UNIVERSITÀ DI PISA

InformaticaUmanistica
SEMINARIO DI CULTURA DIGITALE

Relazione per il Seminario di cultura digitale

Tecnologie linguistico-computazionali per l'evoluzione delle competenze di scrittura

Alessio Miaschi

Matricola: 475963

A.A. 2015-2016

Introduzione

Gli ultimi dieci hanno visto l'affermarsi a livello internazionale di numerose tecnologie linguistico-computazionali per lo studio delle competenze linguistiche di apprendenti la propria lingua madre o una lingua straniera. Nonostante i diversi obiettivi, le varie ricerche sono caratterizzate dalla medesima metodologia: l'uso degli strumenti di annotazione automatica. In questa prospettiva, il testo morfo-sintatticamente e sintatticamente annotato rappresenta il punto di partenza attraverso il quale poter rintracciare informazioni fondamentali per lo studio delle competenze linguistiche. Naturalmente, l'ambito di ricerca è particolarmente ampio: si passa dallo sviluppo di strumenti in grado di valutare la qualità di scrittura di apprendenti una lingua straniera alla creazione di programmi di correzione automatica degli errori.
Un recente caso di studi ha inoltre proposto di utilizzare le informazioni linguistiche estratte automaticamente per monitorare l'evoluzione del processo di apprendimento linguistico di una lingua madre. Tale prospettiva è particolarmente innovativa, poiché parte dal presupposto che le tecnologie linguistico-computazionali possano giocare un ruolo centrale nella valutazione della competenza linguistica di apprendenti una lingua madre e nel tracciarne l'evoluzione nel tempo, incentivando così uno studio diacronico della lingua.

La seguente relazione intende focalizzarsi principalmente su questo nuovo ambito di studi. In particolare, nel primo capitolo si procederà ad una breve illustrazione del processo di annotazione linguistica automatica e dei principali scenari applicativi. Successivamente, verranno riportate le metodologie e i primi risultati dello studio finalizzato al monitoraggio dell'evoluzione delle competenze di scrittura, svolto presso l'*Italian Natural Language Processing Laboratory* (*ItaliaNLP Lab*)[1] dell'Istituto di Linguistica Computazionale "A. Zampolli"[2].
Prima degli esperimenti, sarà però necessario dedicare un capitolo della seguente relazione anche all'illustrazione del *corpus* CItA, ponendo l'accento sulle sue principali caratteristiche distribuzionali e su una delle proprietà che maggiormente lo contraddistingue: l'annotazione degli errori.

[1] http://www.italianlp.it (visitato il 23/12/2016).

[2] http://www.ilc.cnr.it/it (visitato il 23/12/2016).

1. Tecnologie linguistico-computazionali per l'analisi dei testi

Nel corso degli ultimi anni, le tecnologie linguistico-computazionali per l'analisi dei testi sono notevolmente aumentate e il loro crescente utilizzo è stato di incentivo per lo sviluppo e la teorizzazione di nuovi progetti e di nuovi approcci alla materia. Basterebbe una rapida ricerca sul web per rendersi conto della grande varietà dei progetti: estrazione di entità nominate e di relazioni semantiche, monitoraggio delle variazioni fra le lingue, valutazioni sulla leggibilità e altri ancora.
Le tecnologie linguistico-computazionali permettono dunque di accedere al contenuto informativo dei testi attraverso l'individuazione della struttura linguistica sottostante e la sua rappresentazione esplicita[3].

Prima di passare però ai possibili scenari applicativi (e alle rispettive tecnologie), può essere utile ripercorre brevemente i passaggi che permettono l'annotazione della struttura linguistica del testo. In breve, i dati linguistici che vengono raccolti - e che normalmente sono raggruppati in *corpora* - devono passare attraverso una catena di tre processi ben distinti per rendere esplicita, interpretabile ed esplorabile dal computer la struttura linguistica implicita nel testo[4]: la segmentazione in frasi e la tokenizzazione, che permette la frammentazione del testo in frasi e in parole ortografiche, l'annotazione morfo-sintattica, che associa ad ogni token (parola) informazioni relative alla categoria grammaticale che il termine ha nel contesto specifico, e l'annotazione sintattica a dipendenze, che restituisce la struttura sintattica della frase in termini di relazioni di dipendenza.
Oggi, nel campo della linguistica computazionale, lo stato dell'arte nei compiti di annotazione linguistica è rappresentato da sistemi basati su algoritmi di apprendimento automatico.

Una volta estratte le caratteristiche grammaticali di base, è possibile accedere alla struttura linguistica del testo e calcolare così una serie di informazioni statistiche fondamentali, come la distribuzione delle categorie morfo-sintattiche e sintattiche, per la definizione del profilo linguistico e per lo sviluppo delle maggiori tecnologie ad oggi disponibili.

