Word embeddings per il mutamento semantico nel latino e nel greco antico

Giacomo Fidone

LM In Informatica Umanistica (Tecnologie del Linguaggio)
g.fidone1@studenti.unipi.it

Aprile 2023

Abstract

I recenti sviluppi della semantica distribuzionale hanno aperto a nuovi metodi *data-driven* per la linguistica diacronica. Questa relazione intende valutare le possibilità ed i limiti di una rilevazione automatica del mutamento semantico lessicale nelle lingue storiche, con particolare riferimento a casi di studio riguardanti il latino ed il greco antico.

Indice

Abbreviazioni

BERT *Bidirectional Encoder Representations from Transformers*

CBOW *Continuous Bag Of Words*

DSM *Distributional Semantic Model*

GASC *Genre-Aware Semantic Change*

LSC *Lexical Semantic Change*

NLM *Neural Language Model*

NLP *Natural Language Processing*

PPMI *Positive Pointwise Mutual Information*

RSA *Representational Similarity Analysis*

SGNS *Skip-Gram with Negative Sampling*

SVD *Singular Value Decomposition*

1 Introduzione

Uno dei principali temi di ricerca della linguistica diacronica è noto in letteratura con il nome di *Lexical Semantic Change* (LSC), ovvero il tracciamento dei mutamenti semantici nei lessemi di una lingua. Il LSC riveste un ruolo di primo piano nello studio del linguaggio: il lessico è la parte della lingua che rappresenta il mondo, e dal momento che la lingua è un'entità dinamica plasmata dall'uso che ne fanno i parlanti, i mutamenti nel significato lessicale sono correlati a mutamenti nella realtà extra-linguistica.

Non sorprende quindi che lo studio del LSC abbia rappresentato per molto tempo un lavoro oneroso, tipicamente basato su un approccio storico-comparativo e bisognoso di conoscenze enciclopediche riguardanti la cultura e la società dei periodi storici di interesse. I recenti sviluppi nel campo del *Natural Language Processing* nello svolgimento automatico di svariati task linguistici, come ad esempio il *Sentiment Analysis*, il *Part-of-Speech Tagging*, il *Named Entity Recognition*, etc. hanno tuttavia aperto a nuove prospettive metodologiche. In particolare, i modelli semantici distribuzionali (DSM) e la generazione di *word embeddings* offrono nuovi strumenti *data-driven* a supporto della semantica e della pragmatica storica.[1]

Il trattamento computazionale del LSC, nella sua piuttosto recente storia, ha però principalmente riguardato lo studio di lingue oggi in uso – con un evidente sbilanciamento a favore della lingua inglese. La spiegazione di questa preferenza va al di là di un semplice interesse nel monitorare mutamenti semantici più recenti: per ottenere risultati non sub-ottimali, molti metodi computazionali richiedono una enorme quantità di dati linguistici, non di rado già preparati e annotati con modalità *gold-standard*.

Eppure, sono più evidenti i vantaggi che un approccio quantitativo porterebbe con sé nel contesto di lingue storiche, dove l'intuizione soggettiva del linguista risulta spesso ostacolata dalla scarsità dei dati linguistici, dalla frequente mancanza di fonti di supporto (ad esempio lessici antichi), o da caratteristiche intrinseche della lingua (ad esempio un alto grado di polisemia). L'evidenza empirica messa a disposizione dai DSM può quindi svolgere un ruolo di primo piano nella validazione delle ipotesi della linguistica diacronica o persino suggerire pattern di mutamento semantico inattesi e non prevedibili *a priori*.

In questa relazione verrà valutato l'uso di *word embeddings* per la rilevazione automatica del LSC limitatamente al latino e al greco antico. La ragione di questa scelta è stata in parte suggerita: una delle difficoltà nella generazione di *word embeddings* di lingue antiche riguarda proprio la reperibilità di sufficienti quantità di dati, ulteriormente esacerbata dalla necessità di considerare corpora diacronici. Non si tratta quindi di una ragione interamente riconducibile ad un interesse intrinseco: il latino ed il greco antico – soprattutto il primo, vista la sua millenaria egemonia come *lingua franca* della scienza e della cultura – sono

[1] Per una panoramica completa dell'applicazione di approcci distribuzionali al mutamento linguistico il lettore può riferirsi a Kutuzov *et al.* (2018) e a Tahmasebi *et al.* (2018).

proprio tra le lingue storiche di cui disponiamo della maggiore quantità di dati (Sprugnoli *et al.* 2020).

Questo primato è del resto deducibile dalla letteratura sull'argomento: se ancora poche sono le ricerche che sono state condotte sul latino e sul greco antico, rare sono quelle che propongono tentativi su diverse lingue storiche. La rilevazione automatica del LSC in lingue antiche è in effetti uno dei temi di frontiera della linguistica computazionale, che è stato affrontato solo negli ultimi anni e che, ad oggi, rimane ancora in buona parte inesplorato.

A conclusione di queste note introduttive offro qualche indicazione sul contenuto delle sezioni che seguono. In primo luogo verranno introdotti i vantaggi dei *word embeddings* per la rilevazione automatica del LSC e verranno brevemente descritti i principali modelli semantici distribuzionali (DSM), nonché i limiti della loro applicazione a lingue storiche. A seguire, verrà valutato l'impiego di diversi DSM sul latino e sul greco antico per la rilevazione automatica del LSC mediante la considerazione di specifici casi di studio, e verrà in particolare approfondito il problema della disambiguazione tra mutamento semantico e forme di polisemia non diacronica. Seguiranno infine alcune considerazioni conclusive su possibili ulteriori linee di ricerca.

