Valutazione e misurazione automatizzata della produzione scritta

Evaluation and automatic assessment of written composition

CARLOTTA CATERINA BORGHI

La ricerca intende confrontare e correlare le misure di valutatori addestrati con alcune misurazioni automatizzate di un testo scritto da studenti in ingresso nella scuola secondaria di II grado. Ulteriore obiettivo dell'indagine è l'individuazione delle variabili linguistiche e di natura socio-culturale che influenzano i risultati nella produzione scritta.
La ricerca, che ha concluso la fase di *try-out* nell'ottobre 2010, si basa sulla prova 9 dell'indagine internazionale IEA-IPS (consigli ad un coetaneo su come scrivere un tema). I testi prodotti dagli studenti sono stati valutati da esperti formati secondo il metodo IEA-IPS (che prevede la valutazione olistica e dei tratti principali) ed elaborati da un sistema automatizzato *Eulogos* che fornisce dati sulla leggibilità, sull'appartenenza del corpus al Vocabolario di Base e su aspetti formali del testo.

The research described in this paper aims to compare and to correlate the evaluation by a group of trained rathers through automatic assessments of a written essay made by a sample of 6th grade's students. The survey also wants to understand how particular linguistic skills, students' cultural background or socio-economic status can affect the achievements of the written compositions.
The research concluded the try-out phase on October 2010 and is based on the 9th test of the IEA Written Composition international survey (suggestions to a peer about how to write a good essay). The written compositions have been evaluated according to the IEA methodology (that includes a holistic valuation and valuations of the main traits) and processed by the automatic system Eulogos, that provides data on the readability, on the belonging of the corpus to a Basic Vocabulary and on textual features.

Parole chiave: valutazione, misura automatizzata, produzione scritta, abilità linguistiche, caratteristiche testuali, variabili socio-culturali

Key words: evaluation, automatic assessement, written composition, linguistic skills, textual features, cultural backround

1. Il quadro di riferimento

I percorsi di ricerca sperimentale quantitativa hanno una collaudata tradizione di rilevazione delle abilità linguistiche attraverso le prove di comprensione del testo, mentre meno numerosi sono gli studi sulla valutazione o sulla misura della produzione scritta. La scarsa diffusione di ricerche sulla produzione linguistica è determinata probabilmente dalla difficoltà di individuare adeguati strumenti di misurazione di carattere quantitativo e dall'alto grado di eterogeneità della valutazione della produzione, che può avere un ampio margine di errore.

Tra le misurazioni di competenze linguistiche, ricoprono certamente un ruolo importante le indagini internazionali promosse dallo IEA (International Association for the Evaluation of Educational Achievement), ente fondato nel 1959 con lo scopo di indagare la possibilità di svolgere indagini empiriche comparate per poter fornire utili indicazioni per le politiche nazionali in campo educativo.

È del 1984 l'indagine sulla produzione scritta (IPS) o Written Composition Study, compiuta in sedici Paesi su un campione di studenti di tre popolazioni: studenti dell'ultimo anno della scuola primaria, della scuola media inferiore e superiore. L'Italia ha aggiunto un campione di studenti sedicenni per poter confrontare i risultati con gli altri Paesi, in cui l'obbligo scolastico comprende dieci anni di studi (Lucisano,1984; Corda Costa,Visalberghi, 1995; si veda anche rassegna bibliografica di Lucisano, Siniscalco, 1994). L'indagine è stata ovviamente occasione di confronto tra pedagogisti e linguisti di diversi Paesi per concordare una definizione teorica dell'area della produzione scritta, le tipologie di prove e i criteri di valutazione[1]. Le tipologie di scrittura proposte erano molto eterogenee tra loro e spesso differenti da quelle della tradizione scolastica italiana, che tende a privilegiare testi narrativi e non pragmatici. L'indagine prevedeva nove tipologie di prove (1. Stesura di un messaggio informativo; 2. Riassunto; 3. Ristesura di una storia; 4. Composizione descrittiva; 5. Composizione narrativa; 6. Composizione persuasiva; 7. Composizione riflessiva; 8. Composizione libera; 9. Lettera di consigli), in alcuni casi declinati in modo diverso a seconda delle popolazioni considerate (Corda Costa,Visalberghi, 1995). L'indagine aveva richiesto un'importante riflessione metodologica sulla valutazione dello scritto.