Gli scenari applicativi

Dati gli strumenti di annotazione e una serie di misure per descrivere il profilo linguistico dei vari *corpora*, i possibili scenari applicativi sono molteplici. Si va da tecnologie per la classificazione del genere testuale a strumenti in grado di riconoscere la lingua madre di uno scrivente o, addirittura, di identificare l'autore del testo. A questo proposito, un approccio interessante riguarda tutte quelle metodologie e quei tool sviluppati per il calcolo della leggibilità di un testo. L'obiettivo, in questo caso, è di utilizzare gli strumenti di annotazione linguistica e le misure di leggibilità (come, ad

[3] S. Montemagni, *Tecnologie linguistico-computazionali per il monitoraggio della lingua italiana*. "Studi Italiani di Linguistica Teorica e Applicata" (SILTA), anno XLII, numero 1, 2013, pp. 145-172.

[4] A. Lenci, S. Montemagni, V. Pirrelli 2005, Testo e computer, Carocci, Roma, pag. 211.

esempio, la formula Flesch-Kincaid[5]) per definire la complessità di un testo e, di conseguenza, il suo grado di comprensibilità.
Per quanto riguarda la lingua italiana, il primo e unico strumento in grado di calcolare la leggibilità di un testo è rappresentato da READ-IT[6], sviluppato dall'*Italian Natural Language Processing Laboratory* e concepito per fornire un indice di leggibilità avanzato basato su analisi linguistica multi-livello del testo.

Monitoraggio dell'evoluzione delle competenze di scrittura

Le metodologie e i sistemi per il monitoraggio dell'evoluzione delle competenze di scrittura, che verranno illustrate nei capitoli successivi di questa relazione, rientrano a pieno titolo tra le moderne tecnologie che sfruttano i sistemi di analisi linguistica automatica. Utilizzando strumenti di NLP (*Natural Language Processing*) e set di informazioni linguistiche (distribuzioni di categorie grammaticali, errori o misure di complessità lessicale) è infatti possibile investigare la qualità linguistica dei testi per poi tracciare l'evoluzione delle competenze di scrittura. Tale compito, oltre a rappresentare un ulteriore stimolo per lo studio linguistico e per lo sviluppo delle tecnologie in sé, può fornire un'importante supporto per la realizzazione di strumenti computazionali e di metodologie sempre più raffinate. Si pensi alla possibilità, in ambito scolastico, di monitorare automaticamente le prove degli studenti per permettere a tutti di sviluppare al meglio le proprie competenze linguistiche e di scrittura.

In particolare, gli esperimenti che verranno illustrati nei capitoli successivi rientrano a pieno titolo in questo ambito di studi e sono stati effettuati sfruttando algoritmi di apprendimento automatico a partire da un *corpus* di temi scritti da studenti italiani raccolti nel corso dei primi due anni di scuola secondaria di primo grado[7].

---

[5] J. P. Kincaid, R. Lieutenant, R.P. Fishburne, R. L. Rogers, B. S. Chissom, *Derivation of new readability formulas for Navy enlisted personnel*, Research Branch Report, Millington, TN: Chief of Naval Training, 1975, pp. 8-75.

[6] F. Dell'Orletta, S. Montemagni, G. Venturi, *READ-IT: assessing readability of Italian texts with a view to text simplification*, in "Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2011)", Edimburgo, UK, 30 luglio 2011, 2011, pp. 73-83.

[7] S. Richter, A. Cimino, F. Dell'Orletta, G. Venturi, *Tracking the Evolution of Written Language Competence: an NLP–based Approach*. In "Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it), 3-4 December 2015, Trento", Italia, 2015, pp. 31-35.

2. Il corpus CItA

Come già accennato nel capitolo precedente, i sistemi linguistico-computazionali basati su algoritmi di apprendimento automatico richiedono grandi quantità di dati per poter essere eseguiti. Questi dati vengono normalmente raggruppati in *corpora*, collezioni di testi scritti o orali prodotti in contesti comunicativi reali e conservati in formato elettronico[8].

Allo scopo di promuovere uno studio diacronico sull'evoluzione delle abilità di scrittura, è stato preso in esame un corpus di 1352 prove scritte da 156 studenti di sette diverse scuole secondarie di primo grado di Roma[9]. In particolare, 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 per ogni studente sono state raccolte due tipologie di prove scritte: le tracce assegnate dai docenti nei due anni scolastici e due prove comuni (una per anno scolastico) relative alla percezione dell'insegnamento della scrittura.

Inoltre, al *corpus* è stato associato un questionario contenente 34 quesiti relativi al contesto biografico, socio-culturale e socio-linguistico degli studenti[10].

