Univerzita Karlova v Praze

Filozofická fakulta

Ústav románských studií

Filologie – Románské jazyky

Leontýna Bratánková

Le collocazioni Verbo + Nome in apprendenti di italiano L2

vedoucí práce doc. Pavel Štichauer, Ph.D.
doc. Stefania Spina

2015

Il lavoro ha preso in analisi da un punto di vista integrato, quantitativo e linguistico, le collocazioni V+N_{Ogg} prodotte dagli apprendenti di italiano L2 a livello intermedio-avanzato con lo scopo di descrivere la competenza collocazionale degli informanti, valutandola in termini quantitativi rispetto all'uso dei nativi ed individuando le eventuali variabili alle quali questa potrebbe essere soggetta.

Lo studio è stato condotto a partire da un *corpus* di apprendenti di italiano L2 (CAIL2) realizzato *ad hoc*, il quale ne costituisce quindi la base empirica. Il *corpus* CAIL2 di apprendenti di italiano L2, descritto dettagliatamente nel capitolo quarto, ha un'ampiezza di 237 000 *tokens* ed è costituito dalle produzioni scritte di 400 informanti per un totale di 33 L1 di provenienza.

Le combinazioni verbo-nominali sono state estratte in maniera automatica dal *corpus* e filtrate in base ai criteri di frequenza e di associazione lessicale; ciò ha permesso di individuare le collocazioni empiriche, ovvero tutte quelle combinazioni di parole i cui componenti si sono dimostrati co-occorrere molto spesso (per il presente studio la soglia di frequenza è stata fissta a 10) e che sono risultati essere strettamente associati tra di loro (*Mutual Information* ≥ 3 e *t-score* ≥ 2). Le collocazioni empiriche (Evert 2009) sono quindi le collocazioni intese in base alla definizione frequentista datane da J.R. Firth (1957) e ampliata successivamente dalla scuola neo-firthiana, ovvero le combinazioni di parole che ricorrono insieme più spesso di quanto ci si potrebbe aspettare valutando le frequenze individuali dei loro componenti in un dato *corpus* (Jones, Sinclair 1974).

A partire dalla formulazione del principio idiomatico di elaborazione e comprensione del linguaggio (*idiom principle*, Sinclair 1991), in base al quale il parlante opera preferibilmente e primariamente la co-selezione di due o più parole, considerando la loro precedente e regolare co-occorrenza, piuttosto che selezionarle singolarmente, sulla base di una scelta aperta (*open choice principle*), le collocazioni frequenti sono entrate nel raggio di interesse dalla psicolinguistica (Ellis 2002; Wray 2002; Hoey 2005). Ne è conseguito che le sequenze formulaiche in generale (costruzioni, *clusters*, *n-grams*, collocazioni, ecc.) sono diventate il centro delle teorie *usage-based* di acquisizione di L1 e L2 (Bybee 1998; Wray 2002; Ellis 2003; Tomasello 2003; Goldberg 2006), secondo le quali la frequenza di occorrenza di determinate strutture nell'*input* sarebbe fondamentale per la definizione dei meccanismi di elaborazione linguistica, e degli approcci connessionisti dell'acquisizione e dell'elaborazione linguistica (Rumelhart, McClelland 1986; Elman 1990; Christiansen, Chater 1999) per i quali la frequenza determina cosa e quanto i parlanti apprendono ed eventualmente rappresentano nel loro lessico mentale (Conklin, Schmitt 2012).

Uno dei vantaggi più rilevanti di cui è portatore il linguaggio formulaico nell'ambito dell'acquisizione di lingue seconde, è il fatto di agevolare i parlanti nell'essere più fluenti e naturali nella produzione linguistica (Götz 2013). Un'idea questa che è stata delineata per la prima volta da Pawley e Syder (1983) e convalidata dall'evidenza empirica di Dechert (1983), il quale ha trovato che l'*output* parlato di un'apprendente tedesca di inglese fosse più fluente e naturale quando questa faceva ricorso al linguaggio formulaico; Dechert per primo usò la definizione "isole di affidabilità" riferendosi alle sequenze formulaiche e suggerì con tale termine quanto queste fossero necessarie per una elaborazione efficace del linguaggio in tempo reale.

Individuare induttivamente le sequenze formulaiche a partire da un *corpus* sulla base della frequenza e dell'associazione reciproca tra gli elementi nell'ambito della ricerca sulle lingue seconde ha il vantaggio di far emergere le combinazioni che risultano prefabbricate per il parlante o lo scrivente ma che potrebbero anche non essere conformi all'uso nativo.

Nell'analisi quantitativa sulla lingua degli apprendenti condotta nel capitolo quinto del presente lavoro abbiamo stimato la formulaicità valutando la portata delle collocazioni empiriche nel *corpus* CAIL2; i risultati sono stati interpretati alla luce dell'uso nativo per mezzo di un *corpus* di controllo il quale ha fornito i parametri indispensabili per la valutazione dei dati degli apprendenti di italiano a nostra disposizione; si tratta del *corpus* SCUOLA costituito dalle produzioni scritte dei ragazzi delle scuole medie inferiori e superiori. Il *corpus* SCUOLA è una sezione del *Perugia corpus* (PEC), un *corpus* di riferimento dell'italiano contemporaneo dell'ampiezza di oltre 26 milioni di parole elaborato all'Università per Stranieri di Perugia.

Da una prima valutazione quantitativa dei due *corpora* relativa all'incidenza nei testi delle collocazioni empiriche (le combinazioni verbo-nominali con frequenza ≥ 10, *Mutual Information* ≥ 3 e *t-score* ≥ 2), e quindi alla formulaicità che li caratterizza, è emerso che le produzioni scritte degli apprendenti presentano una quantità maggiore in termini di *tokens* di combinazioni lessicali frequenti e strettamente associate rispetto a quelle dei nativi; tuttavia, valutando gli indici di varietà lessicale, abbiamo osservato che le collocazioni empiriche prodotte dagli apprendenti risultano meno varie rispetto a quelle prodotte dai nativi. Ne consegue che le produzioni del *corpus* CAIL2 sono più formulaiche, ovvero più ricche di collocazioni empiriche, ma anche più conservative rispetto a quelle native in quanto vi è stato usato più spesso un gruppo più ristretto, meno vario, di collocazioni empiriche.

