Studi di Glottodidattica 2011, 2, 51-72 ISSN: 1970-1861

IL LESSICO MENTALE BILINGUE E GLI SPAZI SEMANTICI DISTRIBUZIONALI: LE SIMILARITA' TRA VERBI IN L1, IN L2 E NEI CORPORA.

MARIANNA BOLOGNESI
Università di Torino;
Siena Italian Studies.

Abstract

Gli spazi semantici distribuzionali sono modelli computazionali di lessico basati su corpora, in grado di reperire informazioni semantiche dai contesti di occorrenza delle parole e offrire mappe lessicali dove gli elementi sono tanto più vicini quanto più sono simili. L'analisi parte dall'osservazione delle occorrenze delle parole, contestualizzate nei testi che compongono i corpora, e genera similarità tra gli elementi sul piano paradigmatico. Pur senza inferire in maniera diretta e forzata un parallelo tra i meccanismi che soggiacciono a tali modelli e i meccanismi cognitivi che ci permettono di memorizzare le parole nel lessico mentale, questi strumenti offrono un potenziale applicativo molto alto, in quanto consentono di investigare in maniera estensiva le strutture e le dinamiche proprie della lingua viva. Nello studio proposto si confrontano le similarità semantiche tra verbi estratte attraverso l'implementazione di modelli dstribuzionali, con le similarità percepite da parlanti madrelingua e apprendenti stranieri.

Keywords: lessico mentale, bilinguismo, corpus linguistics.

Introduzione

La corpus linguistics è una disciplina applicativa che sta apportando sempre più contributi scientifici allo studio del linguaggio attraverso diversi metodi, sviluppati appositamente per analizzare grandi quantità di dati linguistici e per far luce sui meccanismi di strutturazione ed uso della lingua. I contributi variano dalla creazione e messa a disposizione di corpora veri e propri, sui quali poi fondare analisi linguistiche di diverso tipo, all'implementazione di metodi di analisi più specifici, volti a creare modelli di lessico, ontologie, thesaura, vocabolari e altri strumenti. La giovane età di questa disciplina, purtroppo, fa sì che a volte i risultati raggiunti con questi metodi non riescano a varcare la soglia dell'applicatività vera e propria in campi come la glottodidattica, da un lato perché manca ancora un po' la fiducia nell'importanza dei risultati raggiunti e raggiungibili attraverso l'analisi dei corpora, e dall'altro perché, essendo una disciplina ancora altamente tecnica e specializzata, è spesso difficilmente accessibile da parte di scienziati che non sono specificamente "addetti ai lavori".

Tuttavia, l'importanza di studiare la lingua basandosi sull'effettivo uso che se ne fa di essa, e attraverso materiali contestualizzati e autentici, è un obiettivo sempre più condiviso anche all'interno della comunità scientifica che si dedica all'insegnamento linguistico: l'utilizzo di materiali originali, recenti e aggiornati permette, infatti, di fornire agli apprendenti, oltre che un accesso alla lingua dell'uso, anche una finestra storica sempre aggiornabile, attraverso la quale le varietà della lingua possono venir messe a confronto. Attraverso l'analisi dei corpora è possibile, infatti, osservare le tendenze più recenti relative all'uso della lingua, non ancora istituzionalizzate o riconosciute ufficialmente, come dimostrò Sinclair (1987) con lo sviluppo del progetto COBUILD, mostrando come l'analisi dei contesti di occorrenza di determinate entrate lessicali riveli significati e usi delle parole che non sempre combaciano con quelli descritti nei dizionari tradizionali[1]. Questo approccio empirico e quantitativo è in grado di portare alla luce proprietà non solo semantiche ma anche pragmatiche e situazionali legate al tipo di testi in cui le entrate lessicali vengono utilizzate, soddisfacendo le esigenze contemporanee che spesso spingono un apprendente ad avvicinarsi ad una lingua straniera: il bisogno effettivo di comunicare in un'altra lingua e di conoscere quelle sfumature, esperibili dall'uso, che le parole acquisiscono in base a determinati contesti, in ambienti multiculturali, settoriali e all'avanguardia.

Nel panorama di applicazioni scientifiche relative all'analisi dei corpora, sono stati proposti diversi modelli di lessico, che sebbene non abbiano la presunzione di riprodurre esattamente il funzionamento della mente umana in relazione all'acquisizione e alla strutturazione di informazioni semantiche, proponendo modelli che ricalcherebbero in tutto e per tutto il lessico mentale, sono comunque in grado di riprodurre risultati simili a quelli che produrrebbero parlanti madrelingua o apprendenti, in relazione allo svolgimento di determinati compiti linguistici. Tra questi modelli, i Word Space Models (WSM) rappresentano una famiglia di applicazioni molto recenti, che sta riscuotendo un discreto successo all'interno della comunità scientifica, per la qualità dei risultati raggiunti e la

[1] Collin's Cobuild dictionary fu il primo dizionario a dare il senso di "homosexual" alla parola *gay* (e a datare come antiquato il senso di "lively and cheerful"), basandosi sulla frequenza d'uso di questa entrata lessicale nella lingua inglese, informazione estrapolata dal corpus che oggi ci sembra evidentissima ma ai tempi non lo era; un altro esempio di informazione estratta dal corpus e presente in Cobuild è costituito dal verbo *see*, usato frequentemente nel linguaggio discorsivo come sinonimo di *understand*, in frasi come *I see*, o *you see*, oltre che nel senso tradizionale di "vedere".

plausibilità dei principi teorici su cui si basano. L'implementazione di questi modelli, che appaiono come spazi semantici di tipo distribuzionale, si fonda su una metafora concettuale, espressa da Sahlgren (2006:19) nei seguenti termini: "Meanings are locations in a semantic space, and semantic similarity is proximity between the locations." I modelli ottenuti mostrano dunque spazi semantici, dove il parametro vigente è quello della similarità, e dove significati simili sono anche percepiti come vicini, grazie a una metafora concettuale esprimibile nei termini "similarity-is-proximity". La somiglianza semantica tra significati si ottiene grazie al paragone tra i contesti di occorrenza delle parole, in quanto parole con simili proprietà distribuzionali e dunque con simili contesti di occorrenza e simili argomenti, presenterebbero anche forti legami semantici[2] (Sinclair 1987; Miller, Charles 1991; Rubenstein, Goodenough 1965).

Osserviamo l'esempio seguente, relativo alle parole *libro*, *romanzo*, *sale*, *zucchero*. Intuitivamente diremmo che le prime due parole sono simili tra loro e le seconde due sono simili tra loro. Vediamo, dunque, che distribuzione presentano le quattro parole, cioè in che tipo di contesti vengono tipicamente utilizzate e con che tipo di argomenti.

[2] Lenci (2009) a questo proposito si sofferma sull'ipotesi che entrambe le direzioni dell'affermazione siano valide. In particolare, il fatto che distribuzioni simili possano generare similarità semantiche tra due parole, potrebbe essere la spiegazione per l'uso di metafore, analogie e sensi figurati, fenomeni inerentemente cognitivi e diffusissimi nell'uso della lingua.

a_ *leggere un* libro;
scrivere un libro;
prestare un libro;
un libro *recente*;
un libro *censurato*;
un libro *di cento pagine;*

b_ *leggere un* romanzo;
scrivere un romanzo;
prestare un romanzo;
un romanzo *recente*;
un romanzo *censurato*;
un romanzo *di cento pagine*;

c_ *rovesciare* il sale;
passare il sale;
*discio*gliere il sale;
il sale *accumulato*;
il sale *è bianco*;
il sale *nelle diete*;

d_ *rovesciare* lo zucchero;
passare lo zucchero;
*discio*gliere lo zucchero;
lo zucchero *accumulato*;
lo zucchero *è bianco*;
lo zucchero *nelle diete*;

I pattern sintattici e i tipi semantici degli argomenti che appaiono insieme alle parole *libro* e *romanzo* in (a) e in (b) sono uguali, così come lo sono i contesti di *sale* e *zucchero* in (c) e (d). Questo fa sì che in un modello WSM le parole *sale* e *zucchero* siano considerate molto più simili, e quindi più vicine, rispetto, ad esempio, a *sale* e *libro*.
Come sottolinea Lenci (2009), grazie all'ipotesi distribuzionale si individuano proprietà paradigmatiche tra le parole, analizzando la loro occorrenza in contesto, sul piano sintagmatico. Sul piano cognitivo, questo corrisponderebbe a un modello del lessico mentale in cui i significati non sono organizzati come le definizioni dei sensi di un dizionario, ma piuttosto come rappresentazioni contestuali, del tutto dipendenti dal tipo di contesto che si prende in considerazione. Il fatto che le rappresentazioni lessicali che emergono da questo modello siano strettamente legate al contesto ha suscitato non poche critiche all'interno della comunità scientifica, che hanno portato in primo piano l'argomento relativo alla questione nota come 'symbol grounding problem' (Harnad 1990): le rappresentazioni lessicali di parole che si basano su contesti nelle quali ricorrono altre parole non sono ancorate a entità reali e cognitivamente percettibili, come invece recenti teorie 'embodied' assumono (Lakoff e Johnson 1980, Barsalou 1999, Glenberg e Robertson 2000). Per questo motivo gli spazi semantici distribuzionali, basati su corpora, non sarebbero in grado di comprendere situazioni nuove che si basano su conoscenze pragmatiche derivate dal contatto con il mondo, che invece gli esseri umani interpretano senza problemi, affidandosi sulle affordances dei referenti trattati, come ad esempio l'utilizzo di neologismi verbali di origine denomiale, molto frequenti in inglese (Glenberg, Robertson 2000). Se, secondo molti studiosi, i modelli semantici distribuzionali non sono, dunque, candidati ottimali per proporre un'esaustiva teoria del significato, in quanto si basano sull'uso della lingua viva e contestualizzata, ma non ancorano quest'ultima all'esperienza, cioè non legano in maniera indissolubile la cognizione alla percezione, è scientificamente condivisa l'importanza fondamentale di questi modelli, come potenti strumenti di controllo per testare, analizzare, confermare o discreditare teorie cognitive e del significato, attraverso analisi rigorose, estensive e approfondite delle produzioni verbali (per una rassegna si veda Perfetti 1998).
I tipi di Word Spaces ad oggi implementati sono molti e diversi, ad esempio: Latent Semantic Analysis (Landauer, Dumais 1997), Hyperspace Analogue to Language (Lund, Burgess 1996), Random Indexing (Karlgren, Sahlgren 2001), Distributional Memory (Baroni, Lenci 2008). Essi si fondano sulla stessa idea di base, ma sono implementati con algoritmi diversi e vengono utilizzati per scopi diversi, sia in lessicografia che in linguistica cognitiva, per indagare il rapporto tra l'input linguistico e le sue rappresentazioni semantiche. Uno dei metodi più popolari utilizzati per valutare questi modelli è stato quello di confrontare la loro maniera di strutturare gruppi di sinonimi ai dati forniti da apprendenti di inglese L2, raccolti