	Centro	Periferia	Totale
Riflessivo	49	18	67
Narrativo	21	10	31
Descrittivo	2	1	3
Espositivo	4	6	10
Argomentativo	7	6	13
Totale	83	41	124

Tabella 1. Distribuzione delle tipologie dei temi.

Com'è possibile osservare nella Tabella 1, nel corso dei due anni scolastici i docenti hanno chiesto agli studenti di produrre diverse tipologie di prove scritte, raggruppabili in cinque macro-categorie testuali: riflessiva, narrativa, descrittiva, espositiva e argomentativa. Le diverse distribuzioni seguono abbastanza fedelmente l'approccio all'insegnamento della scrittura da parte degli insegnanti: la stesura di un tema narrativo è da considerarsi la più semplice, mentre la produzione di

[8] Enciclopedia Treccani, voce *Corpora*, http://www.treccani.it/enciclopedia/corpora-di-italiano_(Enciclopedia-dell'Italiano) (visitato il 16/12/2016).

[9] A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi, *Il ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: primi risultati*, In Italian Journal of Computational Linguistica (IJCoL), vol. 1, n. 1, 2015, pp. 99-117.

[10] A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi, *CItA: an L1 Italian Learners Corpus to Study the Development of Writing Competence*. In Proceedings of 10th Edition of International Conference on Language Resources and Evaluation (LREC 2016), 23-28 maggio, Portorož, Slovenia, 2016, pp. 88-95.

testi a carattere espositivo o argomentativo è assai più complessa e richiede competenze linguistiche e discorsive particolarmente avanzate.
Inoltre, la Tabella 1 mette in risalto un'altra importante caratteristica: gli insegnanti delle scuole del centro città tendono ad assegnare un numero più alto di prove scritte rispetto ai colleghi delle scuole di periferia.

Analisi della struttura linguistica

Da un primo studio, effettuato esclusivamente sulle 240 prove comuni, sono state estratte alcune informazioni particolarmente interessanti per il monitoraggio dell'evoluzione delle competenze di scrittura. Si è notato, infatti, che la lunghezza del testo e la lunghezza media dei periodi variano in modo statisticamente significativo nel passaggio dal primo al secondo anno scolastico: se nel primo anno gli studenti tendono a scrivere prove più lunghe, nel secondo i testi sono più brevi e contengono periodi più corti. Inoltre, si nota un evidente aumento della punteggiatura (da 9,70% a 10,60%) e dell'uso dei verbi modali (da 1,09% a 1,81%), a discapito però di un modo verbale semplice come l'indicativo. Tale variazione, più che indicare una diminuzione della complessità sintattica del testo, può essere interpretata come un indice di un maggiore ordinamento lineare del contenuto.
Dall'indagine sulla variazione del lessico, invece, emerge che gli alunni nel corso dei due anni apprendono nuove parole, diminuendo l'utilizzo di termini appartenenti al Vocabolario di Base e producendo prove lessicalmente sempre più ricche.
Infine, comparando questi primi risultati come le variabili di sfondo, si è notato che il lavoro della madre influisce significativamente sulla lunghezza dei testi e sul lessico utilizzato, che esiste una forte correlazione tra chi dedica più tempo alla lettura e la lunghezza delle prove scritte e che la variabile territoriale (la distinzione fra Centro e Periferia) gioca un ruolo particolarmente significativo nella variazione di alcune della caratteristiche morfo-sintattiche e sintattiche che sono state estratte.

Annotazione degli errori

Una delle caratteristiche principali del *corpus* è che le prove scritte sono state annotate manualmente dagli insegnanti con le diverse tipologie di errori. L'annotazione degli errori rappresenta un compito particolarmente complesso, data la mancanza di una tassonomia da poter applicare nella *corpus annotation* e data l'arbitrarietà stessa della definizione di errore. Per questo motivo, è stato necessario predisporre un nuovo schema di annotazione, a partire dalla definizione di italiano neo-standard di Gaetano Berruto[11]. Tale schema è stato suddiviso in quattro macro-classi principali: errore grammaticale, errore ortografico, errore lessicale e omissione.

[11] Definito come l'italiano parlato realmente in tutta Italia nei punti in cui si discosta dalla lingua delle grammatiche. Wikipedia, voce *Italiano neostandard*, https://it.wikipedia.org/wiki/Italiano_neostandard (visitato il 17/12/2016).

In accordo al formato di annotazione definito da Hwee Tou Ng et al., la struttura per la codifica degli errori è la seguente:

> [...] scapparono al piano di sopra e dal <M t="200" c="buio">buglio</M> <M t="113" c="spuntò">spuntarono</M> un esercito [...]

Dove i tag di apertura <M> e di chiusura </M> delimitano l'errore, l'attributo *t* (*type*) ne identifica la tipologia e l'attributo *c* (*correction*) riporta la forma corretta.

