LO SVILUPPO LONGITUDINALE DELLA FRASEOLOGIA IN APPRENDENTI CINESI DI ITALIANO L2

Uno studio preliminare su alcune categorie di errori

Stefania SPINA

ABSTRACT • *The longitudinal development of phraseology in Chinese learners of Italian l2: a preliminary study on some categories of errors*. This study is a preliminary investigation of the development of phraseology in Chinese learners of Italian. Based on an error-annotated longitudinal learner corpus, it aims at analysing three different categories of lexical combinations, in order to understand to what extent phraseological errors are affected by time and by the use of either the open-choice or the idiom principle (Sinclair 1991). The analysis shows that L2 phraseology learning can be slow and uneven. As formulaicity is pervasive in language, research on phraseological errors can provide invaluable support for language teaching and learning.

KEYWORDS • L2 Italian; Learner Corpus Research; Phraseology; Error Analysis

1. Introduzione: la *Learner corpus research* e gli studi sulle unità fraseologiche

A partire dalla fine degli anni 80 del secolo scorso, il filone di studi denominato *Learner corpus research* (LCR) è emerso dall'area di ricerca preesistente della Linguistica dei corpora, con l'obiettivo esplicito di applicare i suoi metodi allo studio empirico dell'acquisizione di una L2, basandosi su dati prodotti da apprendenti e organizzati in *learner corpora*. Come è stato osservato, questo nuovo approccio è diventato nel tempo "a truly interdisciplinary field at the crossroads between corpus linguistics, second language acquisition, language teaching and natural language processing" (Granger, Gilquin, Meunier 2015:3).

Da allora, nonostante molto lavoro resti ancora da fare, la LCR si è affermata come un campo di studi fertile e denso di sviluppi positivi, che sta man mano affinando metodologie e strumenti di analisi. In particolare, due tendenze metodologiche si stanno delineando in questi ultimi anni: da un lato, l'adozione di metodologie miste, che integrino i dati forniti dai corpora con altri dati, di tipo sperimentale (Durrant, Siyanova-Chanturia 2015; Callies 2015); dall'altro, l'uso di tecniche statistiche più mature, che consentano di effettuare predizioni più accurate e di delineare linee di tendenza più affidabili a partire da tali dati (Gries 2015).

Alla LCR si riconosce, in ogni caso, il merito di avere prodotto un insieme consistente di studi e di ricerche che hanno dato un contributo rilevante alla ricerca acquisizionale, consentendo inoltre di integrarla con metodologie di analisi fondate su solide basi quantitative. Tali risultati sono stati raggiunti soprattutto nel campo dell'acquisizione dell'inglese L2, attraverso l'analisi di dati prevalentemente scritti, spesso elicitati tramite la produzione di testi argomentativi o descrittivi, raccolti in contesti spesso legati all'apprendimento in classe (Granger 2008); si è trattato per la maggior parte, inoltre, di studi di tipo trasversale più che

longitudinale, che hanno analizzato soprattutto apprendenti di livelli medio-avanzati (Gilquin 2015).

Una parte consistente di questo insieme di ricerche ha riguardato l'uso delle combinazioni di parole da parte degli apprendenti: tre decenni di analisi di dati estratti da *learner corpora* hanno infatti evidenziato il ruolo centrale della dimensione fraseologica nella produzione linguistica dei parlanti non nativi (per una sintesi, si vedano Granger, Meunier 2008 e Ebeling, Hasselgård 2015).

Le unità fraseologiche possono essere definite come combinazioni lessicali in cui ognuna delle parole che le costituiscono co-occorre con le altre con livelli diversi di probabilità, che non possono essere interamente previste dal significato delle singole parole, ma sono caratterizzate da un grado variabile di convenzionalità e di cristallizzazione dovuta all'uso (Evert 2005). Il verbo *fare*, ad esempio, in italiano può co-occorrere con moltissimi nomi (*amicizia*, *buio*, *caso*, ecc.), e in queste combinazioni perde gran parte del suo significato letterale; l'aggettivo *marchiano*, invece, co-occorre quasi esclusivamente con il nome *errore*, assumendo il significato di "molto grande".

Le unità fraseologiche, dunque, sono elementi lessicali complessi, dotati di proprietà semantiche e sintattiche presenti in modo variabile: il grado di trasparenza/opacità semantica, ad esempio, può variare secondo un continuum di massima trasparenza (*parcheggiare la macchina*), parziale opacità (*prendere una decisione*) o opacità totale (*tirare le cuoia*). Proprio per questa loro estrema eterogeneità, numerosi sono stati i tentativi di fornirne una classificazione, che tenga conto delle loro peculiarità strutturali e semantiche. Tra i moltissimi esempi di classificazione, in ambito italiano è da notare quello di Masini (2012), che distingue le tre categorie di parole sintagmatiche, collocazioni e combinazioni preferenziali, sulla base dei tre parametri di fissità sintagmatica, fissità paradigmatica e familiarità: le parole sintagmatiche si caratterizzano per la presenza di tutti e tre i parametri (*prendere freddo*, *anno accademico*); le collocazioni non sono fisse sintagmaticamente (*aprire un conto*, ma *il conto è stato aperto da*…); le combinazioni preferenziali non sono caratterizzate da fissità né sintagmatica né paradigmatica, ma solo da un grado di familiarità più alto (*gravemente malato*, ma anche *seriamente malato*, ecc.). Nell'ambito della LCR, invece, una delle classificazioni più accreditate è quella di Granger, Paquot (2008), che distinguono tre grandi categorie di frasemi, quelli con funzione referenziale (le espressioni idiomatiche, le collocazioni, i composti come *fine settimana* o *luna di miele*, ecc.), quelli con funzione testuale (i connettivi, gli avverbi e le congiunzioni complesse, come *nella misura in cui*, *in modo che*, *d'altra parte*, ecc.), e quelli con funzione comunicativa (routine discorsive istituzionalizzate che rispondono a particolari atti linguistici, come le formule di saluto o di cortesia).

Nel corso degli ultimi tre decenni, un numero crescente di studi nell'ambito della ricerca acquisizionale e psicolinguistica basata su *learner corpora* ha dimostrato la centralità di questa dimensione fraseologica, da un lato nelle attività di processamento e di comprensione da parte degli apprendenti (Siyanova-Chanturia 2015b), dall'altro in quelle di produzione (Granger, Meunier 2008; Granger, Paquot 2008). Sul versante della produzione, in particolare, è stato ripetutamente evidenziato che la competenza fraseologica ha un ruolo determinante nel rendere i comportamenti linguistici degli apprendenti simili a quelli dei nativi, in modo particolare sul piano della fluenza (Ellis et al. 2008; Ellis, Simpson-Vlach 2009), che migliora sensibilmente con una maggiore competenza nel campo della fraseologia: la disponibilità di combinazioni lessicali già pronte, e, al tempo stesso, la capacità di tali combinazioni di servire da modello per la produzione di sequenze ulteriori, in parte simili, a loro volta memorizzabili come elementi lessicali unici (Ellis et al. 2015), fornisce ai parlanti nativi – come a quelli non nativi – blocchi di discorso già pronti, e contribuisce al tempo stesso a migliorare la fluenza (Pawley, Syder 1983).

In questo contesto, una parte consistente della ricerca basata su *learner corpora* ha lavorato alla disponibilità di misure di associazione in grado di operazionalizzare l'estrazione da corpora e la classificazione delle unità fraseologiche in tipologie differenti, anche allo scopo di quantificarne alcuni parametri. Tra le misure di associazione più utilizzate, la *mutual information* (MI: Church, Hanks 1990; Durrant, Schmitt 2009) misura prevalentemente la forza di associazione tra i due componenti lessicali della combinazione (assegna un valore più alto alle combinazioni i cui componenti tendono ad occorrere più spesso in associazione che separatamente, e, di conseguenza, sono più strettamente associate); il *t-score*, che è la misura più simile a quella della frequenza (Church, Hanks 1990), perché enfatizza in particolare la frequenza della combinazione e degli elementi che la compongono; la *lexical gravity* (Gries 2010), che misura la diversificazione delle due parole, basandosi sulla frequenza dei *types*; e il *DeltaP* (Gries 2013), che considera i modi diversi in cui il primo componente della combinazione attrae il secondo (*DeltaP1*), e viceversa (*DeltaP2*).

1.1. *Learner corpus research* e studi sulla fraseologia dell'italiano L2

> Gli studi sulle unità fraseologiche nell'italiano di apprendenti che possano essere inseriti nel filone della LCR, per ricorso sistematico a dati raccolti secondo criteri stabilmente codificati, organizzati in corpora di apprendenti (Guilquin 2015), e per metodi statistici affidabili usati per analizzarli (Gries 2015), sono relativamente scarsi (Spina 2017)[1].

Siyanova-Chanturia (2015a), ad esempio, a partire da un piccolo corpus longitudinale scritto di apprendenti cinesi di livello elementare, analizza l'evoluzione nell'uso delle combinazioni nome + aggettivo prodotte in italiano, ed evidenzia come, dopo alcune settimane di studio dell'italiano, gli apprendenti producano combinazioni più frequenti e più strettamente associate.

Siyanova-Chanturia e Spina (2015) è invece finalizzato a verificare la corrispondenza tra l'intuizione soggettiva sulla frequenza di 80 combinazioni nome + aggettivo di un campione di parlanti nativi italiani da un lato, e di uno di apprendenti cinesi dall'altro, e, insieme, la correlazione tra l'intuizione sulla frequenza di parlanti nativi e non nativi e i dati estratti da corpora. L'analisi dei dati rivela che i giudizi intuitivi di nativi ed apprendenti sono in buona parte correlati, anche se sono influenzati in modo diverso da altri fattori, come ad esempio la lunghezza delle parole che compongono le combinazioni, e che c'è una sostanziale correlazione tra i giudizi dei parlanti e la frequenza rilevata nei corpora.