Il metodo IEA IPS comprende la valutazione olistica dell'elaborato e la valutazione dei tratti principali, ovvero la qualità del contenuto, l'organizzazione e la presentazione del contenuto, lo stile e l'adeguatezza del registro e l'uso della lingua, distinto a sua volta in grammatica, lessico, ortografia, impaginazione e calligrafia; viene inoltre prevista un'eventuale valutazione della reazione emotiva del correttore di fronte alla prova. Le aree di valutazione si riferiscono alle competenze cognitive, sociali, linguistiche e motorie, secondo lo schema riportato in tabella 1.

1 Sintesi delle premesse teoriche necessarie all'indagine si trovano in Purves,Takala, 1982.

Valutazione olistica dell'elaborato		Valutazione globale	
Valutazione dei tratti principali	Competenza cognitiva	Qualità del contenuto	
		Organizzazione e presentazione del contenuto	
	Competenza sociale	Stile e adeguatezza del registro	
	Competenza linguistica	Uso della lingua e aspetti formali	Grammatica
			Lessico
			Ortografia
			Impaginazione
	Competenza motoria		Calligrafia
Aspetto emotivo		Eventuale reazione del valutatore	

Tab. 1: Le aree di valutazione (Lloyd Jones, 1977)

Per la valutazione di ciascuna di queste aree esistono indicazioni specifiche sperimentate e declinate nelle istruzioni per la correzione. Le valutazioni prevedono un punteggio minimo di 1 e uno massimo di 5, con soglia di sufficienza 3; le prove sostenute da alunni con evidenti disagi o handicap, le prove fuori traccia, le prove indecifrabili o consegnate in bianco non sono considerate valutabili.

Ogni testo è esaminato da almeno due valutatori, incrociati con sistema di controllo, previo addestramento. La fase di valutazione vera e propria, infatti, è preceduta da un importante momento di training e di confronto sulle valutazioni condotte su un gruppo limitato di prove. L'addestramento consente di limitare la percentuale di valutazioni prive di accordo o di consenso[2].

La diffusione e il consolidamento delle conoscenze informatiche ha aperto, d'altra parte, numerose prospettive di ricerca sulla lingua: ne sono nati nuovi ambiti disciplinari, quali la linguistica computazionale o l'analisi statistica dei dati testuali, così come sono stati definiti per la lingua anglosassone dei sistemi automatizzati di valutazione.

De Mauro si sofferma sul rapporto tra linguaggio storico-naturale ed informatica, apprezzando il potenziale sostegno reciproco delle discipline: "l'informatica è in grado di aiutarci nella difficile operazione che Calvino chiamava di *spiazzamento*: badare a quel che diciamo o scriviamo, mettendoci dal punto di vista del destinatario e delle sue conoscenze". La complessità semiotica delle lingue storico-naturali impedisce che il rapporto tra linguistica

2 I risultati dell'indagine internazionale si trovano in Alan Purves (1992); considerando i risultati dei singoli Paesi sono stati individuati 11 fattori per saturazione, riferiti alla famiglia dello studente, alle sue letture e ai suoi compiti, al rapporto con i mass media e alla preparazione e all'aggiornamento dei docenti (Schick, De Masi, Green, 1992). I risultati dell'indagine in Italia si possono invece trovare in Lucisano, 1988 e Lucisano, Benvenuto, 1991; la metodologia elaborata per l'indagine è stata inoltre alla base di alcune ricerche, come la ricerca condotta per conto dell'IRRSAE del Molise (IRRSAE Molise, 1993). Per quanto riguarda in particolare l'analisi dei risultati della prova 9 si veda Fabi e Pavan De Gregorio (1988).

e informatica sia lineare. In particolare la sinonimia non è calcolabile e richiede complessi processi di disambiguazione.

> a me pare che una lingua storico-naturale sia un insieme sufficientemente plastico per poter ammettere che una parte dei suoi usi possa essere piegata a conformarsi ai requisiti di un sistema calcolabile diventando, dunque, analizzabile a questa stregua. […] dalle aree più disparate giunge la franca ammissione della irriducibilità della produzione e ricezione di un testo a processi lineari o comunque integralmente calcolabili. Dobbiamo altresì all'informatica la percezione di quanto tuttavia di automatico e automatizzabile è insediato in questi processi. Con la loro parziale simulazione, l'informatica cessa di essere un mero complemento e ausilio e diventa per i linguisti una fonte non rinunziabile di indicazioni preziose per intendere la problematica natura più che sistemica, più che algoritmica del linguaggio storico-naturale degli esseri umani[3].

In area anglosassone sono stati elaborati diversi sistemi di rilevazione automatica delle caratteristiche dei testi, che hanno mostrato valori di alta correlazione con la valutazione umana.

