Luoghi fantastici e dove collocarli: una prospettiva informatico umanista.

Seminario di Cultura Digitale
Corso di Laurea Magistrale in Informatica Umanistica
Ludovica Pannitto, mat. 491094

Indice

Introduzione

> Marco Polo descrive un ponte, pietra per pietra.
> – Ma qual è la pietra che sostiene il ponte? – chiede Kublai Kan.
> – Il ponte non è sostenuto da questa o quella pietra, – risponde Marco, – ma dalla linea dell'arco che esse formano
>
> Italo Calvino, Le città invisibili

I seminari di cultura digitale incarnano lo *Speakers' Corner* del nostro corso di laurea: un luogo insieme fisico e virtuale in cui non è raro incontrare idee, ma soprattutto persone disposte a condividerle, lungo un tracciato che conserva il grande pregio di accogliere, senza pregiudizi, questa o quell'altra disciplina. Perché solo attraverso il dialogo è possibile muoversi, e solo muovendosi non si affonda mai.

Questo carattere irrequieto ed inclusivo delle digital humanities è ciò che a volte rende difficile tracciarne i confini. Escluse le più classiche, consolidate e forse prototipiche applicazioni, le aree di interesse si insinuano in tutti i campi della conoscenza. Oltre il trattamento dei contenuti culturali si scatena un'esplosione di combinazioni che mettono insieme conoscenza e organizzazione della conoscenza, realtà e rappresentazione.

Si ha allora la sensazione che la cultura digitale, un po' come il ponte di Marco Polo, non stia tanto in una definizione, in questa o in quella pietra, ma nella forza che le tiene tutte insieme. È per questo che questa relazione si colloca un po' ai margini dell'idea più prototipica di cultura digitale, esplorando un ambito ai margini di varie discipline (linguistica, geografia, scienze cognitive, informatica). L'idea di cultura digitale risiede allora proprio nell'ibridazione, nel bisogno che il geografo come il linguista o l'informatico possano avere vicendevolmente bisogno di comprendere gli strumenti dell'altro, strumenti fisici e strumenti concettuali.

Ho provato ad offrire una breve panoramica su alcune applicazioni della linguistica computazionale all'ambito geografico: è possibile inferire la collocazione geografica di un nome a partire dall'utilizzo linguistico di quel nome? Quello che leggiamo su New York ci basta per sapere che è più vicina a Philadelphia che a Los Angeles?

L'idea è impiegare tecniche ispirate alla teoria distribuzionale del significato, una teoria empirista che ha trovato nel periodo storico contemporaneo un campo fertile di proliferazione.
È stata infatti recentemente applicata con successo a task di natural language processing quali la creazione automatica di thesauri, l'espansione di query per i motori di ricerca, il paraphrasing, l'estrazione di informazione o la sentiment analysis.

Le applicazioni recenti non si limitano tuttavia all'ambito dell'analisi linguistica o del natural language engineering: lavori come quello presentato in Rodda et al. (2016), dove si utilizzano metodi algebrici per indagare eventi culturali come il cambiamento semantico, mostrano che la teoria semantica distribuzionale può fornire un valido supporto investigativo anche ad ambiti che non hanno tradizionalmente fatto uso di strumenti quantitativi o di teorie linguistico-cognitive.

Il campo si presenta piuttosto vasto, e necessiterebbe di un'ampia trattazione teorica non solo per introdurre adeguatamente le tecniche utilizzate ma anche per evidenziare problematiche che devono essere prese in considerazione durante studi del genere. Senza pretesa di esaustività, ma solo per dare un'idea degli ambiti che tali tematiche e problematiche toccano, ho deciso di evitare una lunga presentazione del tema e di lasciare spazio ai luoghi oggetto degli esperimenti, integrando volta per volta con i temi interessanti.

Luoghi Lontani

2.1 Cosa sappiamo di un luogo?

Montello and Freundschuh (1995) distinguono vari modi tramite i quali si può acquisire conoscenza geografica riguardo all'ambiente: questo può avvenire tramite esperienza diretta dell'ambiente (ad es. attraverso la locomozione o la vista), attraverso informazione figurativa statica (ad es. diagrammi, dipinti o fotografie) o dinamica (ad es. animazioni, video), o attraverso descrizioni verbali.
Nella cultura contemporanea l'informazione scritta è tuttavia preminente per quanto riguarda l'acquisizione di conoscenza: Louwerse and Zwaan (2009) partono da questa considerazione per indagare l'importanza dell'acquisizione di conoscenza geografica attraverso i testi che leggiamo.

Vari studi hanno investigato l'effetto del linguaggio nella formazione delle relazioni spaziali (Franklin and Tversky (1990); Taylor and Tversky (1992); Ferguson and Hegarty (1994)), confermando la capacità di costruire mappe spaziali da descrizioni verbali o percorsi espliciti. Molti degli studi, anche dal punto di vista computazionale, si sono tuttavia concentrati su ambienti **locali** (Canter and Tagg (1975); Vasardani et al. (2013)), ignorando l'impatto del testo sulla conoscenza geografica su scala **globale**.

La ricerca sulla costruzione soggettiva di rappresentazioni geografiche è anch'essa varia: studi su esperimenti di categorizzazione hanno proposto il ricorso a categorie intermedie quali quelle di stato o regione (Friedman and Montello (2006)), l'uso di euristiche come rotazione e allineamento (Tversky (1981)), o il ricorso a bias di vario tipo, che comprendono credenze, conoscenze geopolitiche o socio-culturali (Friedman et al. (2002)).

Louwerse and Zwaan (2009) propongono per studi di questo genere un'**ipotesi di prossimità linguistica** (*città che sono collocate insieme sono discusse insieme*) in continuità con la prima legge della geografia di Tobler (1970), che ipotizza un bias di prossimità nei giudizi dei parlanti: *everything is related to everything else, but near things are more related than distant things.*

Simili considerazioni si sono dimostrate valide per rappresentazioni mentali soggettive di altre proprietà di entità geografiche, come ad esempio la popolosità di città.
Esperimenti condotti da Simon (1999), Goldstein and Gigerenzer (2002) sulla stima della popolosità di città hanno mostrato che cittadini americani fornivano migliori giudizi se testati su città tedesche che su città statunitensi. I ricercatori spiegano questi risultati attribuendo ai soggetti quella che chiamano **recognition heuristic**: un dato inaccessibile, in questo caso la taglia della popolazione, sarebbe riflettuto da una variabile mediatrice (ad esempio la frequenza con cui una

città viene menzionata nelle news), e il mediatore influenzerebbe la probabilità di riconoscimento.

La variabile mediatrice potrebbe essere una rappresentazione statica o dinamica del fenomeno, compreso il linguaggio stesso.

2.2 USA, Cina, Medio Oriente

Una serie di studi hanno testato l'ipotesi di prossimità linguistica. Tra quelli che si sono occupati del fenomeno su scala globale, ho scelto l'indagine di Louwerse and Zwaan (2009) per gli Stati Uniti d'America e quella di Louwerse et al. (2012) per Cina e Medio Oriente.

Entrambi gli studi, analoghi nei metodi, si propongono di valutare la correlazione tra latitudine e longitudine delle 50 città più popolose di una data area e le coordinate stimate a partire da informazione testuale o giudizi umani.
Mentre in Louwerse and Zwaan (2009) il materiale utilizzato proviene da articoli di giornale del *Wall Street Journal*, *New York Times* e *Los Angeles Times*, in Louwerse et al. (2012) viene testata la possibilità di utilizzare materiale non giornalistico, e viene dunque utilizzato un corpus composto da saggistica, romanzi e libri in generale.
In tutti i casi il materiale raccolto non riguarda in particolare argomenti geografici né sono presenti significative porzioni di testo con descrizioni del territorio.