Con l'obiettivo di valutare le differenze riscontrate nelle produzioni dei nativi e degli apprendenti in maniera più dettagliata ed individuare quale fosse il gruppo di collocazioni più saliente per gli apprendenti, abbiamo classificato le collocazioni empiriche estratte dai due *corpora* in base alla loro forza collocazionale; per questo motivo le abbiamo disposte su una scala decrescente relativa ai punteggi delle due misure di associazione lessicale utilizzate (associazione forte, moderata e debole).

Le produzioni dei nativi del *corpus* SCUOLA si sono distinte, rispetto a quelle degli apprendenti, sia per la ricchezza delle collocazioni caratterizzate da alti punteggi di *Mutual Information*, compresi tra i 14 e i 7 punti, (+ 11,5% di *types*; + 4,4% di *tokens*) che per varietà, quest'ultima stimata sulla base del confronto tra gli indici di varietà lessicale; la presenza maggiore di collocazioni caratterizzate dai valori alti di MI si è dimostrata essere un tratto peculiare che caratterizza lo scritto dei nativi rispetto a quello degli apprendenti.

Gli apprendenti hanno prodotto una quantità superiore di collocazioni con alti valori di *t-score* (+ 7,4% di *types*; + 14% di *tokens*) rispetto ai nativi, ma l'indice di varietà è risultato essere più basso. Ciò vuol dire che gli apprendenti ricorrono più spesso rispetto a quanto facciano i nativi alle collocazioni caratterizzate dagli stessi indici quantitativi di *t-score* ma, essendo questo gruppo meno vario, ne consegue che usano ripetutamente un gruppo di collocazioni più esiguo.

Proprio nel gruppo di collocazioni con i punteggi più alti di *t-score* (17 > *t-score* ≥ 5) abbiamo identificato quelle che negli studi sulla lingua degli apprendenti di inglese (Granger 1998; Lorenz 1999; Kaszubski 2000; Laufer, Waldman 2011) sono state definite "isole di affidabilità" (*islands of reliability*), ovvero gli usi ripetuti e quantitativamente superiori (*oversuse*) rispetto ai nativi di un piccolo repertorio di collocazioni, le quali farebbero sentire gli appredenti più sicuri nella fase di *output*.

I nostri risultati hanno confermato per la lingua italiana quanto era già emerso per la lingua inglese (Lorenz 1999; Durrant 2008; Durrant, Schmitt 2009), ovvero che i nativi fanno ampio uso delle combinazioni lessicali fortemente associate (con punteggi alti di *Mutual Information*) rispetto agli apprendenti, mentre questi ultimi fanno ricorso alle combinazioni molto frequenti, con valori alti di *t-score*. Anche lo studio psicolinguistico di Ellis *et al.* (2008) ha sottolineato l'importanza delle combinazioni molto frequenti, con punteggi alti di *t-score*, per gli apprendenti e la salienza delle combinazioni con punteggi alti di *Mutual Information* per i nativi: proprio la presenza elevata delle combinazioni con quest'ultima caratteristica conferirebbe alle produzioni dei nativi quel tratto di idiomaticità tipica dei madrelingua e, al contrario, la loro presenza moderata nelle produzioni degli apprendenti sarebbe il motivo per cui queste sembrano perdere in naturalezza.

Dopo aver messo a confronto le produzioni dei nativi e degli apprendenti sul terreno della formulaicità ed aver individuato i tratti quantitativi caratterizzanti ciascuna varietà, abbiamo focalizzato l'attenzione sul *corpus* CAIL2 anche da un altro punto di vista.

Abbiamo studiato le produzioni scritte degli apprendenti in base a due variabili sociolinguistiche: 1) il tempo di studio dell'italiano; 2) il tempo di permanenza in Italia (l'esposizione all'*input*). A tal proposito sono stati individuati quattro gruppi di apprendenti per la prima variabile (gruppo A = 1-6 mesi di studio dell'italiano; gruppo B = 7-11 mesi; gruppo C = 12-32 mesi; gruppo D = ≥ 36 mesi) e cinque gruppi per la seconda (gruppo E = 0-1 mese di permanenza in Italia; gruppo F = 2-3 mesi; gruppo G = 4 mesi; gruppo H = 5 mesi; gruppo I = ≥ 6 mesi).

Siamo quindi andati a valutare la portata delle collocazioni empiriche nelle produzioni scritte dei singoli gruppi e le abbiamo messe a confronto per valutare eventuali differenze.

Le collocazioni empiriche, infatti, proprio in ragione del fatto che sono le combinazioni più frequenti e più strettamente associate, rappresentano le collocazioni tipiche di una data varietà linguistica o di un dato campione di linguaggio: individuarne la presenza e metterne a confronto la quantità in due *corpora* o in diverse sezioni dello stesso *corpus* (come nel caso dei gruppi individuati in base alle due variabili del presente studio) permette di valutarne l'incidenza sul testo e di formulare delle ipotesi sulle sue caratteristiche.

I risultati dell'analisi quantitativa relativi alla prima variabile hanno messo in evidenza che il gruppo D ha prodotto il numero minore (in termini di *tokens*) di collocazioni empiriche rispetto agli altri gruppi. Ciò vuol dire che gli apprendenti di lingua italiana da più di 36 mesi hanno fatto ricorso nei loro testi ad una quantità minore di collocazioni tipiche, cioè diffuse trasversalmente nelle produzioni degli apprendenti a tutti i livelli di competenza.