I più diffusi e accreditati sistemi di valutazione automatizzata della produzione scritta per la lingua inglese sono:

- il Project Essay Grade (PEG) (Page, 1966, 1994);
- l'Intellingent Essay Assessor (IEA) nelle sue declinazioni Latent Semantic Analysis (LSA) (Landauer, Foltz, 1998) e Latent Semantic Indexing (LSI) (Landauer 2003), integrato con il sistema Semantically Enhanced Latent Semantic Analysis (SELSA) (Kanejiha, Kumar, Prasad, 2003).
- l'Educational Testing Service's Electronic Essay Rater (E-RATER), diffuso presso l'Educational Testing Service (ETS) (Burstein, 1998; Rudner, Gagne 2001)
- Il Bayesian Networks (Mc Callum, Nigam, 1998; Rudnes, Liang 2002)[4].

Il sistema PEG rappresenta un primo tentativo di correzione automatizzata, concettualmente più semplice, ampiamente utilizzato nelle ricerche, mostrando alte correlazioni (r 0.87 su 20 variabili; r 0.50-0.66 su altre variabili). Il sistema si dimostra però datato nella possibilità di fornire solo punteggi di misurazione relativa tra prove dello stesso tipo e nella necessità di essere ricalibrato per ogni tipologia di prova. Page, ideatore del sistema PEG, usa un modello di regressione con elementi superficiali del testo (lunghezza del testo, lunghezza e numero delle frasi, numero delle parole, punteggiatura, numero di pronomi relativi o di altri connettivi) come variabili indipendenti e il punteggio del testo come variabile dipendente.

L'approccio di Landauer è un modello fattoriale che privilegia i contenuti dei testi, ovvero l'aspetto semantico e informativo. LSA e LSI sono, infatti, tecniche automatizzate e statistiche per confrontare le parole di un testo. Una sintesi di ricerche effettuate utilizzando il sistema IEA ha dimostrato buone correlazioni con le correzioni umane (Chung, O'Neil, 1997). Il sistema SELSA aggiunge, invece, all'approccio LSA/LSI qualche informazione sintattica (Kanejiha et al., 2003).

3 De Mauro, 1994, pp. 114-115 e 118.

4 Gli approcci dei sistemi di misurazione automatizzata vengono descritti e analizzati in alcuni articoli comparativi e in alcune ricerche che riportano risultati di positiva correlazione (Wresch, 1993; Page, 1994; Whittington, Hunt, 1999; Rudnes, Liang, 2002; Millet, 2006).

Burstein usa un modello di regressione considerando gli elementi di contenuto come variabili indipendenti, mostrando un'alta correlazione tra variabili grammaticali e semantiche da una parte e le valutazioni globali dei testi (nel 92% dei casi predice il punteggio esatto o un punteggio molto vicino, con margine r 0.1). Dei sistemi analizzati, l'E-RATER è al momento il più diffuso, trovando una sua particolare declinazione anche per la misurazione dello scritto di studenti di madrelingua non inglese. Viene definito come un sistema ibrido poiché usa, tra le sue variabili, strutture del discorso (come il sistema PEG), analisi semantica (come il sistema LSA, IEA) e variabili sintattiche. Per misurare le variabili sintattiche E-RATER conta per esempio il numero dei complementi, delle subordinate, dei pronomi relativi e dei verbi modali, misurandone la presenza per frase e per testo.

Il sistema Bayesian Network prevede due modelli applicativi nella classificazione dei testi: il modello multivariato Bernoulli, nel quale in ogni testo viene computata l'assenza o la presenza di variabili calibrate, e il modello multinominale, più adatto per analisi più complesse e in presenza di un più ampio vocabolario. Il sistema Bayesian Network raggiunge un valore predittivo in circa l'80% dei testi.

2. Il disegno della ricerca

La ricerca in una prima fase si configura come ricerca per correlazioni, e si propone di individuare le inferenze e le relazioni tra la produzione scritta e altre abilità linguistiche misurate (comprensione del testo scritto, conoscenze morfologiche verbali, conoscenze lessicali). Mira inoltre a confrontare ed eventualmente correlare le valutazioni dello scritto con alcune misurazioni automatizzate. In questo modo, si vorrebbero, da un lato, identificare le variabili linguistiche che influenzano i risultati nella produzione scritta, dall'altro stabilire delle relazioni tra le misurazioni delle abilità linguistiche misurate e altre variabili di sfondo, quali il profilo socio-culturale e linguistico della famiglia di provenienza, la tipologia dell'istituto superiore frequentato dallo studente, aspetti territoriali e voto di licenza media. I dati presentati in questo articolo riportano le prime analisi effettuate relativamente a questo solo aspetto della ricerca.