A partire dai testi è stato costruito uno spazio vettoriale (per le fasi di costruzione dello spazio si veda il paragrafo 2.4) tramite una particolare tecnica, detta *Latent Semantic Analysis*, che permette di portare alla luce regolarità latenti presenti nei dati. Dallo spazio è stata poi ricavata una matrice 50x50 contenente la similarità (nello specifico, il valore di coseno come descritto nella sezione 2.4.3) per ogni coppia di città prese in esame.
Tramite una tecnica matematica (*Multidimensional Scaling*, brevemente descritta nella sezione 2.4.3) è stata poi ottenuta una collocazione bidimensionale delle 50 città. Questo ha permesso di effettuare un'analisi di correlazione tramite regressione bidimensionale (Friedman and Kohler (2003)) tra le coordinate *predette* e le coordinate *reali*.

Le correlazioni, riportate in tabella 2.1 si dimostrano soddisfacenti.
L'estensione dell'analisi a luoghi non familiari come la geografia della Cina o del Medio Oriente e a generi testuali diversi dagli articoli di giornale dimostra inoltre che questa proprietà di *collocazione* è indipendente dalla lingua o dalla varietà linguistica.

Avendo a disposizione informazione localizzata per qualche simbolo, la localizzazione o, in un certo senso, il significato degli altri può essere indotto grazie all'organizzazione della rete di simboli. Le regolarità del linguaggio che vengono messe in atto durante il processing cognitivo fanno sì che, conoscendo la collocazione di una città, il parlante sia in grado di dedurre la collocazione geografica degli altri simboli.

Bisogna comunque osservare che ciò presuppone la localizzazione del primo o dei primi simboli, e la capacità di discriminare all'interno della rete quali sono i simboli da localizzare.
Inoltre è bene tenere a mente che le predizioni avvengono in questo modo in termini relativi e non secondo coordinate assolute, introducendo i problemi discussi nella sezione 4.2.

	r	p	n
Cina	0,57	<0,001	48
Cina – random Monte Carlo	0,13 (SD 0,06)	0,37	48
Medio Oriente	0,53	<0,001	50
Medio Oriente – random Monte Carlo	0,13 (SD 0,07)	0,37	50
USA – Wall Street Journal	0,529	<0,01	50
USA – New York Times	0,277	<0,05	50
USA – Los Angeles Times	0,427	<0,01	50
USA – Human Estimates	0,562	<0,001	50

Tabella 2.1: La tabella mostra i valori di correlazione, ottenuti tramite regressione bidimensionale, tra le coordinate predette e i valori reali di latitudine e longitudine.
Per l'area cinese e per il Medio Oriente, i valori sono stati confrontati con la correlazione ottenuta attraverso una simulazione random Monte Carlo (1000 run).
Per gli USA invece, sempre in Louwerse and Zwaan (2009) il risultato viene confrontato con la correlazione ottenuta tra le coordinate stimate da parlanti e le reali coordinate geografiche. Ulteriori analisi mostrate nello studio confermano la presenza di bias verso le aree di interesse dei giornali e rispetto alla frequenza con cui le città compaiono nel testo.

2.3 Ipotesi Distribuzionale

La giustificazione linguistica dello studio appena descritto sta nell'ipotesi secondo la quale il significato di una parola viene acquisito tramite l'esperienza linguistica che viene fatta dello stimolo. La semantica distribuzionale è una branca dello studio del significato che assume una tale prospettiva empirista, basata sull'assunzione che la distribuzione statistica delle osservazioni in contesto sia significativa per caratterizzare il loro contenuto semantico.

Le fondamenta teoriche risiedono nella **Distributional Hypotesis**, parafrasabile come segue: *lessemi con simili proprietà distribuzionali hanno significati simili.*

Precursori teorici della teoria semantica distribuzionale si riscontrano nel tardo Wittgenstein (1953), nel distribuzionalismo di Harris (1954), e nella più popolare postulazione di Firth (1957): *You shall know a word by the company it keeps.*

La storia dell'ipotesi distribuzionale inizia in effetti fuori dal campo della semantica, nella proposta del metodo distribuzionale di Harris per l'analisi fonologica e morfologica, al fine di fornire alle ipotesi linguistiche una salda base metodologica.

Rientrati nel campo della semantica, il significato di un lessema risiede quindi nella relazione che intercorre tra i lessemi che hanno una relazione sintagmatica con il lessema in oggetto.

Similmente, la psicologia comportamentista ha portato avanti una visione associazionista del significato basata sull'idea che associazioni o co-occorrenze di stimoli costituiscano una fonte primaria per l'apprendimento (Deese (1966)).
Molti studi, a partire da Rubenstein and Goodenough (1965), hanno mostrato la correlazione tra giudizi di similarità e sovrapposizione dei contesti linguistici.

Nelle scienze cognitive, uno dei principali sostenitori dell'ipotesi distrubuzionale di Harris è stato George Miller, utilizzando il distribuzionalismo come un metodo empirico di investigazione delle similarità semantiche.

In Miller and Charles (1991) si legge *a word's contextual representation [...] is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts. [...] Two words are semantically similar to the extent that their contextual representations are similar.*

Il contesto preso in considerazione è tipicamente il contesto linguistico, per motivi pratici e teorici: il contesto linguistico è facile da estrarre, avendo a disposizione corpora della varietà interessante, e attraverso il contesto linguistico è possibile investigare il ruolo della distribuzione statistica nella formazione del significato.

Viste le diverse anime che si sono affacciate sull'ipotesi distribuzionale, Lenci (2008) distingue tra una versione **debole** dell'ipotesi distribuzionale, come metodo empirico per l'analisi semantica, e una versione **forte**.

In questa versione forte l'ipotesi distribuzionale trova la sua realizzazione come un'*ipotesi cognitiva* sulla forma e origine delle rappresentazioni semantiche.
La frequenza con cui il parlante incontra i lessemi in differenti contesti porta alla formazione di una rappresentazione mentale come astrazione dai contesti più significativi dove il lessema è stato utilizzato. In questo modo il comportamento distribuzionale ha valore esplicativo rispetto al contenuto semantico presente a livello cognitivo.
Nelle parole di Miller and Charles (1991), *What people know when they know a word is not how to recite its dictionary definition – they know how to use it (when to produce it and how to understand it) in everyday discourse.*

È stato tuttavia notato che conoscere il significato richiede più che conoscere il comportamento linguistico di un lessema. Oltre alla capacità di navigare la rete di relazioni concettuali che legano i lessemi, parte della conoscenza di un concetto risiede nella capacità di mappare tali entità nel mondo reale. Questo problema, noto come *problema del grounding* (paragrafo 4.2), è ben noto in letteratura ma esula dagli scopi di questa breve introduzione. È inoltre strettamente legato alle critiche portate ai modelli distribuzionali dagli approcci così detti *embodied* della cognizione: teorie secondo le quali la natura delle rappresentazioni concettuali è dipendente dal sistema sensomotorio.
A questo proposito si stanno facendo avanti modelli semantici multimodali in ambito semantico distribuzionale (Feng and Lapata (2010); Bruni et al. (2014)): sono modelli che non si limitano alla rappresentazione distribuzionale di input linguistico, ma a questa affiancano informazione derivata da altre fonti, mettendo così in relazione distribuzioni derivate da contesti linguistico-testuali con informazioni di tipo più prettamente percettivo.

2.4 Distributional Semantics Computational Cookbook

Il framework più popolare di implementazione dell'ipotesi distribuzionale per l'analisi semantica sono i *Distributional Semantic Models (DSMs)*.

La rappresentazione distribuzionale di un item lessicale è un vettore n-dimensionale, le cui componenti sono features distribuzionali che rappresentano le co-occorrenze con contesti linguistici.
Data l'ipotesi distribuzionale, c'è una relazione tra la similarità distribuzionale dei lessemi e la similarità algebrica dei vettori costruiti, e dunque è possibile calcolare la prima misurando la seconda.

Le prime implementazioni computazionali dell'ipotesi distribuzionale furono sviluppate all'inizio degli anni '60 del '900, applicate all'information retrieval o alla costruzione di thesauri per la traduzione automatica.
Il *Vector Space Model*, introdotto in information retrieval da Salton et al. (1975), consiste nel rap-

presentare una collezione di documenti tramite una matrice termine - documento, dove le righe corrispondono a termini (per esempio lessemi) e le colonne a documenti dove questi appaiono. Ogni entrata della matrice tiene traccia in questo modo delle occorrenze di un termine in un documento. La matrice così costruita è stata inizialmente impiegata per calcolare la similarità tra documenti, secondo l'ipotesi per cui sono simili quei documenti che contengono in modo statisticamente rilevante gli stessi termini.
La stessa costruzione è stata poi impiegata per valutare la similarità tra termini.

