MODELLI DISTRIBUZIONALI DEL LESSICO

METODI COMPUTAZIONALI PER L'ANALISI SEMANTICA

di Alessandro Lenci

In questo contributo verranno presentati alcuni metodi di semantica computazionale che rappresentano aspetti del significato lessicale attraverso le distribuzioni statistiche delle parole estratte da corpora testuali. Questi modelli sono in grado di offrire un'alternativa credibile ai modelli classici della rappresentazione del significato, permettendo di affrontare un ampio numero di compiti semantici, quali l'identificazione di sinonimi, la classificazione di relazioni semantiche, il riconoscimento di analogie, l'acquisizione del lessico, ecc.

In this paper, I will present recent computational semantic methods that represent salient aspects of lexical meaning with corpus-derived statistics about word distributions. These models can represent a viable alternative to classical methods for semantic representation. In fact, they have been proven to be able to tackle a wide range of semantic tasks, such as synonym identification, semantic relation classification, analogy recognition, lexical acquisition, etc.

La semantica distribuzionale è una famiglia di approcci all'analisi del significato (con particolare attenzione alla dimensione lessicale) nati in linguistica computazionale e nelle scienze cognitive. Tali modelli condividono una prospettiva empiristica e assumono che la distribuzione statistica delle parole nei contesti linguistici giochi un ruolo chiave nel caratterizzarne il comportamento semantico [1]. Al di là di questa assun-

[1] Cfr. A. Lenci, *Distributional semantics in linguistic and cognitive research*, in A. Lenci (ed.), *From context to meaning. Distributional models of the lexicon in linguistics and cognitive science*, special issue of the "Italian Journal of Linguistics" 20/1 (2008), pp. 1-31.

zione condivisa, i modelli di semantica distribuzionale differiscono per le tecniche matematiche e computazionali impiegate per estrarre e modellare le statistiche di co-occorrenza delle parole dai *corpora*, per le proprietà semantiche che cercano di rappresentare distribuzionalmente e per la stessa caratterizzazione dei contesti linguistici usati per determinare lo spazio combinatorio dei termini lessicali. Nonostante queste differenze, tuttavia, un'analisi più ravvicinata e meno superficiale della galassia delle semantica distribuzionale ci permette di scoprire dietro le apparenti divergenze un modello generale del significato lessicale, un modello che formula ipotesi specifiche e sperimentalmente verificabili sul formato delle rappresentazioni semantiche e sul modo in cui vengono costruite. Nelle sezioni che seguono verranno illustrati i principi di base della semantica distribuzionale, come esempio significativo delle potenzialità che le più recenti tecniche di analisi matematica e computazionale del testo possono offrire per la ricerca linguistica e cognitiva.

1. Semantica distribuzionale

Nel paradigma distribuzionale della rappresentazione semantica, il lessico viene concepito come uno spazio metrico i cui elementi – le parole – sono separati da distanze che dipendono dal loro grado di similarità semantica. Quest'ultima viene misurata attraverso le distribuzioni statistiche di co-occorrenza delle parole nei testi, assumendo come principio epistemologico fondamentale la cosiddetta ipotesi distribuzionale, secondo la quale due parole sono tanto più semanticamente simili, quanto più tendono a ricorrere in contesti linguistici simili [2]. L'ipotesi distribuzionale è strettamente correlata alle *discovery procedures* tipiche della tradizione strutturalista [3], e più in generale è l'erede diretta di una tradizione associazionista e combinatoria che assume come chiave fondamentale per esplorare le proprietà paradigmatiche del les-

[2] Cfr. G. A. Miller, W. G. Charles, *Contextual correlates of semantic similarity*, "Language and Cognitive Processes" 6 (1991), pp. 1-28.

[3] Cfr. Z. Harris, *Mathematical Structures of Language*, Wiley, New York 1968.

sico la ricostruzione dei rapporti sintagmatici che intercorrono tra i suoi elementi nei contesti linguistici. Tale modello trova la sua caratterizzazione più icastica nelle parole del linguista inglese J. R. Firth: *You shall know a word by the company it keeps* [4]. Sul piano cognitivo, questo corrisponde a un modello del lessico mentale in cui i significati non sono organizzati come le definizioni dei sensi di un dizionario, ma secondo rappresentazioni contestuali: *an abstraction of information in the set of natural linguistic contexts in which a word occurs* [5].

