SLI
SOCIETÀ DI LINGUISTICA ITALIANA

LINGUE IN CONTATTO / CONTACT LINGUISTICS

ATTI DEL XLVIII CONGRESSO INTERNAZIONALE DI STUDI
DELLA SOCIETÀ DI LINGUISTICA ITALIANA (SLI)

Udine, 25-27 settembre 2014

a cura di
RAFFAELLA BOMBI E VINCENZO ORIOLES

BULZONI EDITORE ROMA 2016

Volume pubblicato con il contributo
del Dipartimento di Studi Umanistici dell'Università degli Studi di Udine

ISBN 978-88-6897-029-1

© 2016 by Bulzoni Editore S.r.l.
00185 Roma, via dei Liburni, 14
http://www.bulzoni.it
e-mail: bulzoni@bulzoni.it

INDICE

FELICE DELL'ORLETTA, SIMONETTA MONTEMAGNI, GIULIA VENTURI
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR)

Esplorazioni computazionali nello spazio dell'interlingua: verso una nuova metodologia di indagine

1. INTRODUZIONE

L'innovativa prospettiva di ricerca in materia di apprendimento di una lingua seconda aperta da Larry Selinker ha portato alla definizione di interlingua come «a separate linguistic system based on the observable output which results from a learner's attempted production of a TL norm» (Selinker, 1972: 214). All'interno di questa prospettiva, due sono gli aspetti fondamentali: l'identificazione dei dati rilevanti per lo studio del processo di apprendimento e la formulazione di una «psycholinguistic theory of second-language learning» (Selinker, 1972). Gli studi si sono pertanto focalizzati da un lato sulla definizione di metodi di osservazione delle produzioni degli apprendenti, dall'altro sulle procedure di analisi dei dati diagnostici raccolti, sulle modalità cioè di individuazione e descrizione delle caratteristiche dell'interlingua. Nel primo caso l'obiettivo è quello di rilevare l'interlingua e i suoi stadi di evoluzione, come ad esempio le modalità *task-based* adottate in attività di elicitazione dell'interlingua a partire da produzioni orali. Nel secondo caso gli studi mirano a definire una teoria del processo di apprendimento attraverso la formulazione di ipotesi sulle regole di funzionamento della lingua utilizzata dall'apprendente la lingua seconda. A prescindere dall'obiettivo di indagine e dall'approccio adottato, nel campo degli studi acquisizionali sull'interlingua è ben noto il fatto che in questo ambito «il lavoro da fare è ancora enorme, dal momento che non tutti i fenomeni sono stati studiati e anche per i fenomeni studiati si pone il problema della generalizzabilità dei risultati fin qui raggiunti» (Lo Duca, 2004).

Il presente contributo intende porsi in questo contesto e proporre un innovativo approccio alla identificazione delle caratteristiche linguistiche che aiutano a definire l'interlingua. Tale approccio consiste nella ricostruzione del profilo linguistico di corpora di produzioni scritte da apprendenti una lingua seconda basato su strumenti di trattamento automatico del linguaggio. L'intento è qui quello di suggerire come questa strategia di analisi possa offrire una soluzione metodologica al problema della reperibilità su ampia scala e della gene-

ralizzabilità di tratti utili a caratterizzare l'interlingua. Partendo dal presupposto che lo studio dell'interlingua consiste in un processo di ricostruzione del sistema di regolarità osservabili nelle produzioni di apprendenti, l'obiettivo è quello di suggerire come questo processo induttivo possa essere approssimato dalla capacità degli strumenti linguistico-computazionali di indurre un modello linguistico da collezioni di testi linguisticamente annotati basandosi sulla distribuzione di un'ampia gamma di caratteristiche linguistiche.

Tale approccio linguistico-computazionale si propone non solo come un metodo diagnostico finalizzato alla ricostruzione delle caratteristiche dell'interlingua ma anche come un metodo euristico finalizzato all'analisi di caratteristiche dello spazio dell'interlingua poco esplorate perché difficilmente attingibili attraverso un'analisi manuale delle produzioni.

2. LO SPAZIO DELL'INTERLINGUA E IL COMPITO DI NATIVE LANGUAGE IDENTIFICATION

L'ipotesi di partenza è che sia possibile condurre uno studio di alcuni dei tratti caratterizzanti l'interlingua a partire dall'analisi delle peculiarità linguistiche di corpora di produzioni scritte di apprendenti L2 che condividono una L1. Questa ipotesi prende le mosse da un lato dagli esiti dei numerosi studi in materia di interlingua, dall'altro dai risultati raggiunti dagli studi più recenti in materia di identificazione automatica della lingua madre (*Native Language Identification*) condotti in ambito linguistico-computazionale.

Ad oggi, gli studi in materia di interlingua sono concordi nell'affermare che, a prescindere dalla teoria linguistica assunta, «è impossibile negare che la prima lingua abbia un ruolo, anche abbastanza importante, nel determinare le produzioni interlinguistiche degli apprendenti» (Pallotti, 1998: 59). Si tratta del fenomeno del *transfer* che tuttavia non è il solo ma «è *uno* dei processi che contribuiscono a dare forma all'interlingua» (Pallotti, 1998: 60). Ne sono concreta espressione i «fossilizable linguistic phenomena» di cui parla Selinker, cioè «linguistic items, rules, and subsystems which speakers of a particular NL will tend to keep in their IL relative to a particular TL» (Selinker, 1972: 215). Secondo Selinker, tali fenomeni linguistici sono da tenere distinti dai veri e propri *errori* dal momento che essi riemergono nella produzione di interlingua soprattutto nei momenti di maggiore ansia o impegno intellettivo dell'apprendente. Si tratta di strutture fossilizzate nella produzione linguistica di un parlante L2, la cui continua (ri)emergenza è spia del fatto che questo fenomeno di «backsliding» di un apprendente L2 «from a TL norm is not […] either random or towards the speaker's

NL, but towards an IL norm» (Selinker, 1972: 215-216). Sono dunque spia dell'esistenza di quel sistema di regole autonomo che è l'interlingua.

