GABRIELE PALLOTTI
Università di Modena e Reggio Emilia

UNA NUOVA MISURA DELLA COMPLESSITÀ LINGUISTICA: L'INDICE DI COMPLESSITÀ MORFOLOGICA (ICM)

L'articolo presenta una nuova misura per valutare la complessità morfologica dei testi: l'Indice di Complessità Morfologica (ICM). L'indice calcola la varietà media di esponenti morfologici all'interno di un testo, mediante una procedura automatica di comparazione entro e tra campioni estratti casualmente. Viene qui presentato uno studio di validazione, che ha prodotto risultati positivi: la misura risulta essere affidabile e stabile grazie alle molteplici iterazioni di campionamento; non viene influenzata dalla lunghezza del testo; può essere calcolata su sotto-campioni di grandezza variabile tra 5 e 15 forme linguistiche ciascuno. A questo proposito, viene anche proposta un'operativizzazione ancora più semplice del costrutto, denominata CM5a, che si basa solo sulla varietà interna a campioni di 5 forme.

1. *La complessità linguistica*

La nozione di complessità è stata oggetto di considerevole attenzione da parte della linguistica teorica e applicata degli ultimi due decenni. Da un lato, le ricerche tipologiche hanno messo in discussione il dogma della equi-complessità dei sistemi linguistici, avviando un dibattito sulla possibilità di confrontare diverse lingue e stabilire quali sono più o meno complesse. D'altro lato, la linguistica applicata ha trattato la complessità linguistica come variabile dipendente, mettendola in relazione con variabili indipendenti quali il tempo di apprendimento della seconda lingua o il tipo di attività comunicativa (task) in cui sono coinvolti i parlanti.
In questo saggio si proporrà una definizione operativa del costrutto della complessità morfologica dei testi, che risulta essere ancora largamente inesplorato sia in ambito teorico che applicato. Dopo avere passato in rassegna

le principali ricerche precedenti, si fornirà una definizione dell'Indice di Complessità Morfologica (ICM), mostrando come esso possa essere calcolato manualmente e con l'ausilio di un programma informatico dedicato. L'indice così proposto sarà oggetto di uno studio di validazione, per determinarne la tenuta su testi di diverso tipo e lunghezza e in lingue come l'italiano e l'inglese, che sono piuttosto differenti sul piano della morfologia flessiva. Si discuteranno infine prospettive per la ricerca futura e possibili ambiti di applicazione della misura proposta.

Il termine 'complessità' è in sé complesso, e sono necessari alcuni chiarimenti preliminari. Già a livello dizionariale l'aggettivo 'complesso' ha due sensi fondamentali:

1. Che presenta difficoltà per la comprensione o l'orientamento, dovute a profondità od oscurità di concetti oppure a una molteplicità di elementi o di aspetti ....
2. Risultante di due o più parti interdipendenti (contrario di *semplice*). (Devoto-Oli, 2000).

Il primo senso implica difficoltà, sforzo cognitivo per un soggetto umano, il secondo senso descrive invece la complessità intrinseca, strutturale, di un oggetto. Nella discussione tipologica sulla complessità delle lingue, questa distinzione viene formulata in termini di complessità "objective vs agent-related" (Dahl 2004) o "absolute vs relative" (Miestamo 2008). In linguistica si propone di considerare anche un terzo senso di complessità, strettamente legato al primo, che ha a che fare con la difficoltà di apprendimento di un sistema linguistico: secondo Trudgill (2001: 371), "linguistic complexity ... equates with 'difficulty of learning for adults'"; altri parlano di "L2 acquisition difficulty" (Szmrecsanyi, Kortmann 2009) o "outsider complexity" (Kusters 2003).

Questa polisemia è problematica quando si cerca di dare una definizione operativa ed esplicita di un costrutto scientifico, ed è per ciò che in questa sede si opterà per un solo senso, che corrisponde alla complessità strutturale. Naturalmente, come già appare dalla definizione dizionariale, questa 'molteplicità di elementi o di aspetti' (complessità strutturale) può essere responsabile della 'difficoltà per la comprensione o l'orientamento' (complessità cognitiva), ma piuttosto che usare lo stesso termine per designare la causa e l'effetto, preferiremo riservare 'complessità' alle caratteristiche strutturali dell'oggetto e 'difficoltà' agli effetti che esse possono produrre sulla sua elaborazione cognitiva.

Una seconda distinzione concettuale importante riguarda gli oggetti stessi di cui si vuole stabilire la complessità. In ambito linguistico, possiamo

infatti parlare della complessità di un sistema linguistico, nel senso dell'insieme dei suoi elementi e delle sue regole (la *langue* di Saussure), e la complessità di un singolo testo, in quanto specifico atto di *parole*. L'indice di complessità morfologica proposto in questo saggio si riferisce alla complessità dei testi e non a quella dei sistemi linguistici. Inoltre, la nozione di complessità può applicarsi a tutti i livelli di analisi linguistica, dalla fonologia alla semantica: in questa sede tratteremo solo della complessità morfologica, e in particolare di quella inerente alla flessione verbale.

2. *Precedenti definizioni operative per quantificare la complessità morfologica*

Per gran parte del XX secolo, l'equi-complessità delle lingue è stato un assioma praticamente indiscusso. Si è trattato in qualche modo di una salutare reazione alle ideologie linguistiche del XIX secolo, che portavano a classificare le lingue in base alle loro "qualità" intrinseche, producendo gerarchie molto discutibili e spesso eurocentriche. Tuttavia, come ogni assioma, nessuno ha mai tentato di dimostrarlo e le sue fondamenta paiono essere essere tanto ideologiche quanto quelle del suo predecessore. Negli ultimi anni, alcuni linguisti hanno cominciato a porre la domanda quasi tabù che consiste nel chiedersi se una lingua possa dirsi più complessa di un'altra. Trudgill (1989) ha osservato che le lingue che nella loro storia sono state apprese da un gran numero di parlanti non nativi, come l'inglese, tendono ad essere più semplici di quelle parlate in piccole comunità isolate. McWhorter (2001) ha esplicitato ulteriormente questa affermazione, affermando nel titolo di un suo articolo che "The world's simplest grammars are creole grammars". Questo punto di vista è stato ulteriormente sviluppato nel volume *Language interrupted* (McWhorter 2007), in cui si discutono diverse lingue, tra cui i creoli come caso estremo, la cui trasmissione intergenerazionale è stata interrotta da un gran numero di apprendenti non nativi, portando alla nascita di sistemi linguistici "più semplici". Negli anni successivi c'è stata un'esplosione di pubblicazioni sul tema, con decine di articoli e diversi volumi (tra cui si possono ricordare Dahl 2004; Miestamo et al. 2008; Sampson et al. 2009).