Modelli classici (o modelli matriciali) estendono e generalizzano il *Vector Space Model* dell'information retrieval, da cui deriva l'uso di matrici di co-occorrenze per rappresentare informazione distribuzionale.

Come in tutte le ricette che si rispettino, elenchiamo una serie di ingredienti necessari alla costruzione di un DSM, e i passaggi necessari per la realizzazione:

Ingredienti

- un insieme di elementi target T, ovvero i lessemi per cui il DSM fornisce una rappresentazione contestuale
- un insieme di contesti C, con cui i target co-occorrono
- una funzione di peso dei contesti W, per distinguere ciò che è statisticamente rilevante da ciò che non lo è
- una matrice M di dimensioni $|T| \times |C|$, per tenere traccia delle co-occorrenze
- una funzione di riduzione di dimensionalità $R : M \to M'$
- una misura S di similarità tra vettori in M'

Preparazione

- a partire da un corpus si estraggono le co-occorrenze degli item lessicali con i contesti linguistici (paragrafo 2.4.1)
- a partire dalle frequenze di co-occorrenza, gli item si rappresentano algebricamente attraverso vettori distribuzionali (paragrafo 2.4.2)
- tramite la similarità tra vettori distribuzionali si misura la similarità semantica tra item lessicali (paragrafo 2.4.3)

2.4.1 Dal corpus alla matrice

La creazione del modello necessita di una serie di step linguistici che includono la scelta del corpus e il suo *preprocessing*, la selezione degli item target e la definizione dei contesti linguistici.

Per quanto riguarda la scelta del corpus due sono i parametri da tenere in considerazione: da un lato il tipo e la varietà linguistica di cui il corpus rappresenta un campione rappresentativo, dall'altro la sua dimensione. A causa della distribuzione zipfiana degli item linguistici (Zipf (1935)), infatti, la quantità di materiale a disposizione è un parametro potenzialmente problematico per la realizzazione di validi modelli distribuzionali.
La disponibilità dei dati è una questione particolarmente rilevante per le applicazioni dei modelli semantico distribuzionali: se infatti sono ormai disponibili grandi corpora *general-purpose* che permettono la costruzione di modelli stabili, corpora di ambito specifico hanno tipicamente dimensioni più ridotte, rendendo complessa l'estrazione di fenomeni linguisticamente rilevanti.

id	form	lemma	pos	feats	b/i	chunk type	chunk role	head	dep
1	Il	Il	RD	MS	B	N_C	DET	2	DET
2	danno	danno	S	MS	I	N_C	POTGOV	6	SUBJ_PASS
3	non	non	B	NULL	B	BE_C	PREMODIF	6	NEG
4	poteva	potere	V	S3II	I	BE_C	MOD	6	MODAL
5	essere	essere	V	F	I	BE_C	AUX	6	AUX
6	sottovalutato	sottovalutare	V	MSPR	I	BE_C	POTGOV	0	ROOT

Tabella 2.2: La tabella mostra l'analisi della frase *Il danno non poteva essere sottovalutato* fino al livello di analisi sintattica a dipendenze.

	bite	buy	drive	eat	get	live	park	ride
bike	0	9	0	0	12	0	8	6
car	0	13	8	0	15	0	5	0
dog	0	0	0	9	10	7	0	0
lion	6	0	0	1	8	3	0	0

Tabella 2.3: Nella matrice sono riportate sulle righe gli item lessicali e sulle colonne i contesti selezionati.

I dati linguistici devono essere poi tokenizzati ed eventualmente sottoposti a pipeline di analisi che possono includere livelli di crescente complessità (un esempio in tabella 2.2): lemmatizzazione, PoS tagging, analisi sintattica, Named Entities Recognition...
Tali analisi sono computazionalmente costose e richiedono la disponibilità di risorse specifiche di supporto. Introducono inoltre possibili errori, che vanno spesso a inficiare proprio i dati interessanti. È dunque necessario valutare il livello di analisi richiesto, a seconda della strategia scelta per l'identificazione degli item e la scelta dei contesti.

Occorre poi definire il tipo di contesto interessante per l'applicazione considerata. Come detto in precedenza, i modelli matriciali nati nell'information retrieval sono modelli di tipo *bag of words*, in cui l'intero documento è considerato come contesto dell'item.
Per l'analisi linguistica e linguistico-cognitiva tuttavia la scelta più comune è quella di utilizzare gli stessi lessemi come contesti.
L'insieme di contesti C è dunque tipicamente l'insieme degli n lessemi interessanti (selezionati a partire da una lista, o escludendo determinate parti del discorso o viceversa selezionandone solo alcune) più frequenti.

Avendo a questo punto a disposizione un insieme di lessemi target e un insieme di lessemi contesto, il passo decisivo per la costruzione della matrice di co-occorrenze è appunto definire una relazione di co-occorrenza.
Si differenziano essenzialmente relazioni basate su distanze lineari sul testo (collocati *window-based*), in cui l'item viene considerato se occorre all'interno di una finestra di contesto il cui span viene definito a priori, e relazioni basate su dipendenze sintattiche, in cui il contesto è legato al target da una relazione sintagmatica indipendentemente dalla distanza lineare.

A questo punto le istanze delle coppie distribuzionali target-contesto individuate devono essere conteggiate per ottenere la frequenza di co-occorrenza dei target con i contesti linguistici. I dati possono essere rappresentati come una matrice, del tipo esemplificato nella tabella 2.3.

2.4.2 Ottenere lo spazio

I dati raccolti fino a questo momento risultano sparsi e poco affidabili. Una serie di passaggi matematici permettono di ottenere uno spazio vettoriale più affidabile a partire dalla matrice di co-occorrenze.

	bite	buy	drive	eat	get	live	park	ride
bike	0	0,5	0	0	0	0	1,09	1,79
car	0	0,8	1,56	0	0	0	0,18	0
dog	0	0	0	2,01	0	1,65	0	0
lion	2,75	0	0	0	0,26	1,01	0	0

Tabella 2.4: La tabella riporta la matrice in cui le frequenze sono state pesate tramite Positive Pointwise Mutual Information (PPMI).

Come già menzionato, la distribuzione zipfiana degli item nei testi introduce un bias di frequenza nei dati raccolti: in questo modo item lessicali molto frequenti finiscono per avere vettori distribuzionali mediamente più simili tra loro rispetto a quelli meno frequenti. La stessa frequenza dei contesti non permette inoltre di portare alla luce le co-occorrenze più significative. Nella tabella 2.3 notiamo infatti che $f(dog, get)$ è maggiore di $f(dog, eat)$, ma *eat* è sicuramente un contesto più significativo di *get* per caratterizzare il significato di *dog*.

Sono presenti in letteratura numerose funzioni di peso delle frequenze che permettono di mitigare il bias citato. Tali misure valutano uno score di associazione tra il target e il contesto, come funzione non solo della loro frequenza ma della distribuzione generale dei contesti nella matrice.

La matrice così trasformata (un esempio è mostrato in tabella 2.4) è costituita da **vettori espliciti**, ad alta dimensionalità e sparsi[1].
L'alta dimensionalità dello spazio induce un problema per la valutazione delle similarità: i vettori risultano tutti circa equidistanti, e questo rende difficoltoso l'individuazione di vettori simili. Inoltre, i vettori espliciti non riescono a catturare il fatto che alcuni contesti sono a loro volta simili o fortemente correlati.