2 Spazi vettoriali

Un trattamento computazionale del LSC può essere formalmente descritto come segue. Dati n corpora $[C_1, C_2, ..., C_n]$, ciascuno contenenti testi afferenti a distinti momenti temporali di arbitraria granularità $[1, 2, ..., n]$, il compito consiste nel misurare automaticamente la differenza tra il significato S_k di una parola al tempo k ed il significato della parola S_j al tempo j per qualche k, j tali che $1 \leq k \leq n$, $1 \leq j \leq n$, $k \neq j$. Questa definizione richiede la disponibilità di una qualche misura quantitativa del significato S_k (S_j) di una parola al momento k (j).

Prima dell'avvento dei *word embeddings*, un approccio comune in linguistica computazionale consisteva nel considerare la semplice frequenza relativa dei termini lessicali quale misura del loro significato. E' stato infatti dimostrato che una variazione nella frequenza d'uso di un lessema può essere indicativa di un mutamento semantico in atto. Ad esempio, la maggiore diffusione della parola inglese «server» a partire dagli anni '60-'70 segnala l'emergenza del significato afferente alla sfera tecnologica (oggi prevalente) accanto a quello pre-esistente di «persona che serve».[2]

Se il mutamento semantico è spesso causa di una corrispondente variazione nelle abitudini linguistiche, l'inverso non è necessariamente vero. La frequenza d'uso di un lessema può in effetti rispondere a ragioni esclusivamente socio-culturali, e non può pertanto costituire da sola un indicatore affidabile del LSC.

[2] Analogamente, una diminuzione nella frequenza d'uso può segnalare la perdita di un significato.

Il parziale successo di questo approccio può essere tuttavia ricondotto alla sua capacità di approssimare un'intuizione oramai tradizionale della linguistica storica: mutamenti nel significato di una parola corrispondono a mutamenti nel suo uso. Il problema è che l'uso di una parola riguarda non solo la sua generale diffusione all'interno della lingua, ma soprattutto i particolari contesti linguistici in cui occorre.

E' qui che la linguistica diacronica incontra la semantica distribuzionale, ed in particolare la generazione di *word embeddings*, rappresentazioni vettoriali in grado di codificare proprietà distribuzionali di un input linguistico. Il fondamento della semantica distribuzionale è infatti l'equazione tra proprietà semantiche e proprietà distribuzionali di un parola. L'origine di questa idea è eterogenea (Harris 1954, Firth 1957), ma viene solitamente ricondotta alla formulazione della cosiddetta «ipotesi distribuzionale», secondo cui esiste una correlazione tra il significato di una parola e quello dei suoi «collocati» – le parole che occorrono nello stesso contesto linguistico. In breve: parole che occorrono in contesti simili hanno significato simile.

La semantica distribuzionale offre quindi una misura quantitativa del significato lessicale che tiene conto delle intuizioni teoriche della linguistica diacronica riguardanti il mutamento semantico. Per questa ragione, è stato osservato (Boleda 2020) come i *word embeddings* e le loro relazioni geometriche nello spazio vettoriale possano costituire una valida evidenza empirica al modellamento automatico del LSC.

2.1 Modelli Semantici Distribuzionali

Nel corso degli ultimi anni sono stati proposti diversi metodi computazionali per l'estrazione delle proprietà distribuzionali delle parole, noti con il nome di modelli semantici distribuzionali (*Distributional Semantic Models* o DSM). Ovviamente la letteratura sull'argomento è molto ampia ed un trattamento esaustivo dei diversi DSM andrebbe al di là delle possibilità e degli interessi di questa relazione. Tuttavia, in funzione della successiva considerazione dei modelli che sono stati impiegati per il latino ed il greco antico, è perlomeno necessario distinguere tra tre generazioni di DSM: quella dei modelli matriciali o *count-based*, quella dei modelli neurali o NLM (*Neural Language Models*) e quella, più recente, dei modelli contestuali.

I modelli matriciali o *count-based* si basano sul semplice conteggio delle co-occorrenze tra lessemi target e collocati. Le frequenze assolute vengono quindi registrate quali entrate di una matrice di co-occorrenza M di dimensioni $m \times n$, dove m è il numero dei lessemi target ed n il numero dei collocati. Per controllare gli effetti della distribuzione zipfiana delle parole in un corpus (Lenci 2018), le entrate vengono pesate con una misura di associazione quale, tipicamente, la *Positive-Pointwise Mutual Information* (PPMI):

$$PPMI(t, c) = \max\left(0, \log \frac{p(t, c)}{p(t)p(c)}\right)$$

Infine, dal momento che la matrice M è tipicamente sparsa e ad alta dimensionalità, la si riduce con un'operazione di fattorizzazione, quale la *Singular Value Decomposition* (SVD):

$$SVD(M) = U\Sigma V^T$$

Diversamente, sia i modelli neurali che i modelli contestuali sfruttano una rete neurale addestrata su un task detto di *language modeling*, che consiste nel predire la parola più probabile dato un certo contesto linguistico (o viceversa) mediante l'applicazione di specifiche funzioni di attivazione. Una funzione comunemente utilizzata è la *softmax*, la quale è in grado di normalizzare un vettore di numeri reali in una distribuzione di probabilità:

$$p(\mathbf{b}|\mathbf{a}) = \frac{\exp(\mathbf{b} \cdot \mathbf{a})}{\sum_{\mathbf{a}' \in C} \exp(\mathbf{a}' \cdot \mathbf{b})}$$

I *word embeddings* sono quindi in questo caso un sotto-prodotto dell'attività di predizione, ovvero le matrici di pesi della rete risultanti dal processo di *training*.