A partire da questa evidenza abbiamo formulato l'ipotesi che il gruppo D potrebbe aver redatto delle combinazioni Verbo + Nome diverse dalle collocazioni tipiche e

che si potrebbe assistere ad un salto qualitativo nei testi degli apprendenti, relativamente alle combinatorie verbo-nominali, soltanto dopo il terzo anno di studio della lingua.
Abbiamo verificato questa ipotesi con l'analisi linguistica presentata nel capitolo sesto.

Per quanto riguarda la seconda variabile, invece, non si è assistito ad una diminuzione delle collocazioni più comunemente utilizzate dagli apprendenti all'aumentare dei mesi trascorsi in Italia. Al contrario, l'andamento nella produzione delle collocazioni tipiche dell'interlingua si è dimostrato altalenante con un comportamento analogo dei due gruppi agli estremi della variabile (gruppo E: 0-1 mese di permanenza in Italia; gruppo I: più di 6 mesi).
È importante sottolineare che, a differenza di quanto emerso per la prima variabile, le osservazioni relative allo studio sulla seconda variabile sono valide soltanto per il campione di dati costituito dal *corpus* CAIL2 e non possono essere estese alla popolazione da esso rappresentata, in quanto le differenze individuate nelle produzioni delle collocazioni da parte dei diversi gruppi di apprendenti non sono risultate essere statisticamente significative.
Nonostante l'andamento quantitativamente altalenante nella produzione delle collocazioni tipiche dell'interlingua è risultato possibile individuare una costante. Infatti, i due gruppi E ed I, pur avendo avuto esposizioni all'*input* della lingua *target* molto diverse, sono assimilabili per il tempo di studio della lingua italiana; inoltre, il gruppo F (2-3 mesi di permanenza), il quale ha prodotto la quantità minore di collocazioni tipiche dell'interlingua, e il gruppo H (5 mesi di permanenza), il quale è ricorso al numero maggiore di queste, differiscono considerevolmente in relazione al tempo che gli informanti dei due gruppi hanno studiato l'italiano.

Se ne deduce che l'incidenza della variabile del tempo di studio dell'italiano sulla quantità delle collocazioni più diffuse prodotte dagli apprendenti del *corpus* CAIL2 è maggiore rispetto a quella del tempo di esposizione all'*input*; si assiste ad una loro diminuzione nei testi degli apprendenti all'aumentare del tempo di studio e, in particolare, ad una diminuzione considerevole dopo i tre anni di apprendimento. Abbiamo ipotizzato che ciò potrebbe denotare un'apertura verso delle combinazioni verbo-nominali diverse, meno diffuse a tutti i livelli di apprendimento, la cui presenza potrebbe agire dal punto di vista qualitativo sulle produzioni scritte degli apprendenti di lingua italiana. Una ipotesi simile può essere verificata soltanto con l'analisi linguistica che, oltre alla natura delle combinatorie prodotte, prenda in esame la loro regolarità e la loro pertinenza nel contesto.

L'analisi linguistica delle collocazioni verbo-nominali individuate in base ai criteri di frequenza e di associazione lessicale sopra esposti è stata condotta nel capitolo sesto. Nella prima parte del capitolo (parr. 6.1, 6.2, 6.3 e relativi sottoparagrafi), le collocazioni $V+N_{Ogg}$ estratte dai due *corpora* di apprendenti (CAIL2) e di nativi (SCUOLA) sono state analizzate alla luce dei criteri linguistici relativi alle collocazioni di lingua italiana individuati da Elisabetta Ježek e Francesca Masini; nella seconda parte (par. 6.4 e relativi sottoparagrafi), invece, le produzioni degli apprendenti sono state studiate da una prospettiva diversa, per mezzo dell'analisi delle concordanze, al fine di verificare le ipotesi formulate a partire dall'analisi quantitativa condotta nel capitolo quinto.

Conformemente all'approcio integrato, quantitativo e linguistico, assunto nel presente lavoro sulle collocazioni abbiamo adottato una definizione che integra le formulazioni elaborate da Ježek (2005) e Masini (2009) nell'ambito della classificazione delle combinazioni di parole della lingua italiana con i criteri di frequenza (fr. ≥ 10) e di associazione lessicale (*Mutual Information* ≥ 3 e *t-score* ≥ 2) tra i componenti delle combinazioni.

Alla luce di questa considerazione e seguendo Ježek (2005) abbiamo definito collocazioni tutte le combinazioni verbo-nominali che, rientrate nei valori soglia di frequenza e di associazione lessicale adottati nel presente lavoro, fossero caratterizzate dalla presenza di una restrizione attivata dal nome ed imposta al verbo, la cui semantica, dunque, si fosse specificata proprio nella sua co-occorrenza con l'Oggetto (per questo si confronti, ad esempio, il significato del verbo *porre* in *porre una domanda* > *chiedere* e in *porre fine* > *mettere*).

Nelle collocazioni V+N_{Ogg}, i verbi sono selezionati dai nomi per esprimere un significato che non hanno quando sono combinati con altre parole, ma che acquisiscono nella combinazione specifica (il *meaning by collocation* formulato da Firth, 1957).

Per questo motivo, Ježek si riferisce alle collocazioni come a delle solidarietà consolidate dall'uso, distinte dalle solidarietà semantiche, o basate su una implicazione sintagmatica di contenuto, nelle quali uno dei due termini è incluso dal punto di vista del contenuto nell'altro e che possono essere più o meno circoscritte in base alla semantica del verbo (se questo ammette più classi di oggetti, come, ad es., il verbo *comprare*, darà vita ad una combinazione dalla restrizione meno circoscritta rispetto ai verbi come *parcheggiare*, *allattare* o *indossare* i quali ammettono una sola classe di oggetti).