Tramite tecniche di riduzione della dimensionalità che vengono collettivamente denominate *Latent Semantic Analysis*, i vettori espliciti vengono trasformati in **vettori impliciti**, densi e a bassa dimensionalità. Questo processo ha quattro obiettivi principali:

- portate alla luce strutture semantiche latenti
- ridurre il rumore
- catturare co-occorrenze di alto ordine
- ridurre la data sparseness

La via più comune per ottenere uno spazio di vettori impliciti è utilizzare una funzione di riduzione dello spazio che opera tramite la fattorizzazione della matrice nel prodotto di più componenti. Alcune tecniche algebriche, come *Singular Value Decomposition*, permettono di ottenere tale fattorizzazione: il risultato della decomposizione viene troncato per ridurre la dimensionalità ed eliminare porzioni rumorose dello spazio.

2.4.3 Misurare la similarità

Ottenuti vettori affidabili, è a questo punto possibile calcolarne la similarità. Questa è definita in termini spaziali come la prossimità di oggetti nello spazio di rappresentazione (Markman (2013)): la similarità tra item lessicali dipende dunque dalla prossimità dei loro vettori distribuzionali.

[1] consideriamo sparso un vettore o una matrice in cui la maggior parte delle componenti è uguale a zero.

In generale una funzione di similarità tra vettori è una funzione $S : T \times T \to R$ tale che, per ogni coppia di lessemi target u e v, $S(\vec{u}, \vec{v})$ sia proporzionale al grado di similarità tra i lessemi di partenza. S rispetta le seguenti condizioni:

- $S(\vec{u}, \vec{v}) \leq 1$
- $S(\vec{u}, \vec{v}) = 1$ sse u e v sono identici
- S è simmetrica, ovvero $S(\vec{u}, \vec{v}) = S(\vec{v}, \vec{u})$

Una misura di similarità così definita può essere impiegata per costruire una **metrica** sullo spazio che restituisca una nozione di distanza tra i punti individuati dai vettori.

Le misure introdotte in letteratura sono numerose, ne riportiamo qui due, in quanto comunemente utilizzate e in particolare impiegate negli esperimenti descritti.

Distanza Euclidea

La distanza euclidea tra due punti corrisponde alla misura del segmento avente come estremi i due punti in questione.

In formula:

$$E(\vec{u}, \vec{v}) = \sqrt{\sum_{i=1}^{n} |u_i - v_i|^2} \tag{2.1}$$

Cosine similarity

Esprime la similarità tra due vettori in termini del coseno dell'angolo che i due vettori individuano.

In formula:

$$S_C(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\|\|\vec{v}\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} {u_i}^2}\sqrt{\sum_{i=1}^{n} {v_i}^2}} \tag{2.2}$$

La misura così definità esprime la similarità tra due oggetti. Per diventare una misura di distanza deve essere opportunamente trasformata tramite la formula $DS_C(\vec{u}, \vec{v}) = 1 - S_C(\vec{u}, \vec{v})$

Multidimensional Scaling

Parte della potenza esplicativa del modello computazionale presentato risiede nel fatto che esistono tecniche matematiche che permettono di derivare una rappresentazione spaziale visiva a partire dalle distanze tra punti.
Una di queste tecniche, largamente impiegata negli esperimenti descritti, prende il nome di *Multidimensional Scaling* (MDS). A partire da una matrice quadrata contenente relazioni tra gli item in oggetto, l'algoritmo assegna a ogni oggetto una collocazione in uno spazio N-dimensionale, 2- o 3- dimensionale per la visualizzazione (Kruskal and Wish (1978)).
Le dimensioni individuate dall'algoritmo sono chiaramente arbitrarie: nel caso in cui la direzione della visualizzazione è rilevante, come lo è per i dati geografici, è dunque possibile invertirle o ruotarle per fare sì che siano allineate alla configurazione convenzionale.

Luoghi Ritrovati

3.1 Geografia Vernacolare

Fattori socio economici hanno aumentato la frequenza con cui le persone senza conoscenza geografica specialistica hanno accesso e interpretano l'informazione geografica riguardante una certa area.
Le discrepanze tra le geografie delle persone locali e quelle legate a dati ufficiali o commerciali danno spesso luogo a situazioni controverse.

La conoscenza geografica umana sembra infatti tollerare e creare sia vaghezza che inconsistenza nelle conoscenze di luoghi e estensioni: spesso sono presenti nomi diversi per riferirsi alla stessa località o porzione di territorio, o stessi nomi ricorrono per distinguere analoghe manifestazioni in luoghi diversi.
L'insieme di questi fenomeni prende il nome di *geografia vernacolare*.

Davies (2013) ha applicato la metodologia descritta nella sezione precedente all'ambito locale, per la collocazione di nomi di luogo vernacolari la cui identificazione era precedentemente sconosciuta.
Gli sforzi precedenti per identificare nomi vernacolari attraverso il web crawling hanno coinvolto la creazione di query specializzate in database strutturati, la cui collocazione era poi effettuata tramite estrazione di coordinate tipicamente da gazetteer (Pasley et al. (2007); Twaroch et al. (2008)). Tecniche che richiedono interviste ai locali per le richieste di informazione sono solitamente considerate troppo dispendiose e non scalabili, nonché suscettibili ad errori. Anche le risorse online collezionate su base volontaria mostrano infatti un bias verso le visioni di pochi e individui tecnicamente letterati e interessati al tema, le cui visioni potrebbero non essere rappresentative dell'intera popolazione, e potrebbero sovrastimare o sottostimare alcune aree.

Per lo studio è stata selezionata un'area di 432 chilometri quadrati, che comprende la città di Southampton e l'area intorno, sulla costa sud dell'Inghilterra. 59 nomi di luogo sono stati usati come *seeds* per recuperare dal web il corpus testuale. Il risultante corpus, che è stato poi ripulito, conteneva 13.7 milioni di token, da cui è stata estratta una matrice di 4062 type rilevanti.

Analogamente a quanto fatto negli studi precedenti, sono stati creati vettori a bassa dimensionalità (300 componenti) tramite *Latent Semantic Analysis*.
La matrice di coseni per i nomi di luogo originali è stata calcolata dallo spazio, e poi da queste sono state derivate le distanze euclidee. La risultante matrice è stata poi processata con *Multidimensional Scaling*.
Le coordinate dei luoghi nella risultante mappa 2D sono state confrontate con coordinate estratte da gazetteers.

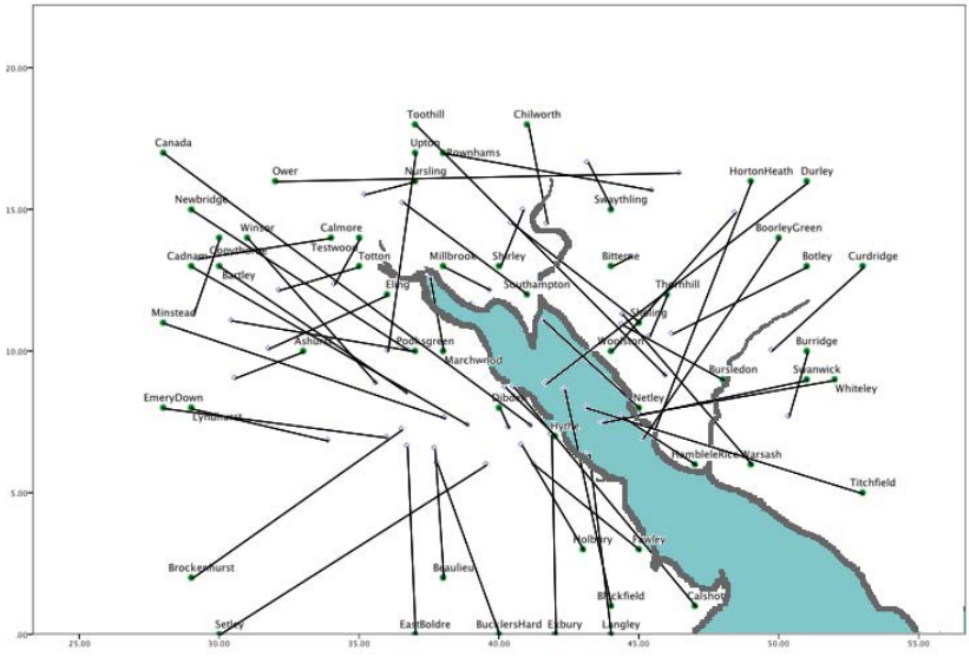

Figura 3.1: Plot tratto da Davies (2013): nella figura si vedono le coordinate nello spazio geografico collegate tramite linee alla posizione predetta a partire dallo spazio semantico.