Nelle pagine seguenti si darà una breve rassegna dei principali studi a carattere generale, per soffermarsi poi su quelli che sviluppano più da vicino l'area della morfologia. Si privilegeranno i contributi che definiscono la complessità in termini strutturali (che sono comunque la maggioranza, almeno nella ricerca tipologica), lasciando da parte quelli che definiscono la complessità in termini di difficoltà cognitiva o acquisizionale.

La maggior parte delle ricerche tipologiche sulla complessità morfologica riguarda il livello dei sistemi linguistici. Un sistema con molte categorie grammaticali, ciascuna rappresentata da una varietà di forme diverse, viene così detto essere più complesso di uno con poche categorie espresse da un piccolo insieme di esponenti ricorrenti. Dahl (2009), per esempio, ha confrontato due lingue geneticamente affini, l'elfdaljan (o övdalian) e lo svedese, concludendo che il primo è più complesso perché ha più casi, generi e declinazioni nominali e paradigmi verbali più ampi, che esprimono non solo il tempo, ma anche la persona e il numero. Baechler e Seiler (2012) hanno calcolato la "inflectional complexity" di cinque varietà di tedesco, definendola come il numero di "markers" (ovvero relazioni forma / funzione, o celle di un paradigma) x il numero di classi flessive. Un approccio leggermente diverso è stato seguito da Shosted (2006), per il quale la complessità morfologica esprime il grado di "inflectional synthesis", operazionalizzata come numero di categorie grammaticali per parola + numero di posizioni in cui esse possono apparire sul verbo (Bickel & Nichols 2005). Dai suoi calcoli emergono differenze notevoli tra lingue, che vanno da un valore di 29 per il Wichita allo 0 del vietnamita.

Szmrecsanyi, Kortmann (2009) sono, per quanto ci risulta, gli unici autori in ambito linguistico-tipologico che hanno tentato di misurare la complessità morfologica non a livello di sistema, ma di realizzazioni testuali. Essi distinguono "synthetic and analytic grammaticity", che risultano rispettivamente dal numero di morfemi legati e liberi, la cui somma produce un indice di grammaticità ("grammaticity") generale. Il loro approccio consiste nell'estrarre gruppi casuali di 1000 parole da corpora orali di diverse varietà di inglese e calcolare la proporzione di morfemi liberi e legati. Il risultato è che varietà vernacolari tradizionali, geograficamente isolate, presentano l'indice di grammaticity più alto, mentre questo tende ad essere più basso in varietà native parlate in zone di alto contatto geolinguistico, e ancora più basso in varietà principalmente apprese come L2, come gli inglesi di Hong Kong o della Giamaica. Gli autori propongono anche un indice di trasparenza, dato dalla percentuale di allomorfi regolari su tutti i morfemi grammaticali legati. Anche su questo parametro, le varietà vernacolari tradizionali tendono ad essere poco trasparenti, mentre le varietà apprese come L2 tendono a esibire la massima regolarità.

Il problema della complessità morfologica dei testi è stato trattato anche nelle ricerche acquisizionali. Il settore degli studi sulla seconda lingua non ha affrontato questo aspetto in modo ampio e sistematico, e i pochi studi che hanno cercato di operativizzare la complessità morfologica hanno fatto ricorso a indicatori piuttosto grossolani, come la frequenza delle forme verbali flesse, il numero di diverse forme verbali o la varietà delle forme

verbali passate (per una rassegna, Bulté e Housen 2012). Oltre a essere parziali, tutti questi indici hanno il problema di esprimere valori assoluti (quantità, frequenze), che, come è noto, sono estremamente sensibili alla lunghezza del testo.

Un approccio che cerca di superare questo problema è stato proposto da Malvern et al. (2004), che hanno lavorato principalmente sull'acquisizione della L1. Questi autori hanno sviluppato una misura di diversità lessicale chiamata D, che è quasi completamente insensibile alla lunghezza del testo, essendo calcolata su piccoli campioni di poche decine di parole estratti casualmente dal testo complessivo. Una misura derivata da D è ID, che sta per Inflectional Diversity, e può essere calcolata automaticamente attraverso il programma Clan (childes.psy.cmu.edu). Tale programma consente infatti di calcolare il valore di diversità lessicale D sia per le forme flesse (per cui *go, going, goes, went* contano come quattro types) che per i lemmi (per cui *go, going, goes, went* contano come un unico type). La diversità flessiva ID risulta semplicemente dalla sottrazione del valore di D relativo ai lemmi dal valore di D per le forme flesse: se la differenza è grande, ciò significa che per ogni lemma si trovano in media molte forme flesse, e dunque si può dire che in quel testo c'è un alto grado di diversità flessionale; al contrario, se il numero di lemmi e di forme flesse è molto simile, ciò significa che i lemmi appaiono sempre o quasi sempre nella stessa forma, dando un basso valore di diversità flessionale.

Per quanto questo approccio sia attraente, soprattutto per la sua semplicità e per la linearità dell'interpretazione, esso ha due importanti limiti. Il primo è che ID è sensibile alla lunghezza dei testi, quando essi sono di lunghezza inferiore alle 200 parole (Xanthos, Gillis 2010). Il secondo, e più grave, è che ID è direttamente legato al valore di D, cioè alla diversità lessicale: se un testo utilizza molti lessemi differenti, questo alzerà il valore di D, ma abbasserà automaticamente quello di ID, perché la maggior parte dei lessemi apparirà in una forma sola. Nel caso teorico estremo di un testo che contenga solo lessemi unici (cioè con un rapporto type/token lessicale = 1), ID sarà uguale a 0 per definizione, indipendentemente dal fatto che i lessemi del testo presentino tutti la stessa flessione o ne abbiamo molte decine.

Questi limiti sono superati con la misura (Normalized) Mean Size of Paradigm, (N)MSP, proposta da Xanthos e Gillis (2010). Anche questo approccio si basa sull'estrazione di campioni casuali dal testo (o dal corpus). Per ciascun campione, e per ciascuna classe di parole (nomi, verbi, aggettivi ecc), si calcola il numero di forme flesse, che corrisponde alla dimensione del paradigma per quella classe di parole entro quel campione. La grandezza media del paradigma (normalizzata perché riferita a campioni di dimensioni costanti in modo da ovviare al problema di testi di lunghezza

differente) si ottiene trovando la dimensione media dei paradigmi in diversi campioni.

