possano anche essere ascrivibili per alcuni aspetti alle caratteristiche che distinguono la lingua scritta da quella parlata. È il caso ad esempio della diminuzione di verbi all'imperfetto. Sebbene infatti essi diminuiscano nel secondo anno in seguito alla diversa traccia, è pur vero che l'uso estensivo dell'imperfetto è una delle caratteristiche distintive del parlato (Masini 2003). Queste diverse distribuzioni possono essere dunque considerate un'ulteriore spia della progressiva riduzione di forme tipiche della lingua parlata verso l'acquisizione di maggiori abilità di scrittura.

5.2 Caratteristiche sintattiche e lessicali

La diversa distribuzione di alcune delle caratteristiche linguistiche rintracciabili a partire dal livello di annotazione sintattica automatica farebbe inizialmente pensare ad una minore complessità delle prove nel secondo anno. Tuttavia, come discusso precedentemente, il dato va letto alla luce della tendenza, nel passaggio dal primo al secondo anno scolastico, verso una forma di scrittura più "canonica". Va in questa direzione ad esempio l'aumento dei complementi oggetto in posizione post–verbale e della conseguente diminuzione di quelli in posizione pre–verbale: nelle prove del secondo anno gli studenti dimostrano di aver acquisito una maggiore propensione per un ordine canonico dei costituenti nella lingua scritta. La diversa distribuzione fa inoltre ipotizzare

Caratteristiche	I anno	II anno	Significatività
Distribuzione di verbi:			
– ausiliari	1,88%	0,98%	0,00
– modali	1,09%	1,81%	0,00
– di modo condizionale	0,14%	0,64%	0,00
– di modo gerundio	1,68%	2,21%	0,00
– di modo indicativo	53,76%	41,86%	0,00
– al tempo imperfetto	31,78%	1,10%	0,00
– al tempo passato	2,21%	0,75%	0,00
– al tempo presente	56,06%	85,78%	0,00

Tabella 6
Distribuzione di tempi e modi verbali e significatività della variazione tra I e II anno.

un uso ridotto da parte degli studenti della dislocazione a sinistra del tema (dunque del complemento oggetto in posizione pre–verbale), caratteristica tipica del parlato.

Caratteristiche	I anno	II anno	Significatività
Distribuzione di relazioni di dipendenza sintattica di tipo:			
– *complement*	8,00%	7,71%	0,00
– *modifier*	16,60%	17,84%	0,00
– *subject*	5,85%	5,00%	0,00
– *subordinate clause*	2,80%	2,41%	0,00
Lunghezza media delle più lunghe relazioni di dipendenza sintattica	9,22	7,80	0,02
Media di proposizioni per periodo	4,00	3,36	0,01
Media di token per proposizione	6,17	6,42	0,02
Distribuzione dei complementi oggetto:			
– post–verbali	80,93%	86,66%	0,00
– pre–verbali	18,31%	13,34%	0,00

Tabella 7
Caratteristiche sintattiche e significatività della variazione tra I e II anno.

Alcuni dei tratti osservati riflettono inoltre quanto avevamo osservato a proposito della lunghezza della frase. Il fatto che le prove del secondo anno abbiamo periodi mediamente più corti di quelli del primo anno influisce ad esempio sul fatto che i periodi del secondo anno contengano relazioni di dipendenza sintattica più corte rispetto alle relazioni di dipendenza delle prove del primo anno[10]. Sebbene dunque tale parametro sia tradizionalmente associato ad una maggiore complessità sintattica (Hudson 1995), la presenza di relazioni di dipendenza mediamente più corte nelle prove del secondo anno potrebbe essere conseguenza di una strutturazione interna del periodo più canonica. I risultati del monitoraggio di questo parametro sintattico ci restituirebbero prove del secondo anno caratterizzate da periodi più corti, più strutturati e con dipendenze sintattiche più corte.

10 La lunghezza delle relazioni di dipendenza sintattica è qui calcolata come la distanza tra la testa e il dipendente (in tokens).