La differenza tra modelli neurali e modelli contestuali riguarda l'architettura della rete (che può essere più o meno «profonda») e il tipo di rappresentazione prodotta. Le reti della prima generazione, di cui l'esempio più significativo è *Word2Vec* (Mikolov *et al.* 2013), hanno una topologia relativamente semplice e generano un *embedding* per ciascuna parola del vocabolario (un *embedding* «non contestualizzato» o «statico»); mentre le reti della seconda generazione, rappresentate dai *transformers* come BERT (Devlin *et al.* 2018), sono reti «profonde» in grado di generare un *embedding* per ciascun *token* di una parola (un *embedding* «contestualizzato» o «dinamico»).

Nonostante i metodi siano molto diversi tra di loro, il risultato è analogo: una matrice densa in cui ciascuna riga costituisce l'*embedding* di una parola (*type* o *token*). Una volta ottenuta, la similarità semantica tra due parole potrà essere misurata come la distanza nello spazio vettoriale tra i rispettivi *embeddings* $\mathbf{w}_1$ e $\mathbf{w}_2$. Una misura comune in semantica distribuzionale è il coseno, in quanto si considerano i vettori normalizzati:

$$cos(\mathbf{w}_1, \mathbf{w}_2) = \frac{\mathbf{w}_1\mathbf{w}_2}{|\mathbf{w}_1||\mathbf{w}_2|}$$

Misure alternative possono essere definite sulla base dell'individuazione dei primi K vicini semantici nello spazio vettoriale, la quali richiedono comunque la preliminare indicazione di una misura della distanza tra vettori.

2.2 DSM e lingue storiche

In teoria non esistono particolari limitazioni riguardo l'utilizzo di un certo DSM per la generazione di uno spazio vettoriale di una lingua storica come il latino o il greco antico. Tuttavia, le scelte di design del modello devono tener conto di alcune difficoltà che valgono anche nel caso generale, ma che nel contesto di una lingua storica risultano enormemente più accentuate. Si tratta infatti, come è noto in letteratura, di difficoltà riguardanti prevalentemente

i dati linguistici per l'addestramento del modello.

Anzitutto, ciascun DSM richiede normalmente la disponibilità di una grande quantità di dati linguistici affinché gli *embeddings* generati godano di una sufficiente affidabilità statistica. I corpora di lingue quali il latino o il greco antico – pur essendo queste le lingue storiche di cui disponiamo della maggior quantità di dati – non riescono a raggiungere le stesse dimensioni dei corpora di una lingua moderna quale l'inglese: se i primi si collocano nell'ordine dei milioni di parole, i secondi eccedono facilmente il miliardo di parole.

Le ragioni non dipendono esclusivamente dalla scarsità intrinseca delle fonti – per cui le lingue storiche possono essere ascritte alla categoria generale delle lingue *low-resource* – ma anche dalla necessità di lavorare con dati pre-processati. Operazioni quali la tokenizzazione, la lemmatizzazione ed il filtraggio delle *stop-words*[3] costituiscono spesso dei requisiti indispensabili alla loro effettiva usabilità, ma altrettanto richiesta è l'annotazione *gold-standard* (effettuata manualmente da operatori esperti) di informazione aggiuntiva dipendente dal task considerato. Inoltre, la scarsità dei dati può essere ulteriormente esacerbata qualora sia necessario (come nel nostro caso) considerare sub-corpora diacronici, ovvero *embeddings* relativi a periodi cronologici circoscritti.

Uno studio condotto da Sahlgren e Lenci (2016) ha fornito indicazioni più dettagliate riguardo alle prestazioni di varie tipologie di DSM su diverse quantità di dati. Come previsto, è stato osservato che al diminuire delle dimensioni del corpus tutti i DSM restituiscono risultati sub-ottimali. Tuttavia, gli autori segnalano anche che al di sotto di una certa soglia i modelli neurali tendono ad esibire prestazioni meno competitive rispetto alle controparti *count-based* (Figura 1).

Una seconda ma non meno importante difficoltà riguarda il bilanciamento del corpus: affinché un *embedding* costituisca una rappresentazione veritiera degli usi di una parola in una lingua, sarebbe necessario che il corpus da cui viene estratto sia rappresentativo della lingua stessa. Il bilanciamento di un corpus è tuttavia un'operazione notoriamente complessa, e diventa quasi proibitiva per una lingua *low-resource* proprio a causa della scarsità dei dati a disposizione. In questi casi un approccio comunemente adottato in linguistica computazionale è quello di affidarsi all'adagio «conoscere che il tuo corpus è sbilanciato è ciò che conta» (Atkins *et al.*, 1992: 6): i *bias* dovuti allo sbilanciamento del corpus possono essere rilevati a patto di avere una pregressa conoscenza della sua composizione e del suo effettivo grado di rappresentività.

Dal momento che la maggior parte delle fonti di lingue storiche come il latino ed il greco antico sono costituite da testi della letteratura, il bilanciamento di un corpus dipenderà soprattutto dalla distribuzione dei vari generi letterari e solo in misura minore da variabili diatopiche o diamesiche. Nel contesto di un tracciamento computazionale del LSC, vedremo

[3] E' stato dimostrato che l'eliminazione delle parole funzionali (il cui ruolo di collocati è superfluo) migliora notevolmente le prestazioni computazionali.

