RIFL (2024) SFL: 172-186
DOI: 10.4396/SFL202302

Osservazioni sulla Filosofia della Linguistica Computazionale. Chiarificazione dei presupposti teorici del Natural Language Processing.

Luca Capone
Università di Pisa
luca.capone@fileli.unipi.it

Abstract The article presents an overview of the literature concerning the relationship between recent Language Models and the concept of meaning. Technical advancements have prompted extensive reflections on the implications of these findings for language and semantics studies. These implications are currently fueling a lively debate among scholars across various disciplines, who are engaging in speculative discussions regarding the nature of meaning and its representation. From a philosophical perspective, the theories of meaning that emerged from such reflections often replicate several misconceptions about the nature of language outlined by Ludwig Wittgenstein in his works. The literature exhibits many of the positions criticised by the Austrian philosopher: a psychological understanding of words comprehension, a logicist interpretation of language functioning, a referentialist view of the meaning of linguistic signs, and so forth. This article endeavors to clarify these misunderstandings by drawing upon classic insights from Wittgenstein's work, in order to avoid the theoretical impasses encountered by scholars when analysing Language Models (LM). The benefits of this approach are twofold. Firstly, the phenomenon of meaning is placed in its natural context, that is language, while avoiding interference from unrelated disciplinary fields (psychology, sociology, logic, cognitive sciences, etc.); secondly, the field of theoretical investigation concerning the performance of LMs is cleared of conceptual confusions and it becomes possible to describe the relationship between meaning and the vector representations of models.

Keywords: Semantics, Wittgenstein, Natural Language Processing, Computational Linguistics, Artificial Intelligence

Received 15 03 2024; accepted 24 06 2024.

0. Introduzione

Il *natural language* processing (NLP) è una disciplina appartenente al *machine learning,* protagonista di notevoli avanzamenti negli ultimi dieci anni. Gli ottimi risultati riscontrati nella replicazione di alcune performance umane legate al pensiero verbale, sono da attribuire principalmente allo sviluppo delle reti neurali profonde (Bengio 2009) e all'introduzione di nuove tecniche di addestramento non supervisionato (Pilehvar e Camacho-Collados 2021). I successi in ambito tecnico hanno alimentato riflessioni sulle

ripercussioni di questi risultati per quanto riguarda gli studi sul linguaggio. Tali ripercussioni sono al momento oggetto di un acceso dibattito, all'interno del quale studiosi appartenenti a diversi ambiti disciplinari avanzano speculazioni sullo statuto del *significato* e della sua *rappresentazione* (Bender e Koller 2020, Bender et al. 2021, Sahlgren e Carlsson 2021, Piantadosi e Hill 2022, Søgaard 2022, Lenci 2023). Da un punto di vista filosofico, le teorie del significato e in generale le concezioni del linguaggio emerse da queste riflessioni riproducono molte delle incomprensioni, circa lo statuto del *significare,* evidenziate da Ludwig Wittgenstein nelle sue opere. La letteratura esibisce molte delle posizioni criticate dal filosofo austriaco: concezione psicologizzante della comprensione delle parole, interpretazione *logicista* del funzionamento del linguaggio, comprensione referenzialista del significato dei segni linguistici ecc.

L'articolo si propone di chiarificare questi fraintendimenti tramite alcuni luoghi classici dell'opera di Wittgenstein e di fornire una risoluzione ai vicoli ciechi teorici incontrati dai ricercatori nell'analisi dei Language Models (LM). I fraintendimenti sulla natura del linguaggio presenti in letteratura sono risolti attraverso una concezione della lingua come *sistemazione sempre aperta di forme significanti* (De Mauro 1969: 128). Questo approccio consente, in primo luogo di inserire il fenomeno del significato nel suo contesto naturale, la lingua, lasciando da parte intrusioni da ambiti disciplinari non pertinenti (psicologia, sociologia, logica, scienze cognitive ecc.); in seconda battuta, di dissodare il terreno di indagine teorica intorno alle performance dei LM, descrivendo il rapporto esistente fra significati e rappresentazioni vettoriali apprese dai modelli.

1. Forma logica e referenti nella letteratura sui LM

A partire dal 2017 (anno di rilascio del modello transformer, Vaswani et al. 2017), i LM hanno suscitato un dibattito piuttosto acceso, a cui hanno partecipato ricercatori appartenenti a varie discipline: linguistica, filosofia del linguaggio, ma anche psicologia, ingegneria, informatica, scienze cognitive ecc. (Potts 2020). Il dibattito ha assunto forme molto diverse, gli studiosi si sono chiesti se i modelli davvero *capiscono* i segni che processano, se quella che manifestano sia davvero *intelligenza*, se sono davvero capaci di comportamento *creativo*. Ovviamente, il problema con tutte queste domande è che non esiste un benchmark, un test, una metrica per decidere se rispondere in un senso o in un altro, o meglio, qualsiasi test prodotto non costituisce una prova incontrovertibile, ma può sempre essere rigettato sulla base di un disaccordo sullo statuto del fenomeno sotto esame (il *pensiero*, il *comprendere*, la *creatività* ecc.). Anche a causa di questa non-verificabilità delle precedenti questioni, diversi studiosi hanno eletto il significato a elemento determinante per la comprensione dell'operato dei modelli (Bender e Koller 2021, Gastaldi 2021, Piantadosi e Hill 2022, Søgaard 2022). Le motivazioni sono facilmente intuibili, in primo luogo non è immediatamente chiaro cosa possa voler dire che una rete *comprende* una frase, una parola, un discorso, o il soggetto di un'immagine. Il concetto di significato consente di porre la stessa domanda in termini più chiari, interrogandosi sullo statuto delle rappresentazioni dei modelli in relazione al concetto di significato linguistico. Allo stesso modo, le capacità di *ragionamento* manifestate dai modelli possono essere ricondotte alla manipolazione di segni e in ultima istanza a relazioni fra rappresentazioni dei significati dei segni, le quali producono determinati risultati. Infine, la lingua e in particolare la semantica possono avere un ruolo importante in ciò che comunemente viene chiamato creatività, sia per quanto riguarda i modelli multimodali, i quali realizzano immagini a partire da prompt testuali interpretabili in maniera non univoca, sia per quanto riguarda i modelli che si muovono esclusivamente sul piano del discorso, impiegando i segni propri delle lingue storico naturali in contesti eterogenei. In breve, le varie domande precedentemente elencate, relativamente alle *qualità intellettuali* dei modelli, possono essere condensate in un'unica questione, ovvero nella questione