Osservando la distribuzione degli errori, si possono evidenziare alcune caratteristiche significative. Anzitutto, gli errori grammaticali e ortografici sono di gran lunga i più frequenti (46,55% e 47,33%), e circa il 22.32% del totale riguarda gli errori ortografici non classificati (appartenenti perciò alla categoria *Other*). Inoltre, la maggior parte degli errori mostra una variazione statistica particolarmente significativa nel corso dei due anni. Ad esempio, per quanto riguarda l'uso dei tempi verbali, si passa da una frequenza del 7.78% nel primo anno, al 15.67% del secondo. Tale variazione potrebbe essere dovuta alla diverse tipologie di temi assegnate dai docenti. Infatti, se durante il primo anno agli studenti era stato chiesto di produrre principalmente temi narrativi, durante il secondo il numero di prove scritte a carattere argomentativo ed espositivo (le più complesse) è aumentato.
Oltre a ciò, è importante notare che la distribuzione di alcune tipologie di errori sembra essere fortemente correlata alle informazioni di contesto raccolte nei questionari. Non è un caso, quindi, che studenti che affermano di leggere molto commettano meno errori di tipo lessicale nel corso dei due anni. Come si può inoltre osservare dalla Tabella 2, la quantità media di errori grammaticali decresce per tutte le scuole del centro città, mentre aumenta per due sedi della periferia.

Scuola		I anno	II anno
Centro	**1**	2,6	0,9
	2	5,2	3,1
	3	15,1	9,3
Periferia	**4**	3,5	8,2
	5	6,4	4,6
	6	5,4	4,6
	7	1,5	2,8

Tabella 2. Media delle occorrenze degli errori grammaticali rispetto alle scuole.

Esiste poi una correlazione statisticamente significativa tra l'area urbana della scuola e la distribuzione di alcune componenti linguistiche, quali le congiunzioni, i sostantivi, le preposizioni articolate e i pronomi personali. Dalla Tabella 3 si evince che gli studenti delle scuole di periferia scrivono usando più congiunzioni e sostantivi, ma meno pronomi personali e preposizioni articolate.

Area	Congiunzioni	Sostantivi	Preposizioni articolate	Pronomi personali
Centro	6,17	18,05	3,1	0,81
Periferia	6,62	19,86	3,06	0,74

Tabella 3. Variazioni di distribuzioni di alcune caratteristiche linguistiche rispetto all'area urbana.

Infine, comparando la media degli errori commessi da studenti nati in Italia e all'estero, notiamo una differenza particolarmente indicativa: nel passaggio dal primo al secondo anno, la produzione di errori da parte di studenti nati in un altro paese decresce notevolmente, sopratutto se comparata a quella degli alunni italiani. Tale variazione, però, non deve sorprendere: nonostante il livello iniziale di capacità grammaticale sia più basso, gli studenti provenienti dall'estero hanno maggiore possibilità di migliorare le proprie competenze linguistiche in un breve lasso di tempo.

3. Esperimenti di monitoraggio linguistico

L'approccio utilizzato per tracciare l'evoluzione delle competenze di scrittura si basa sulla metodologia applicata e discussa nell'articolo *Tracking the Evolution of Written Language Competence: an NLP-based Approach* di Stefan Richter et al. L'idea di base è che, dato un set di temi ordinati cronologicamente scritti dallo stesso studente, un documento d_j dovrebbe avere una qualità linguistica più alta rispetto ad uno scritto in precedenza. Seguendo questo approccio, il monitoraggio dell'evoluzione delle competenze linguistiche può essere visto come un problema di classificazione: dati due documenti d_i e d_j scritti dallo stesso studente, si vuole identificare se $t(d_i) > t(d_j)$, dove $t(d_i)$ indica il tempo in cui è stato scritto il tema.
Di conseguenza, è stato necessario predisporre un classificatore[12] in grado di assegnare ad ogni coppia di documenti (d_i, d_j) due possibili classi: 1 nel caso in cui $t(d_i) > t(d_j)$, 0 altrimenti. Per fare ciò, per ogni coppia di temi è stato estratto un vettore di *features* linguistiche (V_i, V_j) tramite il quale è stato poi possibile costruire l'evento da fornire al classificatore:[13]

$$E = V_i + V_j + (V_i - V_j)$$

Dove $V_i - V_j$ corrisponde alla differenza vettoriale tra le *features* dei due documenti.
Il classificatore è stato sviluppato utilizzando le *Support Vector Machines* (*SVM*)[14] come algoritmo di apprendimento.