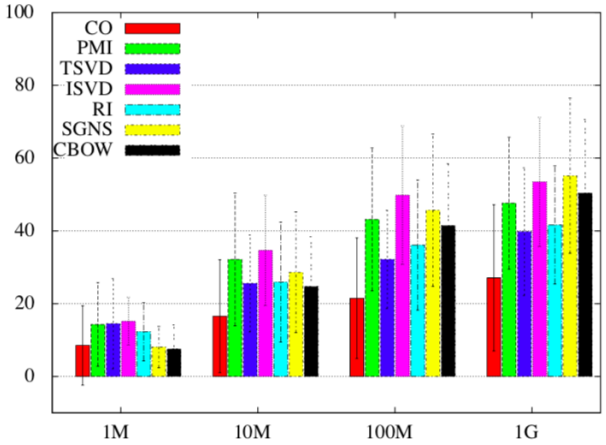

Figura 1: Media dell'accuratezza dei modelli al crescere delle dimensioni del corpus rispetto ai diversi *benchmark* considerati in Sahlgren e Lenci (2016). Si osserva come le prestazioni di SGNS siano le migliori per corpora di grandi dimensioni (1G), ma decrescono significativamente al diminuire dei dati. Il modello che sembra performare meglio indipendentemente dalla quantità di dati è il modello *count-based* ISVD (ovvero basato su SVD in cui si rimuovono le prime 200 dimensioni latenti).

in particolare che la rappresentatività dei corpora rispetto al genere letterario costituisce un possibile ostacolo alla disambiguazione tra fenomeni di polisemia e mutamento diacronico.

3 Word embeddings diacronici

La valutazione delle possibilità e dei limiti di un'implementazione dei DSM per la rilevazione del LSC nel latino e nel greco antico verrà effettuata in questa sezione mediante la considerazione di alcuni esempi applicativi. L'ordine seguito non sarà di tipo cronologico, ma sarà basato prevalentemente sulla complessità del DSM proposto.

Tra i modelli di tipo *count-based*, un caso di studio interessante è quello di Rodda *et al.* (2017), nel quale si usano i *word embeddings* per un duplice scopo: monitorare il LSC del greco antico dall'età pre-cristiana (VII-I sec. a.C.) all'età cristiana (I-V sec. d.C.) e mostrare la plausibilità metodologica di un uso della semantica distribuzionale su una lingua *low-resource*.

Il corpus utilizzato, lemmatizzato dagli autori, è il *Thesaurus Linguae Grecae*, dalle dimensioni relativamente contenute (meno di 26 milioni di token). La tecnica di *embedding* proposta per generare gli spazi vettoriali per ciascuna finestra temporale è *count-based*, dove la matrice di co-occorrenza è pesata con PPMI e successivamente ridotta con SVD. Il confronto tra i due spazi semantici è basato invece sul *Representational Similarity Analysis* (RSA): dopo aver calcolato le matrici di similarità per ciascuno spazio (si considera il coseno come misura di distanza tra vettori), vengono confrontate con una matrice di coincidenza ottenuta calcolando il relativo coefficiente di correlazione di Pearson.

Gli autori dello studio mostrano come i *word embeddings* siano in grado di corroborare l'ipotesi di un ruolo di primo piano del cristianesimo nell'indirizzare il LSC del greco antico. In particolare, alcuni termini lessicali designati ad indicare importanti concetti cristiani sono tra quelli che subiscono la maggiore variazione di significato, soprattutto nella forma di un *narrowing* semantico[4]. Esempi eloquenti sono παραβολή (*parabolé*), che dal generico significato di «confronto» si specializza in «parabola»; ἀγγελος (*àngellos*), che passa dal generico significato di «messaggero» a quello di «angelo»; o lo stesso θεὸς (*theòs*), che se prima designava una qualche divinità, acquisisce in seguito i connotati tipici del dio cristiano. In Figura 2 è possibile apprezzare meglio le relazioni tra i termini lessicali che hanno subito la maggiore variazione di significato ed i corrispettivi vicini semantici nello spazio relativo al periodo d.C.

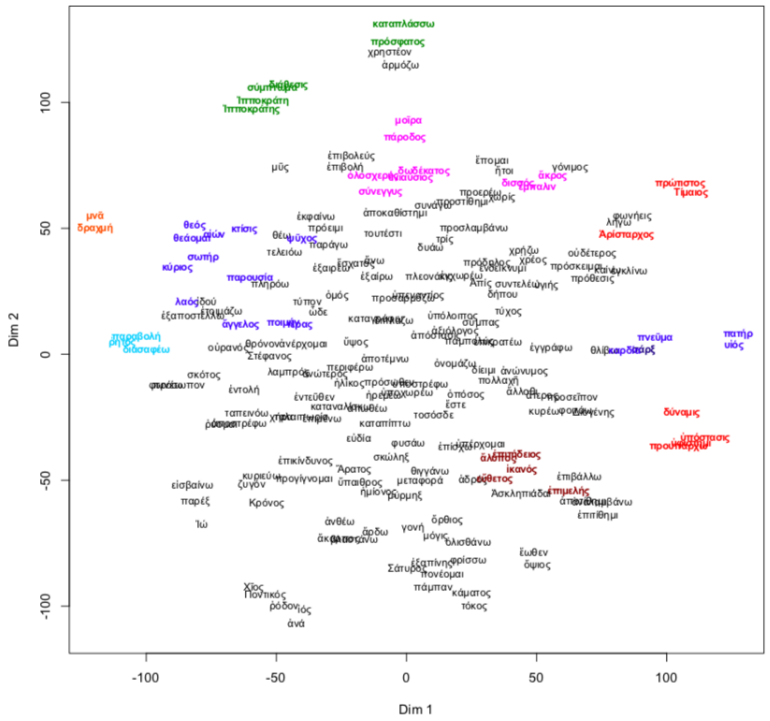

Figura 2: Rappresentazione dello spazio semantico (t-SNE) riferito alle parole d.C. che hanno subito la maggior variazione di significato. La vicinanza delle parole è proporzionale alla similarità semantica. Si osserva, ad esempio, come παραβολή (*parabolé*, in azzurro a sinistra) è vicina a parole afferenti all'esegesi cristiana (mentre nello spazio a.C. era più vicina a termini di pertinenza della geometria). Interessante è anche la vicinanza tra πατηρ (*patér*, «padre») e υιος (*uiòs*, «figlio») con πνευμα (*pnèuma*, in viola a destra), che in origine aveva il significato di «respiro» ma che qui risulta già associata al significato religioso di «spirito».