Nonostante la sua lunga storia, l'ipotesi distribuzionale ha guadagnato nuovo slancio grazie alla disponibilità di *corpora* testuali di grandi dimensioni e di tecniche statistiche e informatiche più sofisticate per estrarre gli schemi distribuzionali dei lessemi. Questo ha consentito di tradurre l'ipotesi distribuzionale in modelli computazionali per la costruzione di spazi semantico – lessicali, che sono stati applicati alla simulazione di aspetti diversi della competenza semantica. Attualmente sono disponibili vari modelli di semantica distribuzionale, i più noti dei quali sono *Latent Semantic Analysis* (LSA) [6], *Hyperspace Analogue to Language* (HAL) [7], e più di recente *Random Indexing* [8]. Le rappresentazioni semantiche basate su spazi distribuzionali sono state utilizzate per modellare la selezione di termini sinonimi [9], il *priming* semantico [10],

---

[4] J. R. Firth, *Papers in Linguistics*, Oxford University Press, London 1957, p. 11.

[5] W. G. Charles, *Contextual correlates of meaning*, "Applied Psycholinguistics" 21 (2000), pp. 505-524; p. 507.

[6] Cfr. Th. K. Landauer, S. T. Dumais, *A Solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge*, "Psychological Review" 104/2 (1997), pp. 211-240.

[7] Cfr. K. Lund, C. Burgess, R. A. Atchley, *Semantic and associative priming in high-dimensional semantic space.* in *Proceedings of the Cognitive Science Society*, Erlbaum Publishers, Hillsdale N.J. 1995, pp. 660-665.

[8] Cfr. J. Karlgren, M. Sahlgren, *From words to understanding*, in Y. Uesaka, P. Kanerva e H. Asoh (eds), *Foundations of real-world intelligence*, CSLI, Stanford 2001, pp. 294-308.

[9] Cfr. Th. K. Landauer, S. T. Dumais, *op. cit.*

[10] Cfr. M. N. Jones, W. Kintsch, D. J. K. Mewhort, *High dimensional semantic space accounts of priming*, "Journal of memory and Language", 55 (2006), pp. 534-552. Il *priming* semantico è un effetto osservato in esperimenti psicolinguistici nei quali ai soggetti viene chiesto di svolgere un compito lessicale, come decidere se una certa stringa di caratteri è una parola oppure no. Il compito è svolto più velocemente se la

il ragionamento analogico [11], i giudizi di similarità semantica [12], e come vedremo più avanti la comprensione della metafora [13]. I modelli distribuzionali sono stati applicati anche all'acquisizione lessicale, simulando l'espansione del vocabolario da parte del bambino attraverso un processo di induzione del contenuto semantico delle parole da statistiche di co-occorrenza nell'input dell'adulto [14].

1.1. Costruire spazi semantici distribuzionali

Il tratto che accomuna le diverse realizzazioni dell'ipotesi distribuzionale è l'assunto che quantificare la similarità semantica tra due parole sia equivalente di fatto a valutare la misura in cui si sovrappongono i contesti linguistici nei quali esse ricorrono. All'interno di questo schema generale, i modelli differiscono per vari parametri, sia rappresentazionali che algoritmici, spesso legati anche alle diverse finalità teoriche o applicative che fanno da riferimento a ciascun modello.

La nozione di *spazio semantico* si basa su una semplice analogia con lo spazio geometrico. Come ciascun punto dello spazio è definito da un vettore di *n* numeri che rappresentano le sue coordinate rispetto a *n* assi cartesiani (le dimensioni dello spazio), così il contenuto seman-

---

parola *target* viene presentata ai soggetti dopo un'altra parola (detta *prime*) a essa semanticamente correlata. Ad esempio, la parola *dottore* è riconosciuta più velocemente se è preceduta dalla parola *infermiera*, piuttosto che dalla parola *torta*. Gli effetti di *priming* semantico sono usati come evidenza comportamentale dei legami associativi che strutturano il lessico mentale.

[11] Cfr. M. Ramscar, D. Yarlett, *Semantic grounding in models of analogy. An environmental approach*, "Cognitive Science" 27/1 (2003), pp. 41-71.

[12] Cfr. S. McDonald, M. Ramscar, *Testing the distributional hypothesis. The influence of context on judgements of semantic similarity*, in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, LEA, Edinburgh 2001, pp. 611-616.