attraverso l'esame scritto TOEFL (Test of English as Foreign Language). In questo test, tra i vari esercizi, viene chiesto di indicare, in un gruppo di parole simili, quella più simile alla parola target (ad esempio data la parola *rusty*, nella frase: *a rusty nail is not as strong as a clean, new one* , si chiede di indicare il suo sinonimo tra *corroded*, *black*, *dirty*, *painted*). Landauer e Dumais nel 1997 hanno paragonato questi dati a quelli estrapolati dai corpora attraverso il loro modello distribuzionale (Latent Semantic Analysis), ottenendo un livello di congruenza tra i risultati pari al 64,38%; in seguito altri modelli hanno raggiunto livelli di correlazione relativi a questo compito anche più alti (Rapp 2003; Jarmasz, Szpakowicz 2003; Bullinaria, Levy 2006; Padó, Lapata 2007; Turney 2008).
Anche nel caso illustrato, il fatto che gli algoritmi implementati siano paragonati ai risultati ottenuti da apprendenti di una lingua, piuttosto che da parlanti madrelingua, può far sorgere alcune critiche, in quanto i procedimenti di memorizzazione e organizzazione dei significati in L1 ed L2, a nostro avviso, non possono essere trattati allo stesso modo. In particolare, paragonando il tipo di similarità semantiche percepite da apprendenti di inglese L2 (i dati del TOEFL) a quelle estrapolate computazionalmente dai modelli, non possiamo essere certi del fatto che l'algoritmo proposto rifletta l'organizzazione della struttura semantica del lessico mentale dei parlanti nativi di una lingua.
Landauer e Dumais osservano che se le rappresentazioni delle parole sono di natura semantica, cioè sono "meanings abstracted and averaged from many experiences", i contesti, sono singoli eventi, cioè "unique combinations of events" (Landauer; Dumais 1997:228), ma che poi non sono più sempre reperibili singolarmente, in quanto si amalgamano e si fondono l'uno nell'altro, mentre ciò che viene ricordato sono le informazioni semantiche estratte dai contesti e riportate sul piano paradigmatico.

Nello studio che si propone, il lessico mentale in L1 e quello in L2 sono messi a confronto, in relazione a un gruppo di verbi (di movimento e di pensiero), sia in inglese che in italiano. L'ipotesi che si vuole testare, attraverso l'implementazione di alcuni word space models, e il confronto tra i dati di origine cognitiva e quelli di origine computazionale, è che le rappresentazioni lessicali in L2 abbiano origine diversa rispetto a quelle in L1, e che questa differenza sia osservabile attraverso il grado di correlazione esistente tra le rappresentazioni lessicali emerse dal modello distribuzionale e quelle emerse, rispettivamente, in L1 e in L2.
Le parole in L1, secondo modelli psicologici tutt'oggi condivisi, sono memorizzate su base semantica e caratterizzate da diversi legami, che le connettono ad altre parole principalmente sul piano paradigmatico; i legami paradigmatici, infatti, sembrano essere i più salienti, in base alle analisi effettuate in esperimenti di associazioni lessicali libere (Meara 2009). Solo quando ci vengono proposte come stimolo parole a bassissima frequenza d'uso, allora tendiamo ad associare ad esse parole che si legano sul piano sintagmatico, formando collocazioni o chunks. In L2, al contrario, la variabilità relativa alle associazioni lessicali è molto più ampia e difficilmente classificabile, sebbene sembri essere caratterizzata dalla presenza di associazioni tra parole che sembrerebbero occorrere insieme in un contesto situazionale più ampio di quello strettamente sintagmatico. Dal punto di vista acquisizionale, si riconosce, inoltre, l'importanza che l'attività di lettura ricopre nei processi di accrescimento delle competenze lessicali in L1 e persino in L2, soprattutto ai primi stadi di apprendimento, stimolando meccanismi di inferenza e interpretazione che si muovono su direzioni inverse complementari, sia bottom-up che top-down (si consulti Cardona 2008 per una rassegna dettagliata). Questi meccanismi si basano su procedimenti di estrazione e inferenza dei significati lessicali a partire dall'analisi dei contesti d'uso delle parole nuove. Il fatto che le rappresentazioni lessicali siano possano essere, almeno inizialmente, di natura contestuale, o episodica, cioè strettamente legate ai contesti (linguistici ed extralinguistici) nei quali le parole sono state incontrate, è un'ipotesi che si affaccia sul panorama scientifico delle teorie legate al significato. Recentemente quest'idea è stata proposta in alcuni studi di carattere sperimentale nei quali si osservavano diversi trattamenti relativi alle parole in L1 e a quelle in

L2, e diversi effetti conseguentemente registrati durante lo svolgimento di compiti linguistici che coinvolgevano il passaggio da L1 a L2 e quello da L2 a L1 (Jiang, Forster 2001; Segalowitz, Almeida 2002; Bolognesi 2010). Inoltre, la natura empirica e percettibile dei significati lessicali è ampiamente analizzata e supportata dalle più recenti teorie cognitiviste (Barsalou 1999), nelle quali si sostiene che essa pervada tutto il sistema cognitivo umano, e di conseguenza comprenda anche il linguaggio, che sarebbe allocato nelle aree cerebrali dedicate alla processazione di informazioni legate alla percezione, quali le aree sensori-motorie.
L'ipotesi che viene sondata in questo studio è che i modelli semantici distribuzionali, nonostante siano basati su contesti esclusivamente linguistici, riescano comunque a produrre rappresentazioni lessicali correlate a quelle di origine umana. Inoltre, si ipotizza che tali rappresentazioni lessicali, in quanto emergenti direttamente dai contesti di occorrenza delle parole stesse, secondo meccanismi di tipo bottom-up, siano eventualmente più simili a quelle prodotte dagli apprendenti stranieri (sia in italiano L2 che in inglese L2), piuttosto che dai parlanti madrelingua, in quanto il lessico mentale in L1 potrebbe essere caratterizzato anche da meccanismi di rielaborazione dei significati lessicali di tipo top-down, che arricchirebbero le rappresentazioni lessicali di proprietà complesse che non sono direttamente derivabili dal contatto con i contesti di occorrenza.
Nei prossimi due paragrafi si riportano le procedure relative all'implementazione dei modelli semantici da confrontare: da un lato quelli con i dati di origine computazionale e dall'altro quelli con i dati di origine cognitiva.

I verbi scelti per l'analisi sono stati selezionati rispettivamente in inglese e in italiano. La scelta è stata effettuata tenendo presente diversi parametri, che in questo paragrafo verranno elencati; in particolare, i verbi non sono stati scelti in una delle due lingue, e poi tradotti nell'altra, ma al contrario sono stati individuati concetti espressi attraverso verbi equivalenti nelle due lingue, che avessero caratteristiche simili e potessero essere paragonabili. Uno dei parametri presi in considerazione è la frequenza di occorrenza dei verbi in ognuna delle due lingue: verbi equivalenti in italiano e inglese, infatti, possono avere frequenze di occorrenza diverse nelle rispettive lingue. Ad esempio, il verbo *fall* in inglese ha una frequenza più alta dell'equivalente italiano *cadere*, perchè supporta molte costruzioni sintattiche che in italiano non esistono, tra cui gli idiomi molto frequenti *fall in love* e *fall asleep*, o alcuni poco frequenti ma comunque non direttamente traducibili in italiano, come ad esempio *fall foul of the law* ('cadere dalla parte sbagliata della legge', cioè al di fuori di essa). A sostegno di questa intuizione, abbiamo consultato il CQS Sketchengine[3], a proposito delle frequenze dei due verbi nei corpora; per questo paragone abbiamo scelto due grandi corpora come ItWac e UkWac (corpora di dati estratti dal web rispettivamente in italiano e in inglese), che contano rispettivamente 1,909,535,703 e 1,565,274,190 occorrenze. Il computo delle occorrenze dei due verbi è il seguente: *fall* 236 295, *cadere* 160 826. Il fatto che *fall* sia più frequente e partecipi a un numero maggiore di pattern sintattici (51 word sketches secondo Sketchengine) rispetto all'equivalente *cadere* (37 word sketches) potrebbe influenzare, plausibilmente, il tipo di rappresentazione mentale delle due parole, da parte di parlanti nativi e di apprendenti. Per questo motivo nel campione di verbi scelti per l'analisi ci sono verbi equivalenti la cui frequenza in italiano e inglese è bilanciata (come ad esempio *ballare*/*dance*,