I corpora

Prima di passare alla fase di classificazione vera e propria, è stato necessario creare una serie di *corpora* per le diverse tipologie di esperimenti, partendo dall'unico corpus di riferimento (*CItA*). Infatti, data una collezione di temi scritti e dato l'evento da fornire al classificatore, le possibili configurazioni temporali sono assai numerose: si passa dal confronto fra tutti gli scritti prodotti nel corso dei due anni da ogni studente, al confronto delle sole prove comuni.
In particolare, per i nostri esperimenti è stato deciso di creare i *corpora* a partire da otto diversi ordini temporali, confrontando per ogni alunno: tutte le prove realizzate nel corso dei due anni, le prove comuni, le prove del primo anno con quelle del secondo, le prove con distanza uguale ad 1[15],

---

[12] Un classificatore è una mappatura da uno spazio (discreto o continuo) di *feature X* a un insieme di *etichette Y*. Wikipedia, Voce *Classificatore (matematica)*, https://it.wikipedia.org/wiki/Classificatore_(matematica) (visitato il 19/12/2016).

[13] In fase di *training*, ad ogni evento è stata associata la classe corrispondente [0,1] in base alla sua stessa conformazione: 0 quando V_i precede V_j, 1 altrimenti.

[14] Metodologie di apprendimento supervisionato che appartengono alla famiglia dei classificatori lineari. Dato un set di eventi (*training set*), ognuno dei quali è stato annotato con una classe, l'algoritmo SVM costruisce un modello che assegna ad ogni classe i corrispettivi eventi. In fase di *testing*, l'algoritmo predice la classe dei nuovi esempi in base alle informazioni estratte nella fase precedente. Wikipedia, Voce *Support vector machine*, https://en.wikipedia.org/wiki/Support_vector_machine (visitato il 19/12/2016).

[15] Ovvero le prove realizzate con distanza minima una tra l'altra (prima prova con seconda, seconda con terza, ecc.).

il primo tema con il penultimo tema (nel singolo anno e nel biennio), il primo tema con la prova comune (nel singolo anno e nel biennio).
La Tabella 4 permette di confrontare gli ordini di grandezza dei *test corpora* per i diversi compiti di classificazione:

	Tutte le prove	**Prove comuni**	**I° anno - II° Anno**	**Distanza = 1**	**Iª - ultima prova (anno)**	**Iª - ultima prova (biennio)**	**Iª - prova comune (anno)**	**Iª - prova comune (biennio)**
Media eventi	2302, 28	168	833,14	181,14	82,85	42	70,85	32,85

Tabella 4. Medie degli eventi dei *test corpora* per i diversi ordini temporali.

Naturalmente, per ogni ordine temporale sono stati generati tanti *corpora* quante sono le scuole prese in considerazione nell'indagine. Di conseguenza, ogni compito di classificazione è stato ripetuto sette volte e, per ognuno di questi compiti, sono stati prodotti un *training set* e un *test set* necessari per l'apprendimento tramite SVM.

Le features

La maggior parte delle *features* utilizzate per i diversi compiti di classificazioni sono state estratte automaticamente dai testi, sfruttando i metodi di annotazione e di analisi linguistico-computazionale citati nel primo capitolo. In particolare, per l'analisi morfo-sintattica è stato utilizzato il POS tagger descritto nell'articolo *Ensemble system for Part-of-Speech tagging* di Felice Dell'Orletta et al[16], mentre per quanto riguarda l'annotazione sintattica si è fatto riferimento al parser DeSR[17].
Dal punto di vista morfo-sintattico, le informazioni linguistiche principali riguardano la densità lessicale (intesa come il rapporto tra i termini pieni e il numero totale di parole in un testo), la distribuzione dei tempi verbali e la distribuzione generale delle parti del discorso. Dal punto di vista sintattico, invece, si è deciso di prendere in considerazione la distribuzione delle dipendenze e delle subordinate, le diverse configurazioni degli alberi sintattici, la lunghezza dei *link* sintattici (intesa come la distanza che occorre tra una testa sintattica e il suo dipendente), e altre caratteristiche minori.
Inoltre, nel vettore di *features* linguistiche sono state inserite anche alcune informazioni base di carattere lessicale: lunghezza media delle frasi e delle parole, composizione del vocabolario interno[18] e *Type/Token Ratio* (rapporto tra parole tipo[19] e numero totale di parole).

---

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

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

[18] Calcolata in riferimento al Grande dizionario italiano dell'uso (Gradit) di Tullio De Mauro.

[19] Due parole (*token*) appartengono allo stesso tipo se sono identiche a prescindere dalla posizione nel testo. Wikipedia, Voce *Linguistica Computazionale*, https://it.wikipedia.org/wiki/Linguistica_computazionale (visitato il 20/12/2016).

Per la prima fase di esperimenti, sono state dunque prese in considerazione 147 *features* linguistiche per ogni tema. Successivamente, è stato deciso di riprodurre gli esperimenti aggiungendo ai vettori precedentemente creati due ulteriori insiemi di caratteristiche linguistiche: la *Words Frequency class* e le statistiche relative all'annotazione degli errori. Tali informazioni saranno meglio descritte nelle sezioni successive.