[13] Cfr. W. Kintsch, *Metaphor comprehension. A computational theory*, "Psychonomic Bulletin & Review" 7 (2000), pp. 257-266.

[14] Cfr. P. Li, I. Farkas, B. MacWhinney, *Early lexical development in a self-organizing neural network*, "Neural Networks" 17 (2004), pp. 1345-1362; M. Baroni, A. Lenci, L. Onnis, *ISA meets Lara. A fully incremental word space model for cognitively plausible simulations of semantic learning*, in *Proceedings of the ACL Workshop on Cognitive Aspects of Language Acquisition*, Praga 29 Giugno 2007, pp. 49-56.

tico di una parola è rappresentato dalla sua posizione in uno spazio definito da un sistema di coordinate, determinato dai contesti linguistici in cui la parola può ricorrere. Formalmente, uno *spazio semantico di parole* è definito dalla quadrupla <*T*, *B*, *M*, *S*> [15]. *T* è l'insieme delle parole *target* che formano gli elementi che popolano lo spazio e di cui questo fornisce una rappresentazione semantica. *B* è la base che definisce le dimensioni dello spazio e contiene i contesti linguistici rispetto ai quali viene valutata la similarità distribuzionale delle parole *target*. *M* è una matrice di co-occorrenza che fornisce una rappresentazione vettoriale di ogni parola in *T*.

Come si è detto prima, alla base dei modelli di semantica distribuzionale risiede l'idea che due parole che tendono a combinarsi con elementi linguistici simili vengono anche a collocarsi in punti dello spazio semantico più vicini rispetto a quelli occupati da parole che invece si distribuiscono in maniera diversa nel testo. Questa assunzione è tipicamente formalizzata rappresentando ogni parola come un *vettore a n dimensioni*, ciascuna delle quali registra il numero di volte in cui la parola compare in un certo contesto definito dalla base *B*. Ogni parola *target* corrisponde, dunque, a una riga della matrice *M*, le cui colonne corrispondono invece agli elementi in *B*. Nel caso più semplice, il valore di una cella della matrice è equivalente alla frequenza di co-occorrenza della parola in un dato contesto.

La *Tabella 1* rappresenta il caso ipotetico in cui la parola *auto* ricorre 7 volte nel contesto di *guidare*, nel quale invece non compaiono mai né *gatto* né *panino* [16].

[15] Cfr. W. Lowe, *Towards a theory of semantic space*, in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, LEA, Philadelphia 2001, pp. 576-581; S. Padó, M. Lapata, *Dependency-based construction of semantic space models*, "Computational Linguistics" 33/2 (2007), pp. 161-199.

[16] In realtà, le celle possono anche contenere misure matematiche più sofisticate (es. mutua informazione, entropia, ecc.) che stimano il grado di salienza statistica della correlazione tra una parola *target* e un dato contesto linguistico.

Tabella 1. Matrice di co-occorrenza tra parole

	guidare	mangiare	aprire	miagolare	topo	gustoso
ministro	6	2	5	0	1	0
gatto	0	3	2	5	7	0
auto	10	0	2	0	0	0
panino	0	7	0	0	0	3

I modelli computazionali di semantica distribuzionale differiscono per la nozione di contesto che adottano, ovvero per il modo in cui viene definita la base *B*. Nella versione più comune, i vettori registrano *co-occorrenze tra parole di un testo*: una parola *target* w_i viene quindi rappresentata come un vettore in cui ciascuna dimensione d_{ij} registra il numero di volte in cui w_i ricorre all'interno di una finestra di n parole prima e dopo la parola w_j, dove n è un parametro fissato empiricamente. In questo caso, il numero delle dimensioni dei vettori è un sottoinsieme delle parole tipo di un testo [17]. In realtà, il modello degli spazi semantici non stabilisce nessun tipo di vincolo per quanto riguarda i tipi di elementi contestuali che formano la base. È possibile, infatti, progettare spazi semantici le cui dimensioni sono determinate da costruzioni linguistiche più astratte, come lemmi, relazioni di dipendenza sintattica, paragrafi, ecc. [18] Ad esempio in LSA, ogni documento di una collezione rappresenta uno specifico contesto e ciascuna dimensione di un vettore di una parola w_i registra il numero di volte in cui la parola ricorre in un certo documento [19].