Sulla variazione di questo parametro linguistico potrebbe inoltre influire, come già discusso, una maggiore precisione dell'annotazione sintattica automatica delle prove del secondo anno. È noto che periodi molto lunghi, tipicamente caratterizzati da lunghe relazioni di dipendenza sintattica, richiedono un maggiore costo di elaborazione umana e computazionale (Miller 1956; Hudson 1995). Nel trattamento di periodi lunghi si generano ambiguità di analisi che si ripercuotono negativamente sulla precisione del processo di annotazione automatica. Sono in particolare dipendenze sintattiche lunghe a influire in modo negativo sui risultati dell'analisi (McDonald e Nivre 2007). Periodi più brevi contengono inoltre meno relazioni di dipendenza sintattica di tipo: complemento preposizionale, sia esso modificatore o argomento e designato come *comp(lement)*[11] nello schema di annotazione a dipendenze adottato[12]; oppure, *mod(ifier)*[13], tipicamente espressione di modificazione nominale o frasale. Entrambi costituiscono luoghi di maggiore ambiguità di annotazione linguistica automatica (McDonald e Nivre 2007). I risultati del monitoraggio automatico della lunghezza e dei tipi di relazioni di dipendenza sintattica vanno pertanto letti alla luce di queste considerazioni sulla precisione degli strumenti di annotazione linguistica automatica.

È inoltre interessante osservare che i periodi più corti contenuti nelle prove del secondo anno, con in media meno proposizioni per periodo[14] (*Media di proposizioni per periodo* nella Tabella 8), contengono proposizioni più lunghe (in termini di token)[15] (*Media di token per proposizione*). Questo ci fornisce ulteriore conferma di come le prove del secondo anno, sebbene più brevi, siano caratterizzate da una organizzazione del contenuto in strutture sintattiche più articolate, cioè in proposizioni più lunghe.

Inoltre, alcune delle caratteristiche sono riconducibili ad alcune delle caratteristiche di base del testo e morfo–sintattiche osservate prima. È il caso ad esempio della distribuzione delle relazioni di dipendenza sintattica che marcano la presenza di una proposizione subordinata, cioè *sub(ordinate clause)*[16], la cui diminuzione trova il corrispettivo nella diminuzione di congiunzioni subordinanti.

Dall'indagine sulla variazione della distribuzione del lessico emerge che gli studenti nel passaggio dal primo al secondo anno apprendono nuove parole diminuendo l'uso di parole che appartengono al Vocabolario di Base (De Mauro 2000), mentre non risulta statisticamente significativa la variazione distribuzionale delle parole rispetto ai tre repertori d'uso (Fondamentale, Alto Uso e Alta Disponibilità). Inoltre, le prove del secondo anno risultano lessicalmente più ricche di quelle del primo anno essendo

11 *comp* si riferisce alla relazione tra una testa e un complemento preposizionale. Questa relazione funzionale sottospecificata è particolarmente utile in quei casi in cui è difficile stabilire la natura argomentale o di modificatore del complemento; esempio: *Fu* ***assassinata*** *da un pazzo*.

12 http://www.italianlp.it/docs/ISST-TANL-DEPtagset.pdf

13 *mod* designa la relazione tra una testa e il suo modificatore; tale relazione copre modificatori di tipo frasale, aggettivale avverbiale e nominale; esempio: ***Colori*** *intensi; Per arrivare in tempo, sono* ***partito*** *molto presto*.

14 In base allo schema di annotazione adottato in questo studio, la media di proposizioni per periodo è stata calcolata come la media di teste verbali (cioè di verbi testa sintattica da cui dipende un token o un sotto–albero sintattico) sul totale di periodi presenti nel testo.

15 La lunghezza della proposizione è stata calcolata come il rapporto tra il numero totale di token della prova e il numero totale di teste verbali della prova.

16 *sub* è la relazione tra una congiunzione subordinante e la testa verbale di una proposizione subordinata; esempio: *Ha detto* **che** *non intendeva fare nulla*.

caratterizzate da un valore di Type/Token ratio[17] maggiore. Questo testimonia una crescita nel tempo delle competenze semantico–lessicali degli studenti.