[3] Sketch Engine è un Corpus Query System, cioè un programma che permette l'interrogazione di corpora in molti modi diversi. Tra le manipolazioni più interessanti, il programma permette di visualizzare le co-occorrenze di una parola organizzate in tipi di pattern a cui una parola partecipa (Word Sketches), e permette di calcolare le differenze tra le parole, in termini di condivisione di pattern sintattici (sketch difference). I concetti di salienza e frequenza in Sketchengine sono così rappresentati: se la frequenza conta il numero di occorrenze, il valore definito *overall* score definisce la salienza, cioè specificità di un pattern per una determinata parola data come input, attraverso la relazione tra la frequenza del pattern in occorrenza con la parola data e la frequenza del pattern in generale. Per esempio, dato un verbo come input, il pattern "SOGGETTO+verbo" inserito avrà probabilmente frequenza alta, ma *overall score* basso, in quanto è un pattern molto frequente, con qualsiasi verbo. D'altro canto, preso per esempio il verbo *to doubt*, il pattern "verbo+ WHETHER" avrà un overall score molto alto, perché la frequenza di occorrenza di questo pattern è molto peculiare per il verbo *to doubt*.

venire/*come*, *odiare*/*hate*, *apprezzare*/*appreciate*), coppie di verbi equivalenti dove l'elemento italiano ha frequenza maggiore rispetto a quello inglese (come ad esempio *protestare*/*protest*, o *rispettare*/*respect*) e coppie di verbi dove l'elemento inglese ha frequenza molto più alta rispetto all'equivalente italiano (ad esempio *march/marciare*, o *roll*/*rotolare*). In questo modo sarà possibile osservare più differenze tra le rappresentazioni dei verbi nelle due lingue, da parte dei diversi gruppi di partecipanti.

Le parole scelte hanno una morfologia molto ben riconoscibile, attribuibile senza tentennamenti all'una o all'altra lingua e identificabile immediatamente con la categoria grammaticale dei verbi. In italiano, infatti, tutti i verbi scelti sono espressi all'infinito e hanno il suffisso *–re*; la vocale tematica è *a*, *e* o *i*, e non sono state utilizzate forme riflessive. Per quanto riguarda le forme verbali inglesi, ognuna è stata preceduta dalla preposizione *to*, in quanto la stessa forma lessicale in inglese può assumere diversi significati e diversi ruoli all'interno della frase. Spessissimo, infatti, in inglese si assiste a slittamenti di categoria grammaticale, e alla composizione di nuove forme lessicali derivate da verbi, nomi o aggettivi (basti pensare alla parola *round*, che può assumere il ruolo di aggettivo, nome o verbo a seconda del contesto). Quando, però, la forma lessicale è preceduta dalla preposizione *to*, i dubbi si dissolvono.

Altri due parametri importanti, utilizzati per selezionare i verbi della lista, sono i seguenti: innanzitutto sono stati esclusi quei verbi che, in una delle due lingue, costituivano una forma derivata da un altro verbo presente nella lista (ad esempio *saltare*/*jump* è stato incluso, ma *saltellare*/*hop* è stato escluso perché in italiano è una forma derivata). In secondo luogo sono stati esclusi i verbi ad altissima frequenza che fossero costituiti da più di una parola (fenomeno frequente per l'inglese, che è una lingua di tipo satellite-frame e codifica esternamente al verbo la direzione del movimento, come in *go in, go out, go up, go down, come in*, eccetera). Per concludere questa preliminare descrizione dei verbi prescelti, si sottolinea che i verbi sono stati bilanciati anche in base al parametro della disponibilità nell'input di parlanti ed apprendenti, e la loro familiarità. Sono stati quindi evitati domini concettuali dei quali i partecipanti (soprattutto gli apprendenti) non avessero alcuna familiarità.

I verbi prescelti per l'analisi sono elencati qui sotto.

DUBITARE – TO DOUBT
SCALARE – TO CLIMB
IDEALIZZARE – TO IDEALIZE
CADERE – TO FALL
ODIARE – TO HATE
DONDOLARE – TO SWING
DESIDERARE – TO WISH
REMARE– TO ROW
BALLARE – TO DANCE
CREDERE – TO BELIEVE
SCIVOLARE – TO SLIP
IMMAGINARE – TO IMAGINE
CAMMINARE – TO WALK
PERDONARE – TO FORGIVE
GUIDARE – TO DRIVE
DIMENTICARE – TO FORGET
CORRERE – TO RUN
CAPIRE – TO UNDERSTAND
PATTINARE – TO SKATE
AMMIRARE – TO ADMIRE
VOLARE – TO FLY
RICORDARE – TO REMEMBER
NUOTARE – TO SWIM
AMARE – TO LOVE
SALTARE – TO JUMP
SPERARE – TO HOPE
VENIRE – TO COME
MEMORIZZARE – TO MEMORIZE
PASSEGGIARE – TO STROLL
NEGARE – TO DENY
GATTONARE – TO CRAWL
GIUDICARE – TO JUDGE
MARCIARE – TO MARCH
DECIDERE – TO DECIDE
SCIARE – TO SKI
APPREZZARE – TO APPRECIATE
ROTOLARE – TO ROLL
RISPETTARE - TO RESPECT
PARTIRE – TO LEAVE
CONSIDERARE – TO CONSIDER
ARRIVARE – TO ARRIVE
PROTESTARE – TO PROTEST

SEGUIRE – TO FOLLOW	INTENDERE – TO MEAN
SOGNARE – TO DREAM	VIAGGIARE – TO TRAVEL
SCAPPARE – TO ESCAPE	SUPPORRE – TO SUPPOSE

Gli abbinamenti tra verbi equivalenti sono stati effettuati inizialmente attraverso la ricerca di ogni singolo verbo nel dizionario bilingue Oxford Paravia 2010. In un secondo momento, ad alcuni parlanti nativi italofoni e anglofoni è stato chiesto di tradurre la lista di verbi espressa nella loro lingua, in una lista di verbi espressa nell'altra lingua; in questo modo le coppie di verbi sono state confermate nelle due direzioni: dall'inglese verso l'italiano e dall'italiano verso l'inglese.

I modelli semantici distribuzionali

Tra i modelli WSM, Distributional Memory (DM) si distingue per la sua portabilità, cioè per la sua capacità di adattarsi a svolgere diversi compiti. Generalmente, nel mondo dei modelli Word Space, la tendenza più comune sembra quella espressa dalla massima "one semantic task – one distributional model"; questo approccio tuttavia non sembra coerente con la metafora che mette in correlazione la mente ai modelli computazionali, in quanto la struttura cognitiva umana (la memoria semantica) pur essendo sempre la stessa, è adattabile a diversi compiti. Il modello DM (Baroni, Lenci 2010), invece, si applica allo svolgimento di diversi compiti, generando spazi di parole distribuzionali diversi a partire dagli stessi principi e dalle stesse unità; i compiti che vengono fatti eseguire a questi modelli variano dal riconoscimento di sinonimi, all'organizzazione di campi semantici in base a giudizi di similarità semantiche (confrontando i dati con quelli raccolti da Rubenstein e Goodenough nel 1965), alla categorizzazione tassonomica di nomi e la ricerca di affinità selettive tra parole. Questi compiti vengono svolti attraverso un procedimento che si basa sull'estrazione triplette di parole, chiamate tuple, strutturate nel seguente modo: una parola, un elemento di connessione e un'altra parola, cioè Word, Link, Word (W1-link-W2). Nel modello utilizzato per questo studio, per ogni verbo dato in input sono stati estratti i due elementi che caratterizzano ogni singola occorrenza, come mostra l'esempio seguente, dove sono riportate alcune delle tuple per il verbo *admire*:

W1	Link	W2	LMI
admire-v	as-1	artist-n	21.8294
admire-v	as-1	author-n	17.3123
admire-v	as-1	example-n	18.5736
admire-v	around-1	world-n	102.6617
admire-v	beyond-1	measure-n	15.7057
admire-v	for-1	ability-n	264.6581
admire-v	for-1	bravery-n	104.9274
admire-v	for-1	integrity-n	54.5303
admire-v	for-1	intellect-n	43.9612
admire-v	for-1	intelligence-n	55.1108
admire-v	in-1	people-n	111.4696
admire-v	in-1	regeneration-n	424.2624
admire-v	obj-1	ability-n	650.3732
admire-v	obj-1	anyone-n	161.0347
admire-v	obj-1	nature-n	102.7680
admire-v	sbj_tr-1	visitor-n	323.9702

Le tuple estratte sono 'pesate', cioè sono associate ad un valore (riportato sulla quarta colonna) che indica una misura di associazione tra gli elementi costituenti. Questa misura di associazione non è altro che una versione modificata della Mutual Information[4]. La formula che definisce questa misura di associazione, chiamata Local Mutual Information, che tende a bilanciare la tendenza della MI a favorire i contesti più idiosincratici di ogni parola (che sono generalmente poco frequenti), è la seguente:

$$\text{LMI} = \text{frequenza osservata (word-link-word)} * \log \frac{\text{freq osservata (word-link-word)}}{\text{frequenza attesa (word-link-word)}}$$

Per l'inglese è stato possibile utilizzare un corpus già predisposto a questo tipo di analisi da Baroni e Lenci (2010), composto da tre diversi corpora: ukWac[5], cioè un corpus del web (circa 1.915 miliardi di occorrenze); il BNC[6], cioè il British National Corpus (circa 95 milioni di occorrenze) e una selezione effettuata da English Wikipedia[7] della metà del 2009 (circa 820 milioni di occorrenze).