Seppure in un contesto di ricerca diverso, gli studi più recenti in materia di identificazione automatica della lingua madre hanno dedicato una medesima attenzione a quei tratti linguistici della L1 che si riflettono nelle produzioni in L2 più che agli errori. Il compito, noto come *Native Language Identification* (da qui in avanti riferito come NLI), si pone come obiettivo la classificazione automatica di produzioni linguistiche di apprendenti L2 sulla base della L1 (Koppel *et al.*, 2005)[1]. Pur differenziandosi rispetto agli algoritmi statistici usati e alle caratteristiche linguistiche selezionate, gli attuali modelli di NLI partono dall'assunto che la distribuzione statistica di un insieme selezionato di tratti linguistici rintracciati all'interno di produzioni scritte di apprendenti L2 giochi un ruolo chiave nell'identificare la L1 sottostante. Se i primi i sistemi di NLI si basavano per lo più su idiosincrasie stilistiche (errori ortografici, lessicali, sintattici, ecc.) del testo rispetto alla norma della L2, oggi il processo di identificazione automatica della L1 è realizzato a partire dall'individuazione nel testo di un'ampia tipologia di tratti che spazia tra diversi livelli di descrizione linguistica (come ad esempio la distribuzione di parole funzionali, sequenze di categorie morfo-sintattiche, ecc.). Un tale approccio permette di indurre dalle produzioni L2 un modello linguistico della collezione di testi in esame definito a partire da un insieme di caratteristiche linguistiche la cui distribuzione viene rintracciata all'interno della collezione testuale. In fase di analisi, il modello linguistico viene usato dai sistemi di NLI per classificare un testo rispetto alla L1 dello scrivente impiegando unicamente le informazioni linguistiche delle produzioni L2. È questo il motivo per cui possiamo vedere questo tipo di approcci come fortemente correlato alla nozione di interlingua.

Ciò che accomuna gli studi sull'interlingua e le ricerche finalizzate allo sviluppo di sistemi di NLI è da un lato l'attenzione allo studio delle regolarità presenti nelle produzioni di apprendenti L2 e dall'altro alcuni interrogativi di ricerca, quali: come individuare e rintracciare queste regolarità? E, quali sono le regolarità più significative? Con l'obiettivo di trovare una risposta a queste domande, sono stati nel tempo costruiti corpora di produzioni di apprendenti L2. Per quanto riguarda la lingua italiana, Andorno & Rastelli (2009) forniscono una

[1] «Identifying an author's native language is a type of authorship attribution problem. Instead of identifying a particular author from among a closed list of suspects, we wish to identify an author class, namely, those authors who share a particular native language.» (Koppel et al., 2005).

rassegna dei corpora ad oggi disponibili che raccolgono produzioni scritte e orali di italiano L2 e che sono usati per condurre studi sull'interlingua. Per quanto riguarda l'ambito di ricerca in materia di sviluppo di sistemi di NLI, sebbene in principio tali sistemi potrebbero essere sviluppati per l'identificazione della lingua madre a partire da qualsiasi L2, sino ad oggi le ricerche in questo ambito sono state condotte esclusivamente a partire da corpora di inglese L2. A oggi due sono i principali corpora usati: il corpus ICLE (International Corpus of Learner English) (Granger *et al.*, 2009) e il più recente corpus TOEFL11 (Test of English as a Foreign Language) (Blanchard *et al.*, 2013) costruito nel 2013 e messo a disposizione dei partecipanti della prima competizione internazionale sul tema *Native Language Identification*[2] (Tetreault *et al.*, 2013).

Collocandosi in questo panorama di studi, il presente contributo intende presentare una innovativa metodologia di indagine finalizzata allo studio dello spazio di interlingua a partire dall'analisi di corpora di apprendenti L2 linguisticamente annotati con strumenti di Trattamento Automatico del Linguaggio. Nei paragrafi che seguono descriveremo gli strumenti utilizzati, i corpora a cui abbiamo fatto riferimento e le caratteristiche linguistiche osservate (vedi § 3). Nel paragrafo § 4 mostreremo come la ricostruzione dei principali tratti linguistici osservati in relazione a ognuna delle L1 esaminate abbiano permesso di condurre un'analisi dello spazio dell'interlingua rintracciato in produzioni L2 rispetto sia ai singoli tratti linguistici considerati sia rispetto alle due L2 qui considerate, inglese e italiano.

A conclusione di questa parte introduttiva, riteniamo sia necessaria una precisazione terminologica. In quanto segue il termine *interlingua* verrà utilizzato per riferirsi ad una configurazione di tratti linguistici rintracciati in produzioni scritte in L2 a partire dalla quale il linguista potrà ricostruire il«sistema governato da regole ben precise» caratterizzato da «un notevole grado di variabilità» legato allo «sviluppo interlinguistico», cioè alle varie fasi di apprendimento dell'interlingua (Pallotti, 1998: 21-22). Ne consegue che all'interno di questo studio l'utilizzo del termine riflette la nostra attenzione alla ricostruzione del profilo linguistico comune agli elaborati scritti di apprendenti che condividono una L1 piuttosto che all'identificazione del grado di sviluppo della loro competenza linguistica in L2.

3. LA METODOLOGIA DI INDAGINE

Sulla base dell'accezione di *interlingua* prima delineata, intendiamo qui introdurre una metodologia di analisi che a partire dall'individuazione di un'ampia

[2] https://sites.google.com/site/nlisharedtask2013/home.

gamma di tratti linguistici in corpora di apprendenti L2 consente di condurre un'indagine su ampia scala delle caratteristiche linguistiche influenzate dalla L1 degli apprendenti. In quanto segue, descriveremo gli elementi fondamentali della metodologia.