Una seconda fase dell'indagine ha una prospettiva descrittiva e ha lo scopo di delineare i profili linguistici in uscita dalla Scuola secondaria di primo grado, ovvero in ingresso nella Scuola secondaria di secondo grado. La descrizione delle abilità linguistiche della popolazione contribuisce a definire il profilo di uscita da un ordine di scuola e dunque di ingresso in un altro ordine di scuola. Questa fase potrebbe avere una ricaduta didattica ed essere utile nella programmazione della didattica di entrambi gli ordini di scuola. Inoltre, la codifica dei contenuti della prova di produzione scritta (prova 9-lettera di consigli indagine IEA-IPS) potrebbe permettere di misurare le unità di contenuto presenti nel testo e di riflettere sulla percezione della metodologia e della didattica della scrittura.

Un ultimo scopo della ricerca, nella sua fase sperimentale, è la costruzione di un complesso e strutturato modello di misurazione e di comparazione dei risultati nella produzione scritta e nelle prove di altre abilità linguistiche. Il modello permetterebbe di definire un algoritmo per la valutazione di prove di produzione scritta; l'algoritmo dovrebbe essere capace di predire il punteggio attribuito ad una prova di produzione scritta con una approssimazione ai giudizi espressi da una giuria di esperti formati secondo il metodo IEA Written Composition. L'ipotesi è che sia possibile fornire una valutazione automatizzata di una prova scritta con un accettabile grado di affidabilità, se il modello considera tutti gli aspetti rilevanti della produzione scritta. La condizione di affidabilità deve essere convalidata da una serie di giurie di valutatori esperti secondo criteri prestabiliti.

Obiettivo trasversale di ciascuna fase di ricerca è la ricaduta didattica.

Come popolazione bersaglio si è scelto di lavorare con studenti del I anno della Scuola secondaria superiore, periodo del percorso scolastico non coinvolto da altre indagini nazionali o internazionali o da esami di Stato. La scelta della popolazione è determinata anche dalla volontà di stabilire con gli Istituti e con i docenti un rapporto di collaborazione, chiaramente svincolato da intenti di giudizio sul lavoro di studenti e docenti.

È necessario che al campione siano proposti degli strumenti adeguati e somministrate diverse tipologie di prove, secondo quanto richiesto dagli obiettivi delineati; la compresenza di diversi obiettivi richiede altresì che vengano predisposte chiare metodologie d'analisi dei dati ottenuti.

Coerentemente con gli obiettivi delineati, gli strumenti comprendono una prova di produzione scritta, un insieme di prove strutturate di abilità linguistiche di tipo tradizionale e un questionario. La complessità della lingua comporta l'impossibilità di misurare ogni ambito di competenza linguistica e obbliga quindi in primo luogo a una scelta delle competenze da misurare. Accanto alla produzione scritta, sono misurate la comprensione del testo, le conoscenze lessicali e morfologiche verbali.

La prova di produzione scritta utilizzata è una delle prove dell'indagine internazionale IEA IPS, la prova 9, la lettera di consigli (*Consigli ad un coetaneo su come scrivere un tema perché sia valutato positivamente dagli insegnanti*), poiché sembra soddisfare diverse esigenze. Innanzitutto pone delle richieste precise e circoscritte, che facilitano il confronto dei testi prodotti, ma, allo stesso tempo,ottiene testi relativamente liberi, su cui è interessante l'analisi automatizzata, statistica di dati testuali o di linguistica computazionale. La scelta è inoltre motivata dal fatto che per la lettera di consigli esiste una codifica dei contenuti elaborata in sede internazionale; la codifica dei contenuti permette di misurare le unità di contenuto presenti nel testo, confrontandole eventualmente con il giudizio dei valutatori. La codifica dei contenuti offre l'opportunità, inoltre, di avviare una riflessione sulla percezione della didattica della scrittura e un confronto con i contenuti prodotti durante la somministrazione IEA a più di venticinque anni di distanza.