del rapporto che le rappresentazioni numeriche apprese dai LM intrattengono col fenomeno del significato. È necessario considerare che, anche in questi termini, il problema è tutt'altro che risolto, dato che non è immediatamente chiaro in che rapporto potrebbe stare una rappresentazione vettoriale (una rappresentazione numerica appresa da un LM) con un significato. Il significato stesso si rivela un qualcosa di non autoevidente, che necessita di essere ulteriormente chiarificato. La linguistica e in particolare la semantica non presentano un'unica risposta relativamente alla natura di questo fenomeno, di conseguenza lo studio del rapporto fra *significato* e rappresentazioni dei modelli ha generato esiti diversi, fra chi nega perentoriamente qualsiasi accesso al significato da parte dei LM e chi invece, più ottimisticamente, ritiene che le rappresentazioni vettoriali possano *in qualche maniera* contenere informazioni salienti per la rappresentazione dei concetti. In breve, la domanda circa il rapporto fra rappresentazioni computazionali e fenomeno del significato ripropone il problema della natura del significato nella sua interezza.
Il dibattito cui è stato fatto cenno ha avuto una delle sue manifestazioni più visibili nel *semantic megathread* (Wolf 2018). Si tratta di un thread di Twitter che ha costituto un punto di riferimento, sullo stato della riflessione sull'avanzamento delle ricerche nel NLP. A titolo esemplificativo di seguito sono riportati rispettivamente, il post che ha dato inizio al thread e uno dei post che ha ottenuto più interazioni (figura 1). Andreas è professore presso il dipartimento di Ingegneria Elettronica e Computer Science del MIT, mentre Bender è prof.ssa di Linguistica Computazionale all'Università di Washington. Nel corso del thread gli autori rimandano a diversi lavori pubblicati per chiarificare le proprie posizioni, tuttavia questi due post sono rappresentativi delle rispettive idee e della concezione egemone relativamente al rapporto fra LM e linguaggio.

Fig. 1 - Post rappresentativi del semantic megathread

In primo luogo viene recepito come problematico, o quantomeno degno di nota, il fatto che i LM siano in grado di operare con i segni in maniera semanticamente pertinente senza ricorrere a una forma logica. Andreas si interroga sull'operato dei modelli (e in generale sul linguaggio) aderendo alla cosiddetta *semantica composizionale*. Di conseguenza i suoi lavori trattano di quelle strutture che, secondo i sostenitori di queste posizioni, costituiscono la condizione di sensatezza delle espressioni. Nello specifico Andreas si è occupato diffusamente del rapporto fra modelli distribuiti e strutture logiche (Andreas 2019, Mu e Andreas 2020, Murty et al. 2022). Nel NLP (e ancor di più in filosofia) la tematica è piuttosto conosciuta, la si può far risalire al dibattito fra simbolisti e connessionisti (Fodor e Pylyshyn 1988). Secondo Fodor, esponente di spicco del paradigma simbolico, i significati sono costituiti da contenuti atomici nella mente dei soggetti, articolabili secondo vari rapporti di ordine logico. La possibilità di articolare il pensiero si basa su questo codice innato (il cosiddetto *language of thought*, LOT). «So then, here's how it goes in your head, according to this version of LOTH [LOT hypothesis], when you intend to make it true that *P*. What you do is, you put into the intention box a token of a mental symbol that *means* that P» (Fodor 1987: 136–137).

Le posizioni simboliche non sembrano aver perso impeto col passare del tempo, tant'è che le si ritrovano nella letteratura recente. Ad esempio Friedman, relativamente all'addestramento basato su corpora dei LM, scrive:

> natural language is not an ideal form for the reliable communication of concepts. Instead, formal logic statements are preferable since they are subject to verifiability, reliability and applicability […] natural language is limited in its capacity for the construction of reproducible knowledge [...] generally relies on an imprecise use of words and their meanings (Friedman 2023: 687).

Affermazioni che fanno dubitare della sincerità dell'autore, considerando il fatto che l'articolo è scritto nel vituperato *natural language* o meglio, in una delle lingue storico naturali, l'inglese. Sulla stessa linea: «we should not expect DL [deep learning] models to illuminate linguistic theory» (Dupre 2021: 617) e «When neural network modeling is integrated with grammatical formalisms […], we may be able to go further in assessing the extent to which grammatical representations can be learned from experience, and what aspects of the grammar must be hard-wired» (Pater 2019: 27).
L'idea secondo cui i contenuti si articolano seguendo strutture innate e universali ha indirizzato pesantemente lo studio dei LM. A partire da queste stesse posizioni, diversi autori (ad esempio Kovaleva et al. 2019) cercano di analizzare i layer interni dei modelli alla ricerca di quelle strutture sintattiche (o logiche) su cui, secondo questi studiosi, dovrebbe basarsi il linguaggio (un'idea ereditata dal generativismo[1]). La natura distribuita delle reti neurali tuttavia frustra queste ricerche, gli autori stessi esprimono perplessità di fronte ai risultati, i quali non consentono di rinvenire le strutture desiderate all'interno dei modelli (Rogers et al. 2020, Kulmizev e Nivre 2021). La forma logica cui si riferiscono questi autori, non sembra rappresentata nei layer interni e dunque non è ciò che consente il funzionamento del modello, al contrario, la regolarità sintattica è ciò che *emerge* dal suo funzionamento interno, ovvero dalle operazioni su rappresentazioni distribuite dei segni della lingua. In questi termini, la forma logica, la sintassi, o la grammatica, risulta non una condizione a priori del linguaggio (o nei termini di Fodor, del pensiero) ma una descrizione a posteriori, frutto di un rapporto riflessivo col discorso. Il punto non è sfuggito a Buder-Gröndahl (2023), il quale giustamente pone l'accento sull'ambiguità del concetto di *rappresentazione linguistica* utilizzato in letteratura. Buder-Gröndahl conclude il suo articolo con una giusta domanda relativa a questi tentativi *logicisti* di spiegazione dei modelli:

> The field of "BERTology" [la branca del NLP che si occupa di spiegare come e perché i LM funzionano, nello specifico i modelli basati su BERT (un tipo di architettura per il *language modeling*)] aims to locate linguistic representations in large language models (LLMs). These have commonly been interpreted as representing structural descriptions (SDs) familiar from theoretical linguistics, such as abstract phrase structures […] This leaves a central question unexamined: are those SDs actually needed for describing the LLM on a high level of abstraction? (Buder-Gröndahl 2023: 28).

L'autore sintetizza efficacemente il discorso fin qui svolto. Nell'ambito del NLP, gli studiosi hanno immaginato le rappresentazioni linguistiche come delle entità iscritte

[1] A proposito dei LM, Chomsky in un'intervista afferma: «it's- um it's true there's been a lot of work on um trying to apply uh statistical uh models to various linguistic problems uh I think there have been some successes, but a lot of failures. […] the successes that I know of are those that integrate statistical analysis with some universal grammar properties, some fundamental properties of language; when they're integrated, you sometimes do get results» (Chomsky 2011).

all'interno dei layer dei LM. Queste rappresentazioni sono state intese in accordo col paradigma linguistico dominante, ovvero come strutture proposizionali (*abstract phrase structures*). A questo punto, la seguente obiezione diventa legittima: per quale motivo viene dato per certo che le rappresentazioni linguistiche teorizzate da una parte (sebbene maggioritaria) della linguistica teorica (ovvero le strutture proposizionali) debbano essere fondamentali per descrivere e spiegare il funzionamento dei LM? Perché si ritiene che i modelli, debbano necessariamente aver appreso e rappresentato al loro interno queste strutture in maniera esplicita e localizzabile[2]? Questo è il primo assunto da analizzare relativamente al modo in cui i LM vengono studiati in letteratura, un'assunzione esemplificata dal post di Andreas.
Parallelamente a questo assunto, se ne trova un secondo, strettamente correlato al primo ed esemplificato dal post di Bender. L'idea sostenuta è che qualsiasi rappresentazione basata esclusivamente sulla *forma* (sulla forma dell'espressione), non può costituire una rappresentazione del significato. Anche in questo caso oltre a Bender, la quale ha esposto questa posizione in più occasioni, si trovano posizioni analoghe in letteratura (Bisk et al. 2020, Merrill et al. 2021). Si tratta fondamentalmente di una riproposizione del *symbol grounding problem* di Harnad (1990). «The futility of learning language from linguistic signal alone is intuitive, and mirrors the belief that humans lean deeply on non-linguistic knowledge» (Bisk et al. 2020). E ancora:

> We argue that the language modeling task[3], because it only uses form as training data, cannot in principle lead to learning of meaning […] We take meaning to be the relation between the form and something external to language (Bender e Koller 2020: 5187).

Ricapitolando, le precedenti posizioni, tradotte in positivo equivalgono a dire che una buona rappresentazione (o almeno, quello che ci si aspetta da una buona rappresentazione) del *linguaggio* deve prendere in considerazione due cose: 1) la forma logica, 2) i referenti. Esiste una teoria del linguaggio che abbraccia questi due elementi, ovvero la *picture theory* del *Tractatus Logico-Philosophicus* (Wittgenstein 1921). In breve, secondo questa teoria le *espressioni* (le proposizioni) stanno per *stati di cose* e i nomi che compongono le espressioni stanno per gli elementi che compongono questi stati di cose (gli oggetti o referenti). Un'espressione può stare per uno stato di cose per via dell'*isomorfismo* logico fra le due. Ovvero la *forma logica* della proposizione *mostra*, *esibisce* la forma di uno stato di cose. Una proposizione significa (è *sensata*), perché rispecchia qualcosa che sta al di fuori della lingua (condividendone la struttura fondamentale) (Wittgenstein 1921: §3). Di conseguenza, una frase costituisce «un modello della realtà così come noi ce la immaginiamo» (Wittgenstein 1921: §4.01). È chiaro come la teoria esemplificata dal post di Andreas sia facilmente associabile a queste posizioni. Pensiero, linguaggio e mondo, si uniformano a una struttura universale che consente il rimando reciproco fra queste tre istanze. Questo nodo teorico è ciò che fa dire ad Andreas che è sorprendente che non ci sia una forma logica dietro al funzionamento dei modelli. Questo stesso grumo concettuale è ciò che dà sostanza alle teorie secondo le quali *devono*

[2] Questa assunzione fa il paio con l'idea mentalista per cui le strutture proposizionali che si suppone spieghino il funzionamento del linguaggio debbano trovarsi materialmente *nelle teste dei parlanti* (Fodor 1993, p. 136–137).
[3] Il task di *language modeling* è il tipo di addestramento a cui sono sottoposti i LM. Il task consiste nel predire il successivo token in una sequenza di input. Le rappresentazioni dei segni prodotte da questo task sono strettamente correlate con la distribuzione di parole all'interno di testi, un qualcosa di radicalmente non *grounded*. Il *language modeling* verrà approfondito nel seguente paragrafo.

esserci delle strutture innate tali per cui è possibile apprendere (o rappresentare) una lingua.
Il *Tractatus* non esprime una teoria semantica originale (cosa di cui l'autore era ben consapevole), bensì può essere descritto come la massima formalizzazione di una concezione del linguaggio piuttosto antica. Così sembra pensarla De Mauro (1969: 98), il quale accosta la teoria linguistica del primo Wittgenstein a una corrente di pensiero che chiama *aristotelismo linguistico*[4].