[4]Il *narrowing* (in opposizione al *broadening*) consiste nella «restrizione» del significato lessicale, ovvero nel passaggio da un significato più ampio ad un significato più specifico.

Ma il confronto tra i *word embeddings* non solo supporta le intuizioni del linguista storico: è in grado di rilevare pattern di mutamento semantico non prevedibili *a priori*, come ad esempio una certa specializzazione del lessico tecnico. Un caso eloquente è quello di υποστασις (*hypòstasis*), che dal significato di «fondazione» passa a quello metafisico di «sostanza».

Uno studio simile condotto sulla lingua latina è quello proposto da Sprugnoli *et al.* (2020) nel contesto del progetto *LiLa: Linking Latin*[5], che si propone di generare uno spazio vettoriale per la lingua latina allo scopo di monitorare mutamenti semantici dall'età classica all'età medioevale.

I corpora selezionati (entrambi lemmatizzati) sono l'*Opera Latina*, che contiene testi di età classica di vario genere letterario (meno di 2 milioni di parole); e l'*Opera Maiora*, contenente testi di Tommaso d'Aquino (4.5 milioni di parole). Nonostante le contenute dimensioni dei corpora, gli autori dello studio propongono come DSM il già citato *Word2Vec* (Mikolov *et al.* 2013), ovvero i due modelli neurali *Skip-Gram with Negative Sampling* (SGNS) e *Continuous Bag of Words* (CBOW). Accanto a *Word2Vec* gli autori considerano anche una sua variante, *FastText* (Bojanowki *et al.* 2017), che consente di integrare informazione morfologica nella rappresentazione vettoriale della parola.[6]

Dalla valutazione degli spazi vettoriali mediante il *benchmark* TOEFL (*Test of English as a Foreign Language*), gli autori mostrano come SGNS riesca a raggiungere ottime prestazioni nonostante l'esigua quantità di dati, con un massimo di accuratezza di 86.91% con *FastText*.[7] Per l'analisi diacronica il confronto tra i due *embeddings* viene effettuato semplicemente calcolando l'intersezione tra i primi K vicini semantici – minore l'intersezione, maggiore il mutamento semantico.

I risultati dello studio mostrano una corrispondenza sostanziale con quelli del precedente. Anche in questo caso alcune evidenze vanno a supporto dell'ipotesi di un ruolo del cristianesimo nel mutamento semantico dei lessemi latini. Un esempio eloquente è il lemma *sacer* («sacro»), che se prima era associato al *pantheon* latino, in Tommaso d'Aquino esibisce già un significato più specificamente ecclesiastico. Altre evidenze vanno invece in direzione di una specializzazione tecnica del lessico, tra cui spicca il caso singolare di *equus* («cavallo»), che riflette in realtà un uso del tutto idiosincratico che Tommaso d'Aquino ne fa nelle comparazioni filosofiche. Esempi ulteriori possono essere trovati in in Sprugnoli *et al.* (2019): come nel caso del greco πνευμα (*pnèuma*), anche il latino *spiritus* si sposta in direzione di un significato religioso; oppure ancora *ordo* passa dal significato sociale di «classe» a quello filosofico di «ordine».

I casi di studio di Rodda *et al.* (2017) e di Sprugnoli *et al.* (2020) mostrano quindi chiari

[5] https://lila-erc.eu/

[6] In *FastText* un *embedding* è ottenuto dalla somma dei vettori associati agli n-grammi di ciascun carattere di cui la parola è costituita.

[7] Alla luce della minore competitività dei modelli neurali in presenza di corpora di ridotte dimensioni (Figura 1), potrebbe essere interessante in questo caso confrontare *FastText*-SGNS con un DSM *count-based*.

esempi di informazione distribuzionale tratta da *embeddings* di tipo *count-based* e neurale che può essere integrata nelle indagini della linguistica diacronica, sia in direzione della validazione di ipotesi esistenti sul LSC (ad esempio il ruolo del cristianesimo nei mutamenti semantici delle lingue antiche) sia in direzione di ipotesi inedite (ad esempio rispetto alla specializzazione tecnica di alcuni lessemi).

3.1 Polisemia non diacronica e modelli bayesiani

Il caso limite rappresentato dall'esempio di *equus* segnala in modo inequivocabile un problema comune agli studi di Rodda *et al.* (2017) e di Sprugnoli *et al.* (2020), già segnalato da Perrone *et al.* (2019): la mancata integrazione di possibili variabili non diacroniche nella variazione semantica dei lessemi, e nella fattispecie il dominio di appartenenza. Lingue come il greco antico ed il latino sono infatti caratterizzate da un alto grado di polisemia che spesso dipende proprio dal genere letterario in cui il lessema occorre (Figura 3).