3.1 *I corpora*

L'approccio messo a punto è stato testato su due corpora di apprendenti due L2 diverse, italiano e inglese. Come corpus di inglese L2 è stato analizzato il corpus TOEFL11 (Blanchard et al., 2013) messo a disposizione dei partecipanti dello *Shared task on Native Language Identification*. Si tratta di una collezione di prove scritte dell'esame TOEFL composta da testi scritti da parlanti nativi 11 diverse L2: arabo, cinese, francese, tedesco, hindi, italiano, giapponese, coreano, spagnolo, telugu, turco. Come riportato nella colonna "TOEFL11 (inglese)" della tabella 1, il corpus contiene 1.100 documenti per ogni L1, uniformemente distribuiti rispetto al livello di competenza linguistica (alta, media, bassa) e alle tracce assegnate[3]. Le lingue sono state scelte dagli organizzatori con l'esplicito intento di includere lingue esemplificative di diverse famiglie linguistiche.

Oltre a questo corpus, abbiamo preso in considerazione una porzione del corpus VALICO (Varietà Apprendimento Lingua Italiana Corpus Online)[4] (Barbera & Marello, 2004) di produzioni in italiano L2. In particolare, abbiamo analizzato i testi disponibili per le stesse L1 presenti nel corpus TOEFL11: cinese, francese, tedesco, giapponese, spagnolo (vedi tabella 1). L'obiettivo era quello di indagare la presenza di strutture linguistiche simili all'interno di produzioni scritte di apprendenti L2 diverse che condividono la stessa L1.

L1	L2	
	TOEFL11 (Inglese)	**VALICO (Italiano)**
Arabo	1.100	--
Cinese	1.100	133
Francese	1.100	346
Tedesco	1.100	341

[3] Le prove sono state svolte a partire da 8 diverse tracce. Riportiamo qui un esempio di traccia: *Do you agree or disagree with the following statement? The best way to travel is in a group led by a tour guide. Use reasons and examples to support your answer*.

[4] http://www.valico.org/

L1	L2	
	TOEFL11 (Inglese)	**VALICO (Italiano)**
Hindi	1.100	--
Italiano	1.100	--
Giapponese	1.100	231
Coreano	1.100	--
Spagnolo	1.100	314
Telugu	1.100	--
Turco	1.100	--

Tabella 1:
Distribuzione dei documenti nei due corpora analizzati rispetto alla L1 e L2.

3.2 *Gli strumenti di Trattamento Automatico del Linguaggio*

L'annotazione linguistica automatica del testo ha rappresentato il prerequisito indispensabile per la ricostruzione del suo profilo linguistico. Pertanto, tutti i testi presi in esame sono stati analizzati con strumenti di annotazione linguistica automatica in grado di rendere progressivamente esplicita l'informazione linguistica contenuta in un testo. Per ogni livello di descrizione linguistica uno specifico componente di analisi identifica in modo automatico la struttura del testo, utilizzando come input il risultato prodotto dal componente precedente. L'identificazione della struttura linguistica del testo, o annotazione, avviene tipicamente in modo incrementale, attraverso analisi linguistiche a livelli di complessità crescente: "tokenizzazione", ovvero segmentazione del testo in parole ortografiche (o tokens); analisi morfo-sintattica e lemmatizzazione del testo tokenizzato; analisi della struttura sintattica della frase in termini di relazioni di dipendenza.

La tabella 2 mostra un esempio del risultato dell'annotazione linguistica del seguente periodo:

L'altro giorno due uomini camminavano sul marciapiede.

		Lemmatizzazione	Annotazione morfo-sintattica			Annotazione sintattica	
Id	Forma	Lemma	CPoS	FPoS	Tratti morfologici	Testa sintattica	Relazione
1	L'	il	R	RD	num=s|gen=n	3	det
2	altro	altro	D	DI	num=s|gen=m	3	mod

		Lemma-tizzazione	Annotazione morfo-sintattica			Annotazione sintattica	
Id	Forma	Lemma	CPoS	FPoS	Tratti morfologici	Testa sintattica	Relazione
3	giorno	giorno	S	S	num=s|gen=m	6	mod_temp
4	due	due	N	N	_	5	mod
5	uomini	uomo	S	S	num=p|gen=m	6	subj
6	cam-minav ano	camminare	V	V	num=p|per=3| mod=i|ten=i	0	ROOT
7	sul	su	E	EA	num=s|gen=m	6	comp_loc
8	marci-apiede	marci-apiede	S	S	num=s|gen=m	7	prep
9	.	.	F	FS	_	6	punc

Tabella 2: Un esempio di annotazione linguistica.

Innanzitutto, il periodo è stato individuato grazie alla fase di segmentazione in periodi. Durante la successiva fase di tokenizzazione, all'interno del periodo sono stati riconosciuti i tokens corrispondenti alle singole forme (colonna *Forma*), identificate univocamente da un numero progressivo (colonna *Id*). La fase di disambiguazione morfo-sintattica ha permesso di associare ad ogni token individuato *i)* la corretta categoria morfo-sintattica (colonna *CPoS* e *FPoS*)[5] associata al token nel contesto specifico, *ii)* i relativi tratti morfologici (colonna *Tratti morfologici*) e *iii)* il lemma corrispondente (colonnna *Lemma*). Ad esempio, la forma 'uomini' (Id=5) viene ricondotta al lemma 'uomo', viene annotata con la categoria sostantivo (S), arricchita con informazione relativa al numero (plurale) e al genere (maschile).