Per quanto riguarda l'analisi dei verbi italiani, è stato utilizzato il corpus "La Repubblica", che comprende le annate del quotidiano "La Repubblica" dal 1985 al 2000. Si tratta di un ampio corpus tokenizzato e taggato con TreeTagger, e lemmatizzato con MorphIt[8], che comprende testi di italiano giornalistico, ed è composto da circa 380 mila tokens. Nonostante le grandi dimensioni, tra i 3960 verbi più frequenti attestati da questo corpus non appare uno dei verbi del nostro campione: il verbo *gattonare*; per questo motivo i verbi utilizzati per l'analisi computazionale sull'italiano sono stati 47.

Una volta scelti i corpora, sono state estratte le tuple per ognuno dei verbi. Per la lingua inglese ne sono state estratte in tutto 920.710. Di esse, in posizione (W1) era rappresentato uno dei 48 verbi e nella terza posizione (W2) una dei 20.080 nomi e aggettivi più frequenti Al centro della tupla era rappresentato il tipo di legame (link) che univa le due parti. I link rappresentati nella lista di tuple per l'inglese erano in tutto 151. Considerando la somma di link+W2 come contesto dei 48 verbi (W1), il numero di contesti diversi (cioè di link+W2) rappresentati ammontava a 242.322; dei tre elementi costituenti di ogni tupla, era evidente che il livello nel quale i numeri salivano enormemente era il terzo livello, quello occupato da W2. Sebbene il ventaglio di possibilità sull'elemento W2 fosse molto ampio, in seguito ad alcuni tentativi di sfoltimento della lista di tuple è stato deciso di conservare tutti i dati estratti per il seguente motivo: il modello distribuzionale non intende emulare la mente umana e ricalcarne gli stessi meccanismi di funzionamento partendo dagli stessi elementi iniziali, dati in input; il modello distribuzionale, piuttosto, aspira a reperire dalle produzioni linguistiche analizzate similarità semantiche che, qualitativamente, potrebbero avere la stessa natura di quelle percepite dai parlanti, cioè similarità semantiche basate sui contesti d'uso delle parole. Essendo, inoltre, così strettamente legate ai contesti d'uso, esse potrebbero assomigliare a quelle percepite in L2 da apprendenti stranieri, piuttosto che a quelle percepite in L1 da parlanti madrelingua, in quanto il lessico mentale e la memoria semantica di questi

[4] Mutual Information (Church, Hanks 1990; Stubbs 1995), una delle misure più usate, misura il rapporto logaritmico tra la frequenza osservata e la frequenza attesa di una determinata co-occorrenza, attraverso una formula in cui si esprime il rapporto (logaritmico) tra la frequenza della co-occorrenza dei due elementi e la frequenza delle singole occorrenze dei due elementi.

[5] http://wacky.sslmit.unibo.it/

[6] http://www.natcorp.ox.ac.uk

[7] http://en.wikipedia.org/wiki/Wikipedia:Database_downloadhttp://en.wikipedia.org/wiki/Wikipedia:Database download

[8] http://dev.sslmit.unibo.it/linguistics/morph-it.php

ultimi potrebbe essere caratterizzata da rielaborazioni più complesse tra domini concettuali, che genererebbero rappresentazioni lessicali non legate in modo così diretto ai contesti d'uso delle parole stesse.

Dopo aver estratto le tuple sia in inglese che in italiano, le due liste sono state ordinate in base al valore della misura di associazione LMI che indica il peso della co-occorrenza del verbo con un determinato contesto. Qui di seguito riportiamo uno stralcio delle prime 30 tuple delle due liste, in inglese e in italiano:

fall-v, in-1, love-n, 234 851,
come-v, into-1, force-n, 186 340,
follow-v, obj-1, link-n, 126 218,
follow-v, obj-1, instruction-n, 83 780,
come-v, into-1, effect-n, 80 998.7,
come-v, into-1, contact-n, 79 085,
fall-v, into-1, category-n, 71 321.1,
walk-v, obj-1, distance-n, 67 816.6,
come-v, sbj_intr-1, people-n, 67 688.5,
follow-v, in-1, footstep-n, 66 992.2,
run-v, obj-1, business-n, 64 085.1,
come-v, as-1, surprise-n, 61 286.8,
leave-v, obj-1, school-n, 61 013.7,
drive-v, obj-1, car-n, 60 393.3,
walk-v, sbj_intr-1, minute-n, 59 441.8,
run-v, obj-1, course-n, 56 138.9,
come-v, sbj_intr-1, time-n, 50 705.,
leave-v, sbj_intr-1, turn-n, 49 137.2,
come-v, sbj_intr-1, name-n, 48 701.3,
follow-v, obj-1, path-n, 46 614.3,
believe-v, sbj_intr-1, people-n, 46 427.4,
consider-v, obj-1, application-n, 46 355.9,
come-v, into-1, existence-n, 44 910.6,
leave-v, obj-1, band-n, 44 425.1,
come-v, at-1, time-n, 44 092.2,
follow-v, obj-1, procedure-n, 43 726.8,
run-v, for-1, year-n, 43 519.5,
come-v, with-1, idea-n, 43 152.2,
leave-v, obj-1, home-n, 42 574.5,
fall-v, into-1, hand-n, 41 392.4,

credere-v, fin_che, essere-v, 55 970.7,
capire-v, fin_che, essere-v, 31 544.3,
venire-v, modadv, fuori-b, 24 562.1,
ricordare-v, fin_che, essere-v, 22 711.3,
correre-v, obj, rischio-s, 22 010.9,
venire-v, modadv, meno-b, 18 671.3,
decidere-v, inf_di, fare-v, 14 099.9,
capire-v, modadv, bene-b, 11 111.2,
saltare-v, modadv, fuori-b, 10 833.3,
correre-v, comp_a, riparo-s, 10 440.3,
venire-v, comp_in, mente-s, 9697.9,
saltare-v, comp_in, aria-s, 9109.65,
arrivare-v, modadv, fino-b, 9051.68,
seguire-v, comp_con, attenzione-s, 8882.86,
sperare-v, fin_che, essere-v, 8496.52,
arrivare-v, subj, notizia-s, 8040.19,
dimenticare-v, fin_che, essere-v, 7877.08,
arrivare-v, comp_a, conclusione-s, 6999.93,
venire-v, comp_a, luce-s, 6733.18,
capire-v, modadv, meglio-b, 6693.2,
venire-v, inf_a, sapere-v, 6681.27,
decidere-v, subj, governo-s, 6614.87,
credere-v, fin_che, avere-v, 6563.98,
arrivare-v, comp_a, punto-s, 6226.62,
partire-v, comp_da, gennaio-s, 6217.49,
venire-v, inf_a, mancare-v, 6073.99,
rispettare-v, obj, regola-s, 5810.27,
capire-v, obj, cosa-s, 5801.05,
cadere-v, comp_in, trappola-s, 5578.57,
credere-v, fin_che, fare-v, 5516.6

Possiamo osservare che nelle posizioni più alte sono presenti quasi esclusivamente verbi di movimento, e gran parte di essi (quasi tutti) sono inseriti in tuple nelle quali il movimento espresso non è letterale ma è metaforico, fittizio o idiomatico. Il fatto che nella lista italiana ci siano più verbi di pensiero rispetto a quelli di movimento è probabilmente dovuto al fatto che il corpus prescelto (e disponibile) per l'analisi effettuata con i verbi italiani era un corpus giornalistico, dove il numero di verbi di pensiero era superiore rispetto a quello di verbi di movimento, proprio per la natura stessa dei testi in esso inclusi.
Le due liste di occorrenze sono state inserite in matrici che riportavano sulle righe i verbi, sulle colonne ogni contesto che appariva nella lista e nelle celle corrispondenti il valore della misura di associazione LMI associata alla tupla, o il valore zero, nel caso in cui la co-presenza di contesto e un verbo non fosse stata osservata nel corpus. Applicando poi la formula generale del coseno, abbiamo calcolato la distanza tra ogni vettore (cioè ogni verbo) e tutti gli altri vettori, ottenendo due matrici quadrate, di 48*48 verbi inglesi e 47*47 verbi italiani. In queste matrici, simmetriche rispetto alla diagonale, i valori dei coseni variavano dal valore massimo (1, riportato sulla diagonale) al valore minimo, pari a 0. Nella tabella che segue riportiamo una versione condensata dei dati relativi ai valori dei coseni nelle matrici, in relazione al tipo di coppia di verbi: due verbi di movimento, due verbi psicologici, oppure uno e uno.

	WSM ENG	WSM ITA
Mot/mot (276 valori in ENG e 257 in ITA)	M = 0,04 SD = 0,05	M = 0,04 SD = 0,09
Psy/psy (276 valori in entrambe le matrici)	M = 0,09 SD = 0,09	M = 0,16 SD = 0,27
Mot/psy (576 valori in ENG e 548 in ITA)	M = 0,02 SD = 0,03	M = 0,01 SD = 0,02

Tabella 1 i valori medi di similarità estratti dai corpora in relazione al tipo di coppia di verbi.
F (2, 1125) = 63. 70; p = .0005

Curiosamente, il coseno medio che indica la distanza tra coppie di verbi psicologici in italiano è molto più alto rispetto a quello inglese e rispetto agli altri valori attribuiti a coppie di verbi omogenei, e la deviazione standard è molto alta rispetto alle altre. Sembrerebbe, insomma, che alcuni valori all'interno del gruppo di coppie di verbi psicologici, abbiano alzato notevolmente la media, e di conseguenza anche la varianza. I clusters lessicali generati da questi modelli, qui riprodotti attraverso Permap[9], confermano tale supposizione.

[9]. Permap (Perceptual mapping) è un programma che si basa su tecniche di Multidimensional Scaling per rappresentare graficamente e geometricamente la struttura nascosta di un insieme di dati organizzati in modo complesso e interconnesso, come ad esempio una matrice di distanze.