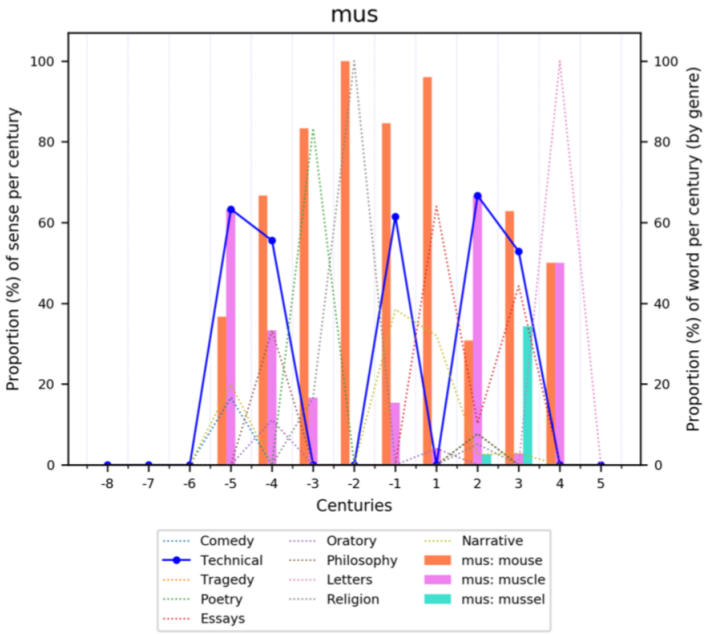

Figura 3: Esempio di variazione semantica della parola latina *mus* in relazione al genere letterario, tratto da Perrone *et al.* (2019). Nel grafico ciascuna linea rappresenta la percentuale di occorrenze della parola nel genere letterario di riferimento e ciascuna barra la percentuale di occorrenze della parola con un certo significato. Si osserva, ad esempio, che la distribuzione di *mus* con il significato di «muscolo» segue in buona parte la distribuzione di *mus* nei testi tecnici, e questa corrispondenza può indicare un potenziale ruolo del genere letterario nell'indurre un significato tecnico sincronicamente presente a quello comune («topo»).

Se quindi i corpora selezionati per ciascun momento temporale presentano distribuzioni diverse rispetto al genere letterario (come nel caso dell'*Opera Maiora*, che contiene solo testi

filosofici), si genera un problema di sbilanciamento che può risultare, congiuntamente all'uso di *embeddings* statici, in casi anomali come quello sopra menzionato di *equus*. In altri termini, gli *embeddings* di una parola in tempi distinti possono risultare diversi non per un mutamento semantico sottostante ma per effetto di un *bias* di rappresentatività.

Questo *bias* viene chiaramente intercettato dagli autori degli studi grazie ad una pregressa conoscenza del grado di sbilanciamento dei corpora, ma al contempo rende meno affidabile l'uso delle evidenze empiriche dei DSM per inferire mutamenti semantici non attesi. Una possibile soluzione potrebbe essere la considerazione di sub-corpora differenziati in base al genere letterario, ma avrebbe l'effetto di produrre un'ulteriore frammentazione dei dati ed una conseguente diminuzione nell'accuratezza dei modelli.

Un approccio che è stato proposto proprio allo scopo di disambiguare tra polisemia non diacronica e mutamento semantico è quello di Perrone *et al.* (2019). Gli autori sviluppano un modello denominato *Genre-Aware Semantic Change* (GASC), ovvero un un *Bayesian Mixture Model* che consente di integrare l'informazione distribuzionale con variabili aggiuntive, quale appunto il genere letterario. Il corpus selezionato è il *Diosiris Annotated Ancient Greek Corpus*, che contiene circa 10 milioni di parole già lemmatizzate e PoS-taggate. Per il task sono state selezionate parole che esibiscono una netta polisemia, ovvero parole aventi un significato astratto accanto ad uno più concreto. GASC restituisce in output, per ciacun token di una parola target, una distribuzione di possibili sensi. La valutazione della distribuzione prodotta da GASC con la *ground-truth* viene effettuata per mezzo di una comparazione *gold-standard*: alcuni esperti hanno precedentemente annotato le parole target del corpus indicando il senso più probabile sulla base del contesto linguistico circostante.

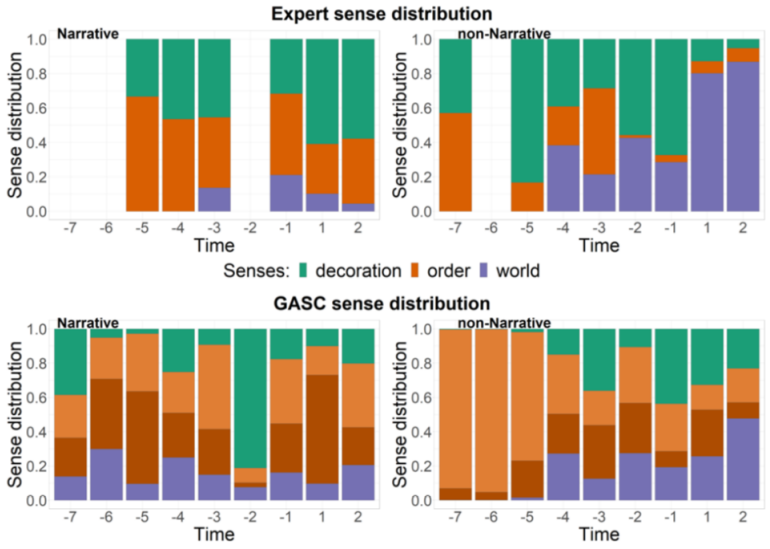

Figura 4: Distribuzione dei sensi della parola κόσμος (*kòsmos*) nei generi «narrativo» e «non-narrativo». Per ciascun genere, si confronta la distribuzione dei sensi assegnata dagli esperti (in alto) con la distribuzione dei sensi assegnata da GASC (in basso).

Le prestazioni di GASC raggiungono lo stato dell'arte della disciplina e consentono di disporre di informazioni più facilmente interpretabili riguardo il mutamento semantico dei termini lessicali. Ad esempio, come si osserva in Figura 4, GASC è in grado di catturare correttamente l'emergenza del senso tecnico di κόσμος (*kòsmos*) come «mondo» nei testi non narrativi a partire dal 400 a.C. In questo caso, la mancata considerazione del genere letterario e la concentrazione dei diversi sensi di κόσμος (*kòsmos*) in un'unica rappresentazione statica avrebbero reso meno chiara la dipendenza del nuovo significato dal contesto tecnico in cui la parola occorre.