[17] Questo sottoinsieme è tipicamente ottenuto selezionando le parole più frequenti all'interno del vocabolario del testo, a esclusione delle cosiddette *stopwords*, ovvero le parole ad alta frequenza appartenenti a classi chiuse come articoli, preposizioni, ausiliari ecc., che non sono significative per determinare le proprietà semantico-distribuzionali dei termini lessicali.

[18] Cfr. M. Baroni, A. Lenci, *Distributional Memory. A General Framework for Corpus-based Semantics*, "Computational Linguistics" (in stampa).

[19] Alcuni metodi distribuzionali rappresentano le parole in uno spazio semantico con un numero di dimensioni ridotto rispetto a quello definito dalla base originaria *B*. Ad esempio, LSA usa il metodo della *Singular Value Decomposition* (SVD) per ridurre i vettori della matrice *M* a vettori tipicamente di 100-300 dimensioni.

L'ultimo elemento che definisce la struttura dello spazio semantico è la metrica *S* che misura la distanza tra i suoi punti nello spazio. Per determinare la posizione di due parole, è necessario comparare i loro vettori rispetto a tutte le dimensioni che li costituiscono. Maggiore è il numero di dimensioni nelle quali due vettori presentano valori simili, maggiore è la loro vicinanza nello spazio e – data l'ipotesi distribuzionale – la similarità semantica delle parole corrispondenti. Una delle misure più comuni di vicinanza spaziale tra due vettori è il *coseno* dell'angolo che essi formano [20]. Se due vettori sono geometricamente allineati sulla stessa linea nella stessa direzione, l'angolo tra loro è 0°, e il coseno è 1 (massima similarità); viceversa, se i due vettori sono indipendenti (ortogonali), il loro angolo è vicino a 90° e il coseno di 90° è uguale a 0 (assenza di similarità). Alternativamente, la distanza tra i vettori può essere misurata utilizzando la classica *metrica euclidea*, generalizzata al caso dello spazio *n*-dimensionale [21]. Se i vettori sono normalizzati, il coseno produce un ordinamento di similarità equivalente a quello stabilito calcolando la distanza euclidea: in altri termini, se vogliamo sapere quale tra due parole w_i e w_j siano più *vicine* a una terza parola w_k, la distanza euclidea e il coseno ci forniscono la medesima risposta.

Th. K. Landauer, S. T. Dumais, *op. cit.* affermano che la trasformazione algebrica prodotta da SVD consente di proiettare le parole in uno spazio formato dalle dimensioni semantiche più salienti, *nascoste* nell'originaria matrice combinatoria, che ne rappresenta la manifestazione esteriore. Dal momento che SVD proietta le colonne di *M* che catturano contesti simili sulla stessa dimensione, la riduzione dello spazio semantico permette di catturare relazioni di similarità del *secondo ordine*, ovvero il fatto che due parole (es. *pane* e *pasta*) sono simili non solo perché ricorrono frequentemente nello stesso contesto (es. vicino alla parola *mangiare*), ma anche perché ricorrono in contesti che sono a loro volta simili (es. *mangiare* e *consumare*).

[20] Se due vettori sono normalizzati, il loro coseno è equivalente alla somma dei prodotti delle rispettive dimensioni. Un vettore si dice *normalizzato* se la sua lunghezza è uguale a 1. La *lunghezza* (o *norma*) di un vettore è uguale alla radice quadrata della somma dei quadrati delle sue dimensioni. Per normalizzare un vettore è sufficiente dividere ciascuna delle sue dimensioni per la norma del vettore.

[21] La *distanza euclidea* tra due vettori è uguale alla radice quadrata della somma dei quadrati delle differenze delle loro dimensioni.