Questa misura, pur offrendo diversi vantaggi, presenta alcune limitazioni. La prima è che gli autori non indicano una grandezza precisa per i campioni: in un saggio essi sono di 50 parole (Xanthos et al. 2011), in un altro si confrontano i valori per campioni di 50 e 500 parole, senza indicare su quale dimensione dovrebbe cadere la scelta (Xanthos, Gillis 2010). Come è evidente, la dimensione media del paradigma non varia in base alla lunghezza del testo, ma varia ampiamente in base alla dimensione del campione: in quelli di 500 parole è chiaro che si troverеanno paradigmi più ricchi che in quelli da 50 parole. C'è poi il problema di contare il numero di forme di una determinata classe di parole, ad esempio i verbi, in campioni di N parole: se in un dato campione, ma anche in un testo nel suo complesso, i verbi sono più o meno frequenti che in un altro, questo avrà un impatto sul numero medio di forme per campione, e cioè sul valore di (N)MSP. E infine, l'estrazione casuale di campioni di 50 parole e il calcolo delle forme dei paradigmi per ciascuno di essi richiede procedure informatiche non accessibili all'utente comune, per cui la misura risulta essere difficilmente replicabile.

3. *L'indice di complessità morfologica (ICM)*

Verrà ora proposta una misura di complessità morfologica che ha l'obiettivo di superare le difficoltà degli approcci precedentemente illustrati. Essa si basa su una definizione di complessità puramente strutturale e può essere applicata a testi di dimensione variabile a partire da un centinaio di parole. Come la (N)MSP Xanthos e Gillis (2010) l'approccio prevede l'estrazione di campioni casuali di misura standard, per superare i problemi dovuti alle diverse dimensioni dei testi. Tuttavia, sussistono importanti differenze con l'approccio precedente. In primo luogo, la dimensione dei campioni non è calcolata sul numero di parole, ma sul numero delle forme di una determinata classe di parole, in modo che tutti i campioni contengano lo stesso numero di verbi, o di nomi e così via. In secondo luogo, viene messa a disposizione una procedura informatica che consente a chiunque di calcolare l'indice in automatico, anche con testi di grandi dimensioni e con campionamenti casuali ripetuti per un grande numero di cicli. Infine, la validità della misura può essere dimostrata empiricamente, ed è questo lo scopo principale del presente lavoro.

L'approccio generale seguito per calcolare l'ICM si basa su due passi consecutivi. Il primo passo consiste in un'analisi linguistica che individua

all'interno di un testo gli esponenti morfologici, ovvero le diverse forme che possono assumere i lessemi appartenenti a una certa classe di parole. Il secondo passo consiste in un'analisi matematica che calcola la varietà degli esponenti all'intero dei campioni e tra di essi.

Per quanto riguarda l'analisi linguistica, è importante chiarire che l'ICM si basa sul computo degli esponenti, cioè delle forme flesse che possono assumere le parole. Da questo punto di vista, in *tu cant-i* (2sg.pres.indic) e in *che lui cant-i* (3sg.pres.cong) è presente un solo esponente, *-i*, anche se esso esprime sincreticamente due diverse funzioni grammaticali. D'altro canto, gli allomorfi che esprimono la stessa funzione grammaticale sono contati separatamente, per cui in *arriv-ato*, *pr-eso* e *rim-asto* sono presenti tre diversi esponenti, includendo in essi anche le vocali tematiche di diverse coniugazioni, che contribuiscono ad aumentare la complessità morfologica del sistema. In altre parole, si calcola il numero delle forme, non delle combinazioni forma/funzione, o celle del paradigma.

Questa scelta può parere controintuitiva, ma si basa su precise ragioni. In particolare, nei testi prodotti da apprendenti di prima e seconda lingua non è sempre possibile identificare l'esatta funzione di una forma flessa: ad esempio, in *Maria prendi gatto* come si deve interpretare *prendi*? Come un presente indicativo, durativo/abituale, di seconda persona singolare, cioè con il valore che ha in italiano, o come un presente generico, che copre anche significati progressivi, o persino di passato, in certi contesti? E il fatto che sappiamo che la forma *-i* viene applicata a un referente di terza persona singolare ci porterà a costruire una cella del paradigma -i:3.sg, oppure diremo che si tratta di una sovraestensione del morfema di seconda singolare, e inseriremo questa *-i* nella cella corrispondente?

In breve, nella varietà di apprendimento non è sempre possibile stabilire nettamente le funzioni grammaticali, mentre le forme linguistiche sono in linea di massima più identificabili, ed è per questo che l'ICM, che si prevede possa essere frequentemente alle ricerche acquisizionali, si limita a calcolare la varietà delle forme flessive. Tuttavia, nulla impedisce di utilizzare la procedura matematica ai rapporti forma-funzione, cioè alle celle dei paradigmi, calcolando, per ciascun campione, il numero di stringhe di analisi morfologica quali 'i:2sg.pres.indic' o 'i:3sg.pres.cong'.

Una volta identificati gli esponenti flessivi per ciascuna classe di parole all'interno di un testo, inizia l'analisi matematica della loro diversità (la tab. 1 presenta un esempio fittizio su forme verbali dell'inglese). Dal numero totale degli esponenti si estraggono campioni di N forme, ad esempio 10 nel nostro caso. Per ciascuno di questi campioni, si calcola il numero di diversi tipi flessivi, cioè di diverse forme (min 1 – max 10): nella tabella 1, essi sono ad esempio 6 nel primo campione e 7 nel secondo. Si calcola poi la diversità

media all'interno dei campioni, nel nostro caso (6+7)/2 = 6.5. Viene poi calcolata la diversità tra campioni, identificando, per ogni coppia di campioni, gli esponenti unici, non condivisi (min 0 – max 20). Nel nostro esempio questi sono 6 (*took, is V-ing, came, went, is*). Si calcola poi la media di questi valori di diversità e la si divide per 2; nel nostro caso, 6/2 = 3. Infine, il valore medio di diversità nei campioni viene sommato al valore medio di diversità tra campioni diviso due, si sottrae 1 e si arriva così all'Indice di Complessità Morfologica. Nel nostro esempio si avrà quindi 6.5 + 3 – 1 = 8.5, valore calcolato su campioni di 10 forme ciascuno (ICM10).

ed	ed
ed	came
took	went
was	was
Ø	Ø
Ø	Ø
is V-ing	Ø
are	Ø
are	are
are	is
6	**7**

Tabella 1. Esempio fittizio su 20 esponenti verbali inglesi

Riassumendo, l'indice di complessità morfologica (ICM) viene calcolato mediante la seguente formula:

ICM = (varietà media all'interno dei campioni + diversità media tra campioni/2) – 1

Con campioni di 10 esponenti, la gamma teorica di valori va da 0 a 19: Min 0 = 1+0-1, Max 19 = 10+(20/2)-1.