In uno studio successivo (Perrone *et al.*, 2021) gli autori estendono l'applicazione di GASC alla lingua latina e forniscono altri esempi significativi dei vantaggi del modello bayesiano. Un caso riportato è quello di παράδεισος (*parádeisos*), che veicola il significato originario «giardino» e quello biblico di «paradiso», i quali risultano difficilmente distinguibili per la caratterizzazione stessa del paradiso biblico come giardino fisico. Come riportato in Figura 5, GASC è in grado di rilevare come l'uso di παράδεισος (*parádeisos*) nel senso di «giardino» decade più velocemente nei testi religiosi, ovvero a partire dalla diffusione del cristianesimo nel I sec. a.C.

Figura 5: Distribuzione dei sensi della parola παράδεισος (*parádeisos*) nei generi «non religioso» e «religioso». Il significato di «giardino» è indicato dai colori verde e giallo, mentre quello biblico di «paradiso» dai colori blu e viola. In basso sono indicati i vicini semantici per ciascun significato.

Il limite principale di GASC rimane tuttavia la limitata reperibilità di dati con annotazione *gold-standard*, nonché la difficoltà riscontrata da parte degli annotatori nel ragionare in termini probabilistici ed offrire stime attendibili della distribuzione dei termini secondo il genere letterario.

3.2 Prospettive future: embeddings contestuali

Il problema della disambiguazione tra polisemia non diacronica e mutamento semantico emerge a partire dall'uso di rappresentazioni non contestualizzate del significato, ovvero gli

embeddings di modelli *count-based* o di modelli neurali. Una diversa soluzione per distinguere con sicurezza i mutamenti diacronici da altri fenomeni di polisemia potrebbe quindi essere individuata considerando le rappresentazioni «dinamiche» di un modello contestuale.

Per la rilevazione automatica del LSC, i modelli contestuali hanno ricevuto minore attenzione rispetto alle controparti *count-based* e neurali. Ad esempio, per il Task-1 del SemEval-2020 (*Unsupervised Lexical Semantic Change Detection*)[8], in cui era anche incluso un corpus di lingua latina, è stato osservato che la maggior parte dei metodi erano basati su *embeddings* non contestualizzati (Kutuzov *et al.* 2020) e raggiungevano comunque risultati più competitivi di quelli delle controparti contestualizzate.

Il limite principale nell'impiego di modelli contestuali è di nuovo l'ingente richiesta di dati linguistici, la quale supera di gran lunga quella di un modello neurale o di un modello *count-based*. Sappiamo inoltre che nel caso di lingue storiche questo limite risulta ulteriormente aggravato dalla più difficile reperibilità di sufficienti quantità di dati.

Uno studio di Bamman e Burns (2020) ha tuttavia proposto una versione di BERT per la lingua latina («LatinBERT»). Il modello è stato addestrato su un corpus di 642,7 milioni di token, costruito attingendo a svariate fonti che vanno dall'età classica fino alle produzioni più recenti (ad esempio la Wikipedia latina). Il DSM è stato poi ottimizzato per lo svolgimento di task di diverso tipo, tra cui la disambiguazione del significato lessicale (*Word Sense Disambiguation*) e l'individuazione di vicini semantici «contestuali».

Sarebbe quindi utile valutare l'adozione di questo modello anche per un task di rilevazione automatica del LSC. Le modalità di utilizzo delle rappresentazioni contestuali a questo scopo non sono tuttavia affatto scontate. Un caso applicativo proposto da Kutuzov *et al.* (2020) mostra ad esempio la costruzione di rappresentazioni matriciali formate dagli *embeddings* di ciascuna parola del vocabolario al fine di definire misure quantitative di variazione semantica analoghe al coseno utilizzato per gli *embeddings* statici. Nel tentativo di ottenere rappresentazioni uniche per ciascuna parola del vocabolario, gli autori segnalano come questo approccio finisca per confrontarsi con le stesse difficoltà che riguardavano i modelli *count-based* e neurali. In particolare, le misure di variazione semantica assegnerebbero punteggi elevati a parole che non hanno subito un reale mutamento semantico, ma che esibiscono un alto grado di polisemia rispetto al contesto in cui sono usate.

4 Conclusioni

I *word embeddings* dei modelli semantici distribuzionali costituiscono un valido strumento *data-driven* per la rilevazione automatica del LSC in lingue storiche come il latino ed il greco antico. Nonostante le esigue quantità di dati linguistici a disposizione, i casi di studio considerati mostrano come mediante opportune scelte di design è possibile sviluppare DSM

[8] https://alt.qcri.org/semeval2020/

in grado di raggiungere prestazioni non sub-ottimali e rilevare fenomeni di mutamento diacronico coerenti con le intuizioni del linguista storico.

Lo sbilanciamento dei corpora e l'impiego di *embeddings* statici possono tuttavia ostacolare il riconoscimento di forme di polisemia non dipendenti da variabili diacroniche, rendendo meno affidabile l'uso dell'evidenza distribuzionale per la rilevazione di mutamenti semantici non attesi. Un modello bayesiano come il GASC può rappresentare in questo senso un'alternativa ai metodi di *embeddings* classici, ma la sua valutazione dipende dalla qualità delle stime probabilistiche proposte dagli esperti. D'altra parte, un uso diacronico di *embeddings* contestuali potrebbe lasciare irrisolto il problema e rimane ancora poco praticabile per l'ingente richiesta di dati linguistici.

I limiti strutturali dei DSM non consentono quindi di sostituire lo studio qualitativo del linguista con un approccio puramente quantitativo. Ciononostante, i DSM possono già costituire un valido strumento *data-driven* per supportare con evidenza empirica la ricerca della linguistica diacronica, e nella fattispecie la validazione di ipotesi riguardanti il LSC e l'indicazione di pattern di mutamento semantico non prevedibili sulla base dei metodi comparativi tradizionali. Questa fondazione risulta ancora più vantaggiosa nel contesto delle lingue storiche, dove il dato distribuzionale può compensare le maggiori difficoltà interpretative. Inoltre, dal momento che i mutamenti semantici di lungo periodo sono più facili da tracciare (Kutuzov *et al.* 2020), lingue storiche come il latino ed il greco antico offrono una base maggiore per lo studio distribuzionale dei mutamenti semantici *tout court*.