Masini (2009) ha ristretto ulteriormente il concetto definendo le collocazioni come co-selezioni tra due lessemi in cui l'uso di un determinato termine *x* implica necessariamente la presenza di un termine *y* per esprimere un determinato concetto (ad es., *aprire un conto*) e distinguendole dalle "combinazioni preferenziali" in cui un termine *x* (ad es., *pioggia*) può richiedere preferibilmente l'uso di un termine *y* (*torrenziale/ battente*) perché la combinazione risulti molto più familiare rispetto ad altre combinazioni possibili (*pioggia forte/ intensa*).

Il tratto della familiarità che contraddistingue le combinazioni preferenziali implica il concetto di uso e di frequenza, centrale anche nella formulazione di Ježek; per questo motivo e relativamente all'importanza che la frequenza riveste negli studi basati su *corpora*, abbiamo fatto rientrare nel concetto di collocazione sia le combinazioni in cui il Nome seleziona il Verbo perché lo richiede necessariamente per esprimere un determinato significato (ad es., *seguire un consiglio*), sia i casi in cui la selezione è preferenziale, ha un elemento di convenzionalità e rappresenta il modo tipico di dire una cosa per il campione linguistico studiato (come, ad es., *ripetere l'anno* vs. *rifare l'anno* per il *corpus* SCUOLA). Inoltre, sempre seguendo Ježek (2005), vi abbiamo fatto rientrare anche le costruzioni a verbo supporto (*Vsup*), intendendole come delle collocazioni sbilanciate verso il nome dal punto di vista semantico in quanto il verbo vi ha un significato generico (ad es., *fare un discorso*), e le costruzioni a verbo supporto esteso (*Vsupext*, Cicalese 1999) le quali possono sostituirsi al supporto neutro costituendone delle varianti di registro (*prendere un'infezione* > *contrarre un'infezione*) oppure apportando alcune

informazioni supplementari, delle sfumature di senso aggiuntive rispetto al supporto di grado zero (*fare un affare* > *concludere un affare*).

L'analisi linguistica di tutte le combinazioni verbo-nominali rientrate nei valori soglia di frequenza e di associazione lessicale, esposta nelle prima parte del capitolo sesto, ha riguardato sia la varietà nativa rappresentata dal *corpus* SCUOLA che l'interlingua del *corpus* CAIL2. Nello specifico, dallo studio delle produzioni scritte dei nativi di lingua italiana è emerso che:

1) il 61% delle combinazioni verbo-nominali è risultato essere costituito dalle collocazioni definite in base ai criteri linguistici sopra esposti; il restante 39%, invece, è composto dalle combinazioni ristrette in cui sussite una implicazione sintagmantica di contenuto tra il verbo e il nome, più o meno circoscritte in base alla semantica del verbo (ad es., *ascoltare una canzone, educare un figlio*), e dalle combinazioni libere (ad es., *prendere un libro*).
Da tali risultati abbiamo dedotto che i valori soglia di associazione lessicale utilizzati nel presente lavoro, già adottati negli studi sulle collocazioni basati sui *corpora* di lingua inglese, combinati con valori alti di frequenza, costituiscono dei parametri quantitativi accettabili per l'estrazione delle collocazioni dai *corpora* di lingua italiana;

2) il gruppo delle collocazioni è costituito per il 60% dalle costruzioni a verbo supporto (*Vsup*) e a verbo supporto esteso (*Vsupext*) e per il restante 40% dalle collocazioni formate con un verbo ordinario (*Vord*) il quale, a differenza del verbo coinvolto nelle costruzioni a *Vsup*, non è semanticamente vuoto (ad es., *fare danno*) e non rappresenta nemmeno l'estensione di un verbo a supporto base, come nel caso delle costruzioni a *Vsupext* (ad es., *porre un domanda*), ma è portatore di un significato che emerge nella specifica combinatoria con l'Oggetto (ad es., *prestare attenzione*).
Nel dettaglio delle due misure di associazione utilizzate, abbiamo potuto osservare che, conformemente al fatto che queste tendono ad enfatizzare gruppi diversi di combinazioni lessicali relativamente alla frequenza di occorrenza dei membri nel *corpus*, i valori alti di *t-score* hanno messo in evidenza soprattutto (per il 59%) le costruzioni a verbo supporto, mentre i valori alti di MI hanno fatto emergere prevalentemente (per l'80%) le costruzioni a verbo supporto esteso e le collocazioni con verbo ordinario.

Il gruppo delle collocazioni prodotte dai parlanti nativi è stato analizzato nel dettaglio nel corpo del capitolo sesto. Abbiamo potuto constatare che le costruzioni a verbo supporto (*Vsup*) realizzate sono risultate essere morfosintatticamente eterogenee: oscillano, infatti, tra le combinazioni più coese e parzialmente lessicalizzate (come ad es., *fare amicizia* e *dare voce*) e le combinazioni sintatticamente libere (quali, ad es., *fare una/ la doccia* e *dare un/ il consiglio*) passando per le collocazioni che possono avere una doppia natura in quanto ammettono sia la presenza che l'assenza dell'articolo, un fattore che può influenzarne la semantica (come ad es., *fare parte* e *fare la parte*).
Il paradigma che è stato realizzato più frequentemente dai nativi è *fare* + Nome Predicativo (ad es., *fare la/ una/ [] differenza, fare fronte*, *ecc.*), seguito da *fare* + Nome Deverbale (ad es., *fare l'/ un/ [] appello, fare una/ la/ [] finta, ecc.*); gli altri

verbi utilizzati sono *avere* e *dare* seguiti dai nomi predicativi (ad es., *avere il/ un/ [] bisogno, dare il/ un contributo, ecc.*) e deverbali (ad es., *avere il/ un/ [] rendimento, dare l'/ un'informazione*, *ecc.*); il verbo *dare* è stato utilizzato anche nelle costruzioni causative quali *dare la forza, dare fastidio, ecc.* I restanti verbi utilizzati per la costruzione delle *Vsup* sono *prendere, essere* e *mettere (*ad es., *prendere la/ una decisione, essere sinonimo, mettere fine, ecc*.).