Il risultato dell'annotazione sintattica riportato nelle colonne *Testa sintattica* e *Relazione* della tabella 2 permette inoltre di stabilire che, ad esempio, il

[5] Per ogni token viene riconosciuta la categoria morfo-sintattica generale (CPoS) e eventuali sottocategorie (FPoS). Ad esempio, alla forma (token) 'sul' viene associata la categoria preposizione (E) e viene ulteriormente specificato che si tratta di una preposizione articolata (EA). Allo stesso modo, il token '. 'viene annotato come un segno di punteggiatura (F) di fine periodo (FS). La lista completa delle categorie morfo-sintattiche è consultabile alla pagina http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf; la descrizione dei tipi di dipendenza sintattica è consultabile alla pagina http://www.italianlp.it/docs/ISST-TANL-DEPtagset.pdf.

sostantivo 'uomini' è il soggetto (subj) del verbo 'camminavano', il quale costituisce la testa sintattica della relazione. Questa informazione è riportata nella colonna *Testa* dove è infatti segnalato che la testa sintattica del dipendente 'uomini' ha Id=6, l'Id cioè del token 'camminavano'. In questo caso 'camminavano' ha testa sintattica 0 dal momento che rappresenta il verbo della frase principale, radice (root) dell'albero sintattico dell'intero periodo. La fase di annotazione sintattica a dipendenze permette dunque di fornire una descrizione esplicita dell'intero albero sintattico del periodo analizzato, sotto forma di relazioni di dipendenza che legano i tokens che lo compongono. L'informazione può inoltre essere graficamente visualizzata, come mostra la figura 1 che riporta la struttura sintattica della frase annotata, rappresentata come una serie di nodi lessicali (i singoli tokens), messi in collegamento da archi di dipendenza a loro volta etichettati con il nome del tipo di relazione di dipendenza (gli archi e le etichette graficamente rappresentati).

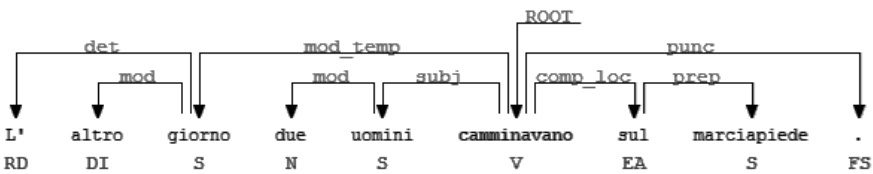

Figura 1:
Un esempio di rappresentazione grafica dell'annotazione sintattica a dipendenze.

In questo studio l'annotazione linguistica è stata realizzata da LinguA (Linguistic Annotation pipeline)[6], una catena di strumenti di Trattamento Automatico del Linguaggio sviluppati in modo congiunto dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa e dall'Università di Pisa. Si tratta di strumenti di analisi linguistica automatica multilingue che combinano algoritmi basati su regole (*rule-based*) con metodi di apprendimento automatico (*machine learning*) e le cui prestazioni rappresentano lo stato dell'arte per la lingua italiana, come testimoniato dai risultati raggiunti nell'am-

[6] http://www.italianlp.it/demo/linguistic-annotation-tool.

bito delle periodiche campagne di valutazione di componenti per il trattamento automatico dell'italiano EVALITA[7].

3.3 *Le caratteristiche linguistiche*

La ricostruzione del profilo linguistico di corpora di apprendenti L2 che condividono la L1 prende le mosse dagli studi condotti da Douglas Biber in materia di *register variation* e finalizzati a dimostrare come uno studio esaustivo delle specificità linguistiche proprie di diverse varietà della lingua sia possibile a partire dall'analisi comparativa della diversa distribuzione d'uso di tratti lessicali e grammaticali rilevanti rintracciati in corpora testuali rappresentativi di diverse varietà (Biber, 1993). Intendiamo qui illustrare come un simile approccio comparativo, possa essere utilizzato anche per ricostruire le caratteristiche specifiche di corpora di L2 mettendo a confronto produzioni di apprendenti con L1 diverse.

I tratti presi in considerazione spaziano tra i vari livelli di descrizione linguistica (lessicale, morfo-sintattico e sintattico) e sono rintracciati in modo automatico a partire dall'output del processo di annotazione automatica del testo. La tabella 3 riporta alcuni dei tratti monitorati.

Tipo di caratteristica	Livello di annotazione linguistica	Caratteristica
Di base	Divisione in frasi e tokenizzazione	Lunghezza media dei periodi (calcolata in tokens) e delle parole (calcolata in caratteri)
Lessicale	Lemmatizzazione e annotazione morfo-sintattica	Sequenze di caratteri e di parole (n-grammi)
		Indice di ricchezza lessicale (type/token ratio)
Morfo-sintattico	Annotazione morfo-sintattica	Distribuzione delle categorie morfo-sintattiche
		Densità lessicale (parole piene/funzionali)
Sintattico	Annotazione sintattica a dipendenze	Distribuzione dei vari tipi di relazioni di dipendenza
		"Valenza" verbale, calcolata come il numero delle dipendenze effettive di una testa verbale
		Caratteristiche relative alla struttura dell'albero sintattico analizzato: - profondità media dell'albero sintattico

[7] http://www.evalita.it/

Tipo di caratteristica	Livello di annotazione linguistica	Caratteristica
		(numero medio di relazioni di dipendenza consecutive all'interno di una frase), - lunghezza media della più lunga relazione di dipendenza, calcolata come la distanza tra la testa e il dipendente (in tokens)
		Caratteristiche relative all'uso della subordinazione: - distribuzione di frasi principali vs. subordinate, - lunghezza media di sequenze consecutive di subordinate
		Caratteristiche relative alla modificazione nominale: - lunghezza media di catene di dipendenza sintattica a testa nominale

Tabella 3:
Alcune delle caratteristiche considerate in fase di monitoraggio linguistico.

L'uso di tali tratti è già stato sperimentato con successo su diverse tipologie di testi: per monitorare la lingua italiana nelle sue varietà diamesiche, diastratiche, diafasiche (Montemagni, 2013), per individuare similarità e differenze tra le produzioni scritte scolastiche di apprendenti l'italiano L1 e L2 e i materiali didattici offerti nella scuola primaria e secondaria (Dell'Orletta *et al.*, 2011a), per monitorare l'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado (Barbagli *et al.*, 2014). Inoltre questi tratti sono stati utilizzati in compiti applicativi per valutare in modo automatico il livello di leggibilità di un testo (Dell'Orletta *et al.*, 2011b), per classificare il genere testuale di un documento (Dell'Orletta *et al.*, 2014), e come menzionato in precedenza per riconoscere in modo automatico la lingua madre a partire da produzioni L2 (Cimino *et al.*, 2013).