Spina (2015) si occupa invece di combinazioni di parole usate in ambito accademico, sulla base di un corpus di interazioni in forum di discussione tra parlanti nativi e non nativi avanzati nel corso di un master universitario. L'analisi dei dati sulle combinazioni lessicali rivela che gli apprendenti utilizzano combinazioni verbo + nome e nome + aggettivo in modo quantitativamente simile ai nativi, e si servono di un insieme ristretto, particolarmente frequente, di tali combinazioni, in misura anche maggiore rispetto ai nativi. Gli apprendenti

[1] I poli di ricerca italiani entro cui si sono sviluppati i pochi studi sull'italiano di apprendenti basati su *learner corpora*, nonché le risorse e i dati necessari per realizzarli, si raccolgono soprattutto intorno alle due Università per Stranieri di Perugia e di Siena, alle Università di Torino e di Pavia, e a Eurac di Bolzano.

tendono dunque a ripetere più spesso le stesse, limitate combinazioni che hanno appreso e che riescono a padroneggiare.

Spina (2010a, 2010b e 2016) sono invece tutti finalizzati ad illustrare il progetto di costituzione di un *Dizionario delle collocazioni italiane per apprendenti* (DICI-A: http://www.dici-a.it), attraverso l'estrazione di unità fraseologiche da un corpus di riferimento dell'italiano (il *Perugia corpus*: Spina 2014), il loro filtraggio attraverso la frequenza, combinata con misure statistiche di dispersione e di MI, e l'attribuzione ad uno di tre livelli di competenza (elementare, intermedio e avanzato; Spina 2016).

Infine, Bagna e Machetti (2008) analizza l'uso di polirematiche in un corpus scritto di apprendenti di livello B1-C2, confrontandole con quelle presenti nel dizionario GRADIT. Konecny et al. (2016) utilizza un corpus scritto di studenti di scuole sudtirolesi di madrelingua tedesca (con livelli di italiano L2 da A2 a C1), costituito per analizzare l'uso di unità fraseologiche. A questo scopo, propone una loro categorizzazione, per facilitare l'analisi del loro uso. Mulhall (2017) analizza l'uso e i pattern ricorrenti delle combinazioni introdotte da verbi di alta frequenza in apprendenti anglofoni di italiano.

2. Motivazione

Gli studi basati su *learner corpora* che hanno indagato le unità fraseologiche hanno rivelato che esse costituiscono uno scoglio per gli apprendenti, anche di livello avanzato (Bestgen, Granger 2014; Ellis et al. 2015; Nesselhauf 2003 e 2005; Wang 2016). Anche a livello internazionale, tuttavia, si registra una scarsità di ricerche che indaghino in modo mirato le aree di maggiore difficoltà di questo fenomeno, con studi incentrati su caratteristiche lessicali, semantiche o sintattiche selezionate e su specifici gruppi di apprendenti, che possano costituire il presupposto, ad esempio, per la creazione di sillabi o materiali didattici direttamente tarati su queste aree (Wanner et al. 2013).

In questo senso, questo studio ha l'obiettivo di dare un contributo all'analisi delle aree più critiche che apprendenti cinesi di livello elementare e pre-intermedio (A1, A2 e B1 del QCER) si trovano a dover affrontare nel campo in specifiche categorie di unità fraseologiche dell'italiano. A tale scopo, sono stati utilizzati dati di tipo longitudinale, raccolti a due riprese a distanza di sei mesi l'una dall'altra, in modo da poter fornire evidenze sia sulle difficoltà più evidenti che riguardano l'uso della fraseologia, sia sul modo in cui il tempo incide su tali difficoltà.

L'approccio scelto per raggiungere questo obiettivo è quello basato sull'analisi degli errori all'interno di *learner corpora* (o *computer-aided error analysis*: Dagneaux et al. 1998); tale approccio prevede che il *learner corpus* venga annotato sulla base di una tassonomia – nel caso dell'annotazione sistematica di tutti gli errori commessi dagli apprendenti – o di uno schema di annotazione – nel caso, come quello presentato in questo studio, dell'annotazione di errori relativi a specifici tratti linguistici (Díaz-Negrillo, Fernández-Domínguez 2006). Uno schema di annotazione consiste in un insieme predefinito di etichette attraverso cui gli errori di produzione vengono associati a determinate categorie; il vantaggio di questo approccio, in rapporto all'analisi degli errori tradizionale (Corder 1967), è duplice: in primo luogo, è basato su uno schema di annotazione predefinito e standardizzato; in secondo luogo, consente di ricercare in modo automatico specifiche categorie di errori, analizzandole in contesto e combinando le ricerche con altri parametri linguistici, a seconda del livello di annotazione del corpus (Callies 2015).

Nonostante i limiti, ripetutamente sottolineati (vedi ad esempio Bley-Vroman 1983), dell'analisi degli errori degli apprendenti come manifestazione della loro interlingua, legati ad esempio alla difficoltà di distinguere errori di esecuzione da errori di competenza, o

all'inevitabile considerazione che l'annotazione di un errore costituisce solo uno dei tanti possibili modi di interpretare i dati di un corpus, l'annotazione sistematica degli errori in *learner corpora* è comunque un modo esplicito e trasparente di classificare ed analizzare le produzioni degli apprendenti che si discostano da un'ipotesi target (Lüdeling, Hirschmann 2015); tale approccio, inoltre, se combinato con altri livelli di annotazione linguistica (per lemma, categoria grammaticale o struttura sintattica), può fare emergere pattern sistematici nell'uso degli apprendenti.

Questo studio preliminare è mirato in particolare ad un'analisi di tipo longitudinale; l'ipotesi che si intende verificare è che la variabile tempo abbia effetti significativi sulla produzione di errori da parte degli apprendenti considerati, e che ne influenzi la quantità e la tipologia. Ciò che ci si aspetta di rinvenire nei dati, in definitiva, è una variazione – nell'uno o nell'altro senso – in almeno alcune tipologie di errori prodotti da almeno una parte degli apprendenti, tra la prima e la seconda raccolta dei dati. Oltre a questo, gli errori sono analizzati anche nell'effetto che su di essi hanno il livello di competenza degli apprendenti e le diverse categorie di combinazioni considerate, distinte anche sulla base del loro diverso grado di cristallizzazione.

L'obiettivo più generale è quello di fornire un quadro preliminare delle diverse tipologie di errore in cui gli apprendenti cinesi incorrono più frequentemente, nel tentativo di individuare delle aree di difficoltà ricorrenti, da poter affrontare con strumenti didattici appositamente progettati.

3. Dati

L'analisi è stata condotta a partire dal LOCCLI (*Longitudinal Corpus of Chinese Learners of Italian*: Spina, Siyanova-Chanturia in preparazione), un corpus longitudinale costituito dalle produzioni scritte di 175 apprendenti cinesi che hanno frequentato un corso di italiano di sei mesi all'Università per Stranieri di Perugia nel 2015. Attraverso un test di piazzamento iniziale, gli studenti sono stati assegnati ad uno dei tre livelli A1, A2 o B1. Ad ognuno dei 175 studenti è stato chiesto di scrivere due composizioni: la prima all'inizio dei sei mesi complessivi di corso, e la seconda alla fine, a sei mesi di distanza dalla prima. Al momento di svolgere la seconda composizione, agli studenti è stato chiesto di scegliere un argomento che non si era già scelto la prima volta. Le tre composizioni vertevano su temi di tipo quotidiano: 1) Cosa penso dell'Italia e degli italiani 2) I miei hobby: cosa faccio di solito nel mio tempo libero 3) La mia ultima vacanza.

Dopo aver raccolto i dati, il corpus risultante è stato annotato per categoria grammaticale, con l'aggiunta di una serie di metadati riguardanti gli apprendenti (l'età, il sesso, il livello di competenza, il tempo di permanenza in Italia alla data della produzione del testo) e il task svolto (la composizione scelta). Nella sua versione finale, il corpus misura circa 97.000 parole.

A partire dal LOCCLI, un suo sottoinsieme è stato selezionato in modo casuale, avendo come unico criterio il bilanciamento per livello QCER e il punto cronologico di raccolta dei dati: sono state quindi selezionate in modo casuale sessanta composizioni, di cui trenta della prima raccolta (all'inizio del corso) di dieci apprendenti per ognuno dei tre livelli QCER considerati, e trenta degli stessi apprendenti, scritte al momento della seconda raccolta, alla fine del corso semestrale di italiano.

Questo campione di sessanta composizioni è stato poi annotato manualmente, sulla base dello schema di annotazione descritto nel paragrafo seguente.

4. Metodo

L'annotazione delle combinazioni è stata effettuata attraverso il software *Brat* (Stenetorp et al. 2012), su una versione dei testi selezionati collocata in un server web, a cui gli annotatori avevano accesso in remoto attraverso un normale browser per la navigazione in Internet.

Le combinazioni considerate in questo studio sono quelle utilizzate all'interno di due tipi di relazioni sintattiche: la relazione "aggettivo modificatore di nome" (e in particolare le categorie NADJ: *anno prossimo*; ADJN *bella città*) e la relazione "verbo + oggetto diretto" (la categoria VN: *fare una gita*, *avere fame*, *prendere l'autobus*). Tali categorie di combinazioni lessicali sono tra quelle usate più frequentemente in italiano (Spina 2010a)

Il campione è stato annotato leggendo da schermo ciascuno dei sessanta testi, ed etichettando ogni combinazione, sia quelle corrette che quelle errate; questo ha consentito di avere, relativamente alle categorie considerate, un quadro globale dell'uso che il campione ne ha fatto a distanza di sei mesi.

Come già sottolineato, l'etichettatura delle combinazioni è avvenuta sulla base di uno schema creato ad hoc, e rappresentato nella tabella 1. Secondo un modello ricorrente, sono state considerate due macro-categorie di errori: gli errori grammaticali e gli errori lessicali (Lennon 1991); in questi ultimi rientrano anche errori di appropriatezza rispetto al contesto d'uso. Della prima categoria fanno dunque parte le combinazioni formalmente errate (Osborne 2008), mentre della seconda possono far parte combinazioni formalmente corrette, ma che non corrispondono ad unità fraseologiche usate comunemente in italiano, o che sono usate in modo non appropriato al contesto (James 1998).

Errori lessicali	Errori grammaticali (aggiunta, omissione, scelta, posizione)
Sostituzione di una parola Combinazione non esistente Diverso significato	articolo modificatore accordo numero

Tabella 1: Lo schema di annotazione degli errori nelle combinazioni lessicali.