Nel gruppo delle costruzioni a *Vsupext* prodotte dai nativi abbiamo potuto osservare che la maggior parte (56%) delle estensioni prodotte dai nativi sono varianti di registro: si tratta di combinazioni come, *assumere la responsabilità, attribuire la colpa, svolgere il compito* che entrano in rapporti parafrastici con le rispettive costruzioni a verbo supporto neutro *prendere la responsabilità, dare la colpa* e *fare il compito*. La restante parte dei *Vsupext* realizzati, al contrario, risulta essere portatrice di diverse valenze aspettuali che conferiscono alla collocazione un significato più specifico rispetto alle corrispondenti combinazioni con il verbo a supporto base. Le varianti aspettuali realizzate dai nativi per mezzo delle costruzioni a *Vsup* sono, in ordine di frequenza, l'incoativa (ad es., *creare il/ un problema, diventare abitudine, ecc.*), la continuativa (ad es., *condurre la/ una vita, correre il/ un pericolo*) e la telica (*conquistare il/ un lavoro*).

Anche le costruzioni a Verbo supporto esteso, come le costruzioni a verbo supporto neutro, possiedono diversi gradi di libertà; si passa dalle costruzioni con il nome non referenziale e con un grado più elevato di lessicalizzazione quali *prendere sonno*, *porre fine, ecc.*, ai sintagmi liberi come *svolgere il/ un compito*, *contrarre la/ una malattia*, *ecc*.

Il gruppo delle collocazioni con verbo ordinario (*Vord*) prodotte dai nativi comprende: a) le combinazioni in cui la base (il nome) seleziona il collocato (il verbo) perché lo richiede necessariamente per esprimere un determinato significato in quanto i due termini hanno instaurato tra di loro una solidarietà basata sull'uso (come ad es., *seguire il/ un consiglio, mantenere la/ una famiglia*, *prestare l'/ [] attenzione, ecc.*); b) le combinazioni preferenziali le quali costituiscono il modo più familiare, nonché il più frequente, per esprimere un dato concetto tra le altre combinazioni semanticamente possibili (come ad es., *suscitare interesse* vs. *accendere interesse*).

Abbiamo inteso una collocazione come preferenziale quando questa è risultata essere più frequente rispetto ad altre semanticamente affini attestate nel *corpus*. Nel presente lavoro qualsiasi affermazione sulla frequenza e sulla familiarità delle combinazioni $V+N_{Ogg}$ in lingua italiana si riferisce soltanto al campione linguisticio a nostra disposizione (il *corpus* SCUOLA) e può non essere valida per l'italiano contemporaneo, i giudizi di frequenza sul quale devono essere condotti a partire da un *corpus* di riferimento.

a) Abbiamo interpretato come semanticamente necessarie le collocazioni in cui la sostituzione del verbo con un quasi-sinonimo o con la corrispettiva forma analitica ha restituito combinatorie che non sono sembrate a chi scrive semanticamente assimilabili alla combinatoria analizzata (come ad es., *cambiare aria/ idea > *sostituire/ modificare (l') aria/ idea, prestare attenzione > *dare in prestito l'attenzione*). Nonostante le diverse sfumature di senso che tali trasformazioni comportano, abbiamo utilizzato tale test poiché ci è sembrato un indicatore utile allo scopo di verificare se il verbo avesse mantenuto il suo significato

primario o se, al contrario, la sua semantica si fosse adattata e specializzata nella combinatoria con l'Oggetto, selezionando uno dei possibili usi figurati del verbo oppure una delle sue accezioni secondarie, per esprimere un significato altro, emergente soltanto dalla co-occorrenza dei due componenti.
Altri esempi di collocazioni di questo tipo prodotte dai nativi sono: *affrontare (la/ una) giornata/ (il/ un) problema/ (la) vita (*fronteggiare/ assalire (la/ una) giornata/ (il/ un) problema/ (la) vita), costruire il futuro (*edificare (il) futuro), evitare (il) contagio/ (il) contatto (*aggirare/ schivare (il) contatto/ (il) contagio), lasciare (il/ un) segno (*cedere (il/ un) segno), mantenere (la/ una) famiglia (*conservare (la/ una) famiglia), porre rimedio (*mettere rimedio), ecc.*

b) Abbiamo classificato come preferenziali tutte le collocazioni per le quali è risultata esistere una combinazione V+N semanticamente affine e attestata nel *corpus* dei nativi con una frequenza più bassa (fr. < 10) rispetto a quella adottata nel presente studio per la definizione delle collocazioni (come ad es., *suscitare interesse* vs. *accendere interesse*). In questa seconda tipologia rientrano quindi le collocazioni con un grado di familiarità maggiore per i parlanti nativi rispetto ad altre combinatorie semanticamente affini ma meno frequenti; queste ultime sono costituite da combinatorie che possono essere marcate stilisticamente (come ad es., *perdere il/ [] senso* vs. *smarrire il/ [] senso*) oppure legate alla specifica dimensione diastratica e diafasica della lingua rappresentata dal *corpus* dei nativi sul quale è stato condotto il presente studio (come ad es., *lasciare il posto* vs. *cedere il posto*) o, infine, da combinatorie sinonimiche il cui uso potrebbe essere legato al cotesto (come ad es., *passare un anno/ un giorno/ un mese* vs. *trascorrere un anno/ un giorno/ un mese*).
Nel caso delle collocazioni preferenziali il verbo mantiene il significato primario e la sua combinazione con l'Oggetto costituisce la variante più frequente ma non l'unica, come nel caso delle collocazioni al punto 1, per esprimere un dato concetto. Come avevamo osservato per le collocazioni con *Vsup* e *Vsupext,* anche le collocazioni con *Vord* sono perlopiù sintatticamente libere, eccezion fatta per le combinazioni parzialmente lessicalizzate come *cambiare vita, lasciare il segno, perdere la voglia, ecc.*
Altri esempi di collocazioni preferenziali prodotte dai nativi sono: *arricchire il lessico/ il linguaggio* vs. *ampliare il lessico, sviluppare il linguaggio; chiedere aiuto* vs. *domandare aiuto; trasmettere emozione* vs. *dare emozione; commettere reato* vs. *compiere reato, ecc.*