L'intera procedura per il calcolo dell'ICM può essere svolta attraverso un programma informatico sviluppato da Vaclav Brezina dell'Università di Lancaster, disponibile all'indirizzo http://corpora.lancs.ac.uk/vocab/ analyse_morph.php. Il programma restituisce dapprima un'analisi morfologica del testo inserito, basandosi sul lemmatizzatore TreeTagger (Schmid 1994),

che consente di identificare, per ciascuna parola, il lemma di base, in modo che sia possibile estrarre dalla forma flessa l'esponente morfologico. La lista di esponenti così ottenuta in automatico può essere modificata liberamente dal ricercatore, per rettificare eventuali errori di codifica. È anche possibile importare una lista di esponenti estratti manualmente o con altre procedure.

Una volta costruita la lista di esponenti morfologici, il programma consente di eseguire la parte matematica dell'analisi, estraendo un numero di campioni casuali di N forme ciascuno e calcolando, per ciascuno di essi, il numero di diversi esponenti che contiene e il grado di diversità rispetto a un altro campione casuale. Questa procedura viene iterata un numero X di volte, per arrivare a stabilire la diversità media intra- e inter-campioni, su cui si basa il calcolo dell'ICM. La grandezza (N) del campione e il numero di iterazioni (X) possono essere definiti dall'utente: nelle prossime pagine si presenteranno dati empirici relativi a diverse opzioni per questi parametri, così da tentare una prima validazione della misura.

4. *Validazione della misura*

Il presente studio, che rappresenta il primo approccio alla validazione dell'ICM, si propone di rispondere alle seguenti domande di ricerca:

a. La misura è sensibile al numero di iterazioni di campionamento?

b. Quali sono gli effetti prodotti dalla scelta di campioni di diverse dimensioni?

c. Quali sono, se ci sono, gi effetti della lunghezza del testo?

d. È possibile utilizzare una misura ancora più semplice?

Per l'analisi è stato costituito un corpus di testi paralleli in italiano e in inglese, appartenenti a diversi generi, così da poter verificare la capacità della misura di rilevare differenze inter- e intra-linguistiche. I testi scelti sono i seguenti:

- Quadro comune europeo di riferimento per le lingue (Consiglio d'Europa, 2001), solo la scala relativa ai descrittori generali di competenza (capitolo 3.3). Codice: QCE.
- Narrazione orale della storia per immagini *Frog where are you?* (Meyer 1969), da parte di due apprendenti dell'italiano e dell'inglese come L2. Codice: FROG.
- La fiaba di Cappuccetto rosso. Codice: CAP.
- Un saggio politico di Chomsky, *What Uncle Sam really wants*

(Chomsky 1992), tradotto in italiano come *I cortili dello Zio Sam*. È stata analizzata la prima sezione, di poco più di 500 parole, che ha dimensioni simili agli altri testi del corpus (CHO1). Per osservare il comportamento della misura in testi più lunghi, sono state analizzate inoltre le prime sei sezioni, in italiano e in inglese (CHO1-6) e, solo in inglese, l'intero testo di 43 pagine (CHO43).

Le dimensioni limitate del corpus sono dovute all'indisponibilità, per il momento, di un analizzatore automatico dei testi in italiano, per cui l'estrazione degli esponenti morfologici in questa lingua è stata eseguita manualmente. È attualmente in corso lo sviluppo del programma che consentirà la codifica automatica anche in italiano e, in un futuro prossimo, anche in francese, tedesco e spagnolo. L'analisi dei testi in inglese è invece stata eseguita automaticamente dal programma, con un intervento di verifica manuale da parte dell'autore: data anche la semplicità morfologica dell'inglese, l'accuratezza di codifica per questa lingua si è rivelata essere molto alta, superiore al 90% dei lemmi analizzati. In questo saggio l'Indice di Complessità Morfologica sarà calcolato esclusivamente sui verbi.
La tabella 2 riassume i dati utilizzati per questo studio.

	parole	**verbi**
Frog story L2 ing	305	35
Frog story L2 it	397	56
Scala QCER ing	371	77
Scala QCER it	436	68
Chomsky sez 1 ing	511	92
Chomsky sez 1 it	586	80
Cappuccetto ing	1,000	216
Cappuccetto it	831	157
Chomsky sez 1-6 ing	4.470	795
Chomsky sez 1-6 it	4474	657
Chomsky totale ing	22.886	3.861

Tabella 2: corpus della ricerca.

Per dare una prima idea della diversa complessità morfologica dei testi presi in esame, riportiamo nella tabella 3 l'ICM calcolato su campioni di 10 verbi ciascuno (ICM10).

QCE-ing	4,05
FROG-ing	6,11
CHO1-ing	6,95
FROG-it	7,51
QCE-it	7,84
CAP-ing	8,85
CAP-it	12,39
CHO1-it	15,25

Tabella 3: valori di ICM10 nel corpus.

Come prevedibile, i testi in italiano presentano in genere un maggiore livello di complessità morfologica sui verbi rispetto a quelli in inglese. Tuttavia, la fiaba di Cappuccetto rosso in inglese ha una maggiore complessità morfologica della narrazione in L2 della Frog story e dei descrittori del QCE in italiano. Inoltre, l'ordine di complessità crescente dei testi è sostanzialmente analogo nelle due lingue, con un'eccezione: mentre in inglese Cappuccetto rosso ha una maggiore complessità verbale del saggio di Chomsky, in italiano l'ordine di questi due testi è rovesciato.

4.1 *Numero di iterazioni di campionamento*

La prima domanda di ricerca riguarda la sensibilità della misura al numero di iterazioni della procedura di campionamento. A titolo esemplificativo, sulla prima sezione e sulle prime sei sezioni del saggio di Chomsky in inglese (CHO1-ing e CHO1-6-ing), sono stati calcolati gli indici di complessità morfologica con campioni di 5 e di 10 verbi ciascuno (MC5 e MC10), con 10 prove successive. In un caso, sono stati estratti solo 10 campioni casuali, nell'altro, ne sono stati estratti 100. I risultati sono riportati nelle tabelle 4 e 5.