I risultati sono mostrati nella figura 3.1. Nonostante alcuni sensibili discostamenti dalle vere collocazioni, nessuna linea attraversa l'estuario di Southampton, suggerendo la presenza di un confine cognitivo nella geografia mentale locale e nel linguaggio scritto.

3.2 La valle dell'Indo

I risultati incoraggianti sugli studi precedenti hanno permesso a Recchia and Louwerse (2016) di azzardare un'applicazione di simili teorie in ambito di ricerca storico-archeologica.

Il corpus utilizzato in questo caso è un corpus di iscrizioni provenienti dagli scavi archeologici della valle dell'Indo.
Le iscrizioni, provenienti da **sigilli** e **matrici** di sigilli (alcuni esempi sono visibili nelle figure 3.2, 3.3, 3.4, 3.5), sono state arricchite con metadati riguardanti il tipo di materiale, il numero di righe o la direzione del testo.
Sono stati creati due diversi insiemi di materiali provenienti dai maggiori cinque siti di scavo dell'area, uno per i sigilli e l'altro per le matrici.

Data la ridotta dimensione di ogni iscrizione, non era ragionevole aspettarsi che le iscrizioni contenessero multipli nomi di luogo. I vettori distribuzionali sono stati allora creati partendo dai simboli che compongono le iscrizioni: questi infatti variano da luogo a luogo, ed è plausibile che artefatti provenienti da siti di scavo vicini contengano simili simboli. Questo è vero anche di artefatti trasportati, nel caso in cui esistessero relazioni, ad esempio commerciali, tra il luogo di produzione e il luogo di ritrovamento.
Da questo punto di vista i sigilli e le matrici hanno proprietà diverse, in quanto i sigilli accompagnavano le spedizioni di beni, mentre le matrici no.

Figura 3.2: Matrici rettangolari e un sigillo in terra cotta con iscrizione (in basso). Fonte: `https://www.harappa.com/slide/seals-and-sealing`

Figura 3.3: Esempio di matrice ritrovato presso il sito di Ghola Dhoro. Fonte: `https://www.harappa.com/goladhoro/goladhoroseal.html`

Ogni simbolo è stato dunque rappresentato come un vettore termine, e ogni iscrizione come un vettore documento. Al fine di predire le locazioni relative ai siti archeologici nella valle dell'Indo, i vettori documento per ognuno dei cinque siti sono stati sommati per creare cinque vettori sito.
I coseni risultanti tra i vettori sito sono stati processati tramite MDS e le coordinate risultanti sono state confrontate con le vere latitudini e longitudini dei siti di scavo.

A partire dai due sotto-corpus sono quindi stati ottenuti due set di coordinate per sigilli e matrici. La regressione ha mostrato correlazioni statisticamente significative per i sigilli ($r = 0.88 p < 0.05$) ma non per le matrici ($r = 0.28 p > 0.7$). I risultati del *Multidimensional Scaling* sono mostrati in figura 3.6.

Recchia and Louwerse si sono poi chiesti se le similarità potessero essere utilizzate per predire l'origine geografica di particolari sigilli. È stata quindi effettuata una classificazione dei vettori tramite l'algoritmo *k-nearest neighbors* (kNN).
I risultati, mostrati in tabella 3.1, si dimostrano ben al di sopra della baseline ottenuta effettuando una classificazione random.

Figura 3.4: Riproduzione in resina di un sigillo: le matrici erano utilizzate per produrre un'impronta positiva, come questa riprodotta in resina a partire dalla matrice originale. I sigilli erano tipicamente fatti di ceramica o argilla e usati per sigillare la corta che chiudeva insiemi di beni.
Numerosi sigilli della valle dell'Indo sono stati ritrovati in città mesopotamiche. Un'ulteriore prova degli scambi commerciali intercorsi con l'antica Mesopotamia sono i caratteri della lingua della valle dell'Indo ritrovati su matrici mesopotamiche.
Fonte: `https://www.harappa.com/seal/7.html`

Figura 3.5: Nonostante la scrittura della valle dell'Indo resti ancora non decifrata, i ricercatori concordano che rappresenti una lingua proto-dravidica. Fonte: `https://www.harappa.com/seal/15.html`

Sebbene lo studio abbia carattere esplorativo, risulta di particolare interesse il fatto che tecniche utilizzate nelle scienze cognitive per *misurare* relazioni geografiche e sociali possano essere usate anche per fornire uno sguardo sull'organizzazione di società passate.

3.3 I nomi propri e la semantica distribuzionale

L'applicazione di tecnologie basate su semantica distribuzionale a aree di nicchia o non prettamente linguistiche deve fronteggiare il problema della scarsità dei dati a disposizione.
Come descritto in 2.4, una considerevole quantità di dati linguistici è un parametro imprescindibile

classifier	measure	Mohenjo-daro	Harappa	Lothal	Kalibangan	mean
baseline	precision	23.8	57.1	14.7	4.4	25
	recall	23.8	57.1	14.7	4.4	25
LSA	precision	**69**	82.3	74.6	41.2	66.8
	recall	65.5	86.4	**71.6**	31.8	63.8
ngram	precision	68.8	**84.8**	**78.8**	**63.6**	**74**
	recall	**73.9**	**87.8**	70.3	31.8	**66**

Tabella 3.1: La tabella mostra in percentuale i valori di *precision* e *recall* per vari tipi di classificatori k-NN. La baseline è ottenuta utilizzando come probabilità di assegnamento la percentuale di materiale proveniente da ogni sito. Gli altri due classificatori sono 1-NN basati sui vettori LSA e sui vettori di n-grammi.

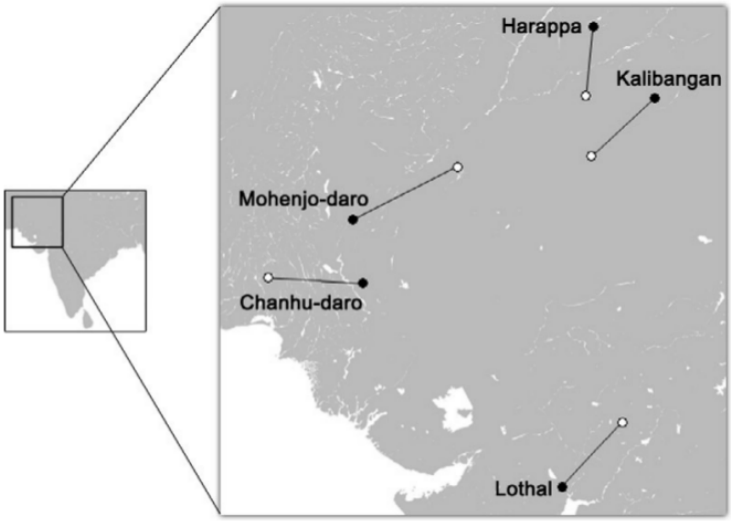

Figura 3.6: I punti neri rappresentano le coordinate geografiche dei cinque maggiori siti archeologici identificati in Mahadevan (1977). I punti bianchi rappresentano le coordinate predette.

per garantire l'affidabilità del modello.
In questi contesti il problema riguarda anche fortemente l'aspetto di preprocessing del corpus a disposizione, in quanto gli strumenti automatici a disposizione per l'analisi linguistica subiscono tipicamente un calo delle prestazioni in contesti di adattamento di dominio.

Esperimenti come quelli descritti in Rodda et al. (2016) o Recchia and Louwerse (2016) soffrono di tale mancanza di dati, e sono sicuramente difficilmente replicabili in situazioni in cui i dati a disposizione non sono digitalizzati o in situazioni storico-culturali in cui i dati semplicemente non esistono (si pensi a lingue antiche o contemporanee poco attestate, ambiti geografici ristretti...).