Primi risultati

Dopo aver predisposto i diversi ordini temporali e i dati linguistici da inserire nei vettori di *features*, è stato possibile passare ad una prima fase di esperimenti. Come già illustrato all'inizio di questo capitolo, per ogni *corpora* sono state generate tutte le coppie di temi possibili dello stesso studente in accordo all'evento $E = V_i + V_j + (V_i - V_j)$. Dal momento che molti algoritmi di apprendimento automatico richiedono che i valori in input appartengano ad un range standard, prima di passare alla fase di classificazione vera e propria è stato deciso di normalizzare tutti i valori delle *features*, scalandoli all'interno dell'intervallo chiuso [0,1].
La Tabella 5 raccoglie una selezione dei risultati più significativi calcolati sui *test set* e in merito all'*F-Score*[20].

	Scuole							**Media**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
	Prove comuni							
F-Score	1,00	0,87	0,89	1,00	1,00	1,00	1,00	0,94
	I° anno - II° anno							
F-Score	0,75	0,69	0,54	0,71	0,65	0,47	0,71	0,63
	Prove a distanza = 1							
F-Score	0,59	0,50	0,55	0,59	0,44	0,49	0,57	0,52
	Iª prova - penultima prova (anno)							
F-Score	0,48	0,57	0,61	0,46	0,50	0,60	0,62	0,54
	Iª prova - penultima prova (biennio)							
F-Score	0,84	0,80	0,57	0,70	0,59	0,57	0,83	0,70
	Iª prova - ultima prova (anno)							
F-Score	0,99	1,00	1,00	1,00	0,97	1,00	0,88	0,98

Tabella 5. Risultati della prima fase di esperimenti.

[20] Unità di misura statistica per il calcolo della precisione di un *test set*. Wikipedia, Voce *F1 Score*, https://en.wikipedia.org/wiki/F1_score (visitato il 20/12/2016).

Com'è possibile osservare dai dati estratti, i risultati più significativi si ottengono negli esperimenti che coinvolgono le prove comuni. Tale risultato è però piuttosto comprensibile. Infatti, se il confronto viene fatto tra un tema qualunque e una prova comune, l'operazione di classificazione si riduce ad un compito di identificazione dell'argomento trattato nel testo (task particolarmente semplice per gli algoritmi di apprendimento automatico). Invece, nel caso in cui il confronto riguarda due prove comuni, è ipotizzabile che le argomentazioni utilizzate negli scritti subiscano un notevole cambiamento nel passaggio tra il primo e il secondo anno scolastico.
Al contrario, i restanti esperimenti sono caratterizzati da punteggi più bassi e da una forte variabilità interna. Ad esempio, nel quinto ordine temporale, si passa da uno 0,84 di precisione per la prima scuola ad uno 0,57 per la sesta. Prevedibilmente, i valori più bassi si ottengono nel compito che riguarda i temi prodotti a distanza minima. È particolarmente difficile infatti determinare l'ordine cronologico di due testi che sono stati scritti in un intervallo di tempo molto ravvicinato.

In generale, questi primi risultati hanno messo in risalto la complessità del compito di classificazione, incentivando l'utilizzo di ulteriori *features* per specializzare le caratteristiche dei singoli documenti e garantire una miglior identificazione dell'evoluzione delle competenze di scrittura.

Words Frequency class e i nuovi risultati

Per la seconda fase di esperimenti, alle 147 *features* di ogni tema è stato deciso di aggiungere un altro set di caratteristiche linguistiche, riconducibili alla nozione di *words frequency class*, ovvero alla media delle frequenze della classe di ogni lemma e forma presenti all'interno di ogni testo. La frequenza della classe viene calcolata tramite la formula seguente:

$$C_{cw} = \lfloor \log_2 \frac{freq(MFL)}{freq(CL)} \rfloor$$

Dove *MFL* (*Most Frequent Lemma*) rappresenta il lemma o la forma più frequente all'interno di un corpus di riferimento e *CL* (*Current Lemma*) corrisponde al lemma o alla forma presa in considerazione.
Per ottenere risultati in grado di rispecchiare al meglio le caratteristiche distribuzionali della lingua italiana, è stato deciso di utilizzare itWAC[21] come *corpus* di riferimento. Composto da quasi due miliardi di parole, itWAC è un *corpus* di testi ricavati con metodi automatici dal web che può essere scaricato liberamente e consultato online tramite un'interfaccia a pagamento[22].
Una volta estratte tutte le frequenze dei lemmi e delle forme dal corpus itWAC (annotato linguisticamente), è stato possibile calcolare la *words frequency class* per ogni tema. Allo scopo di

[21] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora*. In Language resources and evaluation, 43, 3, 2009, pp. 209-231.