Caratteristiche	I anno	II anno	Significatività
Lemmi appartenenti al Vocabolario di Base	83,19%	79,16%	0,00
Distribuzione dei lemmi rispetto ai repertori d'uso:			
Fondamentale	84,37%	83,99%	0,39
Alto Uso	10,84%	10,95%	0,96
Alta Disponibilità	4,79%	5,06	0,20
Type/token ratio (100 lemmi)	0,66	0,69	0,00
Type/token ratio (200 lemmi)	0,55	0,58	0,00

Tabella 8
Caratteristiche lessicali e significatività della variazione tra I e II anno.

5.3 Le caratteristiche linguistiche rispetto alle variabili di sfondo

L'analisi della variazione delle caratteristiche linguistiche rispetto alle variabili di sfondo considerate ha permesso di iniziare a tratteggiare come il composito background personale di ogni studente influisca sull'evoluzione delle sue abilità linguistiche. Sebbene solo uno studio, tutt'ora in corso, sull'intero corpus di produzioni scritte raccolto potrà disegnare l'intero scenario, tuttavia i risultati riportati in questo contributo – per quanto parziali – permettono di trarre alcune preliminari considerazioni.

Ne è emerso, ad esempio, come il lavoro della madre influisca in maniera statisticamente significativa sulla variazione della lunghezza del testo e sul lessico usato nelle prove scritte. Come mostra la Tabella 9, nel primo anno scrive prove più lunghe chi ha la madre che svolge professioni classificate di "Alta professionalità", mentre nel secondo anno le prove più lunghe sono scritte da chi ha la madre che svolge professioni di "Media professionalità". Solo per quanto riguarda le prove del primo anno, è risultato significativo il fatto che gli studenti la cui madre svolge professioni di "Alta professionalità" utilizzano una percentuale maggiore di lessico di Alta Disponibilità.

	Numero di token (I anno)	Numero di token (II anno)	Lessico ad Alta disponibilità (I anno)
Operai e artigiani	313,95	252,76	4,34%
Media professionalità	316,25	284,08	4,55%
Alta professionalità	239,67	202,54	5,30%
Significatività	**0,00**	**0,01**	**0,03**

Tabella 9
Variazione di caratteristiche linguistiche rispetto al lavoro della madre.

17 Misura ampiamente utilizzata in statistica lessicale, la Type/Token ratio consiste nel calcolare il rapporto tra il numero di parole tipo in un testo, il 'vocabolario' di un testo (V_c), e il numero delle occorrenza delle unità del vocabolario nel testo (C). I valori di TTR oscillando tra 0 e 1 indicano se il vocabolario di un testo è poco vario (valori vicini a 0) o molto vario (valori vicini a 1). Considerata la lunghezza media delle prove analizzate (275 tokens le prove del primo anno e 239 tokens quelle del secondo), abbiamo scelto di calcolare la TTR rispetto ai primi 100 e 200 tokens del testo.

Sulla variazione di lunghezza della prova sembrano influire tre variabili di sfondo legate alle abitudini personali degli studenti (vedi Tabella 10). Esiste una correlazione statisticamente significativa tra chi dedica più tempo alla lettura di libri e la lunghezza delle prove scritte nel secondo anno: chi legge di più scrive di più. Al contrario, chi dedica più tempo a giocare a videogiochi on–line e a guardare film scrive prove più brevi.

	Tempo dedicato a leggere libri	**Tempo dedicato a giocare a videogiochi on–line**		**Tempo dedicato a guardare film in TV, al cinema o su DVD**
	nº token II	nº token I	nº token II	nº token I
Per niente	122,50	325,62	254,73	–
Poco	243,55	305,97	284,08	408,40
Abbastanza	235,53	270,81	223,68	300,19
Molto	289,83	207,39	184,86	246,75
Significatività	**0,01**	**0,00**	**0,01**	**0,00**

Tabella 10
Variabili di sfondo che influiscono significativamente sulla lunghezza media della prova in token.