Strettamente connesso a ciò si deve considerare il fatto che studi come quelli qui riportati hanno come oggetto principale di indagine i **nomi propri**: per la semantica distribuzionale questi rappresentano un problema tanto pratico quanto teorico.
Dal punto di vista pratico i nomi propri e le entità nominate in generale sono per definizione proprie del dominio a cui appartengono, e sono quindi dati tipicamente sparsi e difficilmente rappresentabili soprattutto nei casi menzionati sopra. Dal punto di vista teorico la loro rappresentazione nello spazio distribuzionale necessita di alcune attenzioni particolari, in quanto il loro status semantico è diverso da quello dei nomi comuni, ma nella rappresentazione distribuzionale non sono facilmente distinguibili.

Un approccio interessante ad entrambi i problemi è esemplificato in Herbelot (2015): nel lavoro viene evidenziato il fatto che la rappresentazione distribuzionale non incapsula, a prima vista, la definitezza che rende l'individuo - referente unico, trattando di fatto la rappresentazione di nomi-istanze al pari degli altri elementi dello spazio distribuzionale.

Ci si propone di offrire una rappresentazione distribuzionale dei nomi propri dei personaggi di due romanzi, *Orgoglio e Pregiudizio* (circa 13000 token) e *Il vento tra i salici*, un romanzo per bambini di Kenneth Grahame (1908), composto da circa 6000 token.

A causa delle dimensioni dei due romanzi in esame, costruire uno spazio semantico a partire dal solo libro risulterebbe in dati estremamente sparsi e non conformi alla reale distribuzione dei

contesti. Lo spazio viene allora integrato con informazioni estratte dal British National Corpus (BNC, un corpus di British English per un totale di 100 milioni di parole)

Herbelot afferma che, in uno spazio semantico ideale, le distribuzioni dei nomi propri dovrebbero rispettare le seguenti proprietà:

Unicità l'intensione di un nome proprio dovrebbe catturare la sua estensione unica nel mondo di riferimento. Due Smith che si riferiscono a individui separati dovrebbero anche avere intensioni separate, ovvero occupare punti diversi nello spazio distribuzionale.

Istanziazione i nomi dovrebbero essere in una relazione apprendibile rispetto ai concetti che istanziano. Ad esempio Mr Darcy dovrebbe essere chiaramente un'istanza di uomo, persona etc...

Individualità i nomi propri dovrebbero essere distinguibili dai concetti. Assumiamo un mondo con un dodo che si chiama Dolly. Se l'intensione di *Dolly* fosse la stessa di *dodo*, questo renderebbe Dolly unica (perché c'è un solo dodo in quel mondo) ma non rimarcherebbe ma sua individualità - ovvero che lei è un dodo ma non il tipo *dodo*.

La proprietà di *unicità* non può essere soddisfatta che selezionando solo le occorrenze che si riferiscono al nome in oggetto. I metodi sono simili a quelli utilizzati in casi di polisemia.
L'*istanziazione* è testabile utilizzando misure distribuzionali che si sono mostrate ben fondate in task di *hyponymy detection*. Dato un nome distribuzionale, possiamo provare a estrarre i concetti che più probabilmente lo istanziano assumendo che prendano parte a una relazione di inclusione simile a quella di iponimia-iperonimia.
La proprietà di *individualità* non è di facile controllo. Un test potrebbe essere di indagare quanto la distribuzione di un nome interagisce con quei predicati che sono solo applicabili ai tipi (nel caso del dodo, ad esempio, *estinto* o *diffuso*). Sfortunamente ci sono pochi tipi di predicati di questo tipo e dunque la tecnica si applica male a uno studio quantitativo. Alcuni predicati, invece, sono noti per essere non adatti a tipi ma appropriati per individui (es. *ricco* o *povero*). Testare l'accettabilità di tali composizioni tuttavia pone una serie di problemi, dall'inter-annotator agreement alla composizione in campo distribuzionale.
Il metodo proposto da Herbelot è dunque il seguente: si nota che, mentre i contesti più caratteristici di un tipo possono essere estensionalmente esclusivi, quelli di un individuo non dovrebbero esserlo. Per esempio, sia *ricco* che *povero* possono apparire nei contesti di *uomo*, ma solo uno dei due può esserlo nella distribuzione di un individuo.
Le caratteristiche di un individuo dovrebbero quindi essere più coerenti di quelle di un tipo, in quanto gli item associati con un individuo dovrebbero essere in generale più relazionati l'uno all'altro perché quell'individuo non può assumere il range di esperienze assunto da molti membri del gruppo.

Lo studio propone quindi di calcolare la *coerenza* delle 50 caratteristiche più salienti di ogni nome e paragonarla alla coerenza del tipo che istanzia. Come in Newman et al. (2010), la coerenza di un set di parole $w_1, ..., w_n$ è definita come la media della similarità tra ogni coppia:

$$Coherence(w_1...w_n) = mean\{Sim(w_i, w_j), \forall i, j \in 1...n, i < j\} \tag{3.1}$$

Al momento di incontrare *Mr Darcy* nel testo per la prima volta, infatti, il lettore gli attribuisce già le proprietà dell'item lessicale *man*, data la sua distribuzione in un grosso corpus, e poi specializza la rappresentazione venendo a conoscenza dei contesti dove *Mr Darcy* occorre.
Per rispettare la proprietà di individualità c'è però bisogno che le feature che lo distinguono l'individuo dal tipo abbiano più peso.

Si formalizza la distribuzione come segue. Sia N un nome proprio, istanza del tipo K. N ha una distribuzione standard $v(N)$, con m contesti caratteristici $c_1, ...c_m \in C$. Anche K ha una distribuzione $v(K)$ che vive in uno spazio S con b dimensioni $d_1, ..., d_n \in D$, ottenuto da un corpus

Darcy	Elizabeth	Bingley	Jane	Toad	Badger
0.47 gentleman	0.47 moment	0.48 gentleman	0.48 feeling	0.41 animal	0.43 time
0.47 word	0.46 subject	0.48 lady	0.47 sister	0.38 toad	0.43 animal
0.46 manner	0.46 feeling	0.46 sister	0.46 pleasure	0.38 time	0.40 thing
0.46 feeling	0.46 pleasure	0.46 party	0.46 aunt	0.37 way	0.39 friend
0.46 conversation	0.45 house	0.46 answer	0.46 letter	0.36 thing	0.38 toad

Tabella 3.2: Punteggi di inclusione istanza - tipo ottenuti con la misura invCL (Lenci and Benotto (2012)) sullo spazio distribuzionale iniziale.

Darcy	Elizabeth	Bingley	Jane	Toad	Badger
0,97 man	0,97 woman	0,98 man	0,98 woman	0,97 toad	0,97 badger
0,91 girl	0,9 girl	0,91 boy	0,82 girl	0,75 sea	0,72 sight
0,91 face	0,89 eye	0,9 girl	0,82 man	0,74 desert	0,72 dog
0,91 boy	0,88 man	0,88 eye	0,81 other	0,73 rock	0,71 boy
0,9 smile	0,88 face	0,88 face	0,79 eye	0,73 mountain	0,71 fox

Tabella 3.3: Punteggi di inclusione istanza - tipo ottenuti con la misura invCL (Lenci and Benotto (2012)) sullo spazio distribuzionale in cui gli individui sono stati contestualizzati.

di background abbastanza rappresentativo.
Si definisce $v(K)$ in termini di vettori base di S $\{e_{d'}|d' \in D\}$ e una funzione di peso w:

$$\sum_{d' \in D} w(K, d').e_{d'} \tag{3.2}$$

Possiamo contestualizzare $v(K)$ rispetto a ogni contesto in cui il nome appare. Per ragioni di efficienza la contestualizzazione è fatta rispetto a ogni contesto $c' \in C$ di $v(N)$, secondo la seguente funzione:

$$C(K, c') = \sum_{d' \in D} cos(c', d')^p w(K, d').e_{d'} \tag{3.3}$$

Il vettore di N è poi la somma per tutti i contesti caratteristici, ovvero:

$$\sum_{c' \in C} \sum_{d' \in D} cos(c', d')^p w(K, d').e_{d'} \tag{3.4}$$

I risultati si mostrano incoraggianti secondo vari punti di vista: la proprietà di istanziazione viene soddisfatta dal modello (riproduciamo i risultati riportati nell'articolo nelle tabella 3.2 e 3.3). I personaggi di *Orgoglio e Pregiudizio* mostrano inoltre, una volta contestualizzati, valori di coerenza più alti di quelli calcolati sui vettori tipo.