Agli studenti viene somministrato un adattamento della prova di abilità linguistiche (ISFOL, 2010), che comprende:

- Prove per misurare la comprensione del testo scritto: testi, con domande a scelta multipla, per un totale di 27 item e cloze casuale e mirato per 21 completamenti;
- Prove per misurare le conoscenze lessicali in contesto: testi con domande a scelta multipla, per un totale di 22 item;
- Prove per misurare l'uso dei verbi in un contesto dato: coniugazione di 10 forme verbali.;
- Un breve questionario, attraverso cui rilevare le condizioni socio-culturali della famiglia di provenienza, le abitudini linguistiche e le valutazioni di licenza media.

I testi di produzione scritta sono misurati da un sistema automatizzato, secondo il modello GULPEASE, integrato con gli sviluppi realizzati in ambiente *Èulogos*[5].

Le misure delle diverse caratteristiche degli elaborati scritti sono le seguenti:

5 www.eulogos.it

- Indice GULPEASE;
- Numero delle parole utilizzate;
- Numero delle frasi;
- Lunghezza delle frasi;
- Varianza delle parole;
- Vocabolario utilizzato (numero e percentuale del Vocabolario di base- distinto in vocabolario fondamentale, vocabolario ad alta frequenza e vocabolario ad alto uso- e vocabolario non di base).

Il disegno di ricerca contempla, ad integrazione della misurazione automatizzata, anche altre variabili linguistiche in analisi di linguistica computazionale, che considerino aspetti formali più complessi di carattere morfologico o sintattico, per raffinare la misura e completare l'aspetto descrittivo. I dati presentati in questo articolo non comprendono ancora analisi computazionali di natura morfologica e sintattica[6].

I testi prodotti dagli studenti devono ricevere anche una valutazione da una giuria di valutatori esperti secondo criteri prestabiliti. La produzione scritta viene valutata secondo i criteri elaborati nell'ambito dell'indagine IEA-IPS da un gruppo di valutatori esperti secondo le metodologie elaborate e sperimentate sempre in IEA-IPS. La correzione dei testi raccolti nella fase di *try-out* è avvenuta nell'ambito di un'esercitazione di ricerca predisposta nel corso di Laurea di Scienze dell'Educazione e della Formazione.

L'analisi dei dati raccolti con le prove strutturate per le altre abilità linguistiche è avvenuta con Item analisi classica CTT (Classical Test Theory), completata da analisi con modello di Rasch IRT (Item Response Theory). Entrambe le analisi ne hanno confermata la validità.

3. Prime analisi dei dati

Come stabilito nel disegno di ricerca, nel mese di ottobre 2010 si è svolta la fase di *try-out*. Il campione per il *try-out* doveva essere costituito da un minimo di 400 studenti (circa 20 classi), per consentire un'adeguata taratura delle analisi automatizzate dei testi. Sono stati coinvolti dieci Istituti superiori (nel dettaglio 4 Licei, 3 Istituti Tecnici e 3 Istituti Professionali di diverse zone della città) e ventidue classi, per un totale di 471 soggetti utili. Per l'aspetto della ricerca descrittivo dei profili in uscita, sarà inoltre necessario sottrarre gli studenti ripetenti della classe prima.

Le correlazioni tra punteggi nei quattro subtest di abilità linguistica e le valutazioni secondo la metodologia IEA (eccezion fatta per il criterio di valutazione di lessico) sono sempre significative. Sono più alte le correlazioni tra i subtest e le valutazioni su aspetti strutturali e di contenuto (Tab. 2).

6 È attivata una collaborazione con la dott.ssa Montemagni, il dott. Dell'Orletta e la dott.ssa Venturi dell'Istituto di Linguistica Computazionale del CNR di Pisa.

N 452		Val.Gl.	Cont.	Org.	Stile	Gram.	Less.	Ort.	Imp.	Call.
Lettura	R.	,479	,487	,398	,359	,132	-,005	,135	,284	,155
	Sig.	,000	,000	,000	,000	,005	,917	,004	,000	,001
Lessico	R	,431	,424	,391	,363	,123	,079	,147	,278	,139
	Sig.	,000	,000	,000	,000	,009	,094	,002	,000	,003
Verbi	R	,410	,437	,388	,354	,134	,042	,169	,239	,141
	Sig.	,000	,000	,000	,000	,004	,372	,000	,000	,003
Cloze	R	,393	,399	,390	,299	,152	,067	,150	,224	,121
	Sig.	,000	,000	,000	,000	,001	,155	,001	,000	0,01

Tab. 2: correlazione tra subtest e valutazioni IEA

Calcolando un punteggio fattoriale di abilità linguistica con i quattro sub test, si è ottenuta una componente che spiega il 65% della varianza dei punteggi È stata poi calcolata la correlazione tra questo punteggio fattoriale e i criteri di valutazione della produzione scritta. Anche in questo caso, ad eccezione del criterio di lessico, le correlazioni risultano sempre significative e piuttosto alte per la qualità e l'organizzazione del contenuto, oltre che per la valutazione globale (Tab. 3).