Prova	MC5 / 10cicli	MC10 / 10cicli	MC5 / 100cicli	MC10 /100cicli
1	4,670	6,920	4,650	6,930
2	4,510	6,960	4,650	6,950
3	4,670	6,960	4,660	6,970
4	4,680	7,040	4,650	6,930

5	4,670	6,870	4,660	6,960
6	4,620	6,890	4,650	6,940
7	4,660	6,910	4,670	6,930
8	4,610	7,000	4,650	6,920
9	4,660	6,860	4,660	6,930
10	4,530	6,850	4,660	6,950
Media	**4,628**	**6,926**	**4,656**	**6,941**
DS	0,061	0,063	0,007	0,016
CV	0,013	0,009	0,002	0,002

Tabella 4: CM5 e CM10 calcolate su serie di campionamenti casuali con 10 e 100 cicli di campionamento; testo CHO1-ing (511 parole).

Prova	MC5 / 10cicli	MC10 / 10cicli	MC5 / 100cicli	MC10 /100cicli
1	4,690	7,310	4,660	7,300
2	4,680	7,310	4,670	7,300
3	4,680	7,310	4,660	7,290
4	4,640	7,300	4,660	7,310
5	4,650	7,300	4,670	7,310
6	4,690	7,310	4,660	7,300
7	4,640	7,310	4,670	7,300
8	4,680	7,290	4,670	7,290
9	4,680	7,320	4,670	7,300
10	4,660	7,310	4,670	7,300
Media	4,669	7,307	4,666	7,300
DS	0,020	0,008	0,005	0,007
CV	0,004	0,001	0,001	0,001

Tabella 5: CM5 e CM10 calcolate su serie di campionamenti casuali con 10 e 100 cicli di campionamento; testo CHO1-6-ing (4.470 parole)

I valori di MC5 sono ovviamente più bassi di quelli di MC10, dato che si tratta di scale diverse, aventi come massimi teorici rispettivamente 9 e 19. Tuttavia, uno sguardo alle tabelle mostra che diverse prove di campio-

namento casuale non danno risultati significativamente diversi, come indicano i valori decisamente bassi della deviazione standard e del coefficiente di variazione, che consiste nel rapporto tra deviazione standard e media. Esprimendo il coefficiente di variazione come rapporto percentuale tra la deviazione standard (cioè lo scarto medio dalla media) e la media, si può dire che la variabilità tra misurazioni più alta si riscontra su testi brevi (511 parole), con campioni piccoli (MC5) e pochi cicli di campionamento (10), ma anche qui rimane intorno all'1,3%. Nelle stesse condizioni, aumentare i cicli di campionamento a 100 riduce già il margine di variabilità tra misurazioni a uno 0,2% rispetto alla media, che è un valore del tutto trascurabile. Con un testo più lungo, di alcune migliaia di parole, la variazione scende ulteriormente, andando da uno 0,4% per MC5 e 10 soli campionamenti, a meno dello 0,1% in tutti gli altri casi.

Si può dunque affermare che già con 100 cicli di campionamento casuale, che richiedono un tempo di elaborazione informatica dell'ordine di pochi secondi su testi di alcune centinaia o migliaia di parole, si riescono a ottenere valori estremamente stabili e affidabili dell'ICM. Per questa ragione, tutte le analisi delle prossime pagine si baseranno su valori ottenuti con 100 iterazioni di campionamento.

4.2 *Dimensioni dei campioni*

L'ICM si basa sul calcolo della diversità media all'interno e tra campioni di N forme verbali. Ci si può domandare quali siano gli effetti sull'indice dovuti all'uso di campioni di diverse dimensioni, cioè al variare di N. Come si è detto, i valori dell'indice crescono automaticamente con l'aumentare della grandezza del campione: se è vero, come si è visto, che con campioni di 10 verbi, il massimo teorico è 19 (10+[20/2]-1), è evidente che il massimo teorico per campioni di 5 verbi sarà 9 (5+[10/2]-1), mentre per campioni di 15 verbi sarà 29 (15+[30/2]-1). Dunque, lo stesso testo riceverà valori diversi a seconda che se ne calcolino gli indici CM5, CM10 o CM15.

Il punto cruciale è capire se questi diversi indici portano a diverse classifiche nella complessità relativa di una serie di testi, o se sono sostanzialmente equivalenti. Il grafico 1 mostra l'andamento dei valori degli indici calcolati su campioni di 5, 8, 10, 12 e 15 esponenti ciascuno (CM5, CM8, CM10, CM12, CM15), prendendo come base l'andamento di CM10. Come si vede, le linee corrono in modo del tutto parallelo, con due sole eccezioni: CAP-ing risulta meno complesso di QCE-it con CM5, mentre è più complesso con tutti gli altri valori; inoltre, con MC15, CHO1-6-ing è più complesso di FROG-it, mentre l'ordine è invertito con tutti gli altri valori.

Nonostante queste due piccole discrasie (che riguardano comunque te-

sti adiacenti nella scala di complessità crescente), la correlazione tra le misure è altissima (*r* di Pearson sempre superiore a 0.96; *rho* di Spearman sempre superiore a 0.97; $p < 0.001$). Questo si evince anche calcolando il fattore di inflazione della varianza (Variance Inflation Factor, o VIF), che misura il grado di collinearità tra serie di misure. Valori di VIF superiori a 10 indicano che le misure seguono un andamento collineare: con i nostri dati, il VIF tra le diverse serie di misure è sempre più alto di 1000, confermando l'altissimo livello di collinearità.

Si può dunque concludere che la scelta di campioni di grandezza diversa non ha un particolare impatto sull'esito finale relativo alla complessità maggiore o minore di diversi testi.