Questa relazione ha considerato solo una parte dell'ampia letteratura riguardante la rilevazione automatica dei mutamenti semantici. In particolare, la trattazione è stata circoscritta al LSC, ma ulteriori applicazioni di ricerca potrebbero riguardare la specificazione del tipo di mutamento lessicale o altre forme di mutamento semantico. Un esempio eloquente è rappresentato da Rodda *et al.* (2017), in cui si sviluppa un modello *count-based* per valutare il mutamento diacronico nella flessibilità di alcune espressioni poliriematiche dell'epica greca.

I risultati incoraggianti ottenuti sul latino e sul greco antico suggeriscono inoltre un'estensione dei metodi distribuzionali per il LSC ad altre lingue storiche, la quale tuttavia può essere maggiormente ostacolata dalla difficile reperibilità di sufficienti dati linguistici. Alcuni tentativi vanno in direzione di modelli neurali e contestuali: un esempio è rappresentato Sandhan *et al.* (2021), in cui si considerano diversi modelli per la costruzione di uno spazio vettoriale per il sanscrito. Potrebbe tuttavia essere utile una maggiore attenzione all'ottimizzazione di modelli *count-based* alla luce della loro maggiore competitività in presenza di corpora di ridotte dimensioni. Ad esempio, Jiang *et al.* (2018) propongono un diverso metodo di fattorizzazione della matrice di co-occorrenza denominato *PU-Learning* che migliora le prestazioni del modello per lingue *low-resource*.

La rilevazione automatica del LSC è del resto un settore della linguistica computazionale che si è sviluppato solo negli ultimi anni e che sta ricevendo solo oggi maggiori attenzioni grazie al progressivo perfezionamento dei modelli e alla crescente disponibilità di dati annotati. La piena efficacia dei modelli distribuzionali per il LSC soffre ancora della mancanza di riferimenti oggettivi per lo sviluppo e la valutazione dei modelli. A questo scopo, si

richiederebbe una maggiore collaborazione tra gli esperti nel campo umanistico e gli esperti nel campo informatico in un'ottica pienamente bidirezionale: per la creazione di *benchmark* per lo sviluppo e la valutazione di modelli del LSC; e per una sistematica integrazione dell'evidenza distribuzionale nello studio diacronico del significato.

Bibliografia

Atkins, S., Clear, J., Ostler, N. (1992). Corpus Design Criteria. In *Literary and Linguistic Computing*, 1(1): 1–16.

Bamman, D., Burns, P. J. (2020). Latin BERT: A Contextual Language Model for Classical Philology. 10.48550/arXiv.2009.10053

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching Word Vectors with Subword Information. In *Transactions of the Association for Computational Linguistics*, 5:135-146.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. In *Annual Review of Linguistics*, vol. 6:213-234.

Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). BERT: Pre-Traning of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805

Firth, J. R. (1957). Papers in Linguistics 1934–1951. Oxford University Press, London.

Harris, Z.S. (1954). Distributional Structure. In *Word*, 10:146-162.

Kutuzov, A., Øvrelid, L., Szymanski, T., Velldal, E. (2018). Diachronic Word Embeddings and Semantic Shifts: a Survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1384–1397.

Kutuzov, A., Velldal, E., Øvrelid, L. (2020). Contextualized Language Models for Semantic Change Detection: Lesson Learned. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 126–134.

Jiang, C., Yu, H., Hsieh, C., Kai-Wei, C. (2018). Learning Word Embeddings for Low-resource Languages by PU Learning. In *Proceedings of NAACL-HLT 2018* 1024-1034.

Lenci, A. (2018). Distributional Models of Word Meaning. In *Annual Review of Linguistics*, vol. 4:151-171.

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR.*

Perrone, V., Hengchen, S., Palma, M., Vatri, A., Smith, J. Q., e McGillivray, B. (2019). Genre-Aware Semantic Change for Ancient Greek. In *Proceedings of the 1st International*

Workshop on Computational Approaches to Historical Language Change, 56-66.

Perrone, V., Hengchen, S., Palma, M., Vatri, A., Smith, J. Q., e McGillivray, B. (2021). Lexical Semantic Change for Ancient Greek and Latin. In *Computational Approaches To Semantic Change*, 287-310.

Rodda, M. A., Probert, P., e McGillivray, B. (2017). Vector Space Models of Ancient Greek Word Meaning and a Case Study on Homer. In *Traitement Automatique Des Langues*, 60(3), 63–87.

Rodda, M. A., Senaldi, M.S.G., e Lenci, A. (2017). *Panta Rei*: Tracking Semantic Change with Distributional Semantics in Ancient Greek. In *Italian Journal of Computational Linguistics*, 3.

Sahlgren, M., Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantics Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 975-980.

Sandhan, J., Adideva, O., Komal, D., Behera, L., Goyal, P. (2021). Evaluating Neural Word Embeddings for Sanskrit. arXiv:2104.00270

Sprugnoli, R., Moretti, G. (2020). Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas. In *Italian Journal of Computational Linguistics*, 6.

Sprugnoli, R., Passarotti, M., Moretti, G. (2019). Vir is to moderatus as mulier is to intemperans. Lemma Embeddings for Latin. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*, Accademia University Press, Torino.

Tahmasebi, N., Borin, L., Jatowt, A.(2018). Survey of Computational Approaches to Lexical Semantic Change Detection. Language Science Press.