	fattoriale test abilità linguistiche
Qualità del Contenuto IEA	,542**
Valutazione Globale IEA	,532**
Organizzazione IEA	,485**
Registro e Stile IEA	,427**
Impaginazione IEA	,319**
Ortografia IEA	,186**
Calligrafia IEA	,173**
Grammatica IEA	,167**
Lessico IEA	,056

Tab. 3: correlazione tra punteggio fattoriale abilità linguistiche e valutazioni IEA

Allo stesso modo, si sono calcolate le correlazioni tra il punteggio fattoriale della abilità linguistiche ricavate dai quattro sub test e le misure automatizzate della produzione scritta (Tab. 4). La maggioranza delle correlazioni risulta significativa; sono alte le correlazioni con il numero di parole e di frasi, con il numero di parole appartenenti al vocabolario fondamentale e ad alto uso.

N. 465	fattoriale test abilità linguistiche	
	R	Sig.
Indice GULPEASE	-,139	,003
Numero Frasi	,489	,000
Lunghezza Frasi	-,034	,464
Numero Parole	,556	,000
Lunghezza Parole	,282	,000
Varianza Parole (Parole/ParoleNuove)	,344	,000
Numero Vocabolario Fondamentale	,545	,000
Percentuale Vocabolario Fondamentale	-,020	,667
Numero Vocabolario Alto Uso	,517	,000
Percentuale Vocabolario Alto Uso	,209	,000
Numero Vocabolario ad Alta Disponibilità	,240	,000
Percentuale Vocabolario ad Alta Disp.	-,063	,176
Numero Vocabolario di Base	,547	,000
Percentuale Vocabolario di Base	0,09	,053
Numero Vocabolario non di Base	,432	,000
Percentuale Vocabolario non di Base	-,088	,057

Tab. 4: correlazioni tra il punteggio fattoriale di abilità linguistiche e le misure automatizzate della produzione scritta

È stato poi calcolato un punteggio fattoriale per le variabili di valutazione della produzione scritta secondo la metodologia IEA.

Si sono ottenute diverse componenti, la prima delle quali spiega circa il 41% della varianza e la seconda il 15%.

Il primo dei punteggi così ricavati (Fat1IEA) spiega il 41% della varianza. Per questo punteggio assumono un peso rilevante i criteri di valutazione globale, di qualità e organizzazione del contenuto, di registro e stile e di impaginazione. Si tratta in sostanza di quei criteri che considerano soprattutto aspetti strutturali e di contenuto e non aspetti formali. Il criterio di impaginazione, normalmente riconducibile ad aspetti più meccanici, è probabilmente associabile ai criteri strutturali e di contenuto per le particolari caratteristiche della prova scritta, una lettera. Il criterio infatti considera aspetti formali, come il rispetto dei margini, ma anche aspetti strutturali, legati al contenuto, ovvero la presenza di elementi di impaginazione tipici della comunicazione epistolare, come il destinatario o la firma. L'analisi delle componenti del primo dei due punteggi fattoriali permette anche di desumere un forte legame tra la valutazione globale e le variabili strutturali e di contenuto più che le variabili formali.

Il secondo punteggio fattoriale ricavato (Fat2IEA) invece spiega solo il 15% della varianza ed è interessante rilevare che in esso assumono peso i criteri di grammatica, di ortografia e di lessico.

In sostanza si possono distinguere due punteggi, uno con componenti strutturali e di contenuto, l'altro con componenti formali.

La tabella sottostante (Tab. 5) riporta le correlazioni tra le misure automatizzate ricavate in ambiente *Eulogos* e questi due punteggi fattoriali di valutazione, Fat1IEA e Fat2IEA.

Le correlazioni tra le misure automatizzate e il primo punteggio fattoriale Fat1IEA sono significative per quasi tutte le variabili; vanno segnalate, in particolare, il numero di parole e

di frasi, la lunghezza e la varianza delle parole e il numero delle parole appartenenti a diversi vocabolari specifici. Alcune di queste correlazioni sono anche alte, come "il numero delle parole" e "il numero delle parole appartenenti al vocabolario fondamentale e al vocabolario di base".

Le correlazioni tra le misure automatizzate e il secondo punteggio fattoriale fat2IEA sono per lo più significative, tuttavia molto più basse.