È interessante infine far osservare come la variabile territoriale influisca sulla variazioni di alcune delle caratteristiche morfo–sintattiche e sintattiche prese in esame. Esiste infatti una correlazione statisticamente significativa tra l'area urbana della scuola e la distribuzione delle congiunzioni, dei sostantivi e delle preposizioni articolate nelle prove del primo e del secondo anno, nonché dei pronomi personali nelle prove del secondo anno. Gli studenti delle scuole di periferia scrivono usando più congiunzioni e sostantivi (in entrambi gli anni scolastici), meno pronomi personali (variazione significativa solo nelle prove del secondo anno) e nelle prove del primo anno tendono a preferire il complemento oggetto in posizione post–verbale. Analizzati alla luce dei risultati di monitoraggio ottenuti per i due interi anni, questi dati ci permettono di convalidare l'ipotesi iniziale che la collocazione geografica sia fortemente correlata all'evoluzione delle abilità di scrittura degli studenti.

Area urbana	**Congiunzioni**		**Sostantivi**		**Preposizioni articolate**		**Pronomi personali**	**Complementi oggetto pre–verbali**
	I	II	I	II	I	II	II	I
Centro	6,57	5,78	17,52	18,58	2,85	3,35	0,81	82,75
Periferia	7,28	5,96	18,71	21,01	2,61	3,51	0,74	78,49
Significatività	**0,03**	**0,00**	**0,02**	**0,02**	**0,00**	**0,00**	**0,04**	**0,00**

Tabella 11
Variazione nel primo (I) e secondo (II) anno della distribuzione di alcune caratteristiche morfo–sintattiche e sintattiche rispetto all'area urbana.

6. Conclusione e sviluppi futuri

Ad oggi, in Italia non si è ancora affermata un'efficace integrazione delle tecnologie informatiche nei processi di insegnamento e apprendimento nella scuola: quali siano

le potenzialità insite nelle nuove tecnologie rimane un interrogativo aperto. In questo panorama, le tecnologie del linguaggio presentano un forte potenziale innovativo sia dal punto di vista dell'accesso al contenuto testuale sia della valutazione delle strutture linguistiche sottostanti al testo. In questo contributo, abbiamo mostrato in particolare come tali tecnologie possano fornire un valido supporto nel monitoraggio dell'evoluzione della competenza linguistica degli apprendenti.

I risultati ottenuti dall'analisi di un corpus di produzioni scritte nei primi due anni della scuola secondaria di primo grado condotta con strumenti di annotazione linguistica automatica ed estrazione automatica della conoscenza hanno dimostrato come le tecnologie del linguaggio siano oggi mature per monitorare l'evoluzione delle abilità di scrittura. Sebbene ancora preliminari rispetto al più ampio contesto della ricerca in cui si colloca il lavoro descritto in questo articolo, crediamo infatti che le osservazioni che è stato possibile qui proporre mostrino chiaramente le potenzialità dell'incontro tra linguistica computazionale ed educativa, aprendo nuove prospettive di ricerca.

Tra le linee di attività aperte da questo primo lavoro vi è l'utilizzo dell'intero corpus di produzioni scritte raccolto per lo studio e la creazione di modelli di sviluppo delle abilità di scrittura. A questo scopo, tale risorsa è stata arricchita con l'annotazione manuale di diverse tipologie di errori commessi dagli studenti e con la loro relativa correzione e stiamo al momento analizzando come questa ulteriore informazione contribuisca a definire il modo in cui le competenze linguistiche degli studenti mutino ed evolvano nel corso dei due anni scolastici presi in esame (Barbagli et al. 2015). È inoltre in corso la definizione di una metodologia che, sfruttando l'articolazione diacronica della risorsa, permetta di studiare l'evoluzione individuale delle abilità linguistiche di ogni singolo studente quantificando il ruolo svolto dall'evoluzione dei singoli tratti linguistici monitorati in modo automatico (Richter et al. 2015).