> Nella concezione di Aristotele ripresa da Wittgenstein la comunicazione è spiegata affermando che le parole hanno un significato stabile, e tale stabilità è concepita come nascente dal fatto che le parole denotano cose e nozioni "che sono le stesse per tutti". La serie delle parole è legata alla serie delle cose da una corrispondenza biunivoca: in virtù di tale corrispondenza le parole sono "fondamentalmente le stesse per tutti" (De Mauro 1969: 167).

Il ritenere che i contenuti siano elementi atomici dati, con cui comporre espressioni che stanno per fenomeni indipendenti è precisamente ciò che propone una concezione extralinguistica del significato, come quella presentata da Bender. L'obiezione secondo cui un addestramento basato sulla forma significante del discorso non può dar luogo a rappresentazioni del significato delle espressioni, comporta ritenere tali espressioni al pari di etichette, le quali stanno per entità indipendenti, stabili, la cui identità non è in questione, ma che anzi costituiscono la condizione di possibilità (se non la causa) dei significati.
Riassumendo, questi due studiosi, portavoce del dibattito in materia, esemplificano un sentimento comune. Sia che si critichino, sia che semplicemente ci si interroghi sui modelli, l'idea di linguaggio di partenza è vicina a questo nucleo tematico riconducibile al *Tractatus* e ancora più oltre all'aristotelismo linguistico. Un nucleo tematico che, al di là della sua scarsa efficacia nello spiegare il fenomeno della significazione in quanto tale, presenta problemi nel rapportarsi col modo in cui i modelli rappresentano la lingua e replicano le performance linguistiche umane. Come emergerà dal seguente paragrafo, queste posizioni non sono applicabili ai LM per questioni tecniche.

2. LM e rappresentazione vettoriale

Per comprendere il rapporto fra significato e rappresentazione vettoriale dei segni è necessario introdurre alcune nozioni relative al funzionamento dei modelli. Le operazioni interne dei LM, le quali consentono l'elaborazione e la generazione di testo scritto, sono fondamentalmente delle moltiplicazioni fra matrici di vettori. I vettori sono l'elemento chiave del funzionamento delle reti neurali (non solo dei LM). I vettori sono array (liste) di numeri tramite cui i segni vengono rappresentati dal modello. I valori numerici di un vettore costituiscono le coordinate che collocano una parola in uno spazio multidimensionale (tante dimensioni, quanti sono gli elementi del vettore). La disposizione reciproca fra i segni all'interno di questo spazio ne determina il valore, in modo tale che parole con significati simili (o correlati) saranno rappresentate vicine nello spazio vettoriale. Il procedimento che consente l'*apprendimento* di queste rappresentazioni è chiamato *language modeling*. Uno dei primi modelli di rete neurale per il *language modeling*, modello da cui discendono le implementazioni più recenti, è Word2Vec (W2V, Mikolov et al. 2013). Le reti neurali sono essenzialmente delle funzioni parametriche che approssimano distribuzioni di dati (in questo caso, la distribuzione di parole nei testi di addestramento) modificando i propri parametri (i valori interni del modello che

[4] La denominazione non implica che questa sia la teoria del linguaggio di Aristotele (posizione dibattuta, Lo Piparo 2003). Per comodità verrà utilizzata la denominazione proposta da De Mauro.

costituiscono i vettori, le rappresentazioni numeriche dei segni conosciuti dal LM). Affinché questa approssimazione possa aver luogo, il modello necessita di un task di addestramento (qualcosa da predire, in funzione del quale modificare i propri valori interni), questo task è il *language modeling*. L'addestramento comincia con un modello inizializzato con valori casuali, un corpus e un vocabolario di termini (parole o pezzi di parole). Il modello riceve il corpus in *batch* (porzioni di testo). Ogni termine viene codificato tramite un vettore (composto da centinaia di elementi, inizialmente casuali). L'obiettivo del *language modeling* consiste nel predire il termine successivo a quelli della sequenza ricevuta. Questa predizione viene realizzata nella forma di una distribuzione di probabilità su tutti gli elementi del vocabolario, in questo modo il modello identifica i token che completano l'input in termini probabilistici. L'addestramento viene implementato ricercando quei valori (quei parametri interni) che minimizzano l'errore delle predizioni. I valori risultanti costituiscono i vettori (i *word embedding*) che rappresentano il significato dei termini del vocabolario (un vettore per ogni elemento). I vettori possono essere rappresentati graficamente (dopo opportuna riduzione dimensionale). È utile immaginare queste rappresentazioni come degli elementi che si oppongono reciprocamente, in uno spazio di molte dimensioni (figura 2).