4. ESPLORAZIONI NELLO SPAZIO DELL'INTERLINGUA

La metodologia di indagine messa a punto ha permesso di condurre due principali tipologie di analisi finalizzate a ricostruire i principali tratti di interlingua per ognuna delle L1 esaminate, rispetto ai diversi livelli di descrizione

linguistica considerati, e a condurre un'analisi comparativa di produzione L2 inglese e italiano di parlanti la stessa L1.

4.1 *Ricostruzione dei tratti dello spazio di interlingua*

Come introdotto in § 2, in questo lavoro con il termine *interlingua* ci riferiamo all'insieme di tratti linguistici rintracciati in produzioni scritte L2 che condividono una L1. Sulla base di questa precisazione terminologica, il processo di ricostruzione dei tratti dello spazio di interlingua consiste nella ricostruzione del profilo linguistico di una collezione di elaborati di apprendenti L2 che condividono una L1.

La tabella 4 riporta ad esempio la distribuzione di alcune delle caratteristiche linguistiche rintracciate nelle produzioni in inglese L2 di parlanti L1 italiani e cinesi presenti nel corpus TOEFL11.

	Caratteristiche linguistiche	Italiano L1	Cinese L1
Generali	Lunghezza media delle frasi (in token)	12,61	10,02
	Lunghezza media delle parole (in caratteri)	2,20	2,23
Morfo-sintattiche	Distribuzione percentuale di:		
	- determinanti	4,87	4,35
	- preposizioni	5,86	5,27
	- aggettivi	4,10	4,09
	- congiunzioni	1,69	1,47
	- avverbi	3,05	3,28
	- sostantivi	9,94	10,11
	- verbi	8,26	8,33
Sintattiche	Profondità media dell'albero sintattico (numero medio di relazioni di dipendenza consecutive all'interno di una frase)	3,49	2,87
	Lunghezza media delle relazioni (massime) di dipendenza sintattica, calcolata come la distanza tra la testa e il dipendente (in tokens)	4,87	3,90
	Lunghezza media di catene di dipendenza sintattica a testa nominale	0,58	0,56
	"Valenza" media verbale, calcolata come il numero delle dipendenze effettive di una testa verbale	0,85	0,83

Tabella 4: Distribuzione di alcuni tratti nei sotto-corpora di L1 italiano e cinesi del corpus TOEFL11 di produzioni inglese L2.

Come si può notare, le produzioni di parlanti cinesi e italiani si differenziano già per quanto riguarda la lunghezza della frase: i cinesi scrivono frasi con in media 2 tokens in meno rispetto agli italiani. Per quanto riguarda la distribuzione delle categorie morfosintattiche, le prove degli italiani contengono una percentuale maggiore di determinanti, preposizioni e congiunzioni, mentre quelle dei cinesi si distinguono per una percentuale maggiore di sostantivi. Inoltre, gli alberi sintattici prodotti dagli italiani sono in media più profondi e si caratterizzano per relazioni massime di dipendenza sintattica mediamente più lunghe rispetto a quanto succede nelle produzioni dei cinesi. Queste differenze ci permettono di iniziare a formulare delle prime ipotesi su quali sono le caratteristiche linguistiche distintive le produzioni in L2 di parlanti due L1 diverse.

Il modo in cui queste caratteristiche interagiscono tra di loro e come la loro diversa distribuzione permetta di caratterizzare lo spazio di una interlingua sono quesiti ai quali le potenzialità degli strumenti di trattamento automatico del linguaggio permettono di dare una risposta. Grazie a questo approccio induttivo, gli strumenti sono infatti in grado di ricostruire un modello linguistico per ogni gruppo di produzioni L2 che condividono la L1. Come descritto in dettaglio da Cimino e colleghi (Cimino *et al.*, 2013), tale approccio è stato sperimentato con successo in occasione dello *Shared task on Native Language Identification* dove il riconoscimento automatico della lingua madre è stato declinato come un compito di classificazione automatica. La classificazione è realizzata da un algoritmo in grado di predire con quale probabilità un testo L2 è stato scritto da un parlante una determinata L1. A questo scopo sono state usate le produzioni scritte in inglese L2 del corpus TOEFL11 e l'algoritmo di classificazione automatica è stato addestrato sulle caratteristiche linguistiche di 900 prove per ognuna delle 11 L1. In fase di addestramento (*training*) l'algoritmo di classificazione apprende la strategia di classificazione imparando da un cosiddetto corpus *di addestramento*, una collezione di testi nella quale ad ogni testo è stata associata la L1 corrispondente. Durante questa fase di addestramento (*training*) l'algoritmo ha indotto un modello linguistico per ognuna delle L1 basandosi sulla distribuzione statistica delle caratteristiche linguistiche. Ad ognuna di queste caratteristiche è stato associato un peso, che rappresenta il contributo della specifica caratteristica nel predire la L1 di un testo. Terminata la fase di addestramento, in fase di analisi l'algoritmo è stato in grado di riconoscere la lingua madre di un nuovo testo di cui non conosceva la L1 con una precisione generale del 77,9%. L'accuratezza nel riconoscere ognuna delle 11 L1 del corpus TOEFL11 è riportata nella tabella 5.

Arabo	**Cinese**	**Francese**	**Tedesco**	**Hindi**	**Italiano**	**Giapponese**	**Coreano**	**Spagnolo**	**Telugu**	**Turco**
73,8	77,5	83,2	**87,3**	71,1	86,0	78,8	74,2	70,8	76,2	78,0

Tabella 5:
Risultati della partecipazione allo *Shared task on Native Language Identification*.