[22] Enciclopedia Treccani, Voce *Corpora di italiano*, http://www.treccani.it/enciclopedia/corpora-di-italiano_(Enciclopedia-dell'Italiano)/ (visitato il 20/12/2016).

aggiungere ulteriore informazione linguistica, al vettore di *features* iniziali è stato deciso di aggiungere non solo i valori della *words frequency class* globale, ma anche quelli di tre parti del discorso isolate: sostantivi, verbi e aggettivi.

Aggiunte le nuove *features* (8 in totale), si è nuovamente passati alla fase di classificazione. La Tabella 6 ne raccoglie i risultati più significativi:

	Scuole							**Media**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
	Prove comuni							
F-Score	1,00	0,84	0,94	1,00	1,00	1,00	1,00	0,95
	I° anno - II° anno							
F-Score	0,78	0,71	0,59	0,71	0,73	0,46	0,77	0,67
	Prove a distanza = 1							
F-Score	0,56	0,53	0,54	0,54	0,45	0,49	0,51	0,52
	Iª prova - penultima prova (anno)							
F-Score	0,51	0,61	0,57	0,41	0,50	0,64	0,58	0,54
	Iª prova - penultima prova (biennio)							
F-Score	0,86	0,74	0,54	0,85	0,50	0,59	0,83	0,70
	Iª prova - ultima prova (anno)							
F-Score	1,00	1,00	1,00	1,00	1,00	1,00	0,97	0,99

Tabella 6. Risultati della seconda fase di esperimenti.

Osservando la tabella, ci rendiamo subito conto che l'andamento dei risultati non differisce particolarmente da quello della prima fase di classificazione e ciò non ci permette dunque di trarre nuove considerazioni sull'efficienza globale dei compiti selezionati. Tuttavia, in tre esperimenti si ottengono dei risultati leggermente più alti. In particolare, nell'ordine temporale che mette a confronto tutte le prove del primo anno con quelle del secondo, l'incremento medio è di circa quattro punti. Questo dato è particolarmente significativo e ci permette di confermare l'efficacia delle nuove *features*. Infatti, la tipologia di esperimenti soggetta ad un maggiore incremento è composta da circa 833 eventi, un valore particolarmente alto soprattuto se messo a confronto con gli altri compiti. Questa informazione potrebbe dunque suggerire che la *words frequency class* contribuisce a facilitare la classificazione, ma che tale contributo si può riscontrare solo negli esperimenti condotti sui *corpora* più grandi.

L'annotazione degli errori e gli ultimi risultati

Come già annunciato nel capitolo precedente, l'annotazione degli errori e la predisposizione di uno schema di codifica specifico per la lingua italiana rappresenta una delle principali innovazioni del *corpus* CItA.
Per la terza fase di esperimenti è stato dunque deciso di sfruttare tale potenzialità, andando così ad aggiungere alle *features* di ogni tema una serie di informazioni relative alla produzione degli errori. In particolare, ad ogni vettore è stato deciso di concatenare la frequenza di ogni tipologia di errore e la somma totale, ripartita nelle quattro categorie principali: grammatica, ortografia, lessico e omissione.
Nonostante buona parte del lavoro di annotazione fosse già stato compiuto direttamente dagli insegnanti, per quest'ultima fase di classificazione si è preferito ripetere l'operazione di codifica al fine di raccogliere la reale distribuzione del dato linguistico.

	Scuole							**Media**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
	Prove comuni							
F-Score	1,00	0,89	0,94	1,00	1,00	1,00	1,00	0,96
	I° anno - II° anno							
F-Score	0,87	0,76	0,67	0,73	0,77	0,58	0,78	0,73
	Prove a distanza = 1							
F-Score	0,58	0,54	0,55	0,59	0,46	0,53	0,50	0,53
	Iª prova - penultima prova (anno)							
F-Score	0,48	0,57	0,54	0,50	0,69	0,59	0,50	0,55
	Iª prova - penultima prova (biennio)							
F-Score	0,80	0,85	0,48	0,85	0,53	0,70	0,92	0,73
	Iª prova - ultima prova (anno)							
F-Score	1,00	1,00	1,00	1,00	1,00	1,00	0,97	0,99

Tabella 7. Risultati dell'ultima fase di esperimenti.

Osservando i risultati della Tabella 7, si può facilmente notare che la qualità dei risultati è aumentata rispetto agli esperimenti precedenti. Ad esclusione dell'ultimo, infatti, tutti gli altri compiti di classificazione presentano un incremento nei valori di *F-Score*. In particolare, nell'ordine temporale che mette a confronto tutte le prove del primo anno con quelle del secondo, si ottiene un miglioramento di 10 punti rispetto alla prima serie di esperimenti e di 6 punti rispetto alla seconda. Tale incremento è assai significativo, poiché si mostra in netto contrasto con la tesi esposta

nell'articolo di Stefan Richter at al.[23], dove si affermava che l'aggiunta delle informazioni linguistiche relative all'annotazione degli errori non produce un particolare miglioramento nel *task* di classificazione.