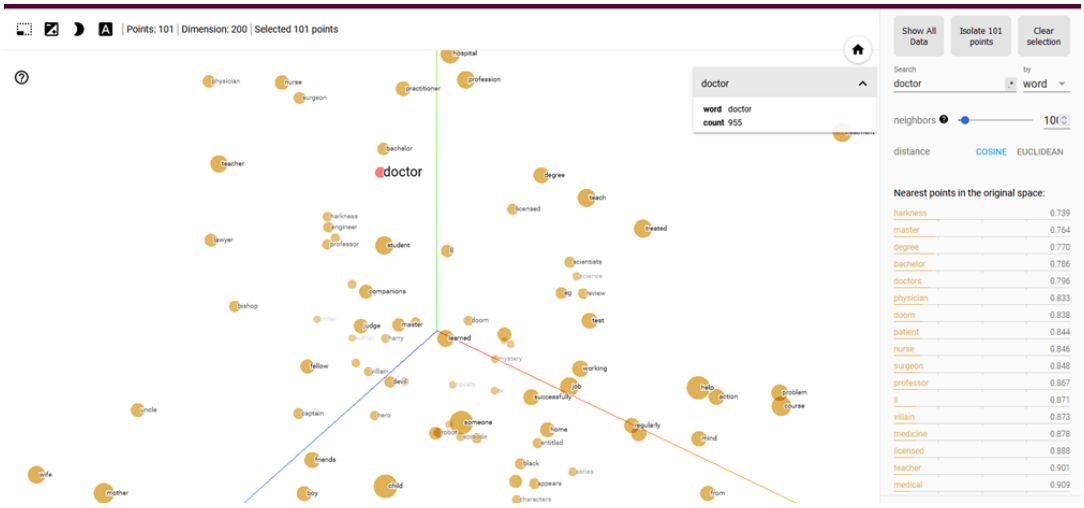

Fig. 2 - Riduzione a tre dimensioni dei 100 embedding (o vettori) più vicini al vettore del token "doctor" (Smilkov et al. 2016)

Da una prospettiva linguistica, la rete si basa sulle relazioni sintagmatiche fra termini del discorso per calcolare la probabilità di un token di essere in relazione di contiguità rispetto alla sequenza di input. Ovvero, la rete è addestrata a predire il token del suo vocabolario che con maggiore probabilità può completare la sequenza di input. Durante l'addestramento, le rappresentazioni dei termini vengono modificate in funzione dell'errore commesso nel calcolo di questa probabilità. Una volta addestrati, i vettori di termini che condividono molti contesti, esibiranno relazioni di associazione all'interno dello spazio vettoriale. Questi rapporti, sintagmatici e associativi (Saussure 1916: 157), sono i rapporti in funzione dei quali i token sono rappresentati come simili o dissimili all'interno dello spazio vettoriale. Ad esempio, gli articoli determinativi, avendo tutti una distribuzione simile (occorrendo negli stessi contesti) finiranno per essere rappresentati in aree ravvicinate. Lo spazio vettoriale non esibisce solo relazioni di *similitudine*, ma colloca i token all'interno di questo spazio secondo svariati rapporti reciproci (figura 3).

RIFL (2024) SFL: 172-186
DOI: 10.4396/SFL202302

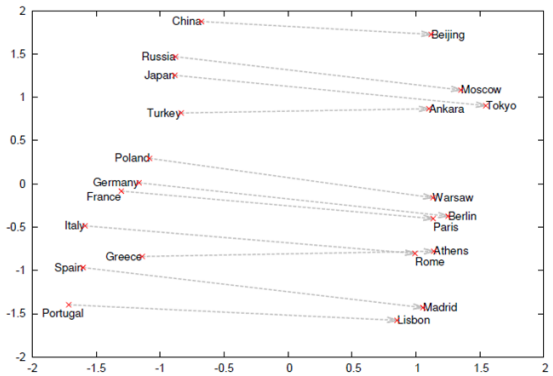

Fig. 3 - Rapporto fra embedding che rappresentano nomi di stati e relative capitali

I vettori (o come solitamente vengono chiamati, *word embedding*) appresi da W2V consentono applicazioni limitate, ciò è dovuto al fatto che sono statici. Ovvero, nel momento di codificare un token, a prescindere dal contesto sintagmatico in cui occorre, questo riceverà sempre la stessa rappresentazione. Questi vettori sono il frutto di una media pesata fra tutte le co-occorrenze all'interno del corpus di addestramento. Di conseguenza, gli *embedding* del modello, sebbene possano costituire una approssimazione del piano del contenuto di una lingua, ancora non possono determinarsi in sensi concreti (l'embedding di *anello*, varrà per *anello nunziale, un anello per domarli, anello di una catena* ecc.). Questo modello è stato successivamente rimpiazzato dall'architettura transformers, la stessa architettura che ha consentito lo sviluppo dei Large LM. Questo nuovo tipo di LM supera le rappresentazioni di W2V, immettendo i vettori statici in una ulteriore rete, la quale contestualizza (modifica) i vettori, in funzione del sintagma in cui occorrono[5] (figura 4). Ciononostante, il principio di base di selezione e contiguità che struttura l'addestramento e la natura relazionale delle rappresentazioni vettoriali, rimangono invariati.

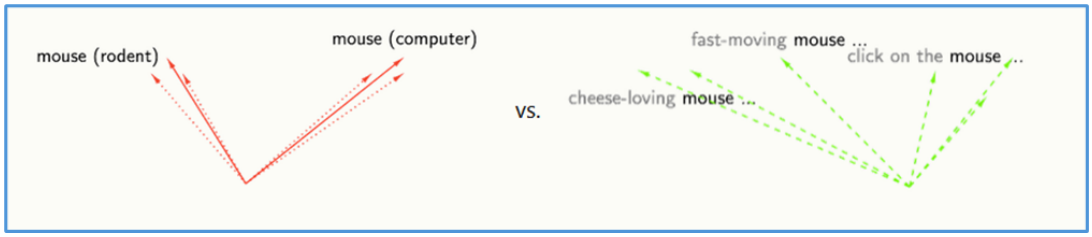

Fig. 4 - Omonimia e polisemia nello spazio vettoriale di un modello basato su transformer (http://ai.stanford.edu/blog/contextual/)