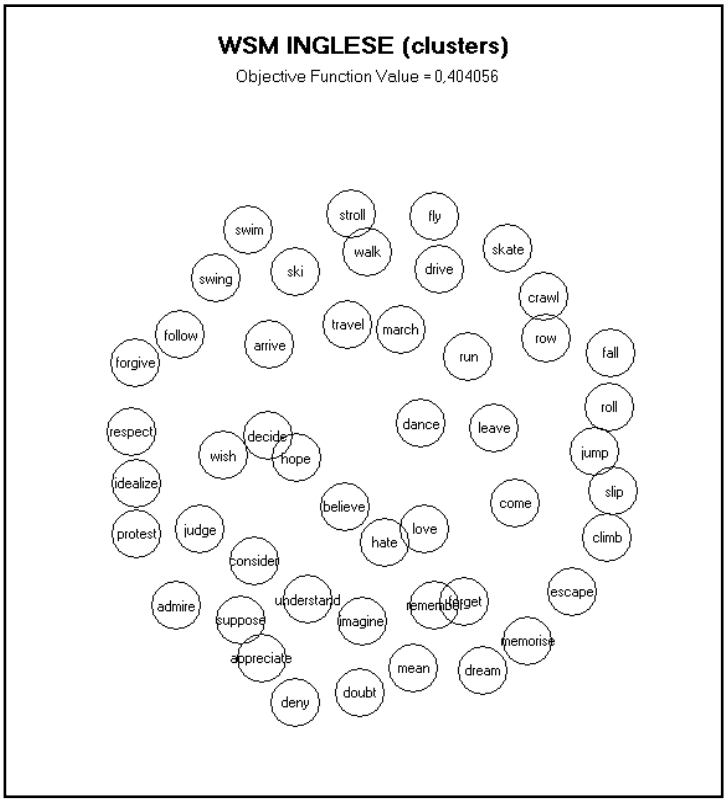

Figura 2. Visualizzazione grafica del modello distribuzionale con i verbi in inglese basata su Permap, un programma che utilizza la tecnica *multidimensional scaling*, dal quale emerge automaticamente la separazione dei due clusters verbali: in alto a destra i verbi di movimento, in basso a sinistra quelli di pensiero.

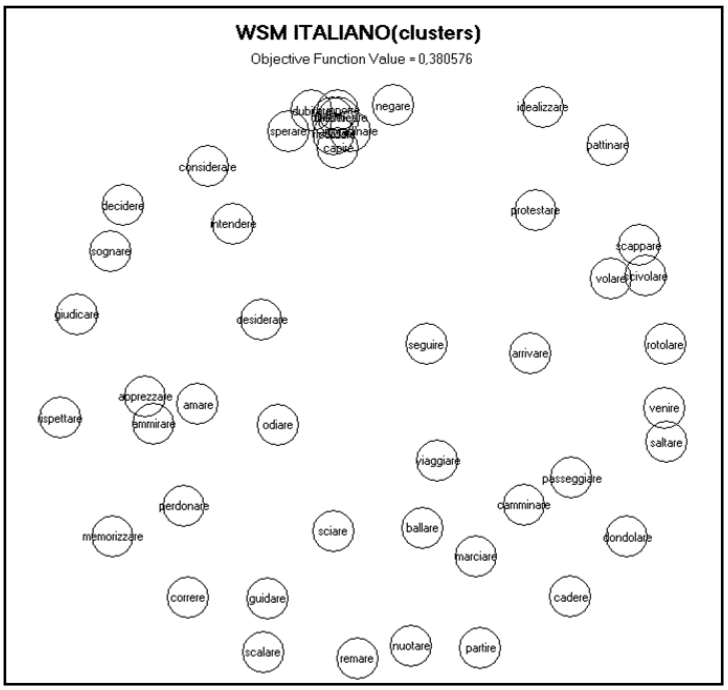

Figura 3 Visualizzazione grafica del modello distribuzionale con i verbi in italiano basata su Permap, un programma che utilizza la tecnica *multidimensional scaling*, dal quale emerge automaticamente la separazione dei due clusters verbali: in alto a destra i verbi di movimento, in basso a sinistra quelli di pensiero.

I dati raccolti dai parlanti

Gli ottanta partecipanti che hanno preso parte alla raccolta dei dati hanno genere e provenienza diversi. L'età dei partecipanti è stata mantenuta come parametro fisso, così come il livello di istruzione: tutti i partecipanti coinvolti erano studenti universitari, maschi e femmine, iscritti ad un corso di laurea triennale, di età compresa tra i 19 e i 23 anni. Metà dei partecipanti (cioè quaranta) erano italofoni e residenti in Italia, l'altra metà erano anglofoni, temporaneamente residenti in Italia per un programma semestrale di studio all'estero. Sia gli studenti italiani che quelli americani provenivano da diversi atenei del loro Paese, e conoscevano la lingua straniera ad un livello basico: quelli italiani l'avevano studiata almeno alle superiori, mentre gli studenti americani stavano studiando l'italiano in Italia da pochi mesi presso un programma di studio full-immersion, e l'avevano studiato per uno o due semestri in America, presso le loro università. Il livello di competenza dichiarato si basava su un'autovalutazione degli apprendenti, ai quali è stato chiesto se conoscessero la lingua straniera in modo "sufficiente", "buono" o "come un madrelingua". Sono stati selezionati i partecipanti che appartenevano alla prima fascia.

Ad ognuno dei partecipanti sono stati somministrati mediamente 5 fogli prestampati, in cima ad ognuno dei quali era riportato un verbo (in L1 o in L2, in ordine sparso), seguito dalla lista di 48 verbi nella stessa lingua dell'intestazione, disposti in ordine sparso. Le istruzioni, date oralmente ai partecipanti, richiedevano che il verbo riportato nell'intestazione del questionario fosse paragonato a ognuno dei verbi che seguivano, e che fosse attribuito un giudizio a ogni coppia di verbi, con un numero da 1 a 7 (scala Likert[10]). In particolare si chiedeva di esprimere la similarità percepita tra i due verbi della coppia, in modo che 7 risultasse essere la similarità massima e 1 la minima (cioè i verbi erano percepiti come molto lontani tra loro). I giudizi raccolti sono stati disposti in quattro matrici quadrate, dove i verbi erano riportati sia sulle righe che sulle colonne, e i valori di similarità nelle celle rispettive celle. Le quattro matrici così compilate sono state nominate nel seguente modo: ENGbyENG (la matrice di verbi inglesi con i dati raccolti da parlanti anglofoni), ENGbyITA (la matrice di verbi inglesi con i dati raccolti da parlanti italofoni), ITAbyITA (la matrice di verbi italiani con i dati raccolti da parlanti italofoni) e ITAbyENG (la matrice di verbi italiani con i dati raccolti da parlanti anglofoni).

Il successivo calcolo della media tra i giudizi raccolti ha reso possibile gli ulteriori sviluppi dell'analisi e l'estrazione delle matrici di distanze tra ogni verbo e gli altri, necessaria al paragone con i dati computazionali. Prima, però, è utile riportare alcune considerazioni relative alla varianza interna ad ogni coppia di verbi. Per ogni coppia di verbi (AB) sono stati raccolti quattro giudizi di similarità: due di essi relativi alla presentazione della coppia nell'ordine (AB) e due di essi nell'ordine (BA). Calcolando la media tra i 4 giudizi purtroppo si perdono le informazioni relative alla varianza interna a questi 4 giudizi, che però ha mostrato interessanti peculiarità. Consideriamo ad esempio una coppia di verbi come *leave-stroll*: essa risulta avere mediamente una distanza pari a 4, secondo i parlanti anglofoni madrelingua, così come la coppia *leave-arrive*. Osservando solamente la media tra i giudizi, diremmo che le due coppie di verbi sono relativamente simili, in quanto la differenza media percepita tra *leave* e *arrive* è uguale a quella percepita tra *leave* e *stroll*. Ciò che sfugge, osservando solo le medie aritmetiche, è che i singoli valori attribuiti alla coppia *leave-arrive* sono stati {1;1;7;7}, mentre quelli attribuiti alla coppia *leave-stroll* sono stati {4;4;4;4}. La devianza rispetto alla media nel primo caso è massima, nel secondo caso è nulla. Quest'ultimo caso si è verificato in poche circostanze, per le seguenti coppie di verbi: LEAVE-ARRIVE (in ENGbyENG); PARTIRE-ARRIVARE (in ITAbyITA); FORGET-MEMORISE (in ENGbyITA); FORGET-REMEMBER (in ENGbyENG).

[10] Questa scala di misura viene tipicamente utilizzata in esperimenti nei quali si richiedono giudizi percettivi a soggetti umani (Field , Hole 2003).

Nei casi osservati, emerge una caratteristica intrinseca ai verbi: i loro significati sono opposti. Osservando le altre coppie di verbi con varianze alte, questa tendenza viene confermata. Se da un lato due verbi dal significato opposto sono simili sul piano paradigmatico e quindi più vicini tra loro, soprattutto in un contesto di tipo paradigmatico come la lista di verbi presentata, dall'altro due verbi dal significato opposto presi in isolamento si dispongono su due poli opposti. In questo caso la metafora della similarità semantica riprodotta attraverso la prossimità geometrica è chiaramente applicata, e trova una sua realizzazione anche il concetto di embodiment, per cui la distanza fisica tra due oggetti viene percepita dal corpo e influenza il modo di concettualizzare e rappresentare la lingua nella mente.