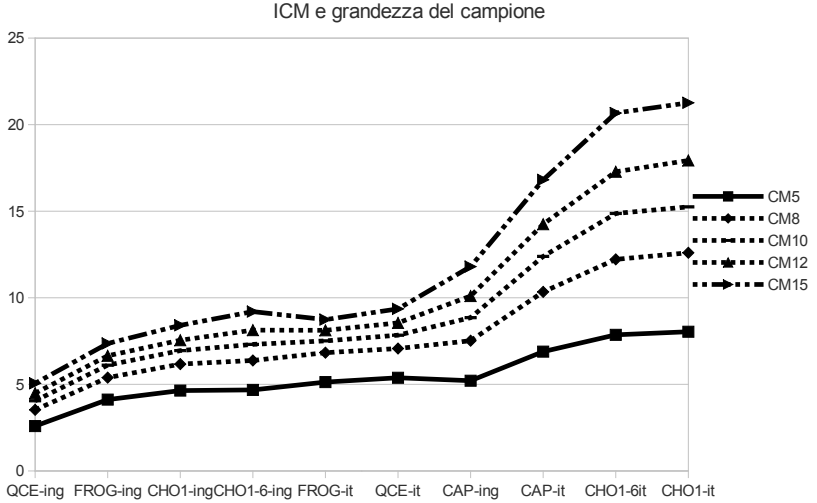

Grafico 1: valori di ICM con campioni di diversa grandezza

4.3 *Complessità morfologica e lunghezza del testo*

L'Indice di Complessità Morfologica si basa essenzialmente sulla logica degli indici type/token di complessità lessicale: i valori più alti si ottengono quando nel testo sono presenti molti tipi di esponente diversi, scarsamente ripetuti; viceversa, la presenza di numerose ripetizioni di una ristretta gamma di tipi porta ad abbassare il valore dell'indice. È risaputo che l'indice type/token di varietà lessicale è significativamente influenzato dalla lunghezza del testo: con testi più lunghi è più probabile che gli stessi tipi lessicali vengano ripetuti più volte, abbassando così il valore del rapporto type/token.

Per ovviare a questo problema sono state proposte diverse soluzioni: quella più semplice da calcolare è l'indice di Guiraud, che mette al numeratore la radice quadrata del numero di token. In tal modo l'effetto della lunghezza del testo viene ridimensionato, ma non è affatto eliminato, così che testi più lunghi hanno comunque indici di Guiraud sistematicamente più bassi di testi brevi. Un'alternativa più soddisfacente è lo Standardized Type/Token Ratio (STTR), che calcola il valore medio dell'indice type/token in sotto-campioni di uguale lunghezza: è la soluzione adottata per la misura (N)MSP di Xanthos e Gillis (2010) e per l'ICM proposto in questo lavoro.

Ci si può dunque chiedere se l'ICM risenta in modo sistematico della lunghezza del testo, che sarebbe un grave limite rispetto alla sua validità. Per rispondere a questa domanda si è calcolato l'ICM in porzioni testuali di lunghezza variabile, ma tratte tutte dallo stesso testo, per controllare la variabilità a livello di tipo testuale.

Il grafico 2 mostra i valori dell'ICM in porzioni di lunghezza crescente ricavate tutte dal saggio di Chomsky in inglese, che ha una complessità morfologica media rispetto all'intero corpus, ed è anche quello per cui è stato possibile calcolare l'indice sull'intero testo di 23.000 parole. Sull'asse delle ordinate sono riportati i valori di complessità morfologica con campioni di tre dimensioni diverse (CM5, CM10 e CM15), sull'asse delle ascisse i valori per porzioni di testo variabili tra 85 parole (con 15 verbi) e 23.000 parole (con 3861 verbi). Come si può vedere, le linee sono sostanzialmente piatte, con un leggero incremento tra 750 e 4500 parole, ma solo per CM 10 e CM15.

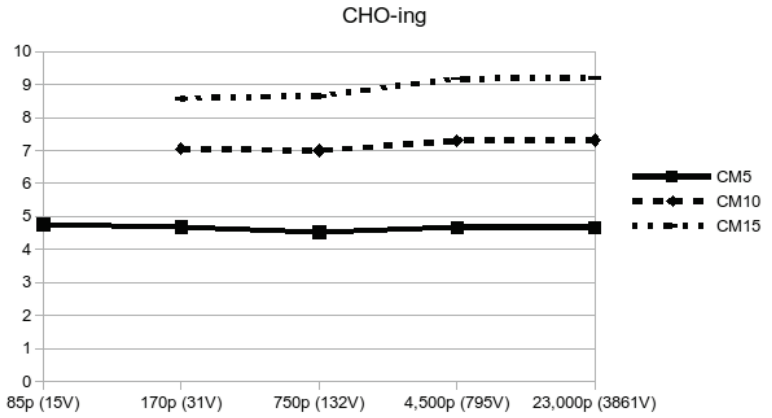

Grafico 2: valori di ICM5, ICM10 e ICM 15 con porzioni di grandezza crescente del testo CHO-ing.

Si è voluto ripetere la verifica con il testo che nel campione ha la complessità morfologica più bassa (il QCER in inglese) e con quello con il valore più alto (il saggio di Chomsky in italiano); i risultati sono presentati nei grafici 3 e 4.

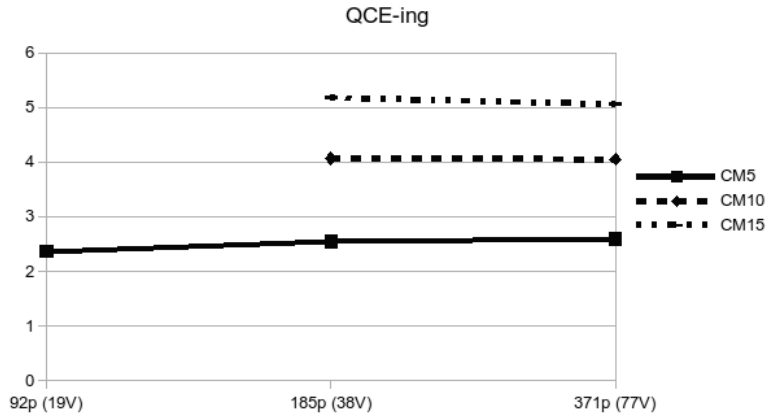

Grafico 3: valori di ICM5, ICM10 e ICM 15 con porzioni di grandezza crescente del testo QCE-ing

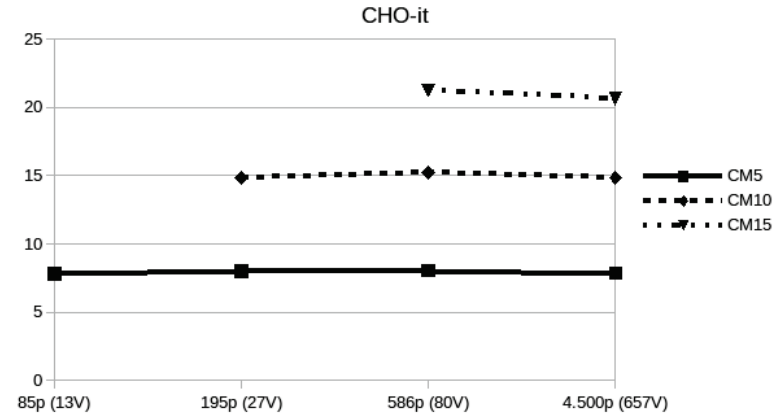

Grafico 4: valori di ICM5, ICM10 e ICM 15 con porzioni di grandezza crescente del testo CHO-it

Anche in questi testi, come si vede, le linee rimangono piatte, anzi MC15 in CHO-it decresce tra le versioni con 580 e 4500 parole, in contro-

tendenza con quanto accade per la versione in inglese dello stesso testo. Si può dunque concludere che le leggerissime variazioni dell'indice non paiono essere sistematicamente correlate alla lunghezza del testo, ma sono probabilmente dovute al caso o a livelli maggiori o minori di complessità morfologica nei diversi campioni estratti (hanno dunque a che fare con i testi in sé, e non con la loro lunghezza).