3. Inapplicabilità delle posizioni riconducibili al *Tractatus*

Ricapitolando, quelli che per i parlanti sono i significati dei termini conosciuti dal modello vengono rappresentati tramite vettori. I numeri che compongono i vettori non hanno un valore positivo, non rappresentano nulla se non il rapporto che un termine intrattiene con gli altri termini del sistema cui appartiene. Di conseguenza, è possibile affermare che i vettori hanno un valore puramente relazionale. L'idea secondo cui il significato corrisponde al valore oppositivo di un elemento in un sistema non costituisce una novità in linguistica:

> all'interno d'una stessa lingua, tutte le parole che esprimono delle idee vicine si limitano reciprocamente: sinonimi come *redouter, craindre, avoir peur* hanno un loro

[5] La procedura, chiamata *attention mechanism*, si svolge operando dei confronti tra vettori appartenenti allo stesso sintagma (Vaswani et al. 2017, Ethayarajh 2019)

> proprio valore solo per la loro opposizione; se *redouter* non esistesse, tutto il suo contenuto andrebbe ai suoi concorrenti (Saussure 1916: 141).

Secondo Saussure, il valore di un segno non è determinato positivamente (dalle caratteristiche positive di un oggetto cui può riferirsi), ma emerge dalla relazione che quel segno intrattiene con tutti gli altri segni del sistema di cui fa parte. Un LM cerca di approssimare questi valori oppositivi computazionalmente, in uno spazio di molte dimensioni. Questo consente alla rete di poter rappresentare molti tipi di relazioni fra segni.

Da questo punto di vista diviene comprensibile perché i LM risultano problematici per gli autori che si rifanno a una prospettiva linguistico-semantica riconducibile a quella del *Tractatus*. Per Andreas, il non trovare una struttura a priori che determina i rapporti fra unità linguistiche è inaspettato. Non solo relativamente al rapporto fra espressioni e stati di cose, ma anche e soprattutto per il fatto che, non essendo data una forma a priori della proposizione, potenzialmente qualsiasi insieme di segni è accettabile come espressione dotata di senso, cosa sostenuta ad esempio da Jakobson (2002: 170, in cui l'autore interpreta la celebre frase "colorless green ideas sleep furiously" (Chomsky 1957, trad. it.: 16), confutandone la presunta insensatezza). Se questo è vero, vuol dire che non sono a priori definibili procedure (forme logiche) per determinare il rapporto fra segni all'interno delle espressioni, non esiste una grammatica che stabilisce le proposizioni appartenenti a una lingua (proprietà che possiede ad esempio il calcolo, chiamata *connessità sintattica*, De Mauro 1982: 89–90). Lo stesso Wittgenstein delle *Ricerche filosofiche*, mette in guardia da simili fraintendimenti, scaturiti dalla grammatica superficiale delle lingue:

> È interessante confrontare la molteplicità degli strumenti del linguaggio e dei loro modi d'impiego, la molteplicità dei tipi di parole e di proposizioni, con quello che sulla struttura del linguaggio hanno detto i logici. (E anche l'autore del *Tractatus logico-philosophicus*.) (1953: I, § 23).

Per quanto riguarda Bender e la concezione referenzialista del significato, la sua posizione fatica nello spiegare com'è possibile che i LM siano in grado di generare testi comprensibili e talvolta di buona qualità, ottenendo punteggi notevoli in differenti benchmark di competenza semantica, senza ricorrere a nessuna realtà extralinguistica in sede d'addestramento. Sempre Wittgenstein, nelle *Ricerche*, critica fortemente la definizione ostensiva (il ricorso ai referenti) come strumento di definizione dei significati.

> La definizione ostensiva spiega l'uso – il significato – della parola, quando sia già chiaro quale funzione la parola debba svolgere, in generale, nel linguaggio. Così, la definizione ostensiva: "Questo si chiama 'seppia'" aiuterà a comprendere la parola se so già che mi si vuol definire il nome di un colore. – E questo si può dire se non si dimentica che alle parole "sapere" ed "essere chiaro" sono connesse questioni di ogni genere (1953: I, § 30).

La concezione secondo cui i significati sono costituiti dai referenti, presenta diversi problemi. Ad esempio, se le espressioni stanno per stati di cose e i nomi stanno per gli oggetti, non si capisce che significati attribuire a diverse parti del discorso, come ad esempio le *stopwords*[6]. Da questo punto di vista la teoria sembra poter rendere conto esclusivamente del significato dei sostantivi.

[6] Nel NLP, articoli, avverbi, congiunzioni, parole ritenute semanticamente poco informative

> Chi descrive in questo modo l'apprendimento del linguaggio pensa, così credo, anzitutto a sostantivi come "tavolo", "sedia", "pane" e ai nomi di persona, e solo in un secondo tempo ai nomi di certe attività e proprietà; e pensa ai rimanenti tipi di parole come a qualcosa che si accomoderà (1953: I, § 1).