L'analisi linguistica ha poi riguardato le combinazioni verbo-nominali estratte in base ai criteri quantitativi di frequenza e di associazione lessicale dal *corpus* di apprendenti CAIL2; ne sintetizziamo le principali osservazioni nei punti seguenti:

1) le collocazioni definite in base ai criteri linguistici costituiscono il 41%, mentre il restante 59% è composto dalle combinazioni lessicali ristrette e dalle combinazioni libere di parole;

2) soltanto un gruppo esiguo di collocazioni (16%) è stato realizzato con il verbo ordinario (*Vord*); tra queste troviamo i verbi *passare* e *trascorrere* selezionati dai nomi indicanti valori temporali quali *tempo* e *giorno*, presenti anche tra le collocazioni prodotte dai nativi insieme alla collocazione [*non*] *vedere (l') ora;* sono presenti anche le due espressioni fisse (*valere la pena* e *rendere conto*) il cui uso si è

cristallizzato in lingua italiana e che vengono perciò adoperate come dei blocchi semantico-lessicali unitari; il gruppo più numeroso (81%) è costituito dalle costruzioni a verbo supporto (ad es., *avere bisogno, fare amicizia, fare festa, prendere il sole, ecc.*), mentre troviamo soltanto una costruzione a verbo supporto esteso (*diventare amico*);

3) i parametri quantitativi per l'estrazione delle collocazioni adottati nel presente studio si sono mostrati sufficientemente accettabili per la lingua dei nativi (ricordiamo che tra le combinazioni V+N_{Ogg} rientranti nei valori quantitativi soglia il 61% era costituito da collocazioni); altrettanto non si può affermare per i dati degli apprendenti in quanto soltanto il 41% delle collocazioni estratte è risultato essere rispondente ai criteri linguistici.

Dall'analisi quantitativa condotta nel capitolo V è emerso che gli apprendenti hanno usato maggiormente, rispetto ai nativi, le collocazioni con i punteggi più alti di *t-score*; per questo motivo, abbiamo identificato con questo gruppo le cosiddette "isole di affidabilità" degli apprendenti, ovvero le combinazioni alle quali questi ricorrono molto spesso nelle proprie produzioni scritte.
In seguito all'analisi linguistica abbiamo potuto constatare che questo gruppo è costituito prevalentemente dalle costruzioni a verbo supporto; le collocazioni V+N di questo tipo, probabilmente per il particolare statuto linguistico che le contraddistingue, ovvero in ragione del fatto che il verbo è vuoto e la componente semantica risiede interamente nel nome, rappresentano uno strumento privilegiato per l'*output* degli apprendenti.

Le collocazioni estratte dal *corpus* CAIL2 in base ai criteri quantitativi di frequenza e di associazione lessicale rappresentano le collocazioni più diffuse nelle produzioni scritte degli apprendenti; come già detto sopra, si tratta delle collocazioni tipiche dell'interlingua in quanto presenti trasversalmente nei gruppi individuati in base alle due variabili del tempo di studio della lingua italiana e del tempo di permanenza in Italia.

Nella seconda parte del capitolo sesto abbiamo verificato l'ipotesi, formulata in seguito all'analisi quantitativa, in base alla quale si potrebbe assistere ad una differenza qualitativa nella produzione delle combinazioni verbo-nominali da parte degli apprendenti di lingua italiana soltanto dopo il terzo anno di studio della lingua.
Con questo scopo abbiamo analizzato come sono state utilizzate nel contesto, per mezzo dello strumento delle concordanze, tutte le combinatorie verbo-nominali individuate a partire da un campione di sostantivi (*problema, lavoro, vita, tempo* e *musica*), scelti nel gruppo dei nomi coinvolti nella formazione delle collocazioni V+N_{Ogg} tipiche.
Nell'analisi abbiamo descritto per ciascun gruppo le irregolarità sintattico-semantiche emerse e segnalato sia gli usi regolari che le combinatorie marcate stilisticamente.
Lo studio ha evidenziato le seguenti tendenze generali:

1) le irregolarità di carattere sintattico e semantico più diffuse in riferimento alle combinatorie V+N_{Ogg} sono le seguenti: dal punto di vista semantico abbiamo riscontrato a) degli usi impropri motivabili con una mancanza nel vocabolario dell'informante (ad es., *frequentare problema*) oppure con la scarsa conoscenza

delle restrizioni di selezione del sostantivo in questione (come ad es., *dire problema*); b) degli slittamenti semantici, ovvero degli usi verbali regolari ma non pertimenti nel contesto scritto (per es., *finire lavoro*); le irregolarità sintattiche più frequenti riguardano c) la dimensione della transitività ovvero gli usi intransitivi di verbi transitivi (ad es., *abituare vita*) e usi transitivi di verbi intransitivi (ad es., *parlare vita*); d) i due piani sintattico e semantico si sono incrociati nelle irregolarità motivabili con le influenze interlinguistiche (della L1 o di altre L2) certe (ad es., *divertirsi vita*) o presunte (es., *faticare lavoro*);