Oltre alle coppie di verbi illustrate, sembra esserci una varianza alta, in entrambe le lingue e per entrambi i tipi di partecipanti, tra le coppie di verbi di cui uno dei due può fungere da supporto per l'altro, creando così un sintagma (o un idioma, o una perifrasi). Il fatto che i due verbi possano ricorrere uno accanto all'altro e creare una collocazione plausibile e di uso abbastanza frequente (come ad esempio *love to travel* o *desiderare correre*) ha fatto percepire ai parlanti una vicinanza intrinseca tra i due verbi. La varianza alta in questo tipo di coppie verbali è probabilmente dovuta a due fattori: innanzitutto può essere che non tutti i partecipanti abbiano considerato queste collocazioni come un indice valido per abbassare la distanza semantica percepita tra i due verbi che le compongono; in secondo luogo non dobbiamo dimenticare che metà dei giudizi relativi ad ognuna delle coppie derivano dalla presentazione dei due verbi nell'altro ordine, quindi non, ad esempio, *fall-love*, ma *love-fall*; e potrebbe darsi che in questi casi non sia "scattato" il meccanismo della collocazione perché i due verbi, letti in quell'ordine, non creavano collocazioni particolari. In particolare questo fenomeno si è manifestato spesso per i verbi *love*, *wish*, *desiderare* e *apprezzare*, in prima posizione (ad esempio in *love-travel*, *wish-escape*, *apprezzare-passeggiare* e *desiderare-ballare*).

Dal calcolo della similarità media percepita tra due verbi, osserviamo che anche in questo caso due verbi omogenei, cioè appartenenti allo stesso dominio concettuale, sono percepiti molto più simili rispetto a due verbi eterogenei, come mostrano i dati riportati nella tabella 2.

M, SD per ogni gruppo di verbi	ENGbyENG	ENGbyITA	ITAbyITA	ITAbyENG
MOT/MOT 276 coppie	M = 0,44 SD = 0,15	M = 0,50 SD = 0,19	M = 0,48 SD= 0,14	M = 0,47 SD = 0,19
PSY/PSY 276 coppie	M = 0,46 SD = 0,17	M = 0,48 SD = 0,19	M = 0,48 SD= 0,16	M = 0,43 SD = 0,20
COPPIE MISTE (576 coppie)	M = 0,28 SD = 0,12	M = 0,23 SD = 0,12	M = 0,27 SD = 0,12	M = 0,21 SD = 0,14

Tabella 2 i valori medi di similarità elicitati dai parlanti in relazione al tipo di coppia di verbi.

In seguito al calcolo della similarità media per ogni coppia di verbi, le 4 matrici ottenute riportavano tipologie di dati confrontabili a quelli rappresentati nei modelli distribuzionali: da un lato le similarità semantiche espresse su una scala da 1 a 7, dall'altro le prossimità geometriche, espresse attraverso i coseni (da zero a 1). Il confronto tra le similarità semantiche che caratterizzano la rappresentazione lessicale di ogni verbo nello spazio distribuzionale e quelle che caratterizzano la rappresentazione lessicale dello stesso verbo negli spazi semantici in L1 e in L2 è stata effettuata attraverso il calcolo della correlazione (coefficiente di Pearson) tra i modelli. Il confronto tra le

correlazioni di ogni verbo nel modello di origine computazionale e in quello di origine cognitiva ha originato interessanti elementi di discussione, in quanto sia per l'italiano che per l'inglese è stato possibile osservare le stesse tendenze. In particolare, sia il modello semantico distribuzionale italiano, che quello inglese, hanno rivelato gradi di correlazione medio-alti con i modelli creati con i giudizi dei parlanti; in alcuni casi, però, le rappresentazioni lessicali dei verbi erano più simili a quelle emerse dai giudizi forniti in L1, mentre in altri casi i giudizi forniti in L2 sembrano essere più congruenti alle rappresentazioni lessicali emerse dal corpus.

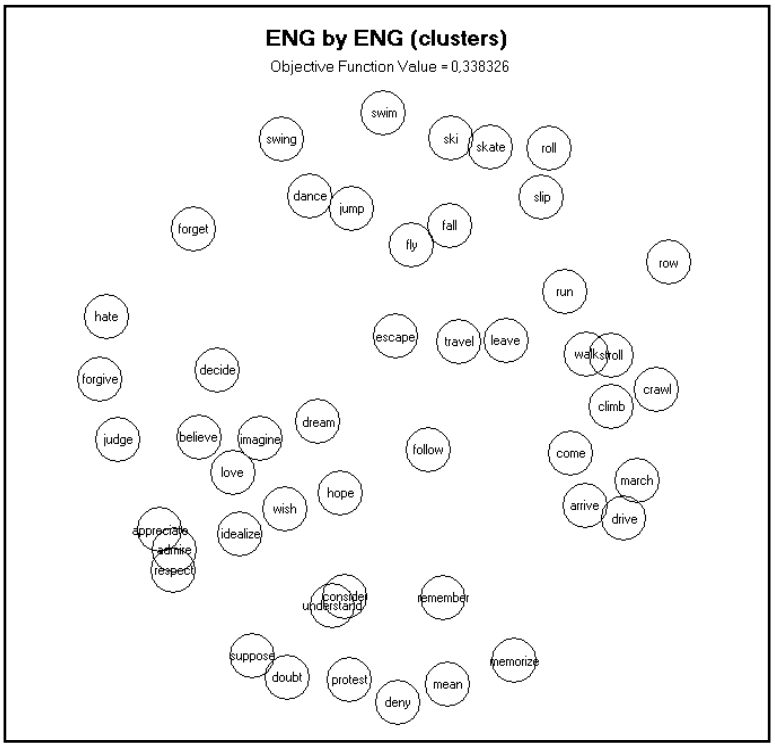

Figura 4 Visualizzazione grafica del modello ottenuto da ENGbyENG basata su Permap, un programma che utilizza la tecnica *multidimensional scaling*, dal quale emerge automaticamente la separazione dei due clusters verbali: in alto a destra i verbi di movimento, in basso a sinistra quelli di pensiero.

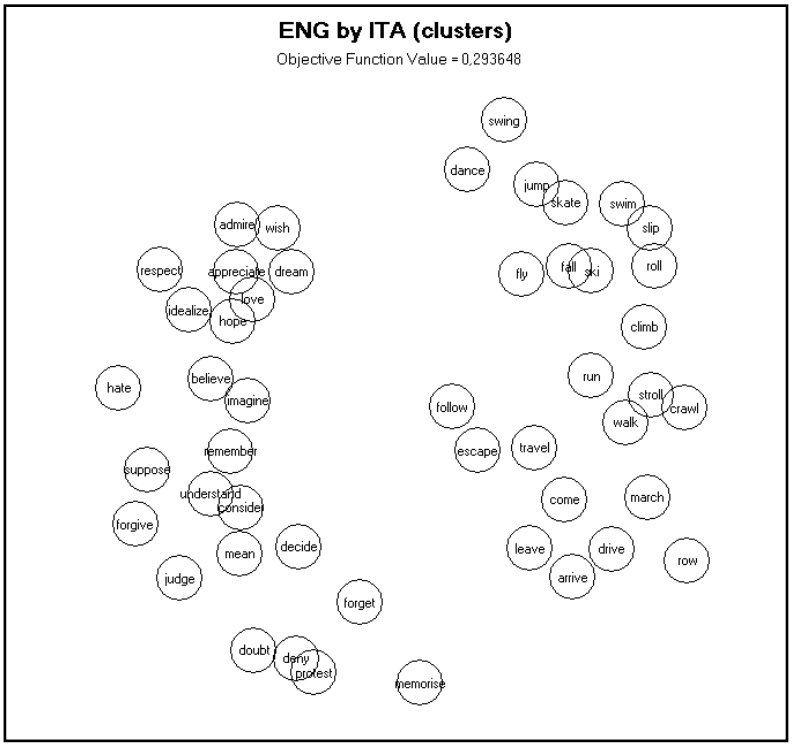

Figura 5 Visualizzazione grafica del modello ottenuto da ENGbyITA basata su Permap, un programma che utilizza la tecnica *multidimensional scaling*, dal quale emerge automaticamente la separazione dei due clusters verbali: in alto a destra i verbi di movimento, in basso a sinistra quelli di pensiero.

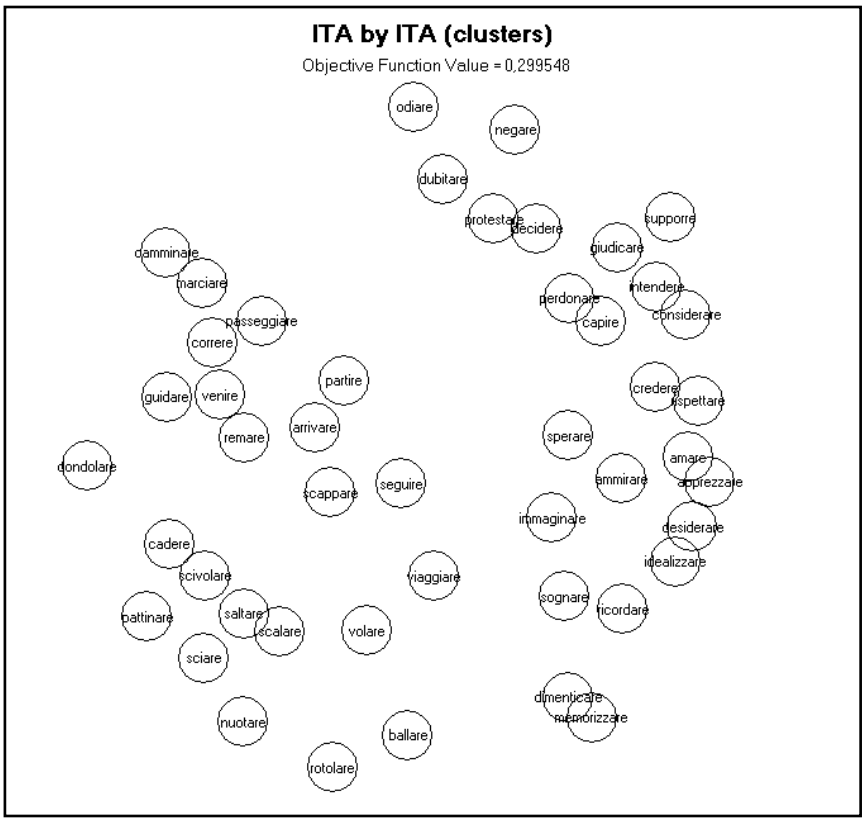

Figura 6. Visualizzazione grafica del modello ottenuto da ITAbyITA, basata su Permap, un programma che utilizza la tecnica *multidimensional scaling*, dal quale emerge automaticamente la separazione dei due clusters verbali: in alto a destra i verbi di pensiero, in basso a sinistra quelli di movimento.