Risulta piuttosto evidente la differenza nelle correlazioni tra le misure automatizzate e i criteri strutturali e di contenuto da una parte e formali dall'altra.

N. 452	fat1IEA variabili strutt./conten		fat2IEA variabili formali	
	R	Sig.	R	Sig.
Indice GULPEASE	-,098	,037	,089	,058
Numero Frasi	,561	,000	-,256	,000
Lunghezza Frasi	-,045	,337	-,068	,149
Numero Parole	,607	,000	-,343	,000
Lunghezza Parole	,123	,009	,094	,046
Varianza Parole (Parole/ParoleNuove)	,460	,000	-,372	,000
Numero Vocabolario Fondamentale	,605	,000	-,341	,000
Perc. Vocabolario Fondamentale	,092	,052	-,106	,024
Numero Vocabolario Alto Uso	,462	,000	-,231	,000
Percentuale Vocabolario Alto Uso	,058	,219	,054	,250
Numero Vocabolario ad Alta Disponibilità	,243	,000	-,182	,000
Perc. Vocabolario ad Alta Disponibilità	-,078	,096	,024	,607
Numero Vocabolario di Base	,605	,000	-,342	,000
Percentuale Vocabolario di Base	,119	,012	-,089	,058
Numero Vocabolario non di Base	,456	,000	-,283	,000
Perc. Vocabolario non di Base	-,119	,012	,089	,059

Tab. 5: Correlazioni tra misure automatizzate e punteggi fattoriali delle valutazioni del testo scritto

Si è calcolato, infine, il punteggio fattoriale delle 16 variabili (6 formali e 10 di vocabolario) ottenute dall'analisi automatizzata in ambiente *Eulogos*. La prima componente che si ottiene (Fat1Eu16) spiega il 37,5% della varianza e in essa assumono molto peso il numero delle frasi, delle parole, la varianza delle parole, il numero delle parole che appartengono al vocabolario fondamentale, di alto uso e di base.

Sono state dunque calcolate le correlazioni tra i singoli criteri di valutazione della produzione scritta secondo la metodologia IEA e il punteggio fattoriale delle misure automatizzate ricavate in ambiente *Eulogos* (Tab. 6)

Il punteggio fattoriale delle misure automatizzate (Fat1EU16) correla in modo significativo con i criteri di struttura e di contenuto (valutazione globale, qualità e organizzazione dei contenuti, registro e stile, impaginazione e calligrafia); le correlazioni sono piuttosto alte in particolare per i primi criteri.

N. 452	Fat1Eu16	
	R	Sig.
Valutazione Globale IEA	,640	,000
Qualità del Contenuto IEA	,680	,000
Organizzazione IEA	,588	,000
Registro e Stile IEA	,402	,000
Grammatica IEA	,006	,900
Lessico IEA	-,092	,050
Ortografia IEA	,002	,961
Impaginazione IEA	,327	,000
Calligrafia IEA	,125	,008

Tab. 6: correlazioni tra criteri di valutazione del testo scritto e punteggi fattoriali delle misure automatizzate

Si è poi proceduto ad analizzare la distribuzione dei punteggi dei subtest, dei criteri di valutazioni IEA e delle misure automatizzate della produzione scritta per le variabili di sfondo di natura socio-culturale, raccolte attraverso il questionario studenti. Vi si farà breve accenno per necessità di sintesi.

Significative sono le distribuzioni per la scelta di tipologia di scuola (liceo, istituto tecnico e istituto professionale), l'anno di nascita, il voto di licenza media, la lingua parlata a casa, la quantità di libri posseduti e il titolo di studio dei genitori; la professione della madre ha distribuzioni significative, mentre non sono significative le distribuzioni per professione del padre.

Conclusioni

Le prime analisi compiute con i dati ad oggi disponibili sul campionamento della fase di *try-out* mettono in luce numerose correlazioni significative, spesso alte, tra i diversi punteggi e valutazioni. Nel caso della produzione scritta si può osservare che i dati ricavati da due metodi di misura e valutazione così differenti presentano numerosissime correlazioni significative. Le correlazioni tra le misure automatizzate e la valutazione globale sono sempre significative, lo sono quasi sempre per la qualità del contenuto, per l'organizzazione del contenuto, il registro, lo stile e, nella maggioranza dei casi, per l'impaginazione.

Questi dati sono piuttosto interessanti perché mostrano correlazioni tra le misure automatizzate e i criteri strutturali e di contenuto, ovvero quelle valutazioni più legate all'intervento umano; in sostanza la correlazione tra le misure automatizzate è significativa proprio con quelle valutazioni che appaiono così distanti da prospettive computazionali o da approcci informatici.