4.4 *Si può semplificare ulteriormente l'ICM?*

Un'ultima domanda di ricerca riguarda la struttura stessa del costrutto di complessità morfologica, così come viene operativizzato dall'ICM. Questo indice risulta infatti dalla somma di due fattori: la diversità media all'interno dei campioni e la diversità media tra campioni. Tuttavia, dato che il programma ripete per un numero elevato di volte (lo standard è 100) la procedura di campionamento sull'intero testo, ci si può chiedere se i due fattori alla fine non misurino in modo diverso la stessa cosa, cioè la varietà media di esponenti morfologici presenti nel testo. Se così fosse, se cioè i due fattori fossero altamente correlati, la diversità media all'interno del campione potrebbe essere presa come unico indicatore della complessità morfologica, rendendo la procedura ancora più semplice.

Nelle prossime pagine vedremo se una versione nettamente semplificata dell'ICM risulta essere correlata con altre versioni dell'indice. Chiameremo questa versione semplificata CM5a, a significare che si tratta di una complessità morfologica calcolata su campioni di 5 verbi e che si riferisce solo al fattore (a), ovvero la diversità media interna al campione.

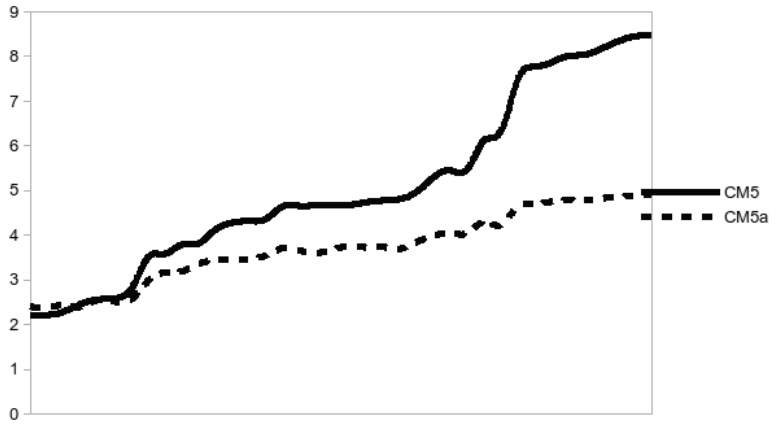

Grafico 5: correlazione tra CM5 e CM5a in testi di complessità crescente

Il grafico 5 mostra la correlazione tra CM5 (l'indice con due fattori) e CM5a (l'indice basato solo sul primo fattore, la diversità interna ai campioni), su tutti i testi e le porzioni di testo del corpus. Come si vede, le due linee hanno un andamento del tutto parallelo, anche se appare evidente che il primo fattore dà un contributo maggiore quando i valori di complessità sono bassi, mentre con valori di complessità più alti il secondo fattore gioca un ruolo sempre più importante. In ogni caso la correlazione è molto forte (r = 0,98, p < 0,001), che significa che l'uso di un indicatore o dell'altro produce sostanzialmente gli stessi risultati in termini di classificazione dei testi per quanto riguarda la loro complessità morfologica.

Ci si può però chiedere se il contributo di CM5a rispetto a CM5 vari in funzione della lunghezza del testo: se così fosse, la scelta di un indicatore rispetto all'altro produrrebbe sistematicamente risultati diversi in testi più o meno lunghi, compromettendo la validità della misura.

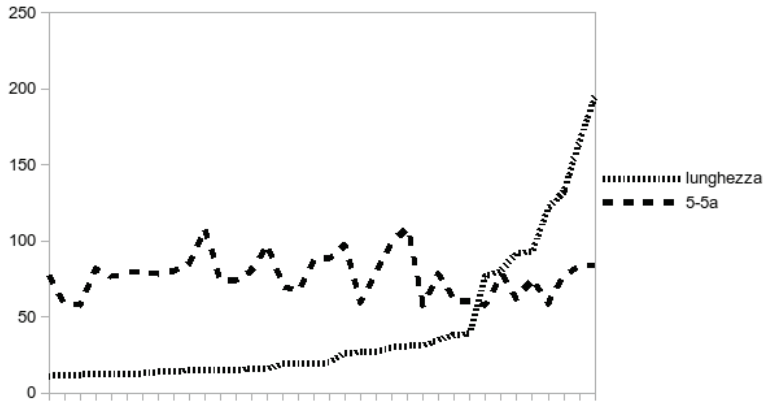

Grafico 6: correlazione tra il contributo di MC5a al valore di MC5 e lunghezza del testo

Il grafico 6 mostra con una linea a tratti larghi il contributo del fattore (a) rispetto all'indice complessivo CM5, e con una linea a puntini la lunghezza dei testi corrispondenti. Appare evidente come il contributo di CM5a a CM5a si alzi e si abbassi in relazione alla maggiore o minore complessità del testo, come si è visto in precedenza, ma non pare avere alcun rapporto con la lunghezza del testo, come conferma anche il test di correlazione (r = 0,006; p = 0,97).

Si può dunque concludere che l'indice CM5a potrebbe essere un'alternativa valida all'ICM complessivo. Se ulteriori ricerche su campioni più

consistenti confermassero questo risultato, CM5a potrebbe essere indicata come la misura definitiva per calcolare la complessità morfologica dei testi, con una serie di vantaggi. Sul piano pratico, l'uso di campioni molto piccoli, di 5 verbi, consentirebbe di applicare la misura anche a testi brevissimi, dai 10 verbi in poi. Sul piano teorico, l'interpretazione del valore di CM5a risulta essere ancora più intuitiva, significando semplicemente la diversità media di tipi morfologici all'interno di campioni di 5 forme.