2. SIGNIFICATI NELLO SPAZIO

Nei modelli semantici distribuzionali, il significato di una parola è totalmente e unicamente definito dalla sua posizione all'interno dello spazio multidimensionale determinato dalla base contestuale. Come afferma Kintsch,

> *Meaning … is a relation among words. In such a relational system, one cannot talk about the meaning of a word in isolation; words have meanings only in virtue of their relations to other words – meaning is a property of the system as a whole.* [22]

I modelli computazionali di semantica distribuzionale adottano così un modello di rappresentazione semantica radicalmente diverso da quello tipico della tradizione linguistica e cognitiva, fondato sull'uso di un metalinguaggio formale costituito da strutture simboliche, quali *frames*, tratti, reti semantiche, ecc. Il vettore assegnato a una parola, infatti, non ha alcun valore semantico intrinseco, ma serve solo a determinarne la posizione nello spazio e la distanza dalle altre parole. Il significato nasce solo dalle configurazioni di punti nello spazio, collocati secondo rapporti proporzionali al loro grado di similarità distribuzionale. La rappresentazione dello spazio semantico che ne deriva assomiglia a quelle mappe delle reti ferroviarie che, pur contenendo solo i punti corrispondenti alle fermate e i tratti di linea che le collegano, nondimeno lasciano intravedere la forma di un paese o di una città, grazie alla posizione reciproca degli elementi che le compongono. Alla stessa maniera delle mappe, gli spazi di parole rappresentano i significati come isomorfismi del secondo ordine [23].

Le dimensioni che formano il vettore non sono, dunque, direttamente interpretabili, né sono associabili a simboli concettuali, corrispondenti ai tratti tipici delle rappresentazioni semantiche tradizionali (es. ANIMATO, CONCRETO, ecc.):

---

[22] W. Kintsch, *Meaning in context*, in. Th. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsch (eds.), *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum, Mahwah NJ 2007, pp. 89-105; p. 91.

[23] Cfr. S. Edelman, *Representation is representation of similarities*, "Behavioral and Brain Sciences" 21 (1998), pp. 449-498.

> *as in a map, the coordinates are arbitrary. North and South are conventionally used for the earth, but the relation of any point to any other would be just as well located by any other set of nonidentical axes. LSA axes are not derived from human verbal descriptions … LSA's theory of meaning is that the underlying map is the primitive substratum that gives words meaning, not vice versa.* [24]

Questo fatto segna anche la differenza principale tra la metafora degli spazi di parole e gli *spazi concettuali* di Gärdenfors [25], che sono invece definiti da dimensioni che impongono una struttura topologica esplicita alla distribuzione dei dati empirici. Secondo il classico esempio di Gärdenfors, la struttura dello *spazio dei colori* è definito da tre dimensioni: tonalità, saturazione e luminosità. Il significato di ciascun termine di colore può dunque essere descritto attraverso un vettore a tre dimensioni i cui valori lo collocano in una specifica posizione rispetto agli assi dello spazio. Al contrario, nei modelli distribuzionali nessuna dimensione del vettore corrisponde a una proprietà o a una dimensione concettuale paragonabile a quelle che definiscono gli spazi di Gärdenfors. Le rappresentazioni vettoriali negli spazi di parole assomigliano piuttosto alle codifiche distribuite tipiche dei modelli connessionistici, nei quali il contenuto informativo è veicolato dalle configurazioni globali della rete neurale, piuttosto che dal valore di attivazione di una specifica unità. Ad esempio Rogers *et al.*, all'interno del paradigma PDP (*Parallel Distributed Processing*), propongono un modello computazionale della memoria semantica in cui

> *The conceptual representations are acquired by the network during learning and are not assigned by the computational modeller … The learned representations are not feature based but are instantiated as points in a high-dimensional space.* [26]

---

[24] Th. K. Landauer, *LSA as a theory of meaning*, in Th. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsch (eds.), *op. cit.*, pp. 3-34; pp. 7-8.

[25] Cfr. P. Gärdenfors, *Conceptual Spaces*, MIT Press, Cambridge MA 2000.

[26] T. T. Rogers, R. M. A. Lambon, P. Garrard, S. Bozeat, J. L. McClelland, J. R. Hodges, K. Patterson, *The structure and deterioration of semantic memory. A neuropsychological and computational investigation*, "Psychological Review" 111/1 (2004), pp. 205-235; p. 231.