Gli errori grammaticali, tradizionalmente considerati più facili da identificare e meno legati all'interpretazione soggettiva (Lüdeling, Hirschmann 2015), coinvolgono quattro elementi su cui l'errore può intervenire (l'articolo, il modificatore, l'accordo o il numero), e quattro operazioni errate compiute dall'apprendente (l'aggiunta non dovuta, l'omissione, la scelta o la posizione). (1)-(7) sono esempi di alcuni tipi di errori grammaticali rinvenuti nel campione del LOCCLI (tra parentesi, le categorie con cui l'errore è stato etichettato e l'ipotesi target a cui è associato. È importante sottolineare che gli errori di tipo ortografico non sono stati presi in considerazione in questo studio):

(1)
Prossima volta vorrei andare a Venezia (articolo-omissione; *la prossima volta*)
(2)
Il secondo giorno, volevamo fare la gita in città (articolo-scelta; *fare una gita*)
(3)
Faccio sempre i viaggi tranne che frequento all'universita (articolo-aggiunta; *faccio sempre viaggi*)
(4)
Secondo me è un molto importante monumento della Roma antica (modificatore-posizione; *un monumento molto importante*)
(5)

Ho visto tanti beli paesaggi naturali (modificatore-scelta; *bei paesaggi*)
(6)
A volte faccio la spese a Ipercoop (accordo-scelta; *faccio la spesa*)
(7)
Ci sono un sacco di vestito bellissimo. (numero-scelta; *vestiti bellissimi*)

Gli errori lessicali, considerati tendenzialmente più soggetti ad interpretazione soggettiva, sono spesso identificati dalla scorretta sostituzione di uno dei due componenti lessicali. Nel caso in cui gli errori riguardino invece la combinazione nella sua globalità, possono essere legati alla creazione di una combinazione del tutto nuova, non esistente in italiano, o all'uso di una combinazione italiana esistente con un altro significato (Wanner et al. 2013). (8)-(10) sono esempi di ciascuna delle tre categorie di errori lessicali:

(8)
Metto la mia valigia (sostituzione di una parola; *faccio/preparo la (mia) valigia*)
(9)
Verso le 15:00, abbiamo fatto il pranzo in un ristorante italiano (combinazione non esistente; *abbiamo pranzato*)
(10)
Abbiamo andare un lungo periodo, noi è stanco (diverso uso; *abbiamo camminato a lungo*)

Le fasi previste dal lavoro di annotazione degli errori sono essenzialmente tre: l'identificazione dell'errore; la scelta dell'ipotesi target da assegnare all'errore; l'interpretazione dell'errore, con la conseguente concreta etichettatura dei dati attraverso il software. In questo processo, un ruolo determinante è svolto dalla scelta dell'ipotesi target, che può a volte risultare un compito molto complesso, sia per la difficoltà di sceglierne solo una tra le diverse possibili a partire dai dati, sia perché ci sono casi in cui risulta arduo anche individuarne una sola. In (11), ad esempio, l'ipotesi target può essere sia *ho preso il treno* che *ho perso il treno*, senza che il contesto riesca in modo definitivo a far decidere per l'una o per l'altra. Il fatto che entrambe le ipotesi portino ad interpretare l'errore come l'omissione di un articolo non impedisce di annotare l'errore, ma tuttavia non elimina il dubbio sul verbo coinvolto. In (12), mentre è evidente l'errore di scelta del partitivo, il contesto non fornisce informazioni sufficienti a formulare una ipotesi target che abbia portato alla produzione della combinazione *fare degli accessori*, che non risulta utilizzata in italiano[2].

(11)
Ho predo terno quasi 2 ore che sono arrivato a Roma
(12)
Nel mio tempo libero anche mi piace giocare con il computer, fare dei accessori con mano

Anche l'ultima fase, quella dell'interpretazione, può presentare problemi non indifferenti all'annotatore: quello più comune è che in una stessa combinazione coesistano errori diversi, anche appartenenti a diverse categorie. In (13) e (14), ad esempio, in due combinazioni di categoria ADJN e VN sono presenti errori legati sia all'omissione di un articolo che all'accordo tra nome e aggettivo.

[2] In casi come questo, è utile ricorrere a corpora italiani di nativi per verificare se e quanto una combinazione è utilizzata: *fare degli accessori* non compare in *Paisà*, un corpus italiano scritto molto esteso (Lyding et al. 2014).

(13)
La mia prima impressione dell'Italia è una paese con lungo storia (*una lunga storia*)
(14)
Siamo andiamo l'albergo per lasciamo nostro bagagli (*i nostri bagagli*)

L'annotazione sul campione del LOCCLI usato in questo studio è stata svolta individualmente da due annotatori esperti, di lingua madre italiana; nei casi dubbi come quelli riportati in (13) e (14) i due annotatori hanno concordato, volta per volta, la categoria a cui attribuire gli errori, sulla base di considerazioni legate alla sua rilevanza all'interno del testo.

Il risultato finale di questo lavoro di annotazione è l'etichettatura di 1042 combinazioni VN, NADJ e ADJN, di cui 373 contengono almeno uno degli errori previsti dallo schema di annotazione descritto sopra.

5. Risultati e discussione

I risultati dell'analisi degli errori nella produzione di combinazioni lessicali delle tre categorie considerate saranno presentati nei tre paragrafi seguenti. Nel paragrafo 5.1. saranno descritte da un punto di vista quantitativo tutte le combinazioni usate dagli apprendenti, con l'obiettivo di verificare la loro evoluzione nel tempo. Nel paragrafo 5.2. saranno invece analizzati gli errori commessi dagli apprendenti, in rapporto alle tre variabili del tempo, della categoria della combinazione e del livello di competenza. Vale la pena notare che le combinazioni prese in esame in questi primi due paragrafi sono ancora considerate prescindendo dal loro grado di cristallizzazione e da quanto i due componenti siano strettamente associati tra loro; dal fatto, dunque, che possano essere considerate vere unità fraseologiche o meno. Nel paragrafo 5.3., invece, verrà considerato e misurato il loro status di unità fraseologiche, e saranno analizzati più in dettaglio gli errori degli apprendenti della categoria NADJ, anche in relazione a tale status.

5.1. Occorrenza e sviluppo longitudinale delle combinazioni lessicali

Il primo dato che emerge dall'analisi è che le diverse categorie di combinazioni sono usate in misure diverse, e producono anche errori in misura diversa. Più in particolare, le combinazioni VN ricorrono con una frequenza quasi doppia rispetto alle altre due messe insieme, e per oltre un terzo sono errate. Le combinazioni meno utilizzate dagli apprendenti cinesi sono le ADJN, che sono al tempo spesso quelle in cui si verificano più errori (il 48%). Le combinazioni NADJ sono invece quelle che vengono sbagliate di meno (solo il 30% delle volte). I due grafici nella Figura 1 riassumono questi dati.

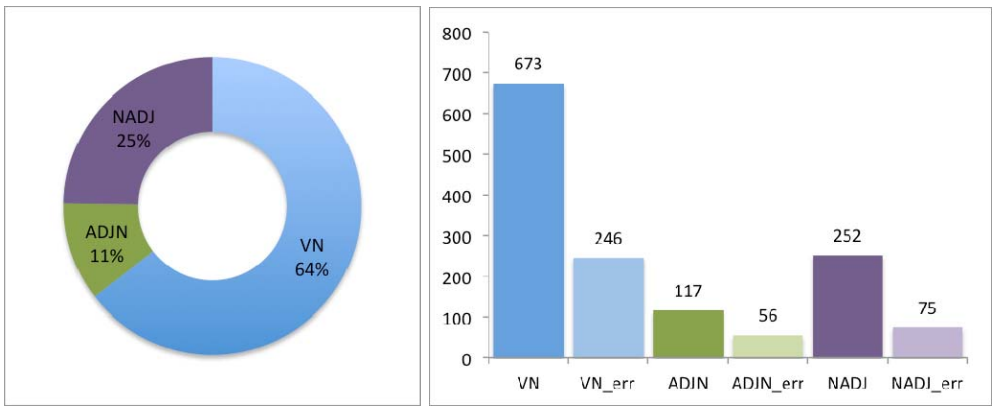

Figura 1: Le percentuali di occorrenza delle tre categorie di combinazioni (a sinistra) e gli errori per tipo di combinazione, presentati accanto al numero complessivo delle combinazioni prodotte (a destra).

La distribuzione complessiva delle diverse categorie di combinazioni in testi scritti prodotti da apprendenti cinesi non è globalmente dissimile, ad esempio, da quella delle stesse categorie di combinazioni prodotte da un campione di studenti italiani di scuola media in composizioni scritte, usato in questa occasione come corpus di controllo: anche in quel caso, le combinazioni VN sono le più usate, seguite dalle NADJ e dalle ADJN. Le differenze si riscontrano invece nel valore percentuale più basso nei parlanti nativi delle VN (48% del totale), e dai valori più alti delle altre due combinazioni. Gli apprendenti sembrano preferire le categorie di combinazioni incluse nelle relazioni sintattiche verbo + oggetto diretto, che realizzano attraverso l'uso di verbi molto frequenti (*ascoltare*, *prendere*, *fare*, *guardare* e *leggere* sono i cinque più frequenti), e che spesso ripetono più volte nel corso delle loro composizioni.

Inoltre, il livello di competenza degli apprendenti ha un effetto significativo sulla produzione di tutte e tre le categorie di combinazioni (un test Anova ha restituito un p-value < 0,01 per le combinazioni VN e NADJ, e = 0.013 per le combinazioni ADJN). La Figura 2 mostra che gli apprendenti di livello B1 sono quelli che più degli altri producono combinazioni VN, e, anche se in misura minore, NADJ (la linea orizzontale scura all'interno dei rettangoli indica il valore medio). Per le combinazioni ADJN – le meno utilizzate - risultano più produttivi gli apprendenti di livello elementare. Gli A2, invece, sono quelli che usano di meno tutte le combinazioni legate alle relazioni sintattiche considerate. In linea di massima, dunque, la frequenza delle categorie di combinazioni lessicali analizzate è correlata positivamente con il livello di competenza degli apprendenti: più elevato è tale livello, più aumentano le combinazioni utilizzate.