Come si può notare, i migliori risultati sono stati ottenuti nel riconoscimento del tedesco e dell'italiano mentre i peggiori per lo spagnolo e l'hindi. A prescindere dai risultati raggiunti, l'obiettivo è qui quello di dimostrare come la metodologia messa a punto sia affidabile per condurre delle esplorazioni su ampia scala finalizzate a ricostruire i tratti caratteristici di interlingua che siano generalizzabili per un gruppo di apprendenti L2. In quanto segue, le caratteristiche linguistiche rintracciate nelle produzioni L2 dei diversi gruppi di parlanti saranno descritte in dettaglio, mettendo in luce le similarità e differenze della loro distribuzione rispetto alle diverse L1 considerate.

4.2 *Dalla singola interlingua a similarità tra interlingue*

L'obiettivo di questo paragrafo è quello di mettere a confronto produzioni L2 di gruppi di parlanti che condividono la L1 rispetto ad alcune delle caratteristiche linguistiche considerate. Saranno inoltre messe in luce similarità e differenze tra produzioni in due L2 diverse, italiano e inglese, scritte da parlanti la stessa L1.

Come si può vedere nei grafici che seguono, a tutti i livelli di analisi emergono similarità interessanti. A partire da una caratteristica di base di analisi del testo come la lunghezza media della frase, espressa in termini di tokens, si può notare nella figura 2 che i testi inglesi L2 prodotti da parlanti L1 arabo, italiano e spagnolo contengono frasi più lunghe (12, 13, 13 tokens rispettivamente) rispetto alle frasi di parlanti L1 giapponese, coreano e turco la cui lunghezza media è di circa 9 tokens. La stessa tendenza si può riscontrare nelle produzioni di VALICO (figura 3). Anche in questo caso le produzioni di madrelingua spagnola contengono le frasi più lunghe (22 tokens), mentre le prove in italiano L2 di giapponesi L1 hanno le frasi più brevi (14 tokens).

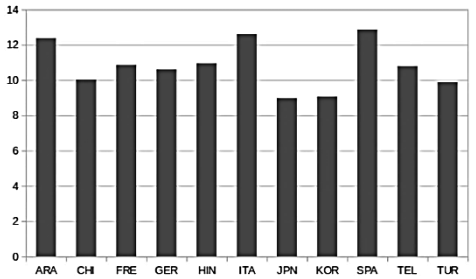

Figura 2: Lunghezza media delle frasi (in tokens) nel corpus TOEFL11.

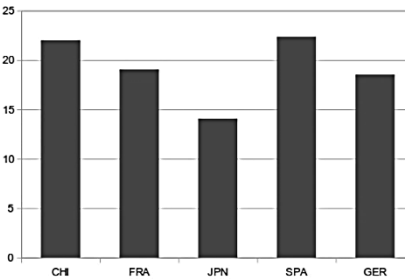

Figura 3: Lunghezza media delle frasi (in tokens) nel corpus VALICO.

Anche rispetto al livello morfo-sintattico di analisi emergono risultati interessanti. L'analisi comparativa della distribuzione percentuale di determinanti, sostantivi e verbi, tre delle categorie morfosintattiche considerate più significative nella letteratura in materia di variazione linguistica, mostra che le produzioni di giapponesi e coreani in inglese L2 sono quelle che contengono meno determinanti (figura 4). Si può notare la stessa tendenza nelle produzioni di giapponesi in italiano L2 (figura 5), sebbene non ci siano nette differenze tra le L1. Se nel corpus TOEFL11 le produzioni di parlanti hindi e telugu sono quelle che contengono la percentuale più alta di sostantivi (circa l'11%), nel corpus VALICO possiamo osservare che francesi e spagnoli hanno la più alta percentuale di sostantivi (24% e 23% rispettivamente) e i giapponesi la più alta percentuale di verbi (circa il 23%). Di conseguenza, le produzioni italiano L2 dei francesi si distinguono per il rapporto sostantivi/verbi più alto mentre quelle dei giapponesi per il rapporto più basso.

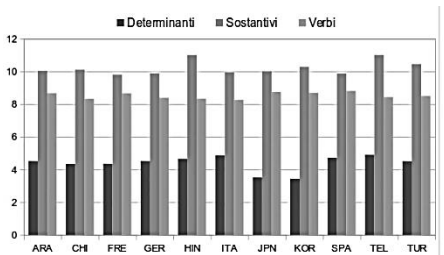

Figura 4: Distribuzione di determinanti, sostantivi e verbi nel corpus TOEFL11.

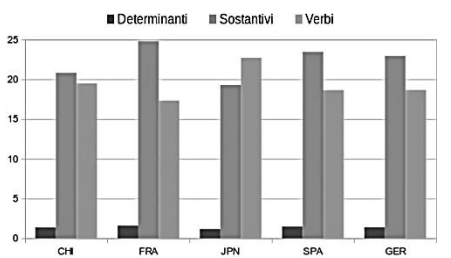

Figura 5: Distribuzione di determinanti, sostantivi e verbi nel corpus VALICO.

Per quanto riguarda l'analisi del profilo sintattico, alcuni dei tratti osservati riflettono in parte quanto avevamo osservato a proposito della lunghezza della frase. All'interno del corpus TOEFL11, le frasi mediamente più lunghe sono quelle di arabi, italiani e spagnoli le cui produzioni si caratterizzano anche per gli alberi sintattici più profondi, con una media di 3,55 relazioni di dipendenza consecutive per frase, 3,49 e 3,57 (figura 6); mentre le produzioni di giapponesi e coreani che hanno le frasi più brevi dell'intero corpus contengono anche gli alberi sintattici meno profondi, con una media di 2,71 e 2,77. Si può osservare un comportamento simile anche nel corpus VALICO (figura 7), dove le produzioni in italiano L2 di giapponesi si contraddistinguono per le frasi più brevi e per gli alberi sintattici meno profondi, con 4,15 relazioni per frase. Sempre in VALICO, le produzioni di spagnoli contengono le frasi con gli alberi sintattici più profondi (una media di circa 6 relazioni consecutive per frase).