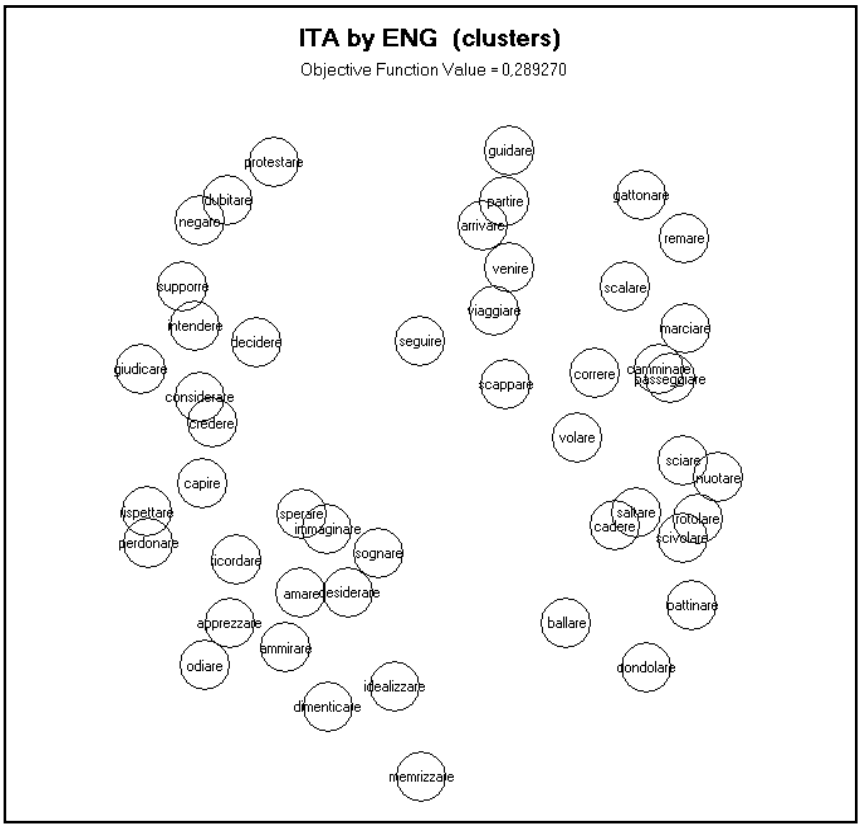

Figura 7. Visualizzazione grafica del modello ottenuto da ENGbyITA, basata su Permap, un programma che utilizza la tecnica *multidimensional scaling*, dal quale emerge automaticamente la separazione dei due clusters verbali: in alto a destra i verbi di movimento, in basso a sinistra quelli di pensiero.

Il confronto tra i dati estratti dai corpora e quelli elicitati dai parlanti

Osservando la tabella 3, dove sono riportati i coefficienti di correlazione per ogni verbo nei diversi modelli, si rileva una tendenza comune, che vede i verbi di pensiero correlarsi in maniera più forte ai giudizi forniti in L2, e i verbi di movimento più correlati ai giudizi in L1. Per visualizzare meglio questa tendenza sono state colorate di scuro le celle della tabella nella quale sono riportati i coefficienti di correlazione dei verbi di pensiero le cui correlazioni sono più alte con quelle in L2, e dei verbi di movimento le cui correlazioni sono più alte con le rappresentazioni lessicali in L1.

Correlazioni (coefficiente di Pearson)	Correlation WSM/ENG-L1	Correlation WSM/ENG-L2	Correlation WSM/ITA-L1	Correlation WSM/ITA-L2
ADMIRE/AMMIRARE	0,58	0,53	0,68	0,61

APPRECIATE/APPREZZARE	0,64	0,60	0,73	0,60
HATE/ODIARE	0,62	0,64	0,74	0,64
HOPE/SPERARE	0,53	0,60	0,51	0,60
BELIEVE/CREDERE	0,68	0,63	0,45	0,61
DECIDE/DECIDERE	0,57	0,63	0,53	0,59
DENY/NEGARE	0,60	0,55	0,45	0,67
CONSIDER/CONSIDERARE	0,69	0,68	0,61	0,68
FORGET/DIMENTICARE	0,40	0,63	0,36	0,51
FORGIVE/PERDONARE	*0,54*	*0,54*	0,50	0,56
DOUBT/DUBITARE	*0,69*	*0,69*	0,67	0,71
DREAM/SOGNARE	0,44	0,45	0,46	0,42
IDEALISE/IDEALIZZARE	0,53	0,56	0,53	0,57
IMAGINE/IMMAGINARE	0,62	0,69	0,41	0,51
JUDGE/GIUDICARE	0,61	0,63	0,53	0,57
LOVE/AMARE	0,48	0,60	0,61	0,53
MEAN/INTENDERE	*0,69*	*0,69*	0,68	0,70
MEMORISE/MEMORIZZARE	0,65	0,73	0,60	0,64
PROTEST/PROTESTARE	0,54	0,51	0,44	0,54
REMEMBER/RICORDARE	0,59	0,66	0,42	0,46
RESPECT/RISPETTARE	0,47	0,52	0,51	0,58
SUPPOSE/SUPPORRE	0,69	0,68	0,59	0,68
UNDERSTAND/CAPIRE	0,57	0,63	0,42	0,54
WISH/DESIDERARE	0,59	0,63	0,71	0,68
JUMP/SALTARE	0,59	0,46	0,52	0,49
LEAVE/PARTIRE	*0,49*	*0,49*	0,51	0,46
MARCH/MARCIARE	0,61	0,53	0,55	0,61
FALL/CADERE	0,52	0,38	0,50	0,49
FLY/VOLARE	0,52	0,43	0,52	0,48
ROLL/ROTOLARE	0,66	0,49	0,71	0,65
ROW/REMARE	0,75	0,55	0,60	0,65
RUN/CORRERE	0,49	0,35	0,48	0,38
SKATE/PATTINARE	0,56	0,47	0,58	0,59
SKI/SCIARE	0,61	0,46	0,48	0,37
SLIP/SCIVOLARE	0,54	0,49	0,55	0,47
STROLL/PASSEGGIARE	0,64	0,61	0,59	0,55
SWIM/NUOTARE	0,67	0,48	0,57	0,51
SWING/DONDOLARE	0,69	0,61	0,68	0,65
TRAVEL/VIAGGIARE	0,54	0,49	0,37	0,42
WALK/CAMMINARE	0,59	0,55	0,73	0,59
ARRIVE/ARRIVARE	0,60	0,53	0,52	0,54
CLIMB/SCALARE	0,57	0,50	0,58	0,48
COME/VENIRE	0,47	0,40	0,43	0,43
FOLLOW/SEGUIRE	0,49	0,42	0,51	0,47
CRAWL/GATTONARE	0,57	0,51	-	-
DANCE/BALLARE	0,59	0,48	0,56	0,52
DRIVE/GUIDARE	0,67	0,59	0,51	0,51
ESCAPE/SCAPPARE	0,45	0,41	0,45	0,36

Tabella 3. I coefficienti di correlazione tra la rappresentazione di ogni verbo nel modello WSM e nei modelli ottenuti con i dati di origine cognitiva.

Possiamo osservare che in generale (soprattutto per la lingua inglese, la cui analisi è stata basata su un corpus più ampio e bilanciato) i modelli WSM sono molto congruenti con le rappresentazioni lessicali delle parole sia in L1 e in L2, in quanto i valori dei coefficienti non scendono mai sotto 0.35. Inoltre, calcolando i coefficienti di determinazione delle correlazioni, al fine di poterne riportare un valore medio, si osserva che la correlazione media tra i verbi di movimento nel modello computazionale e in quello creato con i giudizi degli apprendenti è significativamente più bassa rispetto a tutte le altre correlazioni medie, sia per l'italiano che per l'inglese. Ma perchè proprio i verbi di movimento, sia in italiano che in inglese, costituirebbero un difficile scoglio per l'apprendimento lessicale, e genererebbero di conseguenza rappresentazioni lessicali poco correlate a quelle che emergono dall'analisi dei contesti d'uso, nei corpora? Una possibile risposta a questo interrogativo è da ricercarsi negli usi metaforici, nei quali i verbi di movimento spessissimo vengono utilizzati. Osservando, infatti, le prime tuple estratte per ogni verbo, quelle cioè più frequenti e salienti, con una misura di associazione più alta, la quantità di significati metaforici estratti per ogni verbo è subito evidenziata.