2) dall'analisi dei predicati verbali prodotti in co-occorrenza al campione di sostantivi scelti è emerso che le combinatorie V+N_{Ogg} marcate linguisticamente (come nel caso delle costruzioni a *Vsup*, verso la produzione delle quali, abbiamo visto, gli apprendenti si sono dimostrati particolarmente predisposti e che abbiamo perciò definito "isole di affidabilità") o quantitativamente (fattore desumibile dalle caratteristiche di frequenza e di associazione lessicale che una determinata combinatoria ha nell'*output* dei nativi, nel nostro caso nel *corpus* SCUOLA) sono state utilizzate regolarmente da parte di tutti gli apprendenti che le hanno prodotte. Ciò potrebbe significare che laddove una combinazione lessicale è dotata di particolari caratteristiche riguardanti lo statuto linguistico o quantitativo viene estratta più facilmente dall'*input* proprio in ragione della sua marcatezza. Ciò trova qualche punto di contatto con la teoria di Ellis (2002) in base alla quale avviene una elaborazione più veloce di tutte le sequenze frequenti rispetto a quelle meno frequenti ed è compatibile con i modelli di acquisizione linguistica chiamati *usage-based* (Bybee 1998; Goldberg 2006; Tomasello 2003) e gli approcci connessionisti dell'acquisizione e dell'elaborazione linguistica i quali enfatizzano le proprietà statistiche dell'*input* nell'apprendimento linguistico (Rumelhart, McClelland 1986; Elman 1990; Christiansen, Chater 1999) e per i quali la frequenza determina cosa e quanto i parlanti apprendono ed eventualmente rappresentano nel loro lessico mentale (Conklin, Schmitt 2012);

3) per quanto riguarda la valutazione delle produzioni degli appredenti in base alla variabile del tempo di studio dell'italiano abbiamo potuto constatare che il periodo che va da 1 anno a 3 anni di studio costituisce il momento di massima sperimentazione lessicale (sia in termini di varietà che di regolarità delle combinazioni prodotte); si assite, infatti, ad una tendenza verso la produzione di combinatorie diverse, non presenti negli altri gruppi, che talvolta non trova corrispondenza con la piena padronanza delle strutture sintattico-semantiche.
Dopo il terzo anno di studio si rilevano, invece, usi più stabilizzati con un consolidamento delle costruzioni verbo-nominali messo in evidenza da una diminuzione delle irregolarità sintattico-semantiche e dall'uso improprio di varianti stilistiche.
Ne consegue che l'ipotesi che avevamo formulato a partire dai risultati dell'analisi quantitaviva è stata confermata: dopo il terzo anno di studio della lingua italiana si assite ad una svolta qualitativa nelle produzioni degli apprendenti in quanto viene affinato l'uso delle combinatorie verbo-nominali nel contesto;

4) lo studio della variabile del tempo di permanenza in Italia ha rivelato che il periodo massimo di esposizione all'*input* preso in considerazione nel presente lavoro, ovvero più di 6 mesi (con una media di 18 mesi), non ha inciso sulla quantità

delle irregolarità prodotte le quali, infatti, non diminuiscono all'aumentare del tempo trascorso in Italia ma, al contrario, si distribuiscono con un andamento non lineare; si potrebbe ipotizzare che il tempo massimo di esposizione all'*input* preso in considerazione nel presente studio non sia sufficiente per incidere sulla regolarità delle produzioni verbo-nominali e che dei risultati diversi possano emergere con altri dati relativi agli informanti che abbiano trascorso più tempo in Italia. Inoltre, avendo qui considerato soltanto il tempo di esposizione all'*input*, va aggiunto che per una valutazione esaustiva della seconda variabilie si dovrebbero tenere in considerazione anche altri fattori quali, ad esempio, il tipo e la qualità di esposizione alla lingua a cui gli apprendenti sono stati soggetti durante la loro permanenza in Italia.

Tuttavia, i risultati dell'analisi linguistica confermano quanto era già emerso per la variabile legata al tempo di permanenza in Italia in seguito all'analisi quantitativa dei dati; infatti, relativamente alla produzione delle collocazioni $V+N_{Ogg}$ del *corpus* CAIL2, la variabile del tempo di studio della lingua italiana si era dimostrata essere più incisiva rispetto alla variabile dell'esposizione all'*input*.

I risultati fin qui esposti portano con sé alcune valutazioni e tratteggiano degli sviluppi futuri.

Innanzitutto, la salienza delle combinazioni lessicali con alti valori di *Mutual Information* ($14 > MI \geq 7$) per i nativi emersa dai dati a nostra disposizione è indice al contempo dell'utilità che l'acquisizione delle combinatorie con tale caratteristica potrebbe avere per gli apprendenti di italiano come lingua seconda.

Ciò significa che l'identificazione delle combinatorie lessicali con alti valori di MI in un *corpus* di riferimento, con una comprensibile rivalutazione dei parametri in relazione alla ampiezza della risorsa linguistica dalla quale le combinazioni vengono estratte, potrebbe rivelarsi importante per la creazione di liste di collocazioni da inserire nei materiali didattici per l'insegnamento di italiano L2.

Conoscere le collocazioni italiane con alti valori di *Mutual Information* potrebbe significare per gli studenti di lingua italiana disporre degli strumenti linguistici che, se usati correttamente e in contesti pertinenti, potrebbero conferire al loro *output* una impronta qualitativa importante, guadagnando in naturalezza.

Relativamente alle produzioni degli apprendenti, abbiamo potuto constatare un affinamento nella produzione delle combinatorie verbo-nominali dopo il terzo anno di apprendimento formale della lingua mentre una esposizione alla lingua *target* della durata media di 18 mesi non ha inciso sulla quantità delle irregolarità prodotte. Sia i risultati dell'analisi quantitativa che linguistica hanno dimostrato che la variabile del tempo di studio si è rivelata essere più incisiva della variabile del tempo di esposizione all'*input*. Tuttavia, mentre i risultati legati alla prima variabile sono statisticamente significativi, quelli relativi alla seconda variabile possono essere considerati validi soltanto per il campione linguistico a nostra disposizione.