Anche nella distribuzione delle valutazioni per le variabili di sfondo emerge il ruolo preponderante dei criteri di valutazioni strutturali e di contenuto. Il dato conferma l'aspetto processuale della scrittura, come insieme di comportamenti che conducono alla realizzazione di un testo scritto, tra i quali assumono un ruolo decisivo gli aspetti cognitivi e rielaborativi.

Il disegno di ricerca prevede, d'altra parte, un campionamento ancora più ampio e soprattutto ulteriori misure di linguistica computazionale, di carattere morfologico e sintattico, che consentiranno di raffinare le analisi.

Bibliografia

Burstein J. et al. (1998). Automated scoring using a hybrid feature identification technique. *Proceedings of the annual meeting of the association of computational linguistics.* Montreal: Canada, http://www.ets.org/research/aclfinal.pdf

Chung G. K.W. K., O'Neil H. F. jr (1997). *Methodological approaches to online scoring of essays.* ERIC Document Reproduction Service.

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

De Mauro T. (1994). *Capire le parole.* Bari: Laterza.

Fabi A., Pavan De Gregorio G. (1988). La prova 9: risultati di una ricerca sui contenuti in una prova di consigli sulla scrittura. *Ricerca educativa*, V, 2-3.

IRRSAE Molise (1993). *La produzione scritta nel biennio superiore. Ricerca nelle scuole superiori del Molise.* Campobasso: Lampo.

Kanejiha D., Kumar A., Prasad S. (2003). Automatic evaluation of student's answers using syntactically enhanced LSA. *Proceedings of the NAACL 2003 workshop, association for computational linguistics* (pp. 53-60). Alberta (Canada): Edmonton.

Landauer T. K., Foltz P. W., Laham D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.

Landauer T. K. (2003). Pasteur's quadrant, computational linguistics, LSA, education. *Proceedings of the NAACL 2003 Workshop, Association for Computational Linguistics* (pp. 61-67). Alberta (Canada): Edmonton.

Lloyd Jones R. (1977). Primary trait scoring. In C. R. Cooper, L. Odell, *Evaluating writing: describing, measuring, judging*, Urbana: NCTE.

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

Lucisano P. (1988). La ricerca IEA sulla produzione scritta. *Ricerca educativa*, 2-3, 3-13.

Lucisano P., Benvenuto G. (1991). Insegnare a scrivere: dalla parte degli insegnanti. *Scuola e Città,* 6, 265-279.

Lucisano P., Siniscalco M. T. (1994). Rassegna bibliografica della ricerca IEA. *Cadmo*, II, 5-6, 164-186.

Mc Callum A., Nigam K. (1998). A comparison of event models for Naïve Bayes Text Classification. *Learning for text categorization*, estratto da http://citeseer.nj.nec.com/mccallum98comparison.html.

Millet R. P. (2006). *Automatic holistic scoring of ESL essays using linguistic maturity attributes*. Thesis of Department of Linguistics and English Language. Brigham Young University.

Page E. B. (1966). Grading essays by computer: progress report. *Invitational Conference on Testing Problems*, 87-100.

Page E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62 (2), 127-142.

Purves A., Takala S. (1982). An international perspective on the evaluation of written composition. Evaluation. *Education. An international review series*, 5 (2).

Purves A. (1992). *The IEA study of written composition II: education and performance in fourteen country.* Oxford: Pergamon.

Rudner L., Gagne P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research and Evaluation*, 7 (26).

Rudnes L. M., Liang T. (2002). Automated essays scoring using bayes' theorem. *The Journal of Technology, Learning, and Assessment*, 1, 2, estratto da www.jtla.org.

Schick R., De Masi M. E., Green M. S. (1992). Factors predicting writing performances. In A.C. Purves (Ed.), *The IEA Study of written composition II: education and performance in fourteen country.* Oxford: Pergamon.

Whittington D., Hunt H. (1999). Approaches to the computerized assessment of free text responses. *Proceedings of the Third Annual Computer Assisted Assessment Conference*, 207-219, estratto da http://cvu.strath.ac.uk/dave/publications/caa99.html.

Wresch W. (1993). The imminence of grading essays by computer – 25 years later. *Computer and composition*, 10 (2), 45-58, estratto da http://corax.cwrl.utexas.edu/cac/archiveas/v10/10_2html/10_2_5_Wresch.html.