Luoghi Immaginati

4.1 Nella Terra di Mezzo

Una prospettiva più cognitiva assume lo studio di Louwerse and Benesh (2012): ci si propone di indagare la rappresentazione spaziale mentale che i parlanti costruiscono a partire da fonti di informazione linguistiche, e metterla a confronto con la rappresentazione ottenuta da fonti non linguistiche.
Per fare ciò, Louwerse and Benesh prendono in considerazione il testo de *Il Signore degli Anelli*, il più famoso romanzo di J. R. R. Tolkien. Il testo è di dimensioni adeguate, comprendendo circa mezzo milione di token, ed è ambientato in una regione fittizia sufficientemente particolareggiata, comprendente 32 città.

Anche in questo caso lo spazio distribuzionale è stato ridotto tramite *Latent Semantic Analysis* a 300 dimensioni.
La relazione tra le città è stata stimata tramite il calcolo del coseno tra i vettori corrispondenti e la matrice di coseni 32x32 è stata messa in relazione con la matrice di distanze, ottenuta dalle coordinate 2D delle città sulla mappa.
È stata utilizzata la tecnica di Procruste [1](Schönemann and Carroll (1970)) per tenere in considerazione sia la distanza che la direzione dei punti.
Lo studio ha poi indagato fino a che punto i partecipanti fossero capaci di localizzare le città del *Signore degli Anelli* dopo aver letto il libro o studiato la mappa.
I partecipanti (37) sono stati selezionati e divisi in due gruppi sulla base della loro conoscenza del romanzo: il primo gruppo di partecipanti ha studiato una dettagliata mappa della terra di mezzo per 20 minuti, il secondo gruppo, designato alla lettura del testo, non ha avuto a disposizione alcuna mappa né tempo di studio.
A entrambi i gruppi è stata fornita una mappa muta e la lista di città da posizionare, ed è stato chiesto ai partecipanti di posizionarle sulla mappa.

Le mappe prodotte sono state comparate con la *vera* mappa, ovvero quella allegata alle comuni edizioni del romanzo e approvata dall'autore, e con quella prodotta dallo studio computazionale.
Le correlazioni (mostrate in tabella 4.1) si sono mostrate tutte significative. Inoltre la figura 4.1 mostra come le stime prodotte a partire dalla mappa correlassero meglio con le coordinate autentiche, mentre le stime prodotte a partire dal testo con quelle ottenute tramite LSA.

I risultati suggeriscono che la distanza fisica tra luoghi può essere stimata dalla distribuzione testuale dei lessemi, portando alla conclusione che il linguaggio codifica informazione spaziale.

[1]Procruste, dal greco Προκρούστης, lo stiratore, è il soprannome di un brigante che, nella mitologia greca, aggrediva i viandanti e li straziava battendoli con un martello su di un'incudine a forma di letto scavata nella roccia o metallica, stirandoli se troppo corti o amputandoli qualora sporgessero dal letto.

	r	p
Map-based ∼ authentic	0,81	<0.001
Text-based ∼ authentic	0,77	<0.001
Map-based ∼ LSA	0,36	<0.001
Text-based ∼ LSA	0,39	<0.001

Tabella 4.1: La tabella riporta i valori di correlazione ottenuti confrontando le coordinate autentiche e le coordinate ottenute tramite l'analisi linguistica con le coordinate predette dai due gruppi di partecipanti all'esperimento.

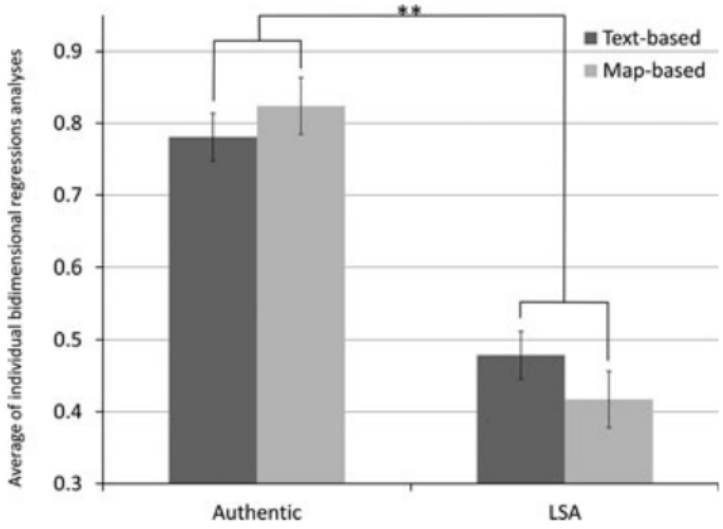

Figura 4.1: Il grafico mostra il valore medio dei coefficienti r di correlazione ottenuti durante l'analisi di regressione bidimensionale. Gli asterischi indicano la significatività ($p < 0.01$)

Le rappresentazioni del layout spaziale derivate dall'input percettivo sono equivalenti a quelle derivate dall'input linguistico. Questo suggerirebbe per gli autori che l'input linguistico è immediatamente trasformato in simulazione percettiva, permettendo che la fonte della mappa cognitiva possa essere diversa, pur mantenendo lo stesso risultato finale.
La conoscenza delle parole e la conoscenza del mondo devono dunque essere collegate.

4.2 Spazio di rappresentazione e *grounding*

Il modello semantico distribuzionale così come la sua implementazione computazionale si basano sull'idea che entità del mondo rappresentano corrispondano a punti in uno spazio di rappresentazione, e che la distanza tra questi punti possa essere usata per rappresentare relazioni tra entità nel mondo rappresentato.
La distanza è coerente con la nozione di senso comune che cose simili siano cognitivamente più vicine di cose dissimili e, come evidenziato in Markman (2013), ipotizzare un isomorfismo tra mondo rappresentato e rappresentazione implica assumere una serie di considerazioni sulle proprietà della rappresentazione.

In particolare bisogna considerare che uno *spazio* di rappresentazione è un concetto matematico ben definito, che obbedisce a tre assiomi **metrici** fondamentali(Tversky (1977); Markman (2013)):

Minimalità $d(x, x) = 0$

Simmetria $d(x, y) = d(y, x)$

Disuguaglianza Triangolare $d(x, y) \leq d(x, z) + d(y, z)$

È stato notato tuttavia che i giudizi di similarità violano sistematicamente questi assiomi. Violazioni della simmetria si riscontrano facilmente pensando a coppie di giudizi del tipo *Quel macellaio è un chirurgo* e *Quel chirurgo è un macellaio*: la relazione tra i due item lessicali *chirurgo* e *macellaio* non è certamente la stessa nei due casi. Similmente casi di violazione della disuguaglianza triangolare si incontrano prendendo in esame situazioni di questo genere: sappiamo che la luna è simile a un pallone, a causa della sua forma, e che la luna può essere simile a un lampione, per la sua lucentezza, ma avremmo molta difficoltà a giudicare nello stesso ordine di similarità un pallone e un lampione. Per ricondurci al caso geografico, è probabile che Verona e Venezia risultino simili nello spazio distribuzionale e al giudizio dei parlanti, ma probabilmente la distanza tra Arzignano, provincia di Vicenza, e Caorle, provincia di Venezia, sebbene geograficamente analoga verrà giudicata molto maggiore.

Il modello distribuzionale, in quanto sfrutta una rappresentazione spaziale del significato, è stato oggetto di varie critiche. La più famosa, legata al problema del *grounding* dei simboli (Harnad (1990)), è stata formulata da Searle (1980) nella celebre argomentazione della stanza cinese[2]:

> Una persona, che non conosce il cinese, è chiusa in una stanza con una grande quantità di materiale in cinese. Nonostante la grande quantità di materiale a sua disposizione, pochi affermerebbero che lui comprenda il cinese.
> Dopo attente letture, sarà in grado di isolare le parole [3] e riconoscere le rispettive collocazioni, e dunque le relazioni che intercorrono tra gli elementi del sistema simbolico a cui è stato esposto, ma, messo davanti ad esempio a raffigurazioni dei referenti dei simboli da lui isolati, come potrà accoppiare ogni simbolo al suo referente con successo?