Tuttavia, nemmeno tutti i sostantivi si riferiscono a cose, ad esempio *giustizia, felicità* ecc. Non è chiaro a quali oggetti extralinguistici debba riferirsi questa classe di parole. Di fronte a questa obiezione, alcuni autori potrebbero correre ai ripari spiegando che la parola *felicità* si riferisce a qualche presunto stato interno. Ma il significato della parola *felicità* non può essere ridotto a un processo fisiologico. È senz'altro possibile, in circostanze specifiche, riferirsi con questa parola a un processo fisiologico, ma non è questo che viene chiamato in causa in una ipotetica spiegazione del significato della parola, come qualsiasi parlante competente può testimoniare. Come specificato nelle *Ricerche,* i criteri per l'utilizzo di termini come *felicità, ricordare, leggere* ecc. poggiano su criteri esterni (1953: I, § 580) ed è a partire da questi criteri che il significato di una parola può essere descritto. Nel caso di *felicità*, il significato non corrisponderà a una definizione del tipo "felicità è quando i tuoi neuroni scaricano in questo modo", ma tirerà in ballo esempi, circostanze specifiche, storie accadute o ipotetiche. Se il significato della parola dipendesse da criteri interni al cervello o alla psiche dei soggetti, il termine sarebbe inutilizzabile, i criteri per la sua applicazione sarebbero inaccessibili ai parlanti. Tra l'altro, non è garantito che si dia un processo fisiologico univocamente identificabile come *processo della felicità.* È probabile che possano essere identificati dei processi fisiologici concomitanti a situazioni in cui è appropriato parlare di *felicità,* ma questi non sono pertinenti per la definizione di un fenomeno semiotico, semmai il contrario, è solo a partire da criteri esterni che qualcuno può essere indotto a ritenere che *qualcosa deve star accadendo dentro* il proprio interlocutore.

> Ciò che neghiamo è che l'immagine del processo interno ci dia l'idea giusta dell'impiego della parola "ricordare". Anzi, diciamo che quest'immagine, con le sue ramificazioni, c'impedisce di vedere l'impiego della parola quale esso è veramente (1953: I, § 305).

Ultimo e più generale problema della concezione referenziale del significato è l'assunzione degli oggetti di esperienza come condizione esclusiva dei contenuti di una lingua. Al di là del fatto banale per cui sono possibili enunciati dotati di senso ma falsi o relativi a entità inesistenti (Eco 1975: 106), nemmeno il significato degli enunciati apparentemente meno problematici può essere semplicemente *estratto* dalla materia dell'esperienza. Il significato di un'espressione come *questa valigia è pesante* non presuppone tanto l'aver saggiato effettivamente l'oggetto, quanto principalmente la disponibilità di un sistema di opposizioni (pesante/leggero) per caratterizzare l'esperienza e ancor più profondamente una struttura (in questo caso sostanza-predicato) che consenta di articolarne la forma. Solo a valle di queste condizioni linguistiche, l'inarticolata e iper-specifica esperienza senso-motoria acquisisce la sua intelligibilità, generalità, comunicabilità e articolabilità[7]. In generale, l'esperienza non può determinare autonomamente e univocamente i contenuti della lingua, al contrario,

[7] È opportuno specificare che non si intende sostenere che l'esperienza non gioca alcun ruolo nell'articolazione dei contenuti della lingua. Semplicemente, la materia dell'esperienza non può determinare in maniera univoca dei contenuti. L'innegabile concretezza dell'esperienza, prestandosi a molteplici pertinentizzazioni possibili, necessita di una ulteriore condizione affinché alcuni (e non altri) tratti vengano pertinentizzati. Queste condizioni sono fornite dalle lingue storico naturali (Eco 1997: 71).

la competenza in una lingua canalizza l'attenzione dei parlanti e indirizza la pertinentizzazione dell'esperienza. Nell'apertura delle *Ricerche,* Wittgenstein fa notare come la descrizione dell'apprendimento linguistico offerta da Agostino, squisitamente referenzialista, debba presupporre surrettiziamente le condizioni linguistiche che dovrebbero invece essere il risultato dell'addestramento.

> Agostino descrive l'apprendimento del linguaggio umano come se il bambino […] possedesse una lingua, ma non questa. O anche: come se il bambino fosse già in grado di pensare, ma non ancora di parlare. E qui "pensare" vorrebbe dire qualcosa come: parlare a sé stessi (1953: I, § 32).

A partire da tutto questo è possibile sostenere che le preoccupazioni, le perplessità e le critiche basate su presupposti teorici referenzialisti, non solo non sono applicabili ai LM, ma poggiano su concezioni problematiche relativamente alla natura del linguaggio, a lungo discusse nel corso della storia della filosofia e in particolar modo nell'opera di Wittgenstein. Resta a questo punto da risolvere la questione di come pensare il rapporto fra rappresentazioni dei modelli e significati.

In base a quanto visto finora è possibile affermare che, relativamente alle rappresentazioni linguistiche dei modelli, non è necessario ipotizzare alcun elemento extralinguistico che ne determini la forma. Tutto ciò che i modelli apprendono lo apprendono tramite la forma dell'espressione del discorso. Questo porta a una conclusione non nuova per quanto riguarda gli studi su Wittgenstein, ma che diventa interessante se rapportata al funzionamento dei LM, la conclusione secondo cui la lingua struttura l'esperienza. Meglio ancora, «la lingua è una sistemazione collettiva dell'esperienza semantica» (De Mauro 1969, p. 197). Wittgenstein esprime questa preminenza (trascendentale) del linguaggio rispetto all'esperienza in diversi luoghi, ad esempio: «che tipo di oggetto una cosa sia: questo dice la grammatica» (1953: I, § 373), «come faccio a sapere che questo colore è rosso? Una risposta potrebbe essere questa: "ho imparato l'italiano"» (I, § 381), «il concetto *dolore* l'hai imparato con il linguaggio» (I, § 384). Di conseguenza, la questione iniziale sul rapporto fra rappresentazioni dei modelli e significati va in parte riformulata. I vettori dei modelli non possono rappresentare *singolarmente* i significati. Affinché un modello possa avere accesso ai significati della lingua, deve aver accesso a quella organizzazione dell'esperienza, a quel sistema di segni che struttura questi significati. La mancanza di questa visione sistemica è uno dei maggiori problemi nella letteratura sui LM.