Non possiamo non sottolineare che il nostro studio ha preso in considerazione soltanto il tempo di permanenza in Italia al momento della raccolta dei dati dichiarata dagli stessi informanti e non ha tenuto conto del tipo e della qualità di esposizione alla lingua che questi hanno avuto.

Per questi motivi, in questa sede non possiamo affermare *tout court* che l'esposizione all'*input* della lingua *target* giochi un ruolo meno importante rispetto all'apprendimento formale nell'acquisizione delle combinatorie verbo-nominali, ma

riteniamo che per una valutazione più profonda altri studi dovrebbero essere condotti sulle produzioni scritte di informanti che hanno trascorso in Italia un periodo superiore ad una media di 18 mesi.

Infine, una valutazione potrebbe essere fatta sui parametri utilizzati nel presente lavoro per l'estrazione delle collocazioni dai *corpora* (frequenza ≥ 10, *Mutual Information* ≥ 3 e *t-score* ≥ 2). Per quanto riguarda la lingua dei nativi, questi parametri quantitativi hanno restituito dei risultati più che accettabili in quanto la maggior parte delle combinazioni con tali caratteristiche è risultata essere rispondente ai criteri linguistici formulati per la definizione delle collocazioni di lingua italiana.

Pur essendo consapevoli del fatto che i parametri quantitativi debbano essere impostati *ad hoc* in base al tipo di risorsa empirica che si ha a disposizione e al tipo di studio che si intende condurre, potrebbe risultare utile abbassare ulteriormente tali valori soglia per cercare di individuare quali sono, orientativamente, i parametri di frequenza e di associazione lessicale in grado di restituire il bacino più ampio di collocazioni da un *corpus.*

Bibliografia:

Bybee, J. (1998), *The emergent lexicon*, in *Chicago Linguistic Society*, 34, pp. 421–435.

Christiansen, M., Chater, N. (1999), "Toward a connectionist model of recursion in human linguistic performance", in *Cognitive Science*, 23, pp. 157–205.

Cicalese, A. (1999) "Le estensioni di verbo supporto: uno studio introduttivo", in Studi Italiani di Linguistica Teorica e Applicata (SILTA), 3, 447-485.

Conklin, K., Schmitt, N. (2012), "The Processing of Formulaic Language", in *Annual Review of Applied Linguistics* (2012), 32, pp. 45–61.

Dechert, H. W. (1983), "How a story is done in a second language", in Faerch, C., Kasper, G. (eds.) *Strategies in Interlanguage Communication*, London, Longman, pp. 175 – 196.

Durrant, P. (2008), *High frequency collocations and second language learning*, Final Thesis Ph.D., University of Nottingham.

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.

Ellis, N. C. (2002), "Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition", in *Studies in Second Language Acquisition*, 24, pp. 143–188.

Ellis, N. C. (2003), "Constructions, chunking, and connectionism: The emergence of second language structure", in Doughty C. J., Long M. H. (Eds.), *The handbook of second language acquisition*, Oxford, UK, Blackwell, pp. 63–103.

Ellis, N. C., Simpson-Vlach, R., Maynard, C. (2008), "Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL", in *TESOL Quarterly*, 42, pp. 375–396.

Elman, J. (1990), "Finding structure in time", in *Cognitive Science*, 14, pp. 179–211.

Evert, S. (2009), "Corpora and collocations", in Lüdeling, A., Kytö, M. (eds.) *Corpus Linguistics: An International Handbook, Volume 2*, Berlin, New York, de Gruyter, pp.

1212-1248.

Firth, J. R. (1957), "A synopsis of linguistic theory, 1930-55", in *Studies in Linguistic Analysis*, Philological Society, Oxford, pp. 1-32, ristampato in F. R. Palmer (ed.) (1968), *Selected papers of J.R. Firth 1952-1959,* Harlow, Longman, pp. 168-205.

Goldberg, A. E. (2006), *Constructions at work: the nature of generalization in language*, Oxford, Oxford University Press.

Götz, S. (2013), *Fluency in Native and Nonnative English Speech*, Amsterdam, John Benjamins.

Granger, S. (1998a), "Prefabricated patterns in advanced EFL writing: collocations and formulae", in Cowie A. P. (ed.), *Phraseology: Theory, Analysis and Applications,* Oxford, Oxford University Press, pp. 145-160.

Hoey, M. (2005), *Lexical priming: A new theory of words and language*, London, Routledge.

Jones, S., Sinclair, J. (1974), "English lexical collocations", in *Cahiers de Lexicologie, 24*, pp. 15-61.

Ježek, E. (2005), *Lessico. Classi di parole, strutture, combinazioni*, Bologna, Il Mulino.

Kaszubski, P. (2000), *Selected aspects of lexicon, phraseology and style in the writing of Polish advanced learners of English: a contrastive, corpus-based approach*, Final Thesis Ph.D., Adam Mickiewicz University, Poznán.

Laufer, B., T. Waldman (2011), "Verb-noun collocations in second language writing: a corpus analysis of learners' English", in *Language Learning*, pp. 648-672.

Lorenz, G. (1999), *Adjective intensification - learners versus native speakers: A corpus study of argumentative writing*, Amsterdam, Rodopi.

Masini, F. (2009), *Combinazioni di parole e parole sintagmatiche,* in Lombardi Vallauri, E., Mereu, L. (eds.), "Spazi linguistici. Studi in onore di Raffaele Simone", Roma, Bulzoni, pp. 191-209.

Rumelhart, D., McClelland, J. (1986), "On learning the past tenses of English verbs", in Rumelhart, D., McClelland, J. (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, Cambridge, MA, MIT Press, pp. 216–271.

Tomasello, M. (2003), *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA, London, UK, Harvard University Press.

Wray, A. (2002), *Formulaic language and the lexicon*, Cambridge, Cambridge University Press.