Come gli stessi autori riconoscono, questo approccio è molto simile a quello dei *word spaces*, in quanto entrambi *extract high-order cooccurrence statistics across stimulus events in the environment* [27].
Nonostante la loro natura intrinsecamente relazionale, le rappresentazioni semantico-computazionali in termini di spazi di parole devono anche essere accuratamente distinte dai modelli relazionali basati su reti semantiche [28] o su reti lessicali stile *WordNet* [29]. Prima di tutto, gli elementi dello spazio sono parole e non entità concettuali o sensi come nelle reti semantiche. Il contenuto semantico di un lessema nasce solo dai suoi rapporti di similarità distribuzionale tradotti in distanze nello spazio. La differenza più sostanziale risiede nelle relazioni che le legano. Le connessioni tra i nodi delle reti semantiche sono, infatti, distinte sul piano *qualitativo* (iperonimia, meronimia, ecc.). Inoltre, sebbene ci siano stati molti tentativi di definire metriche per determinare la distanza tra due nodi concettuali da una rete, questa ha una struttura intrinsecamente discreta. Al contrario gli spazi di parole hanno una struttura continua e puramente *quantitativa*: come nello spazio geometrico, l'unica domanda legittima è *quanto* sono distanti due punti/parole.

3. Spazi di parole e dinamiche semantiche

Un aspetto importante della competenza lessicale catturato dai modelli semantico-distribuzionali è costituito dai giudizi di similarità semantica tra parole, come si evince anche dai dati riportati nella *Tabella 2*. Questa rappresenta una sorta di *distanziometro* per un insieme di nomi italiani: le celle della tabella sono coseni calcolati con *InfoMap*, una variante di LSA addestrata sul *corpus* lemmatizzato *La Repubblica* [30]. Mag-

[27] Ivi, p. 232.

[28] Cfr. M. R. Quillian, *Word concepts. A theory and simulation of some basic semantic capabilities*, "Behavioral Science" 12 (1967), pp. 410-430.

[29] C. Fellbaum, *WordNet. An electronic lexical database*, Cambridge University Press, Cambridge 1998.

[30] Il *Corpus La Repubblica* è una collezione di testi estratti dall'omonimo quotidiano, per un totale di circa 380 milioni di *tokens*.

giore è il valore del coseno, minore è la distanza tra le due parole nello spazio distribuzionale. La tabella mostra come le parole più simili dal punto di vista semantico (es. *cane* e *animale*, oppure *auto* e *aereo*) abbiano effettivamente un coseno più elevato (come evidenziato dai valori in grassetto). L'ipotesi distribuzionale sembra così trovare un'effettiva corrispondenza con le nostre intuizioni semantiche, e la similarità di significato tra due termini lessicali può dunque essere definita e misurata attraverso la loro proiezione in uno spazio costruito su base distribuzionale [31]. Più in generale, le rappresentazioni lessicali basate su spazi distribuzionali sono in grado di modellare vari tipi di evidenza comportamentale correlata alla distanza semantica tra parole (es. errori di interferenza in compiti di riconoscimento di parole, *priming* semantico, ecc.), in maniera più accurata di quanto possano fare modelli del lessico basati su rappresentazioni *simboliche* con reti lessicali, come ad esempio *WordNet* [32].

Al di là di questi risultati, uno dei contributi più interessanti offerto dalla semantica distribuzionale è dato in realtà dalla *rivoluzione copernicana* che essa realizza nel rapporto tra *significato* e *contesto*. Secondo una tradizione consolidata nelle scienze cognitive e in linguistica, rappresentare il contenuto semantico di una parola consiste nella sua proiezione su un'*ontologia di simboli concettuali*.

---

[31] Th. K. Landauer, S. T. Dumais, *op. cit.* hanno inoltre dimostrato sperimentalmente che LSA è in grado di ottenere prestazioni comparabili a quelle di soggetti umani nell'identificazione di parole sinonime. Il test standard per questo tipo di esperimenti è costituito dalla sezione di sinonimi del TOEFL (*Test of English as a Foreign Language*), che comprende 80 parole delle quali si deve individuare il sinonimo più appropriato, selezionandolo tra quattro alternative possibili. Gli autori riportano il 64.4% di accuratezza per LSA, del tutto comparabile all'accuratezza media del 64.5% raggiunta da un gruppo di studenti di inglese come lingua seconda. Modelli LSA più sofisticati e addestrati su *corpora* di maggiori dimensioni possono raggiungere anche un'accuratezza di oltre il 90%: cfr. R. Rapp, *A Freely Available Automatically Generated Thesaurus of Related Words.* in *Proceedings of LREC 2004*, ELRA, Lisbona 2004, pp. 395-398.

[32] Cfr. G. Vigliocco, D. P. Vinson, W. Lewis, M. F. Garrett, *Representing the meanings of object and action words. The featural and unitary semantic space hypothesis*, "Cognitive Psychology" 48 (2004), pp. 422-488.