Il corpus di produzioni scritte così arricchito con l'annotazione relative agli errori commessi dagli studenti apre anche nuovi orizzonti di ricerca ad esempio nello sviluppo di sistemi a supporto dell'insegnamento (Granger 2003) o in altri compiti applicativi perseguiti all'interno della comunità di ricerca internazionale focalizzata sull'uso delle tecnologie del linguaggio in ambito scolastico ed educativo, quali ad esempio la valutazione automatica delle produzioni scritte (Attali and Burstein 2006) o la correzione automatica degli errori (Ng et al. 2013, 2014). Ad oggi tali compiti vengono per lo più realizzati per la lingua inglese e a partire da produzioni scritte di apprendenti l'inglese come lingua straniera (L2). La risorsa messa a punto nell'ambito delle attività qui descritte potrà costituire il punto di riferimento per la realizzazione di compiti simili per la lingua italiana e a partire da produzioni scritte di apprendenti la lingua madre (L1) in età scolare.

Bibliografia

Asquini, Giorgio, Giulio De Martino e Luigi Menna. 1993. Analisi della prova 9. In AA.VV, editori, *La produzione scritta nel biennio superiore. Ricerca nelle scuole superiori del Molise*, IRRSAE MOLISE, Campobasso, Lampo, pagine 77–100.

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

Attali, Yigal e Jill Burstein. 2006. Automated Essay Scoring with e–rater V.2. *Journal of Technology, Learning, and Assessment*, 4(3):1–31.

Attardi, Giuseppe, Felice Dell'Orletta, Maria Simi e Joseph Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, pagine 1–8, Reggio Emilia (Italia).

Barbagli, Alessia, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni e Giulia Venturi. 2015. CItA: un Corpus di Produzioni Scritte di Apprendenti l'Italiano L1 Annotato con Errori. In *Proceedings of the 2nd Italian Conference on Computational Linguistics (CLiC-it)*, Trento, (Italia).

Beaman, Karen. 1984. Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discorse. In Tannen D. e Freedle R., editori, *Coherence in Spoken and Written Discorse*, Norwood, N.J., pagine 45–80.

Biber, Douglas. 1993. Using Register-diversified Corpora for General Language Studies. *Computational Linguistics Journal*, 19(2):219–241.

Bonin, Francesca, Felice Dell'Orletta, Simonetta Montemagni e Giulia Venturi. 2010. A Contrastive Approach to Multi–word Extraction from Domain–specific Corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pagine 3222–3229, Valletta (Malta).

Corbett, Albert T. e John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user–adapted interaction*, 4(4):253–278.

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

Deane, Paul e Thomas Quinlan. 2010. What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, 2(2):151–177.

Dell'Orletta, Felice. 2009. Ensemble system for Part-of-Speech tagging. In *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, pagine 1–8, Reggio Emilia (Italia).

Dell'Orletta, Felice, Simonetta Montemagni, Eva M. Vecchi e Giulia 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, editori, *Percorsi migranti: uomini, diritto, lavoro, linguaggi*, Milano, McGraw-Hill, pagine 319–336.

Dell'Orletta, Felice e Simonetta 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)*, pagine 343–359, Viterbo (Italia).

Dell'Orletta, Felice, Simonetta Montemagni e Giulia Venturi. 2013a. Linguistic Profiling of Texts Across Textual Genre and Readability Level. An Exploratory Study on Italian Fictional Prose. In *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-2013)*, pagine 189–197, Hissar (Bulgaria).

Dell'Orletta, Felice, Giulia Venturi e Simonetta Montemagni. 2013b. Unsupervised Linguistically-Driven Reliable Dependency Parses Detection and Self-Training for Adaptation to the Biomedical Domain. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BIONLP-2013)*, pagine 45–53, Sofia (Bulgaria).

Dell'Orletta, Felice, Giulia Venturi, Andrea Cimino e Simonetta 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)*, pagine 2062–2070, Reykjavik (Islanda).

De Mauro, Tullio. 2000. *Grande dizionario italiano dell'uso (GRADIT)*. Torino, UTET.

Ekanadham, Chaitanya e Yan Karklin. 2015. T-SKIRT: Online Estimation of Student Proficiency in an Adaptive Learning System. In *Proceedings of the 31st International Conference on Machine Learning*, pagine 1–6, Lille (Francia).

Fabi, Aldo e Gabriella Pavan De Gregorio. 1988. La prova 9: risultati di una ricerca sui contenuti in una prova di consigli sulla scrittura. *Ricerca educativa*, 5:2–3.