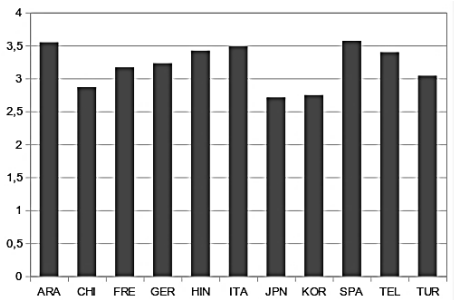

Figura 6: Profondità media degli alberi sintattici nel corpus TOEFL11.

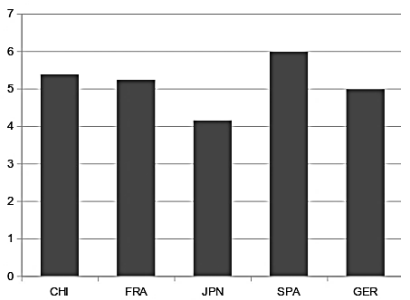

Figura 7: Profondità media degli alberi sintattici nel corpus VALICO.

In linea con questi tratti sintattici, le produzioni di giapponesi sia in inglese sia in italiano L2 contengono frasi con le più brevi relazioni di dipendenza sintattica (con una lunghezza media di 3,45 tokens in TOEFL11 e 5,85 in VALICO; figure 8 e 9) e con le più brevi catene di dipendenza sintattica a testa nominale (con una lunghezza media inferiore a 1 dipendenza in TOEFL11 e uguale a 1 dipendenza in VALICO; figure 10 e 11). I valori più alti di questi due tratti si hanno invece nelle frasi delle produzioni di arabi, spagnoli e italiani nel corpus TOEFL11. Interessante osservare come in questo corpus le produzioni di parlanti hindi L1 si caratterizzino per le catene di modificazione nominale più lunghe, possibile conseguenza dell'elevata percentuale di sostantivi presenti.

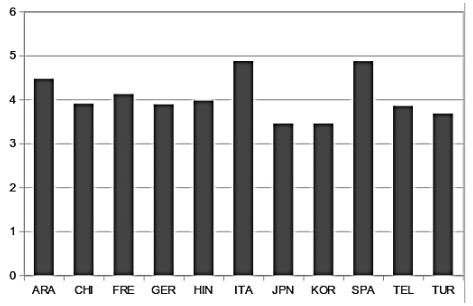

Figura 8: Lunghezza media delle relazioni di dipendenza sintattica nel corpus TOEFL11.

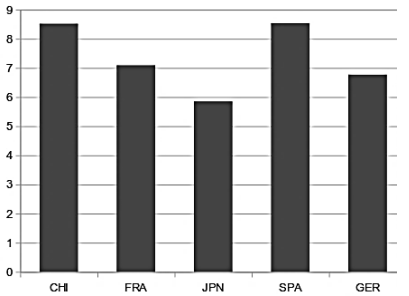

Figura 9: Lunghezza media delle relazioni di dipendenza sintattica nel corpus VALICO.

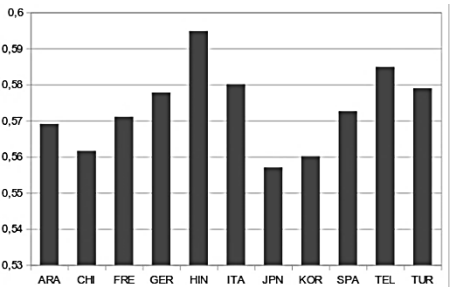

Figura 10: Lunghezza media delle catene di dipendenza sintattica a testa nominale nel corpus TOEFL11.

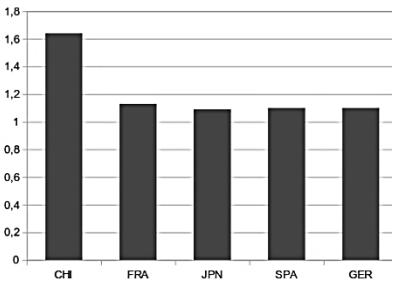

Figura 11: Lunghezza media delle catene di dipendenza sintattica a testa nominale nel corpus VALICO.

Interessante osservare come questi confronti ci permettano di trovare delle similarità tra L1 che appartengono alle stesse famiglie linguistiche: italiano e spagnolo come lingue romanze da un lato, giapponese, coreano e turco come lingue altaiche dall'altro. Da questa prima descrizione comparativa tra produzioni in L2 diverse emerge inoltre chiaramente che, come ci si poteva aspettare, la L2 influenza i valori delle caratteristiche linguistiche monitorate. È il caso, ad esempio, della lunghezza media della frase: la lunghezza delle frasi più lunghe nel corpus di inglese L2, contenute nelle produzioni di arabi, italiani e spagnoli, è pari (se non di poco inferiore) alla lunghezza delle frasi più brevi in italiano L2, quelle cioè prodotte dai giapponesi. Lo stesso è il caso di due tratti sintattici

strettamente collegati: l'altezza media dell'albero sintattico della frase e la lunghezza media delle relazioni di dipendenza sintattica. Rispetto a questi tratti, sebbene ad esempio i giapponesi mantengano la loro peculiare maggiore brevità rispetto ad altre L1, tuttavia quando scrivono in italiano costruiscono frasi con alberi sintattici più alti e con relazioni di dipendenza più lunghe.

5. CONCLUSIONI E SVILUPPI FUTURI

In questo contributo abbiamo illustrato una innovativa metodologia di indagine dello spazio dell'interlingua condotta con metodi e tecniche della linguistica computazionale. La metodologia elaborata ci ha permesso di condurre un'indagine delle caratteristiche linguistiche specifiche di corpora di produzioni scritte di apprendenti L2. Sono stati inoltre messi a confronto i risultati delle indagini nello spazio dell'interlingua di apprendenti due diverse L2, italiano e inglese. Questo ha permesso di rintracciare una serie di similarità e differenze di apprendenti diverse L2 che condividono la L1.