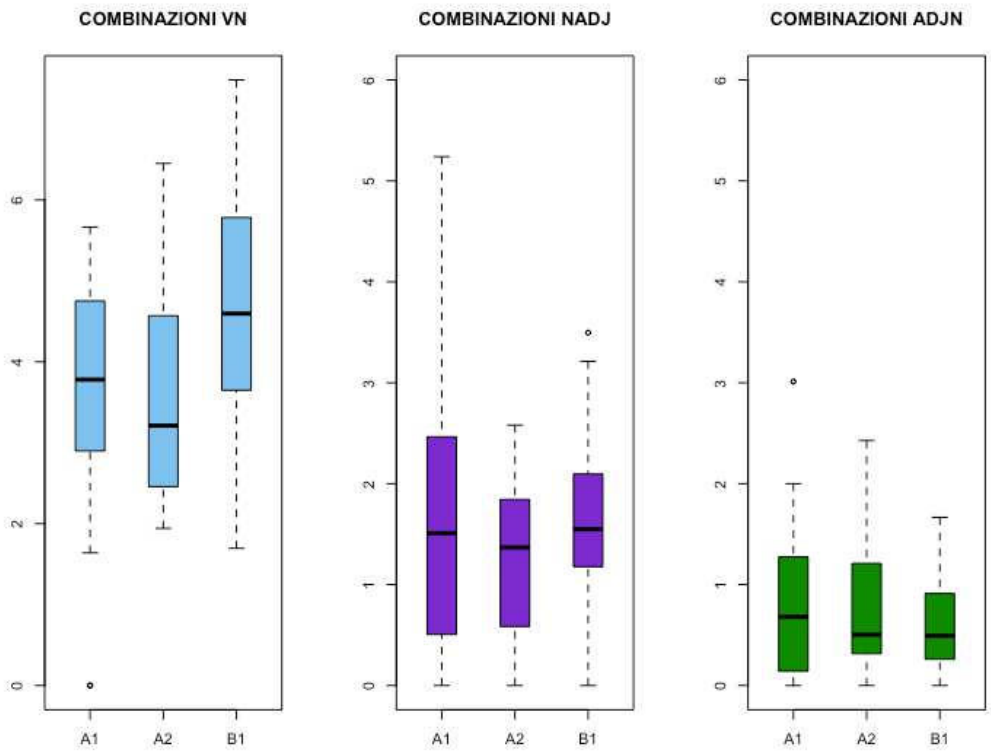

Figura 2: Le tre categorie di combinazioni suddivise per livello QCER (frequenze relative rispetto al numero di tokens).

L'analisi di come la variabile tempo incide sulla produzione di combinazioni lessicali fornisce inoltre dei dati interessanti, e per certi versi sorprendenti (vedi Figura 3): diversamente da quanto emerso in Siyanova (2015a), che analizzava tuttavia dei dati longitudinali con un intervallo temporale minore, gli apprendenti cinesi nei testi della raccolta *b* (la seconda in ordine cronologico) utilizzano significativamente di meno le combinazioni lessicali delle categorie VN e ADJN (un test Anova ha restituito un p-value < 0,01 per le due categorie menzionate). La stessa differenza, sempre nel senso di un decremento longitudinale nell'uso, non è risultata significativa per la categoria NADJ.

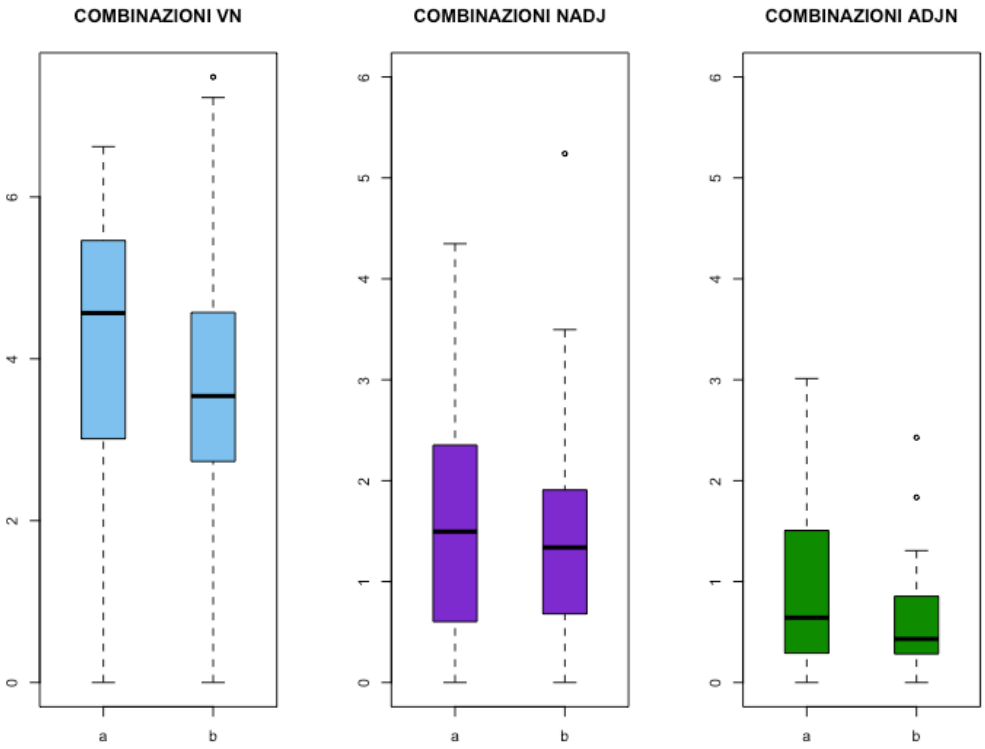

Figura 3: Le combinazioni considerate suddivise per punto di raccolta (frequenze relative rispetto al numero di tokens).

La prima considerazione suggerita dall'analisi dei dati, dunque, è che un arco di tempo di sei mesi, in cui gli apprendenti cinesi vivono in Italia e frequentano un corso intensivo di lingua italiana, influisce sulla loro produzione di determinate combinazioni lessicali, diminuendone significativamente l'uso. Questo sembra indicare che, dopo sei mesi, gli apprendenti si sforzano di produrre testi ottenuti attraverso costruzioni diverse da quelle considerate. Ad esempio, nella sezione *b* del LOCCLI, che contiene i testi della seconda raccolta, aumenta in modo significativo la frequenza del verbo *essere*, in particolare come copula nella sequenza *essere* + (avverbio) + aggettivo o *essere* + nome (*era molto bello*, *è piacevole*, *è stato divertente*, *è una ragazza*, ecc.). A titolo di esempio, (15) e (16) sono due passaggi scritti dallo stesso apprendente di livello A2: il primo fa parte della raccolta *a*, il secondo della raccolta *b*. In (15), il corsivo indica le combinazioni lessicali utilizzate, in (16) gli usi citati di *essere* all'interno di un predicato nominale.

(15)
Mi piace *guardare il film* in computer, soprattutto il film d'amore. Raramente *guardo il film italiano*, perché mi sento molto difficile per me. Quando esco, mi piace *fare spese* con la mia amica
(16)
Queste storie *erano molto romantiche*. Terzo giorno abbiamo andati alla Piazza di Spagna. *È famoso* per il film che si chiama Roman Holiday. E al quarto giorno abbiamo andati al museo Vaticano. *Era meraviglioso*. C'erano tanti tesori d'arte.

Nel caso in cui, dunque, gli apprendenti si trovino nella necessità di usare un elemento lessicale, anche di tipo complesso come quelli su cui verte questo studio, che non conoscono o che hanno difficoltà ad utilizzare, sembrano mettere in atto strategie compensative (Banfi et al. 2008), ristrutturando la frase e sostituendo in parte con costruzioni diverse le combinazioni lessicali che invece all'inizio del corso si sforzavano di produrre.

5.2. Occorrenza e sviluppo longitudinale degli errori

Il quadro si fa più accurato se si prendono in considerazione le combinazioni errate in rapporto a quelle corrette, e la loro evoluzione nell'arco di sei mesi. La differenza tra gli errori commessi nel punto di raccolta *a* e quelli commessi sei mesi dopo, in *b*, è significativa solo nel caso delle combinazioni NADJ (un test chi-quadrato ha restituito un p-value = 0.004), in cui, come mostra la Figura 4 (grafico a sinistra), il numero di errori commessi aumenta. Incrociando questo dato con quelli descritti nel paragrafo precedente, quindi, emerge che, a distanza di sei mesi, gli apprendenti producono una quantità minore di combinazioni VN, NADJ e ADJN, all'interno delle quali, almeno per quanto riguarda le NADJ, aumentano quelle errate.

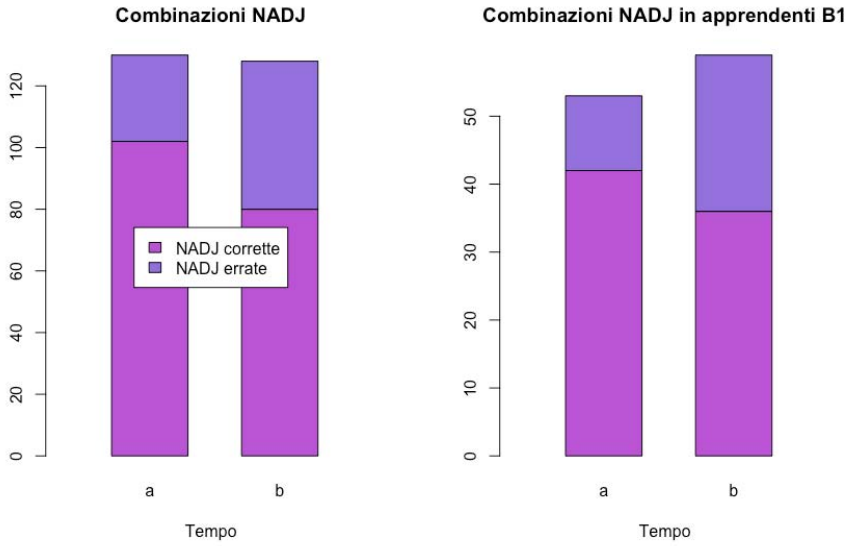

Figura 4: Le combinazioni NADJ corrette ed errate suddivise per punto di raccolta (a sinistra), e le stesse combinazioni prodotte solo dagli apprendenti di livello B1 (a destra).