5. *Conclusioni*

In questo saggio si è presentato un nuovo indice per misurare la complessità morfologica dei testi, cercando di dimostrarne empiricamente la validità. L'indice di complessità morfologica (ICM) presenta i seguenti punti di forza:

- si basa su un costrutto relativamente semplice e di immediata interpretazione;
- può essere calcolato automaticamente grazie a un programma online apposito, di libera consultazione;
- il calcolo automatico consente il campionamento ripetuto di numerosi campioni casuali, consentendo di arrivare a stime affidabili della complessità media del testo;
- può essere applicato anche a testi brevi e brevissimi, fino a un minimo di circa 100 parole;
- è sostanzialmente insensibile alla lunghezza del testo;

Si è anche dimostrato come esso possa essere ulteriormente semplificato attraverso l'uso dell'indice CM5a, che sarà oggetto di ulteriori studi in futuro. Oltre a questi, i progetti di sviluppo includono l'estensione dell'analisi morfologica automatica all'italiano, francese, tedesco e spagnolo. Si provvederà inoltre a migliorare ulteriormente l'affidabilità delle analisi morfologiche automatiche relative all'inglese, benché esse abbiano già raggiunto un ottimo livello. Infine, si cercherà di sviluppare procedure di analisi automatica che consentano il trattamento anche di corpora di parlato, notoriamente difficili da etichettare.

Le applicazioni dell'indice sono molteplici. In primo luogo, esso può essere utilizzato nella linguistica dei corpora come parametro ulteriore per l'analisi testuale, consentendo il confronto tra testi nella stessa lingua e tra diverse lingue, offrendo così una nuova dimensione alla linguistica testuale comparativa, alla stilometria e al calcolo degli indici di leggibilità dei testi.

Un altro ambito di potenziale applicazione riguarda l'acquisizione della seconda lingua: l'ICM può andare ad arricchire la lista di costrutti che vengono utilizzati per studiare la complessità delle produzioni in L2, sia relativamente all'evoluzione nel tempo, sia relativamente alla variazione tra diverse attività comunicative, o task. Fino ad ora sono state esplorate con una certa sistematicità la complessità lessicale e quella sintattica, ma quella morfologica è rimasta sostanzialmente ignorata. Ugualmente si possono ipotizzare applicazioni dell'indice all'acquisizione della lingua materna, per descrivere le traiettorie di sviluppo e stabilire valori normativi per le diverse fasce di età, anche allo scopo di diagnosticare eventuali ritardi o disturbi specifici del linguaggio.

Riferimenti bibliografici

BAECHLER R., SEILER G., 2012, Simplification, complexification, and microvariation: Towards a quantification of inflectional complexity in closely related varieties", in RALLI, A., BOOIJ G., SCALISE S., KARASIMOS A. (a cura di), *Morphology and the Architecture of Grammar. On-line Proceedings of the Eighth Mediterranean Morphology Meeting*, http://morbo.lingue.unibo.it/mm

BICKEL B., NICHOLS J., 2005, "Inflectional synthesis of the verb", in HASPELMATH M., DRYER M., GIL D., COMRIE N. (a cura di), *The World Atlas of Language Structures*, Oxford, Oxford University Press, http://wals.info/chapter/22

BULTÉ B., HOUSEN A., 2012, "Defining and operationalising L2 complexity", in HOUSEN A., KUIKEN F., VEDDER I. (a cura di), *Dimensions of L2 Performance and Proficiency – Investigating Complexity, Accuracy and Fluency in SLA*, Amsterdam, Benjamins.

CHOMSKY N., 1992, *What Uncle Sam Really Wants*, Berkeley, CA, Odonian Press (trad. it *I cortili dello Zio Sam; gli obiettivi della politica estera americana dal Vecchio al Nuovo Ordine Mondiale*, a cura di David Barsamian, Gamberetti, 1996).

CONSIGLIO D'EUROPA, 2001, *Common European Framework of Reference for Languages*, Cambridge University Press (trad. it. *Quadro comune europeo* di riferimento per le lingue, Firenze, La Nuova Italia, 2002)

DAHL O., 2009, "Testing the Assumption of Complexity Invariance: The Case of Elfdalian and Swedish", in SAMPSON G., GIL D., TRUDGILL P. (a cura di), *Language Complexity as an Evolving Variable,* Oxford, Oxford University Press.

DAHL O., 2004, *The Growth and Maintenance of Linguistic Complexity*, Amsterdam, Benjamins.

DEVOTO G., OLI, C.C., 2000, *Dizionario della lingua italiana*. Firenze, Le Monnier.

MALVERN D., RICHARDS B., CHIPERE N. AND DURÁN P., 2004, *Lexical diversity and language development,* Houndmills, Palgrave MacMillan.

MAYER M. ,1969, *Frog, where are you?*, New York, Penguins.

MCWHORTER J., 2007, *Language Interrupted,* Oxford, Oxford University Press

MIESTAMO M, SINNEMÄKI K., KARLSSON F. (a cura di), 2008, *Language Complexity: Typology, Contact, Change,* Amsterdam, Benjamins.

MIESTAMO M., 2008, "Grammatical complexity in a cross-linguistic perspective", in MIESTAMO M., SINNEMÄKI K., KARLSSON F. (a cura di) *Language complexity: Typology, contact, change,* Amsterdam, Benjamins.

SAMPSON G, GIL D, TRUDGILL P (a cura di), 2009, *Language Complexity as an Evolving Variable,* Oxford, Oxford University Press.

SCHMID H, 1994, "Probabilistic Part-of-Speech Tagging Using Decision Trees", in *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

SHOSTED, R. 2006, "Correlating complexity: A typological approach", in *Linguistic Typology,* n. 10.

SZMRECSANYI B., KORTMANN B., 2009, "Between simplification and complexification: non-standard varieties of English around the world", in SAMPSON G., GIL D., TRUDGILL P. (a cura di), *Language Complexity as an Evolving Variable,* Oxford, Oxford University Press.

TRUDGILL P., 1989, "Contact and isolation in linguistic change", in BREIVIK L., JAHR E. H. (a cura di), *Language Change: Contributions to the Study of its Causes,* Berlin, Mouton de Gruyter.

XANTHOS A. *et al.,* 2011, "On the role of morphological richness in the early development of noun and verb inflection", in *First Language*, n. 4.

XANTHOS A., GILLIS S., 2010, "Quantifying the development of inflectional diversity", in *First Language*, n. 2.