Il dibattito contemporaneo riconosce la concorrenza dei due modelli, simbolico e percettivo, e si è concentrato sul ruolo del sistema linguistico formale nella rappresentazione semantica. È stato riconosciuto, ad esempio in Louwerse and Jeuniaux (2010) e Barsalou et al. (2008), che stimolo verbale e percettivo interagiscono e che associazioni linguistiche permettono di eseguire task di processing concettuale.
Louwerse and Jeuniaux (2010) interpretano ciò come evidenza a favore della *Symbol Independence Hypotesis*, formulando che:

- è possibile inferire rappresentazioni semantiche combinando informazioni derivate dal linguaggio con rappresentazioni *grounded* già presenti
- il sistema semantico utilizzato durante il processing cognitivo integra l'informazione proprio in questo modo

Queste osservazioni non implicano infatti che rappresentazioni spaziali non siano adatte a applicazioni il cui oggetto sono stati mentali, ma si propongono di sottolineare la necessità di tenere sempre ben distinti il piano della rappresentazione da quello dell'oggetto di studio.
È possibile che altri tipi di rappresentazione possano catturare meglio alcune proprietà del dominio, ed è comunque sempre necessario interpretare i risultati ottenuti attraverso proprietà dello spazio di rappresentazione e proprietà note del dominio. Integrare risultati ottenuti da varie rappresentazioni può contribuire ad esplorarne i limiti e costruire una migliore spiegazione del fenomeno in esame.

[2] Riportiamo la versione formulata in Louwerse et al. (2012)

[3] sulla possibilità di acquisizione di strutture linguistiche a partire dall'analisi statistica dell'input citiamo ad esempio i lavori di Elman (1990) sulle reti neurali e i modelli connessionisti

Conclusioni

> Ogni città riceve la sua forma dal deserto
> a cui si oppone.
>
> Italo Calvino, Le città invisibili

Gli studi che sono stati riassunti e presentati in questa relazione non sono stati condotti da informatici umanisti, né era probabilmente nelle intenzioni dei ricercatori essere considerati tali. Molti aspetti delle ricerche prese in esame, pertinenti alle discipline da cui nascono, sono stati tralasciati nella trattazione per lasciare spazio a quegli altri aspetti che, forse loro malgrado, entrano a gamba tesa nella rete delle *digital humanities*.

Spero che la trattazione, seppure incompleta e non adeguatamente approfondita, offra l'occasione di dare uno sguardo a temi che, nel corso dei seminari proposti, sono stati presentati da una prospettiva diversa. Penso ai lavori riguardanti la digitalizzazione di materiale epigrafico, la creazione di banche dati e basi di conoscenza, l'annotazione di materiale e, di massima importanza, la standardizzazione e la condivisione di pratiche per tutti questi processi.
Sono i seminari che mi hanno incuriosito maggiormente, forse a causa di qualche anno di liceo classico alle spalle, e sono le aree da cui sono partita per la stesura di questa relazione.
Tuttavia, oltre che un po' di tensione romantica verso ruderi e scartoffie, gli anni del liceo mi hanno lasciato anche un gran bisogno di generalizzazione, e una gran voglia di esplorare, più che spazi fisici, spazi concettuali. Perché questi spazi possano però prendere forma e consistenza sono però necessari vari fattori, non ultimo dei quali è l'utilizzo sensato e l'interpretazione di tipi di analisi quantitative, la presenza delle competenze necessarie per fare sì che, con un po' di presunzione, *i dati parlino da soli*, o diano almeno la loro versione dei fatti.

Bibliografia

Barsalou, L. W., Santos, A., Simmons, W. K., and Wilson, C. D. (2008). Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, pages 245–283.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(2014):1–47.

Canter, D. and Tagg, S. K. (1975). Distance estimation in cities. *Environment and behavior*, 7(1):59–80.

Davies, C. (2013). Reading geography between the lines: Extracting local place knowledge from text. In *International Conference on Spatial Information Theory*, pages 320–337. Springer.

Deese, J. (1966). *The structure of associations in language and thought*. Johns Hopkins University Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Feng, Y. and Lapata, M. (2010). Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99. Association for Computational Linguistics.

Ferguson, E. L. and Hegarty, M. (1994). Properties of cognitive maps constructed from texts. *Memory & Cognition*, 22(4):455–473.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Franklin, N. and Tversky, B. (1990). Searching imagined environments. *Journal of Experimental Psychology: General*, 119(1):63.

Friedman, A., Kerkman, D. D., and Brown, N. R. (2002). Spatial location judgments: A cross-national comparison of estimation bias in subjective north american geography. *Psychonomic Bulletin & Review*, 9(3):615–623.

Friedman, A. and Kohler, B. (2003). Bidimensional regression: assessing the configural similarity and accuracy of cognitive maps and other two-dimensional data sets. *Psychological methods*, 8(4):468.

Friedman, A. and Montello, D. R. (2006). Global-scale location and distance estimates: common representations and strategies in absolute and relative judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(2):333.

Goldstein, D. G. and Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychological review*, 109(1):75.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Herbelot, A. (2015). Mr darcy and mr toad, gentlemen: distributional names and their kinds. In *IWCS*, pages 151–161.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional scaling*, volume 11. Sage.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Lenci, A. and Benotto, G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 75–79. Association for Computational Linguistics.

Louwerse, M., Hutchinson, S., and Cai, Z. (2012). The chinese route argument: Predicting the longitude and latitude of cities in china and the middle east using statistical linguistic frequencies. In *Proceedings of the Cognitive Science Society*, volume 34.

Louwerse, M. M. and Benesh, N. (2012). Representing spatial structure through maps and language: Lord of the rings encodes the spatial structure of middle earth. *Cognitive science*, 36(8):1556–1569.

Louwerse, M. M. and Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, 114(1):96–104.

Louwerse, M. M. and Zwaan, R. A. (2009). Language encodes geographical information. *Cognitive Science*, 33(1):51–73.

Mahadevan, I. (1977). *The indus script: texts, concordance, and tables*. Number 77. Archaeological Survey of India.

Markman, A. B. (2013). *Knowledge representation*. Psychology Press.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Montello, D. R. and Freundschuh, S. M. (1995). Sources of spatial knowledge and their implications for gis: An introduction. *Geographical Systems*, 2(1):169–176.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Pasley, R. C., Clough, P. D., and Sanderson, M. (2007). Geo-tagging for imprecise regions of different sizes. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 77–82. ACM.

Recchia, G. L. and Louwerse, M. M. (2016). Archaeology through computational linguistics: inscription statistics predict excavation sites of indus valley artifacts. *Cognitive science*, 40(8):2065–2080.

Rodda, M. A., Senaldi, M. S., and Lenci, A. (2016). Panta rei: Tracking semantic change with distributional semantics in ancient greek. *CLiC it*, page 258.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Schönemann, P. H. and Carroll, R. M. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, 35(2):245–255.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.

Simon, H. A. (1999). The recognition heuristic how ignorance makes us smart. *Simple heuristics that make us smart*, page 37.

Taylor, H. A. and Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. *Journal of Memory and language*, 31(2):261–292.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240.

Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4):327.

Tversky, B. (1981). Distortions in memory for maps. *Cognitive psychology*, 13(3):407–433.

Twaroch, F. A., Jones, C. B., and Abdelmoty, A. I. (2008). Acquisition of a vernacular gazetteer from web sources. In *Proceedings of the first international workshop on Location and the web*, pages 61–64. ACM.

Vasardani, M., Timpf, S., Winter, S., and Tomko, M. (2013). From descriptions to depictions: A conceptual framework. In *International Conference on Spatial Information Theory*, pages 299–319. Springer.

Wittgenstein, L. (1953). Philosophical investigations (gem anscombe, trans.).

Zipf, G. K. (1935). The psycho-biology of language.