Il livello di competenza degli apprendenti non sembra svolgere un ruolo determinante in rapporto all'evoluzione degli errori nell'arco di sei mesi: l'unico caso in cui si rinviene una differenza significativa negli errori commessi è ancora nelle combinazioni NADJ (vedi Figura 4, grafico a destra), per gli apprendenti di livello B1 (un test chi-quadrato ha restituito un p-value = 0.036). Dall'analisi dei dati emerge quindi che, dopo sei mesi di lezioni intensive di lingua italiana, gli apprendenti cinesi producono meno combinazioni delle tre categorie considerate, e parallelamente commettono, nel caso delle combinazioni NADJ, un numero significativamente maggiore di errori; questo riguarda in particolare gli apprendenti di livello B1.

Infine, come mostrato dalla Figura 5, gli apprendenti commettono globalmente un numero significativamente molto maggiore di errori grammaticali rispetto a quelli lessicali (un

Wilcoxon signed-rank test ha restituito un p-value < 0.01). Alcuni studi precedenti hanno ottenuto in questo senso risultati diversi: Thewissen (2008), ad esempio, ha mostrato che in apprendenti francesi, tedeschi e spagnoli di inglese L2 di livello avanzato tendono a prevalere gli errori di tipo lessicale. I risultati dei due studi, tuttavia, non possono essere messi a confronto senza delle precise distinzioni: al di là della differenza nel livello di competenza degli apprendenti, la L1 e la sua distanza tipologica dalla L2, molto ampia nel caso degli apprendenti cinesi, svolge un ruolo che non si può trascurare (Giacalone Ramat 2011). Ne è una riprova Wang (2016), che, nella sua analisi di combinazioni verbo + nome nell'inglese L2 di apprendenti avanzati svedesi e cinesi, ha riscontrato, coerentemente con questo studio, un'incidenza maggiore degli errori grammaticali rispetto a quelli lessicali negli apprendenti cinesi (ad esempio nelle combinazioni *have* + nome), e una situazione diametralmente opposta per gli apprendenti svedesi.

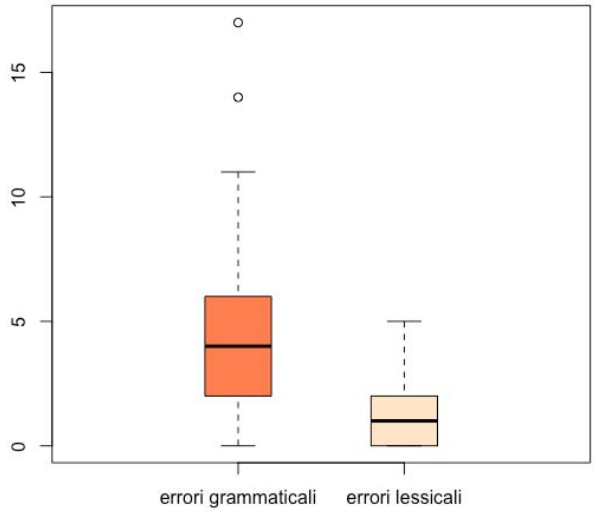

Figura 5: Gli errori grammaticali e lessicali commessi dagli apprendenti nei 60 testi del campione.

La prevalenza degli errori grammaticali su quelli lessicali sembra dunque coerente con gli studi precedenti sugli errori degli apprendenti cinesi. La quantità degli errori delle due tipologie, tuttavia, non sembra essere a sua volta influenzata dalla variabile tempo: sia se gli errori grammaticali e lessicali commessi dagli apprendenti sono considerati tutti insieme, sia se sono suddivisi per livello di competenza, la loro distribuzione non cambia significativamente nell'arco dei sei mesi considerati.

La prima ipotesi che questo studio intendeva preliminarmente verificare, dunque – e cioè che un periodo di tempo di sei mesi può produrre variazioni significative nel numero e nelle tipologie degli errori commessi – risulta confermata solo nel caso delle combinazioni NADJ, che sono usate di meno nei testi della seconda raccolta, in particolare dagli apprendenti con il livello di competenza più elevato, e che sono sbagliate di più. In tutti gli altri casi, l'effetto di sei mesi di permanenza in Italia e di studio intensivo dell'italiano costituiscono una variabile nel complesso poco significativa.

5.3. Analisi delle combinazioni NADJ: unità fraseologiche e combinazioni libere

Procedendo con l'analisi, si è cercato di distinguere, nelle produzioni degli apprendenti, le combinazioni lessicali i cui componenti co-occorrono in modo non cristallizzato e convenzionale, dalle vere e proprie unità fraseologiche. Questo aspetto è molto rilevante per analizzare l'uso che ne fanno gli apprendenti: come si è visto, infatti, la produzione sistematica di una vasta gamma di unità fraseologiche, che caratterizza i parlanti nativi, ha effetti determinanti sulla velocità di processamento e sulla fluenza, e costituisce un aspetto problematico per gli apprendenti anche a livelli di competenza avanzati (Ellis et al. 2015). È dunque importante determinare in che misura gli apprendenti cinesi del campione fanno uso di una reale competenza fraseologica, producendo combinazioni che si comportano come blocchi lessicali unici, o invece si servono di singole parole, che sono volta per volta associate ad altre attraverso regole di combinazione. Seguendo la nota distinzione di Sinclair (1991), si è cercato di individuare i casi in cui gli apprendenti si servono del "principio di scelta aperta" (*open-choice principle*), in base al quale le parole si combinano con altre parole in modo aperto e creativo, avendo come unica restrizione la grammaticalità della frase, da quelli in cui usano il "principio idiomatico", e producono "semi-preconstructed phrases that constitute single choices" (Sinclair 1991:110).

Alla luce dei risultati descritti nel paragrafo precedente, che hanno evidenziato un comportamento sensibile alla variabile tempo da parte delle combinazioni NADJ, l'analisi si è concentrata in questo caso su questa categoria di combinazioni.

Uno dei metodi più consolidati per misurare lo status fraseologico delle combinazioni lessicali è quello di combinare i due valori di frequenza e di forza di associazione (MI), calcolati sulla base di un corpus esterno di riferimento di testi prodotti da parlanti nativi (Siyanova, Schmitt 2008). In tal modo, è possibile discriminare le combinazioni lessicali occasionali, non cristallizzate nell'uso, da quelle realmente fraseologiche, e verificarne, oltre al dato della loro eventuale evoluzione nel tempo, anche le potenziali differenze negli errori cui danno origine. Il vantaggio di combinare due misure diverse risiede nella possibilità di cogliere ed integrare caratteristiche differenti delle combinazioni lessicali: in particolare, l'occorrenza sistematica (data dalla frequenza) e la forza con la quale i due componenti lessicali si attraggono a vicenda (misurata dalla MI). Di conseguenza, seguendo l'approccio già utilizzato ad esempio da Siyanova (2015a), è stata scelta una soglia di fraseologicità articolata su due livelli, e in particolare: (1) su una frequenza nel corpus di riferimento uguale a 6, e (2) su un valore di MI uguale a 3. Come corpus di riferimento è stato scelto *Paisà* (Lyding et al. 2014), che comprende testi italiani scritti estratti dal web, per un totale di oltre 200 milioni di parole, e garantisce dunque un ampio livello di rappresentatività.

Pertanto, qualsiasi combinazione NADJ utilizzata nel campione di 60 composizioni, che nel corpus di riferimento abbia una frequenza ≥ 6 e una MI ≥ 3, è considerata una unità fraseologica, i cui due componenti lessicali sono usati insieme più spesso di quanto ci si aspetterebbe basandosi solo sul caso (Gries 2008). Tutte le combinazioni i cui valori di frequenza e di MI sono invece inferiori alle soglie indicate sono considerate combinazioni libere. Alcuni esempi di unità fraseologiche e di combinazioni libere usate nel campione sono riportati nella Tabella 2.

Combinazioni libere	Frequenza in Paisà	MI in Paisà
giorno finale	18	0,04
cibo buono	3	0,52
costruzione antica	24	0,67
lezione italiana	4	-0,32

Unità fraseologiche		
giorno feriale	319	10,2
opera lirica	674	8,86
scuola elementare	2051	10,53
capelli corti	203	9,06

Tabella 2: Alcuni esempi di combinazioni libere e di unità fraseologiche usate nel campione.

Delle 252 combinazioni NADJ usate nel campione, 126 sono presenti nei testi della raccolta *a* e 126 in quelli della raccolta *b*, quindi esattamente il 50% in ciascuno dei due periodi. Tuttavia, nei testi *a* le unità fraseologiche sono il 58% del totale, contro il 42% delle combinazioni libere, mentre questi valori si invertono nei testi della raccolta *b*.

Il primo dato importante che emerge dall'analisi, dunque, è che, dopo sei mesi trascorsi in Italia frequentando un corso intensivo di lingua italiana, gli apprendenti cinesi producono un numero significativamente inferiore di unità fraseologiche NADJ (un test chi-quadrato ha restituito un p-value = 0,032), con una frequenza e una forza di associazione, quindi, superiori ai valori soglia previsti. La Figura 6 visualizza questo dato.

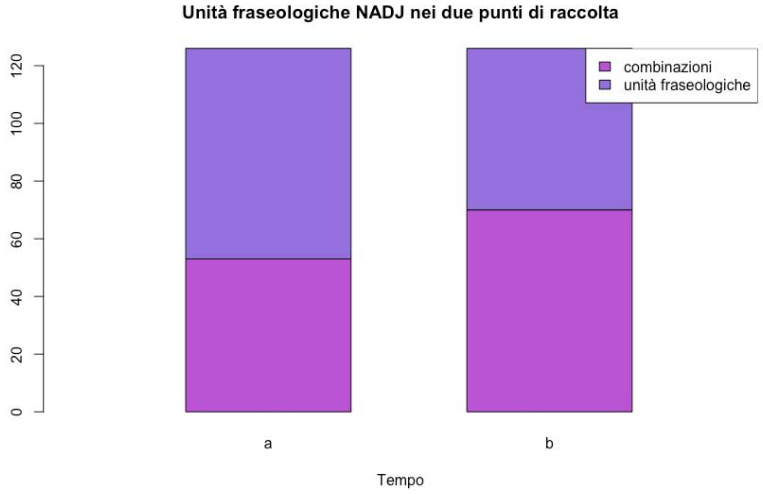

Figura 6: La differenza quantitativa tra combinazioni libere e unità fraseologiche NADJ nei due punti di raccolta *a* e *b*.