Nel lessico mentale degli apprendenti stranieri, sia di italiano che di inglese, i significati metaforici dei verbi di movimento, che nelle distribuzioni dei WSM sono trattati esattamente allo stesso modo rispetto ai significati letterali, non sono altrettanto salienti e quindi presenti. Per questo motivo, gli utilizzi metaforici (molto frequenti in L1 e nei contesti di occorrenza) che contribuiscono a creare il significato complesso dei verbi di movimento e di conseguenza le rappresentazioni lessicali riportate nei modelli distribuzionali, fanno in modo che siano proprio i parlanti madrelingua a generare giudizi di similarità tra verbi che assomigliano maggiormente a quelli estratti dai modelli WSM.

Conclusioni

Concludendo, attraverso un'analisi empirica complessa, che ha seguito parallelamente due metodi sperimentali diversi, per poi mettere a confronto i risultati ottenuti, è stato possibile osservare alcune dinamiche relative al reperimento di informazioni semantiche dai contesti d'uso delle parole stesse. A questo proposito, in relazione ai domini concettuali dei verbi di movimento e dei verbi di pensiero, le similarità semantiche percepite da parlanti madrelingua e apprendenti stranieri risultano correlarsi in maniera discretamente forte con le similarità paradigmatiche che emergono dall'implementazione dei modelli distribuzionali creati con Distributional Memory. In particolare è stato possibile osservare che le rappresentazioni lessicali sembrano essere maggiormente correlate a quelle che emergono dai dati raccolti con apprendenti stranieri, sia in italiano che in inglese. Questo potrebbe portare a pensare che l'affidamento sul contesto per il reperimento di informazioni semantiche è un meccanismo che caratterizza i primi stadi della memorizzazione delle entrate lessicali, e di conseguenza le rappresentazioni delle parole nel lessico mentale che sono state recentemente apprese. In seguito, invece, le stesse rappresentazioni lessicali potrebbero subire rielaborazioni semantiche basate su processi cognitivi di tipo top-down, che provocherebbero un leggero distacco dalle rappresentazioni iniziali, più legate ai contesti di occorrenza e d'uso. Questo fenomeno, comprovato dall'osservazione dei coefficienti di correlazione relativi ai verbi di pensiero, non sarebbe altrettanto evidente con i verbi di movimento in quanto questi ultimi tendono ad apparire molto spesso in contesti che ne modulano significati non letterali, i quali a loro volta non sono salienti e quindi facilmente acquisibili da parte degli apprendenti stranieri, nonostante essi siano molto frequenti nell'input.

Dal punto di vista delle applicazioni glottodidattiche, questo tipo di analisi ha mostrato come l'affidamento sul contesto (anche propriamente linguistico) sia effettivamente molto peculiare nel processo di apprendimento di una lingua straniera, in quanto le rappresentazioni lessicali delle parole in L2 nella mente degli apprendenti generano livelli di correlazione alti quando vengono confrontate con le rappresentazioni lessicali che emergono negli spazi semantici creati attraverso

analisi distribuzionali dei contesti di occorrenza. I modelli WSM offrono, dunque, uno strumento potente per l'analisi dei contesti di occorrenza delle parole, che può contribuire a far luce sui meccanismi cognitivi che caratterizzano l'apprendimento e il reperimento di informazioni semantiche dai contesti linguistici. Inoltre, dallo studio presentato, emerge la questione relativa agli utilizzi non letterali di molti verbi di movimento, fenomeno presente in modo massivo all'interno della lingua d'uso, ma spesso non altrettanto presente nei manuali di didattica delle lingue. Gli utilizzi non letterali dei verbi di movimento, così come il linguaggio non letterale in generale, sottintendono sia concettualizzazioni che strutture linguistiche che possono variare da lingua a lingua e da cultura a cultura; costituendo tutt'oggi uno scoglio non ancora affrontato in modo metodico, nel panorama della glottodidattica. La tendenza a non affrontare la suddetta questione è probabilmente legata al fatto che i significati metaforici vengono tradizionalmente considerati come secondari, minori e più difficili. Ciò nonostante, essi fanno parte del nostro linguaggio quotidiano, e dal momento che spesso sfuggono all'equivalenza lessicale che ci porta a tradurre mentalmente le parole da L2 a L1, costituiscono un aspetto del lessico mentale complesso e strettamente legato alla concettualizzazione che soggiace alla L2. Tra gli sviluppi futuri legati all'analisi delle dinamiche dell'apprendimento e strutturazione dei significati nel lessico mentale bilingue, lo studio del linguaggio non letterale in L2 potrebbe rivelare interessanti fenomeni cognitivi. Allo stesso modo, l'utilizzo di strumenti basati su corpora, utili per modellare porzioni di lessico, apre nuove frontiere all'analisi: in particolare gli studi più recenti stanno iniziando a prendere in considerazione corpora non solo di produzioni verbali, ma anche di immagini, al fine di investigare i meccanismi di estrazione di informazioni semantiche da input non verbali. Concludendo, nell'ottica di perseguire uno studio scientifico del lessico mentale, necessario ad attuare metodi ed approcci didattici sempre più efficaci e naturali, si auspica dunque che le prospettive e le questioni sollevate in questo studio possano contribuire a generare nuove ricerche e nuovi sviluppi nell'ambito della glottodidattica.

Bibliografia

Baroni M, Lenci A., 2010, "Distributional Memory: A general framework for corpus-based semantics", in *Computational Linguistics*, 36, 4, pp. 1-49.
Barsalou L.W., 1999, "Perceptions of Perceptual Symbols" in Behavioral and Brain Sciences, 22, 4, pp. 637-660.
Bolognesi M., 2010 "Il lessico mentale bilingue: i legami semantici e quelli episodici", in *Studi di glottodidattica*.
Bullinaria J.A., Levy J.P., 2007, "Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study", in *Behavior Research Methods*, 39, pp. 510-526.
Cardona M., 2008, "L'abilità di lettura e lo sviluppo della competenza lessicale", in *Studi di Glottodidattica*, 2, pp. 10-36.
Church K.W., Hanks P., 1990, "Word association norms, *mutual information* and lexicography", in *Computational Linguistics*, 16 , 1, pp. 22-29.
Field A., Hole G., 2003, *How to design and report experiments.* London, Sage.
Forster K.I., Jiang N., 2001, "The nature of the bilingual lexicon: experiments with the masked priming paradigm", in Nicol J.L (a cura di) *One Mind, Two languages: Bilingual Language Processing*, pp. 72-83, Oxford, Blackwell.
Glenberg A.M., Robertson D.A., 2000, "Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning", in *Journal of Memory and Language*, 43, pp. 379-401.
Harnad S., 1990, "The Symbol Grounding Problem", in *Physica D*, 42, pp.335-346.
Jarmasz M., Szpakowicz S., 2003, "Roget's thesaurus and semantic similarity", in Angelova G., Bontcheva K., Mitkov R., Nicolov N. (a cura di) *Proceedings* of *Conference on recent advances in Natural Language Processing*, pp. 212–219.
Lakoff G., Johnson M., 1980, *Metaphors We Live By.* Chicago, University Press.
Landauer T.K., Dumais S.T., 1997, "A Solution to Plato's Problem: the Latent Semantic Analysis. Theory of Acquisition, Induction and Representation of Knowledge", in *Psychological Review*, 104, 2, pp. 211-240.
Lenci A., 2009, "Spazi di parole: metafore e rappresentazioni semantiche", in *Paradigmi*, 27, pp. 83-100.
Lund K., Burgess C., 1996, "Producing high-dimensional semantic spaces from lexical co-occurrence", in *Behaviour Research Methods*, 28, pp. 203–208.
Meara P., 2009, *Connected Words: Word Associations and Second Language Vocabulary Acquisition,* Amsterdam, Benjamins.
Miller G.A., Charles W.G., 1991, "Contextual Correlates of Semantic Similarity", in *Language and cognitive Processes*, 6, 1, pp. 1-28.
Oxford Paravia, Il dizionario Italiano-Inglese, Inglese-Italiano , 2010, Torino, Paravia.
Padó S., Lapata M., 2007, "Dependency-based construction of semantic space models", in *Computational Linguistics*, 33, 2, pp. 161-199.
Perfetti C. A., 1998, "Two basic questions about reading and learning to read", in Reitsma P, Verhoeven L. (a cura di), *Problems and interventions in literacy development*, pp. 15-47.
Rapp R., 2003, "Word sense discovery based on sense descriptor dissimilarity", in *Proceedings of the 9th Machine Translation Summit*, pp. 315–322.
Rubenstein H., Goodenough J., 1965, "Contextual correlates of synonymy", in *Communications of the Association for Computing Machinery*, 8, 10, pp. 627-633.
Rubenstein H., Goodenough J., 1965, "Contextual correlates of synonymy", in *Communications of the Association for Computing Machinery*, 8, 10, pp. 627-633.

Sahlgren M., 2006, *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in Highdimensional Vector Spaces*, Ph.D. dissertation, Department of Linguistics, Stockholm University.
Sahlgren M., Karlgren J. ,2001, "Vector-Based Semantic Analysis Using Random Indexing for Cross-lingual Query Expansion", in Peters C., Braschler M., Gonzalo J., Kluck M. (a cura di) *Evaluation of Cross-Language Information Retrieval Systems*, pp. 169-176, Darmstat, Springer.
Segalowitz N., De Almeida R.G., 2002, "Conceptual representation of verbs in bilinguals: Semantic field effects and a second language performance paradox", in *Brain and Language*, 81, pp. 517-531.
Sinclair J., 1987, *Introduction to the Collins Cobuild English Language Dictionary*. London, Collins.
Stubbs M., 1995, "Collocations and semantic profiles: On the cause of the trouble with quantitative studies" in *Functions of Language*, 2, 1, pp. 23-55.
Turney P.D., 2008, "A uniform approach to analogies, synonyms, antonyms, and associations", in *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 905-912.