Tabella 2. Coseni tra parole in uno spazio distribuzionale

animale	**0.53**						
sentimento	0.04	0.14					
odio	-0.01	0.05	**0.52**				
auto	0.22	0.10	-0.04	-0.07			
aereo	0.15	0.03	0.03	-0.03	**0.25**		
presidente	-0.03	-0.01	-0.02	0.04	-0.005	0.03	
ministro	0.04	0.07	-0.06	-0.03	0.002	0.02	**0.16**
	cane	**animale**	**sentimento**	**odio**	**auto**	**aereo**	**presidente**

Nell'ambito della rappresentazione della conoscenza e della linguistica computazionale, un'*ontologia* è la specificazione in un linguaggio formale di un sistema di categorie concettuali [33]. Nel caso specifico della descrizione lessicale, le ontologie sono sistemi di simboli che *rappresentano* il contenuto semantico dei lessemi. Significati diversi della stessa parola vengono rappresentati con elementi differenti dell'ontologia, mentre l'architettura del sistema di concetti si fa carico delle relazioni inferenziali tra i sensi delle parole. Il tratto più caratterizzante di questo tipo di rappresentazione è che i significati sono modellati come entità essenzialmente *indipendenti dal contesto*, ricordando da vicino la classica opposizione tra *competenza* e *uso*, tipica del paradigma generativo in linguistica. La maggior parte degli approcci simbolici alla rappresentazione del significato assumono, infatti, una parallela dicotomia tra la competenza semantica di una parola e il suo uso nei contesti linguistici.

Il significato lessicale è ovviamente soggetto a processi di acquisizione, modulazione e cambiamento, ma questi aspetti sono comunque indipendenti dal modo in cui viene rappresentata l'informazione semantica, per poi usarla e applicarla in contesto. Una delle conseguenze principali di tale paradigma è l'intrinseca difficoltà delle ontologie a modellare quei processi dinamici che si realizzano nel momento in cui i

[33] Cfr. Th. R. Gruber, *Toward principles for the design of ontologies used for knowledge sharing*, "International Journal of Human-Computer Studies" 43 (1995), pp. 907-928; N. Guarino, *Formal ontology in information systems*, in N. Guarino (ed.), *Formal Ontology in Information Systems. Proceedings of FOIS'98*, IOS Press, Amsterdam 1998, pp. 3-15.

significati si inverano nei concreti contesti testuali. I sensi delle parole sono infatti realtà multidimensionali, dai confini incerti e spesso sottodeterminati, che malamente si prestano a riduzionistiche proiezioni su sistemi di simboli concettuali non sufficientemente adeguati a rappresentarne la complessità strutturale interna e l'intrinseca variabilità. Come sottolineato in Pustejovsky [34], una rappresentazione lessicale soddisfacente deve essere in grado di rendere conto della natura proteiforme del lessico e delle sue dinamiche, che dipendono dai rapporti che si vengono a instaurare tra i lessemi sull'asse sintagmatico.

Nelle rappresentazioni semantiche in termini di ontologie di simboli concettuali, la funzione del contesto è di natura essenzialmente *discriminativa*. Esso agisce, infatti, come *fattore di disambiguazione*, consentendo la selezione all'interno del repertorio dei sensi di una parola del particolare significato appropriato a una specifica situazione di uso. La metafora degli spazi di parole ribalta questa prospettiva assegnando al contesto un ruolo *costitutivo* del significato:

> *To know the meaning of a word is to know if a given context is a member of the set satisfying the word's contextual representation.* [35]

Il contenuto informativo di una parola viene così a essere direttamente radicato nei contesti linguistici, dai quali emerge e dai quali viene plasmato. Ne consegue un modello diverso di rappresentazione semantica, *sensibile al contesto* (*context-sensitive*) e intrinsecamente dinamica, che offre nuove prospettive per reimpostare il rapporto tra rappresentazione del significato e modellazione delle sue dinamiche.

.

[34] Cfr. J. Pustejovsky, *The Generative Lexicon*, The MIT Press, Cambridge MA 1995.

[35] W. G. Charles, *Contextual Correlates of Meaning*, "Applied Psycholinguistics" 21 (2000), pp. 505-524; p. 507.