Come è emerso anche dall'analisi delle combinazioni NADJ prodotte da tutti i 175 apprendenti inclusi nel LOCCLI (Siyanova & Spina in preparazione), gli apprendenti, in una fase cronologicamente successiva del loro percorso di apprendimento dell'italiano, producono testi basati di più su sequenze lessicali ottenute attraverso l'applicazione di regole di combinazione, che non servendosi di unità fraseologiche cristallizzate. Questa tendenza dà luogo all'impiego, quantitativamente preponderante nelle composizioni scritte nel secondo punto di raccolta, di combinazioni come *quadro bello*, *fascini diversi*, *oggetti locali*, che non sono usate dai parlanti nativi italiani in modo così frequente da renderle convenzionali e cristallizzate nell'uso.

Questo risultato è coerente con quelli descritti in Bestgen, Granger (2014): man mano che sono esposti all'uso delle due parole che compongono le unità fraseologiche in un insieme di contesti che si amplia nel tempo, gli apprendenti sperimentano nuovi usi di ciascuna in contesti d'uso inediti e sempre più differenziati, producendo quindi combinazioni meno strettamente

associate e meno cristallizzate. Il punto di raccolta dei dati *b*, situato a sei mesi di distanza dall'inizio del loro percorso guidato e intensivo di studio dell'italiano e di contatto con parlanti nativi, sembra dunque coincidere con una fase di sperimentazione, che porta gli apprendenti da un lato a servirsi di un numero più ridotto di unità fraseologiche, che hanno appreso ed usano in modo stabile, e dall'altro a cercare di sfruttare un meccanismo produttivo e creativo dell'apprendimento (Chini 2005): l'obiettivo diventa ricombinare sequenze già apprese in modo nuovo, analizzandole nelle loro parti e inserendo al loro interno elementi lessicali diversi, producendo dunque frequentemente combinazioni libere invece che unità fraseologiche. Oltre che ad un meccanismo fisiologico di ricombinazione di elementi noti, tuttavia, questo ricorso maggiore al principio di scelta aperta potrebbe anche essere dovuto ad un input non sufficiente, in particolare nel contesto della classe, relativo alle combinazioni lessicali di tipo fraseologico.

Le composizioni della raccolta *a* presentano dunque una probabilità maggiore di contenere unità fraseologiche strettamente associate. Il passo successivo è stato quello di verificare se c'è una differenza nella distribuzione degli errori nelle due diverse tipologie di combinazioni, libere e fraseologiche.

La Figura 1 (a destra) ha già mostrato che le combinazioni NADJ sono, tra le tre considerate, quelle che producono al loro interno meno errori (solo il 29%). Separando le combinazioni libere dalle unità fraseologiche, possiamo verificare che, sul totale delle 252 combinazioni lessicali NADJ prodotte, 108 (il 43%) sono unità fraseologiche corrette. Questo risultato è complementare a quanto riportato ad esempio da Siyanova, Schmitt (2008) per studenti di livello avanzato: gli apprendenti, fin dai livelli più bassi, sono in grado di usare correttamente una percentuale rilevante di unità fraseologiche.

Mettendo a confronto, all'interno delle sequenze NADJ, le combinazioni corrette con quelle che contengono uno qualsiasi degli errori previsti dallo schema di annotazione, emerge una differenza di distribuzione molto significativa tra le combinazioni libere e le unità fraseologiche (un test chi-quadrato ha restituito un p-value < 0,01): come mostrato nella Figura 7, gli errori ricorrono in misura molto maggiore tra le combinazioni libere. L'uso congiunto di due elementi lessicali meno strettamente associati (le combinazioni libere), e dunque il ricorso al principio di scelta aperta, aumenta in modo significativo la probabilità, per gli apprendenti, di commettere errori; al contrario, il ricorso ad unità fraseologiche con un grado più elevato di associazione riduce la probabilità di errore. Questo dato è in parziale contraddizione con quello descritto in Nesselhauf (2005) sulle combinazioni lessicali verbo + nome nell'inglese L2 scritto di germanofoni di livello avanzato: in quel caso, le combinazioni libere producevano un numero minore di errori rispetto alle unità fraseologiche. C'è da osservare, in ogni caso, che incidono su questa differenza sia il diverso livello e la L1 degli apprendenti studiati, sia le diverse modalità di classificazione delle combinazioni.

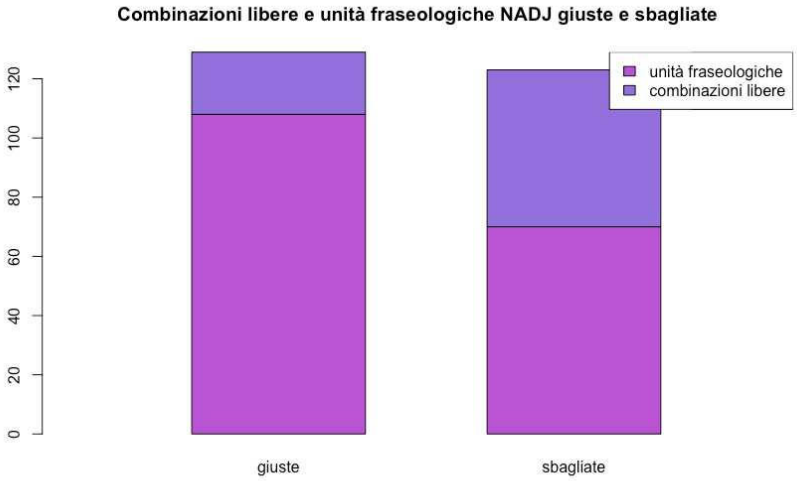

Figura 7: La distribuzione degli errori nelle unità fraseologiche e nelle combinazioni libere NADJ.

Un altro dato importante riguarda la distribuzione dei tipi di errore all'interno delle due tipologie di combinazione NADJ: gli errori lessicali, infatti, occorrono in misura percentualmente più che doppia (il 20% contro il 9%) nelle combinazioni libere. Si tratta per lo più di errori dovuti alla creazione di combinazioni che non sono presenti nel corpus di riferimento *Paisà* (*giornata scorsa, sole biondo, parete digitale, edificazione romantica*). La limitata competenza lessicale e fraseologica porta quindi gli apprendenti ad impiegare combinazioni che hanno gradi diversi di non convenzionalità: da quelle semplicemente poco probabili (con frequenza e MI molto basse), ma in ogni caso utilizzate dai parlanti nativi, anche se molto di rado (*via piccola, oggetti locali*), a quelle "creative", che associano due elementi lessicali che i parlanti nativi non usano insieme, e che proprio per questo gli annotatori hanno individuato come errori. Ad esempio, (17) è un errore lessicale di sostituzione, la cui ipotesi target potrebbe essere "una cosa piacevole" oppure "una cosa bella", prodotto da un apprendente di livello B1 nella composizione scritta per la raccolta di dati *b*:

(17)
Gioco con amici è una cosa felice

Un altro dato interessante riguarda la tipologia di errore che coinvolge la posizione del modificatore rispetto al nome, esemplificato da (18). Questo errore, verosimilmente dovuto a fenomeni di transfer (Banfi 2003), la cui analisi non rientra negli obiettivi di questo articolo, si verifica tuttavia nel campione solo nelle combinazioni libere, e copre il 14% dell'intera categoria degli errori grammaticali.

(18)
Infine, ho tronato gli spagnoli ragazzi sono non belli di Italiani ragazzi

Le unità fraseologiche, invece, sono associazioni di parole coese e con un grado più elevato di cristallizzazione, nelle quali la posizione rispettiva dei due costituenti lessicali genera più difficilmente errori, in quanto esse tendono ad essere acquisite e memorizzate come un unico blocco lessicale (Sinclair 1991). In esse, gli errori grammaticali sono invece per la maggior parte errori di accordo (63%), o di omissione, aggiunta non dovuta e scelta dell'articolo

(26%; anche gli errori nell'uso dell'articolo sono verosimilmente dovuti a fenomeni di transfer dalla L1; cfr. Wang 2016). (19) è un esempio di omissione dell'articolo, che non è stato incorporato nella preposizione *in*, riferito ad una unità fraseologica che è invece utilizzate in maniera corretta dal punto di vista lessicale e semantico, nonché nella posizione rispettiva dei suoi due componenti.

(19)
Il fine settimana, faccio spesso la spesa in centro commerciale

Per la quantità di errori commessi dagli apprendenti, dunque, e per la distribuzione dei tipi di tali errori, le unità fraseologiche differiscono in modo significativo dalle combinazioni libere: la forza di associazione che le caratterizza le rende infatti degli elementi lessicali utilizzati come blocchi unici, delle "islands of reliability" (Dechert 1983:184) che si prestano di meno a produrre errori. Gli apprendenti, infatti, nell'utilizzarli, si servono del principio idiomatico e devono confrontarsi con un carico minore di regole di combinazione da applicare (Siyanova-Chanturia 2015b, Siyanova-Chanturia, Martinez 2015).

È interessante, inoltre, esaminare se e in che modo il livello di competenza degli apprendenti incide a sua volta sugli errori commessi nelle due diverse tipologie di combinazione. Dai dati emerge che non c'è una differenza significativa tra gli errori commessi rispettivamente nella produzione di combinazioni libere e unità fraseologiche NADJ da apprendenti di livello A1, A2 e B1: le percentuali di combinazioni corrette prodotte nei tre livelli restano praticamente invariate (unità fraseologiche: A1 - 81%; A2 - 84%; B1 – 84%; combinazioni libere: A1 - 51%; A2 - 53%; B1 - 53%). Il livello di competenza, dunque, non sembra incidere sull'effetto che le due diverse tipologie di combinazioni hanno sulla quantità e sul tipo di errori.

L'ultima fase dell'analisi dei dati è consistita, infine, nel verificare se la variabile tempo influenza la quantità e il tipo di errori commessi nei due tipi di combinazioni (combinazioni libere e unità fraseologiche) della categoria NADJ. Sia il confronto tra la quantità globale di errori, di tutte le tipologie previste dallo schema di annotazione, sia quello tra errori grammaticali ed errori lessicali, non ha prodotto risultati significativi. Il tempo, dunque, non ha effetti rilevanti sul modo in cui le due tipologie di combinazioni condizionano il numero e il tipo di errori commessi dagli apprendenti.