Frantzi, Katerina, Sophia Ananiadou e Hideki Mima. 2000. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, Springer–Verlag.

Givón, Thomas. 1991. Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language*, 15(2):335–370.

Granger, Sylviane. 2003. Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal*, 20:465–480.

Hudson, Richard A. 1995. Measuring syntactic difficulty. Manuscript, University College, London disponibile alla pagina http://www.phon.ucl.ac.uk/home/dick/difficulty.htm

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

Lubetich, Shannon e Kenji Sagae. 2014. Data–Driven Measurement of Child Language Development with Simple Syntactic Templates. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pagine 2151–2160, Dublino (Irlanda).
Lucisano, Pietro. 1984. L'indagine IEA sulla produzione scritta. *Ricerca educativa*, 5:41–61.
Lucisano, Pietro. 1988. La ricerca IEA sulla produzione scritta. *Ricerca educativa*, 2:3–13.
Lucisano, Pietro e Guido Benvenuto. 1991. Insegnare a scrivere: dalla parte degli insegnanti. *Scuola e Città*, 6:265–279.
Masini, Andrea. 2003. L'italiano contemporaneo e le sue varietá. In I. Bonomi, A. Masini, S. Morgana e M. Piotti, editori, *Elementi di Linguistica Italiana*, Roma, Carocci, pagine 15–86.
McDonald, Ryan e Joakim Nivre. 2007. Characterizing the errors of data–driven dependency parsing models. In *Proceedings of the the EMNLP-CoNLL*, pagine 122–131, Praga (Repubblica Ceca).
Montemagni, Simonetta. 2013. Tecnologie linguistico–computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, XLII(1):145–172.
Mortara Garavelli, Bice. 2003. Strutture testuali e stereotipi nel linguaggio forense. In P. Mariani Biagini, editori, *La lingua, la legge, la professione forense. Atti del convegno Accademia della Crusca (Firenze, 31 gennaio-1 febbraio 2002)*, Milano, Giuffrè, pagine 3-19.
Miller, George A.. 1956. The magical number seven, plus or minus two: some limits on pur capacity for processing information. *Psycological Review*, 63:81–97.
Ng, Hwee T., Siew M. Wu, Yuanbin Wu, Christian Hadiwinoto e Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pagine 1–12, Sofia (Bulgaria).
Ng, Hwee T., Siew M. Wu, Ted Briscoe, Christian Hadiwinoto, Raymond H. Susanto e Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pagine 1–14, Baltimore (Maryland).
Petersen, Sarah E. e Mari Ostendorf. 2009. A machine learning approach to reading level assessment. In *Computer Speech and Language*, 23:89–106.
Piech, Chris, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas e Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. *ArXiv e-prints:1506.05908 2015*, pagine 1–13.
Purvues, Alan C. 1992. *The IEA Study of Written Composition II: Education and Performance in Fourteen Countries vol 6*. Oxford, Pergamon.
Richter, Stefan, Andrea Cimino, Felice Dell'Orletta e Giulia Venturi. 2015. Tracking the Evolution of Language Competence: an NLP–based Approach. In *Proceedings of the 2nd Italian Conference on Computational Linguistics (CLiC-it)*, 2–3 December, Trento, Italy.
Rigo, Roberta. 2005. *Didattica delle abilità linguistiche. Percorsi di progettazione e di formazione insegnanti*. Armando Editore
Roark, Brian, Margaret Mitchell e Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pagine 1–8, Praga (Repubblica Ceca).
Rouhizadeh, Masoud, Emily Prud'hommeaux, Brian Roark e Jan 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*, pagine 709–714, Atlanta (Georgia, USA).
Sagae, Kenji, Alon Lavie e Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05)*, pagine 197–204, Ann Arbor (Michigan, USA).
Schwarm, Sarah E. e Mari 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)*, pagine 523–530, Ann Arbor (Michigan, USA).
Vu, Thuy, Ai T. Aw e Min Zhang. 2008. Term Extraction Through Unithood and Termhood Unification. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pagine 631–636, Hyderabad (India).