La metodologia messa a punto ci ha permesso di suggerire una soluzione al problema della reperibilità su ampia scala dei tratti caratterizzanti lo spazio dell'interlingua e della loro generalizzabilità per un gruppo di apprendenti, problema ben noto nell'ambito degli studi in materia di interlingua (cfr. Lo Duca, 2004). La possibilità di generalizzare le caratteristiche rintracciate nei corpora L2 è legata alle ricadute applicative che abbiamo messo in luce. Grazie alle potenzialità degli strumenti di trattamento automatico del linguaggio che abbiamo impiegato, la metodologia di indagine si è infatti rivelata affidabile come punto di partenza per riconoscere in modo automatico la lingua madre a partire dalle caratteristiche linguistiche proprie degli elaborati di un gruppo di apprendenti L2 che condividono la L1.

Numerose sono le prospettive di ricerca che potrebbero prendere le mosse dai risultati raggiunti in questo studio, sebbene essi siano ancora preliminari. Innanzitutto, la metodologia di indagine dello spazio di interlingua potrebbe essere ulteriormente specializzata per prendere in considerazione tratti linguistici non ancora monitorati, la cui selezione dovrebbe essere guidata da studiosi esperti nello studio dell'interlingua. Inoltre, come già accennato, tale metodologia potrebbe rappresentare un ausilio per gli studi di linguistica acquisizionale come supporto metodologico al linguista impegnato nella raccolta dei fenomeni caratterizzanti l'interlingua e nella loro generalizzazione. L'approccio messo a punto si propone infatti non solo come un metodo "diagnostico" finalizzato alla ricostruzione delle caratteristiche dell'interlingua ma anche come un metodo

euristico finalizzato all'analisi di caratteristiche poco esplorate perché difficilmente attingibili attraverso un'analisi manuale delle produzioni.

Infine, la metodologia di monitoraggio delle caratteristiche specifiche di un gruppo di apprendenti L2 potrebbe essere impiegata in ambito didattico come ausilio all'insegnante nella personalizzazione della sua azione educativa. Essa potrebbe essere cioè d'aiuto per indirizzare l'impegno degli insegnanti verso bisogni specifici di apprendenti una L2 che condividono una L1 così come per monitorare l'evoluzione del processo di apprendimento la L2, tenendo traccia dell'evoluzione delle competenze linguistiche nelle varie fasi di sviluppo dell'interlingua.

RIFERIMENTI BIBLIOGRAFICI

Andorno, Cecilia & Stefano Rastelli (2009), *Corpora di Italiano L2: tecnologie, metodi, spunti teorici*, Perugia, Guerra Edizioni.

Barbagli, Alessia, Lucisano Pietro, Dell'Orletta Felice, Montemagni Simonetta & Venturi Giulia (2014), *Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado*, in Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it), 9-10 December, Pisa, Italy.

Barbera, Manuel & Carla Marello (2004), *VALICO (varietà di apprendimento della lingua italiana Corpus Online): una presentazione*, in «Itals, Didattica e linguistica dell'italiano a stranieri», anno II numero 4 2004, 7-18.

Biber, Douglas (1993), *Using register-diversified corpora for general language studies*, in «Computational Linguistics Journal», volume 19(2), 219-241.

Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill & Martin Chodorow (2013), *TOEFL11: A Corpus of Non-Native English*, Educational Testing Service, Research Report ETS RR–13-24, http://www.ets.org/Media/Research/pdf/RR-13-24.pdf.

Cimino, Andrea, Felice Dell'Orletta, Giulia Venturi & Simonetta Montemagni (2013), *Linguistic Profiling based on General-purpose Features and Native Language Identification*, in Proceedings of Eighth Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta, Georgia, June 13, 207-215.

Dell'Orletta, Felice, Simonetta Montemagni & Giulia Venturi (2011b), *READ-IT: assessing readability of Italian texts with a view to text simplification*, in Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT'11), Edimburgo, UK, 30 luglio, 73-83.

Dell'Orletta, Felice, Simonetta Montemagni, Eva Maria Vecchi & Giulia Venturi (2011a), *Tecnologie linguistico-computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria*,

in Giovanni Carlo Bruno, Immacolata Caruso, Manuela Sanna, Immacolata Velleco (a cura di), Percorsi migranti: uomini, diritto, lavoro, linguaggi, Milano, McGraw-Hill, 319-336.

Dell'Orletta, Felice, Simonetta Montemagni, Giulia Venturi (2014), *Assessing document and sentence readability in less resourced languages and across textual genres*, in Thomas François and Delphine Bernhard (a cura di), *Recent Advances in Automatic Readability Assessment and Text Simplification, Special issue of «International Journal of Applied Linguistics»*, 165:2, John Benjamins Publishing Company, 163-193.

Granger, Sylviane, Estelle Dagneaux & Fanny Meunier (2009), *The International Corpus of Learner English, Handbook and CD-ROM*, version 2, Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Koppel, Moshe, Jonathan Schler & Kfir Zigdon (2005), *Determining an author's native language by mining a text for errors*, in Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05), 624-62.

Lo Duca, Maria Giuseppina (2004), *Sulla rilevanza per la glottodidattica dei dati di acquisizione di lingue seconde*, in Anna Giacalone Ramat (a cura di), *Verso l'italiano*, Roma, Carocci, 254-304.

Montemagni, Simonetta (2013), *Tecnologie linguistico-computazionali e monitoraggio della lingua italiana*, in «Studi Italiani di Linguistica Teorica e Applicata (SILTA)», Anno XLII, Numero 1, 145-172.

Pallotti, Gabriele (1998), La seconda lingua, Milano, Bompiani.

Selinker, Larry (1972), *Interlanguage*, in «International Review of Applied Linguistics», 10, 209-31.

Tetreault, Joel, Daniel Blanchard & Aoife Cahill (2013), *A report on the first native language identification shared task*, in Proceedings of the Eighth Workshop on Building Educational Applications Using NLP, Atlanta, GA, USA, June. Association for Computational Linguistics.

Wong, Sze-Meng Jojo & Mark Dras (2009), *Contrastive analysis and native language identification*, in Proceedings of the Australasian Language Technology Workshop (3-4 December 2009, Sydney), Australasian Language Technology Association, 53-61.