6. Conclusioni

Questo studio preliminare ha indagato l'uso di combinazioni lessicali previste dalle relazioni sintattiche "aggettivo modificatore di nome", (NADJ e ADJN), e "verbo + oggetto diretto" (VN). L'analisi ha utilizzato un campione del corpus longitudinale LOCCLI, in cui 60 composizioni di apprendenti cinesi di livello A1, A2 e B1 (30 per ognuno dei due punti di raccolta *a* e *b*, e 10 per ognuno dei tre livelli di competenza) sono state etichettate manualmente sulla base di uno schema di annotazione creato ad hoc per l'analisi degli errori nella fraseologia.

Una prima analisi è stata condotta su tutte le combinazioni lessicali delle tre categorie selezionate, ed ha evidenziato che la categoria che sembra più critica per gli apprendenti cinesi è quella ADJN, che è al tempo stesso la meno utilizzata e quella in cui si verifica il maggior numero di errori. Inoltre, la frequenza di tutte e tre le categorie di combinazioni diminuisce dal primo al secondo punto di raccolta: dopo i primi sei mesi di permanenza e di studio in Italia, gli apprendenti fanno ricorso a costruzioni in parte diverse da quelle analizzate.

In secondo luogo, sono stati analizzati gli errori commessi dagli apprendenti, sia in modo aggregato che suddivisi per tipi di errore, sulla base dei parametri del livello di competenza e

del punto di raccolta longitudinale. Per quanto riguarda i due sottoinsiemi di errori considerati, risultano significativamente più numerosi quelli grammaticali; questa differenza, tuttavia, non si evolve in modo significativo nell'arco di tempo di sei mesi considerato. La categoria di combinazioni che risulta influenzata nell'occorrenza di errori dai due parametri del tempo e del livello di competenza è quella delle NADJ: nel tempo, infatti gli errori in questa categoria aumentano in modo significativo, e gli apprendenti interessati da questo incremento sono in particolare quelli di livello B1.

Infine, un'analisi più specifica è stata condotta sulla categoria di combinazioni che più delle altre si è dimostrata sensibile alla variabile tempo: quella delle combinazioni NADJ. In questo caso, è stato introdotto un filtro a due livelli sulle combinazioni analizzate, per separare quelle più convenzionali e strettamente associate (le unità fraseologiche) dalle altre (combinazioni libere). I due livelli selezionati come soglia sono i valori di frequenza ≥ 6 e di MI ≥ 3. L'analisi dei dati ha evidenziato una diminuzione significativa nella raccolta *b* delle unità fraseologiche NADJ rispetto alle combinazioni libere: dopo sei mesi, gli apprendenti sperimentano nuovi contesti in cui usare produttivamente nuove combinazioni a partire da quelle che hanno già appreso, e si trovano di conseguenza ad utilizzare combinazioni meno convenzionali e cristallizzate, e dunque meno strettamente associate. L'analisi degli errori nella categoria NADJ ha inoltre evidenziato che essi ricorrono significativamente molto di più nelle combinazioni libere, e in misura maggiore sotto forma di errori lessicali o di posizione dell'aggettivo. Il grado di cristallizzazione e la forza di associazione delle combinazioni NADJ, dunque, incidono di più sugli errori commessi rispetto al livello di competenza degli apprendenti e al tempo che hanno trascorso in Italia.

Tutti i fenomeni analizzati concorrono a fare luce sul processo di acquisizione, da parte di apprendenti cinesi di livello elementare e pre-intermedio, di tre categorie di combinazioni lessicali. Tale processo, documentato dall'uso in un corpus di produzioni scritte, è influenzato da diversi fattori, che lo rendono sicuramente complesso, articolato e tutt'altro che lineare. Il percorso di apprendimento è spesso scandito da progressi che si dispongono a vari stadi di competenza (che coincidono, nei tipici contesti educativi, con avanzamenti di livello) e lungo la linea del tempo, con una correlazione positiva tra tempo trascorso nello studio in contesti formali e aumento della competenza nella lingua studiata (Meunier 2015). Questo schema del percorso di apprendimento, seppure non in termini di linearità perfetta, è stato verificato ad esempio in campo morfosintattico (Bettoni, Di Biase 2005; Giacalone Ramat 2011; Vyatkina 2013). L'acquisizione delle combinazioni lessicali, invece, sembra non seguire uno sviluppo così lineare (Bartning, Forsberg 2006).

L'analisi degli errori nelle combinazioni lessicali svolta in questo studio contribuisce ad avvalorare questa tesi: lo sviluppo della competenza fraseologica (Paquot 2018) procede in modo molto meno regolare e rettilineo, perché è più sensibile alla riutilizzazione produttiva dell'input a cui gli apprendenti sono esposti e, di conseguenza, al loro stile comunicativo individuale (Meunier 2015). All'interno del percorso di apprendimento, il periodo di sei mesi considerato sembra costituire, per le combinazioni lessicali, una fase di sperimentazione, in cui gli apprendenti di tutti e tre i livelli si sforzano di utilizzare produttivamente le combinazioni già apprese, creando a partire da esse nuove combinazioni, e commettendo in questo tentativo un numero maggiore di errori (Laufer, Waldman 2011).

L'indicazione forse più interessante deriva tuttavia dall'analisi degli errori commessi nelle combinazioni NADJ: il fatto che gli errori siano legati soprattutto all'uso di combinazioni libere da un lato avvalora quanto appena affermato, e dall'altro suggerisce che le unità fraseologiche strettamente associate, una volta apprese, vengono usate in modo efficace e producono meno errori. Da ciò deriva anche un'indicazione chiara per la concreta pratica didattica: l'insegnamento formale della fraseologia andrebbe senza dubbio favorito, attraverso la

creazione di sillabi specifici, attività e materiali didattici espressamente progettati, e strumenti di valutazione dedicati alla verifica della competenza fraseologica (Paquot 2018). Un apparato didattico di questo tipo, funzionale all'apprendimento delle unità fraseologiche, non sembra ancor utilizzato in modo sistematico per la didattica dell'italiano L2.

Questo studio, infine, ha certamente delle limitazioni, in particolare nell'ampiezza del campione considerato: l'annotazione degli errori è una procedura manuale e complessa, e per questo richiede tempo. I dati annotati vanno sicuramente ampliati a tutto il LOCCLI, e la ripetizione delle analisi già svolte in questa sede con un campione più vasto potrà convalidare i risultati ottenuti. La numerosità e differenziazione delle variabili coinvolte, inoltre, insieme al fatto che i dati analizzati sono prodotti per due volte dallo stesso campione di apprendenti ad uno stesso intervallo di tempo, suggerisce di utilizzare metodologie più avanzate e più adatte a questo tipo di dati, come ad esempio alcune tecniche statistiche multifattoriali (Möller 2017), in grado di fornire inferenze sugli effetti di predittori diversi su una stessa variabile dipendente, come gli errori commessi dagli apprendenti.

La dimensione fraseologica, che è ubiqua e pervade ogni altro aspetto dell'analisi linguistica (Meunier 2012), riveste un'importanza centrale anche nel campo dell'apprendimento. La ricerca attraverso *learner corpora* sull'acquisizione della fraseologia può contribuire in modo rilevante ad evidenziare aree critiche e a fornire un supporto prezioso per identificare i tratti fondamentali che differenziano la produzione degli apprendenti da ciò che è considerato la norma cui tali produzioni devono tendere (Granger 2011).

RIFERIMENTI BIBLIOGRAFICI

Bagna, C., Machetti, S. (2008). *Le polirematiche nel continuum di competenza nativo/non nativo: povertà, vaghezza, creatività linguistica*, in M. Barni, D. Troncarelli, C. Bagna (eds.), *Lessico e apprendimenti. Il ruolo del lessico nella linguistica educativa*, Milano, FrancoAngeli, pp. 87-98.

Banfi, E. (2003), *Italiano L2 di cinesi: percorsi acquisizionali*, Milano, Franco Angeli.

Banfi, E., Piccinini, C., Arcodia G.F. (2008). *Quando mancano le parole: strategie di compensazione lessicale di sinofoni apprendenti Italiano L2*, in M. Barni, D. Troncarelli, C. Bagna (eds.), *Lessico e apprendimenti. Il ruolo del lessico nella linguistica educativa*, Milano, FrancoAngeli, pp. 247-259.

Bartning, I., Forsberg, F. (2006), *Les séquences préfabriquées à travers les stades de développement en français L2*, in *Actes du 16e congrès des romanistes scandinaves*, Department of Language and Culture, Roskilde University.

Bestgen, Y., Granger, S. (2014), *Quantifying the development of phraseological competence in L2 English writing: An automated approach*, in "Journal of Second Language Writing", 26, pp. 28–41.

Bettoni, C., Di Biase, B. (2005), *Sviluppo obbligato e progresso morfosintattico*, in "ITALS", 2(1), pp. 27-48.

Bley-Vroman, R. (1983), *The comparative fallacy in interlanguage studies: The case of systematicity*, in "Language Learning", 33, 1, pp. 1–17.

Callies, M. (2015). *Learner corpus methodology*, in S. Granger, G. Gilquin, F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press, pp. 35-56.

Chini, M. (2005), *Che cos'è la linguistica aquisizionale*, Roma, Carocci.

Church, K.W., Hanks, P., (1990), *Word association norms, mutual information, and lexicography*, in "Computational linguistics", 16, 1, pp. 22-29

Corder, S.P. (1967), *The significance of learner's errors*, in "International Review of Applied Linguistics in Language Teaching", 5(1–4), pp. 161–70.

Dagneaux, E., Denness, S., Granger, S. (1998), *Computer-aided error analysis*, in "System", 26, 2, pp. 163-74.

Dechert, H. (1983). *How a Story is Done in a Second Language*, in C. Faerch, G. Kasper, (eds), *Strategies in Interlanguage Communication*, London, Longman, pp. 175-195.

Díaz-Negrillo, A., Fernández-Domínguez, J. (2006), *Error tagging systems for learner corpora*, in "Revista Española de Lingüística Aplicada", 19, pp. 83–102.

Durrant, P., Schmitt, N. (2009), *To what extent do native and non-native writers make use of collocations?*, in "International Review of Applied Linguistics", 47, 2, pp. 157-177.

Durrant, P., Siyanova-Chanturia, A. (2015). *Learner corpora and psycholinguistics*, in S. Granger, G. Gilquin, F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press, pp. 57-78.

Ebeling, S., Hasselgård, H. (2015). *Learner corpora and phraseology*, in S. Granger, G. Gilquin, F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press, pp. 207-230.