In conclusione, i risultati ottenuti dalle tre fasi di esperimenti ci permettono di trarre ancora due considerazioni. Anzitutto, come si è potuto facilmente osservare nelle pagine precedenti, più ampio è il lasso temporale tra due temi scritti, maggiore è la precisione raggiunta dal classificatore. Ciò è dovuto principalmente al fatto che la crescita della qualità di scrittura è fortemente correlata all'arco temporale. Inoltre, il basso risultato che si ottiene nel confronto fra la prima e la penultima prova scritta del singolo anno, potrebbe suggerire che quest'ultimo gruppo di temi è stato assegnato ad una tipologia linguistica più complessa.

[23] S. Richter, A. Cimino, F. Dell'Orletta, G. Venturi, *Tracking the Evolution of Written Language Competence: an NLP–based Approach*. In "Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it), 3-4 December 2015, Trento", Italia, 2015, pp. 31-35.

Conclusioni

Abbiamo visto come le tecnologie linguistico-computazionali possano rappresentare un potente strumento per l'analisi automatica del linguaggio e, nel nostro caso specifico, per il monitoraggio delle competenze linguistiche. In particolare, abbiamo potuto osservare che tali tecnologie possono essere sfruttate anche per promuovere uno studio diacronico del linguaggio. Da questo punto di vista, i risultati estratti nelle diverse fasi di sperimentazione sono promettenti e, in alcuni casi specifici, particolarmente accurati. Naturalmente, il margine di miglioramento è ancora ampio e gli studi sono ancora in corso. A questo proposito, un futuro caso d'indagine potrebbe consistere in una fase di *feature selection*, ovvero di selezione ed estrazione delle features più significative, in modo da osservare eventuali miglioramenti nei risultati e, soprattutto, per stilare una classifica delle proprietà linguistiche più rilevanti. Inoltre, confrontando le probabilità con le quali il classificatore predice una determinata classe e le risposte date dagli studenti all'interno dei questionari si potrebbero estrarre informazioni significative.
In generale, l'auspicio è che l'utilizzo di tali risorse e i rispettivi risultati possano costituire un nuovo punto di riferimento per la realizzazione di compiti linguistico-computazionali per lo studio, sincronico e diacronico, della lingua italiana e non solo.

Bibliografia

- A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi, *Il ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: primi risultati*, In Italian Journal of Computational Linguistica (IJCoL), vol. 1, n. 1, 2015, pp. 99-117.
- A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi, *CItA: an L1 Italian Learners Corpus to Study the Development of Writing Competence*. In Proceedings of 10th Edition of International Conference on Language Resources and Evaluation (LREC 2016), 23-28 maggio, Portorož, Slovenia, 2016, pp. 88-95.
- A. Lenci, S. Montemagni, V. Pirrelli 2005, Testo e computer, Carocci, Roma, pag. 211.
- F. Dell'Orletta, *Ensemble system for Part-of-Speech tagging*. In Proceedings of Evalita '09 (Evaluation of NLP and Speech Tools for Italian), Reggio Emilia, 2009.
- F. Dell'Orletta, S. Montemagni, G. Venturi, *READ-IT: assessing readability of Italian texts with a view to text simplification*, in "Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2011)", Edimburgo, UK, 30 luglio 2011, 2011, pp. 73-83.
- G. Attardi, F. Dell'Orletta, M. Simi, J. Turian, *Accurate Dependency Parsing with a Stacked Multilayer Perceptron*. In Proceedings of Evalita '09 (Evaluation of NLP and Speech Tools for Italian), Reggio Emilia, 2009
- H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, J, Tetreault, *The CoNLL-2013 Shared Task on Grammatical Error Correction*. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, 2013, pp. 1-12.
- J. P. Kincaid, R. Lieutenant, R.P. Fishburne, R. L. Rogers, B. S. Chissom, *Derivation of new readability formulas for Navy enlisted personnel*, Research Branch Report, Millington, TN: Chief of Naval Training, 1975, pp. 8-75.
- M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora*. In Language resources and evaluation, 43, 3, 2009, pp. 209-231.
- S. Montemagni, *Tecnologie linguistico-computazionali per il monitoraggio della lingua italiana*. "Studi Italiani di Linguistica Teorica e Applicata" (SILTA), anno XLII, numero 1, 2013, pp. 145-172.
- S. Richter, A. Cimino, F. Dell'Orletta, G. Venturi, *Tracking the Evolution of Written Language Competence: an NLP–based Approach*. In "Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it), 3-4 December 2015, Trento", Italia, 2015, pp. 31-35.