Ellis, N.C., Simpson-Vlach, R.C. (2009). *Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education*, in G. Gilquin (ed.), "Corpora and Experimental Methods. Special issue of Corpus Linguistics and Linguistic Theory", 5, 1, pp. 61–78.

Ellis, N.C., Simpson-Vlach, R.C., Maynard, C. (2008), *Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL*, In "TESOL Quarterly", 42, 3, pp. 375–96.

Ellis, N., Simpson-Vlach, R., Römer, U., O'Donnell, M., Wulff, S. (2015). *Learner corpora and formulaic language in SLA*, in S. Granger, G. Gilquin, F. Meunier (eds), *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press, pp. 357-378).

Evert, S., (2005), *The Statistics of Word Cooccurrences: Word Pairs and Collocations*, Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Giacalone Ramat, A. (2011), *Il Quadro teorico*, in A. Giacalone Ramat (ed.), *Verso l'italiano. Percorsi e strategie di acquisizione*, Roma, Carocci, pp. 17-26.

Granger, S. (2008). *Learner Corpora in Foreign Language Education*, in N. Van Deusen-Scholl, N.H. Hornberger (eds), *Encyclopedia of Language and Education*. Volume 4. *Second and Foreign Language Education*, New York, Springer, pp. 337-351.

Granger, S. (2011). *From phraseology to pedagogy: Challenges and prospects*, in T. Herbst, S. Faulhaber, P. Uhrig, (eds.), *The Phraseological View of Language. A Tribute to John Sinclair*, Berlin, Mouton de Gruyter, pp. 123–46.

Granger, S., Gilquin, G., Meunier, F. (2015). *Introduction: Learner corpus research – past, present and future*, in S. Granger, G. Gilquin, F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press, pp. 1-6.

Granger, S., Meunier, F. (eds.) (2008), *Phraseology: an interdisciplinary perspective*, Amsterdam, John Benjamins.

Granger, S., Paquot, M. (2008). *Disentangling the phraseological web*, in S. Granger, F. Meunier (eds.), *Phraseology. An interdisciplinary perspective*, Amsterdam, John Benjamins, pp. 27-49.

Gries, S. Th. (2008). *Phraseology and linguistic theory: A brief survey*, in S. Granger, F. Meunier (eds.), *Phraseology. An interdisciplinary perspective*, Amsterdam, John Benjamins, pp. 3-25.

Gries, S. Th. (2015). *Statistical methods in learner corpus research*, in S. Granger, G. Gilquin, F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press, pp. 159-181.

Gries, S. Th., Mukherjee, J. (2010), *Lexical gravity across varieties of English: an ICE-based study of n-grams in Asian Englishes*, in "International Journal of Corpus Linguistics", 15, 4, pp. 520-548.

Gries, S. Th. (2013), *50-something years of work on collocations: what is or should be next…*, in "International Journal of Corpus Linguistics", 18,1, pp. 137-165.

Gilquin, G. (2015). *From design to collection of learner corpora*, in S. Granger, G. Gilquin, F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press, pp. 9-34.

James C. (1998), *Errors in Language Learning and Use*. London, Longman.

Konecny, C., Autelli, E., Abel, A., Zanasi, L. (2016). *Identification, Classification and Analysis of Phrasemes in an L2 Learner Corpus of Italian*, in *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, Tradulex, Geneva, pp. 533-542.

Laufer, B., Waldman, T. (2011), *Verb–noun collocations in second language writing: A corpus analysis of learners' English*, in "Language Learning", 61, 2, pp. 647-72.

Lennon, P. (1991), *Error: Some problems of definition, identification, and distinction*, in "Applied Linguistics", 12(2), pp. 180–96.

Lüdeling, A., Hirschmann, H. (2015). *Error annotation systems*, in S. Granger, G. Gilquin, F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press, pp. 135-158.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A., Pirrelli, V. (2014). *The PAISÀ Corpus of Italian Web Texts*, in *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, Gothenburg, Association for Computational Linguistics, pp. 36-43.

Masini, F. (2012), *Parole sintagmatiche in italiano*, Cesena, Caissa.

Meunier, F. (2012), *Formulaic language and language teaching*, "Annual Review of Applied Linguistics", 32, pp. 111–29.

Meunier, F. (2015). *Developmental patterns in learner corpora*, in S. Granger, G. Gilquin, F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press, pp. 379-400.

Möller, V. (2017). *A statistical analysis of learner corpus data, experimental data and individual differences: Monofactorial vs. multifactorial approaches*, in P. de Haan, S. van Vuuren, R. de Vries (eds), *Language, Learners and Levels: Progression and Variation. Corpora and Language in Use*, Louvain-la-Neuve, Presses universitaires de Louvain, pp. 409-439.

Mulhall, C. (2017). *Collocation Patterning and Italian High Frequency Verbs: Corpus Evidence of L1 English Speakers*, in E. Corino, C. Onesti (eds), *Italiano di apprendenti. Studi a partire da VALICO e VINCA*, Perugia, Guerra, pp. 89-101.

Nesselhauf, N. (2003), *The Use of Collocations by Advanced Learners of English and Some Implications for Teaching*, in "Applied Linguistics", 24, 2, pp. 223-242.

Nesselhauf, N. (2005), *Collocations in a Learner Corpus*, Amsterdam, John Benjamins.

Osborne, J. (2008). *Phraseology effects as a trigger for errors in L2 English: The case of more advanced learners*, in F. Meunier, S. Granger (eds), *Phraseology in Foreign Language Learning and Teaching*, Amsterdam, John Benjamins, pp. 67–83.

Paquot, M. (2018), *Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations*, "Language Assessment Quarterly", pp. 1-15.

Pawley, A., Syder, F.H. (1983). *Two puzzles for linguistic theory: Nativelike selection and nativelike fluency*, in J.C. Richards, R.W. Schmidt (eds.), *Language and Communication*, London, Longman, pp. 191–225.

Sinclair, J. McH. (1991), *Corpus, Concordance, Collocation*. Oxford , Oxford University Press.

Siyanova-Chanturia, A. (2015a), *Collocation in beginner learner writing: A longitudinal study*, in "System", 53, pp. 148-160.

Siyanova-Chanturia, A. (2015b), *On the 'holistic' nature of formulaic language*, in "Corpus Linguistics and Linguistic Theory", 11, 2, pp. 285–301.

Siyanova-Chanturia, A., Martinez, R. (2015), *The Idiom Principle Revisited*, in "Applied Linguistics", 36, 5, pp. 549–569.

Siyanova-Chanturia, A., Schmitt, N. (2008). *L2 learner production and process- ing of collocation: A multi-study perspective*, in "The Canadian Modern Language Review / La Revue canadienne des langues vivantes", 64(3): 429–58.

Siyanova-Chanturia, A., Spina, S. (2015), *Investigation of Native Speaker and Second Language Learner Intuition of Collocation Frequency*, in "Language Learning", 65, 3, pp. 533–562.

Siyanova-Chanturia, A., Spina, S. (in preparazione), *Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study*.

Spina, S. (2015). *Phraseology in Academic L2 Discourse: The Use of Multi-words Units in a CMC University Context*, in E. Castello, K. Ackerley, F. Coccetta (eds), *Studies in Learner Corpus Linguistics. Research and Applications for Foreign Language Teaching and Assessment*, Bern, Peter Lang, pp. 279-294.

Spina, S. (2010a). *The Dictionary of Italian Collocations: Design and Integration in an Online Learning Environment*, in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, D. Tapias (eds), *Proceedings of the Seventh conference on International Language*

Resources and Evaluation (LREC'10), Malta, European Language Resources Association, pp. 3202-3208.

Spina, S. (2010b). *The Dici Project: towards a Dictionary of Italian Collocations integrated with an online language learning platform*, in S. Granger, M. Paquot (eds), *eLexicography in the 21st century: New Challenges, New Applications*, Louvain-la-Neuve, Presses Universitaires de Louvain, pp. 273-282.

Spina, S. (2014). *Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione*, in R. Basili, A. Lenci, B. Magnini (eds), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it (Vol. 1)*, Pisa, Pisa University Press, pp. 354–359.

Spina S. (2016). *Learner corpus research and phraseology in Italian as a second language: The case of the DICI-A, a learner dictionary of Italian collocations*, in Begoña Sanromán Vilas (a cura di), *Collocations Cross- Linguistically. Corpora, Dictionaries and Language Teaching*, Mémoires de la Société Néophilologique de Helsinki, tome C, pp. 219-244

Spina, S. (2017). *Learner Corpus Research and the acquisition of Italian as a second language: the case of the Longitudinal Corpus of Chinese Learners of Italian (LoCCLI)*, presentato alla 4th Learner Corpus Research Conference, 5-7 ottobre, Bolzano, Eurac Research.

Thewissen, J. (2008). *The phraseological errors of French-, German- and Spanish-speaking EFL learners: Evidence from an error-tagged learner corpus*, in *Proceedings of the 8th Teaching and Language Corpora Conference (TaLC8), 3–6 July 2008*, Lisbon, Associação de Estudos e de Investigação Científica do ISLA, pp. 300–306.

Stenetorp, P., Pyysalo, S., Topić G., Ohta, T., Ananiadou, S., Tsujii, J. (2012), *brat: a Web-based Tool for NLP-Assisted Text Annotation*, Proceedings of the Demonstrations Session at EACL 2012, Avignon, Association for Computational Linguistics.

Vyatkina, N. (2013), *Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus*, in "The Modern Language Journal", 97, 1, pp. 11-30.

Wang, Y. (2016). *The Idiom Principle and L1 Influence. A contrastive learner-corpus study of delexical verb+noun collocations*, Amsterdam, John Benjamins.

Wanner, L., Ramos, M.A., Vincze, L., Nazar, R., Ferraro, G., Mosqueira, E., Prieto, S. (2013). *Annotation of collocations in a learner corpus for building a learning environment*, in S. Granger, G. Gilquin, F. Meunier (eds.), *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead*, Louvain-la-Neuve, Presses universitaires de Louvain, pp. 493–503.

STEFANIA SPINA • is Associate Professor at the Department of Human and Social Sciences, Università per Stranieri di Perugia. She does research in Corpus Linguistics, Learner Corpus Research, CMC, Discourse Analysis and Computational Linguistics. One of her current project is "Readability of Texts for L2 learners."

E-MAIL • stefania.spina@unistrapg.it