Nelle *Ricerche filosofiche,* a partire da un'analisi di alcune proposizioni ordinarie, Wittgenstein fa emergere la natura sistemica della significazione. Una volta stabilito che l'esperienza non può autonomamente fornire contenuti semantici, ma semplicemente si presta a molteplici (e in linea di principio innumerevoli) pertinentizzazioni possibili, la lingua, intesa come sistema che organizza contenuti, deve intervenire per mettere un ordine (uno degli ordini possibili) all'interno dell'indeterminazione dell'esperienza. Di conseguenza, un'espressione come *questa valigia è pesante* significa non perché semplicemente un parlante sta *sentendo* la pesantezza di un determinato oggetto, ma perché il sistema della lingua distingue certi contenitori adibiti al trasporto di beni di consumo da altri (un container, un pacco postale ecc.), perché prevede che gli oggetti possano essere *genericamente* caratterizzati in base al loro peso come leggeri o pesanti e perché il contenuto identificato dall'espressione *valigia* tende a pertinentizzare certe proprietà (peso, dimensioni, capienza), a scapito di altre (aerodinamicità, resistenza alle alte temperature) che comunque avrebbero potuto essere poste in evidenza in altre circostanze specifiche. In questi termini, risulta più chiaro in che modo il significato di un enunciato, non venendo direttamente causato dall'esperienza, dipenda dal sistema di

relazioni messo in gioco dalla lingua. In altre parole, prima di poter sensatamente chiedere il nome di qualcosa, ma anche prima di poter pronunciare (o comprendere) un giudizio, molto deve essere già pronto a livello di struttura linguistica (Wittgenstein 1953: I, § 30). Nella parabola del pensiero di Wittgenstein, la nozione di sistema è sempre presente in sottotraccia, assumendo però aspetti radicalmente diversi. Nel *Tractatus* il sistema corrisponde alla struttura logica che consente la relazione tra espressioni e stati di cose, il sistema ha in questo caso uno statuto *trascendente,* indipendente dalle lingue storico naturali. Diversamente, nelle *Ricerche* la nozione di sistema assume l'aspetto di una *grammatica trascendentale,* la quale dipende dall'apprendimento di una specifica lingua storico naturale, articola l'esperienza dei parlanti e ne canalizza l'esplorazione del mondo.

A questo punto, per chiarire del tutto il rapporto fra semantica e rappresentazioni vettoriali è necessario integrare l'idea generale di sistema emersa dall'opera dell'ultimo Wittgenstein con una sua descrizione più esplicita e formale. La linguistica strutturale, nella sua formulazione saussuriana (ricostruita da De Mauro), fornisce questa integrazione. I LM, basandosi sui rapporti fra termini all'interno del discorso sono in grado di apprendere rappresentazioni oppositive del significato dei segni, collocandoli in uno spazio vettoriale. Lo spazio vettoriale costituisce una modellizzazione del piano del contenuto di una lingua (Capone 2021, Gastaldi 2021) come descritto da Saussure nel *Corso di Linguistica Generale* (Saussure 1916: 142). Ovvero, a partire da un'analisi distribuzionale degli elementi significanti (o assunti come tali), un LM approssima le articolazioni che definiscono il piano del contenuto. In questo modo il modello è in grado di ritagliare un continuum, lo spazio vettoriale, in accordo con le unità del piano del significato di una lingua storico-naturale.

4. Conclusioni

Il presente lavoro ha cercato di fornire una panoramica degli studi sul rapporto fra significato e LM. Questo scenario viene efficacemente esemplificato dalla parabola *terapeutica* tracciata dal passaggio dal primo al secondo Wittgenstein. Ovvero, il *Tractatus*, descrivendo il linguaggio come il rivestimento esterno di una struttura indipendente già da sempre data, fornisce una teoria semantica insoddisfacente, che riduce il significato al mero riferimento denotativo. Gli stessi risultati sono conseguiti dagli studiosi contemporanei, i quali cercano tracce di strutture logiche all'interno dei modelli, relegando la questione semantica al problema del riferimento (del rapporto fra strutture sintattiche e stati di cose). Nella trattazione è stato illustrato che, se l'analisi parte dall'assunto secondo cui i significati sono determinati dall'incontro con gli oggetti o gli stati di cose, non è possibile capire in qual modo i LM riescono a formulare le rappresentazioni del significato dei segni di una lingua. Una volta posto il problema in termini referenzialisti, l'unica soluzione possibile è negare qualsiasi relazione fra significato e rappresentazioni vettoriali, rinunciando di fatto a spiegare le performance esibite dai LM. Per evitare questo esito, si è rivelata necessaria una chiarificazione dei termini del discorso, in questo caso, una chiarificazione del concetto di *significato*. Una volta sgomberato il terreno dai fraintendimenti derivati da una concezione referenzialista della significazione, l'introduzione di un approccio strutturalista nello studio dei LM ha consentito di superare le difficoltà, delineando i termini del rapporto esistente fra rappresentazioni vettoriali apprese dai LM in fase di addestramento e significato.

References

Andreas, Jacob (2019), «Measuring compositionality in representation learning» in *arXiv*: 1902.07181.

Bender, Emily, e Koller, Alexander (2020), «Climbing towards NLU: On meaning, form, and understanding in the age of data» in *Proceedings of the 58th Annual Meeting of the ACL,* pp. 5185–5198.

Bender, Emily, et al. (2021), «On the dangers of stochastic parrots: Can language models be too big?» in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* pp. 610–623.

Bengio, Yoshua (2009), « Learning deep architectures for AI » in *Foundations and Trends in Machine Learning*, 2(1), pp. 1-127.

Bisk, Yonatan, et al. (2020), «Experience grounds language» in *arXiv*: 2004.10151.

Buder-Gröndahl, Tommi (2023), «The ambiguity of BERTology: What do large language models represent?» in *Synthese*, 203(1).

Capone, Luca (2021), «Which theory of language for deep neural networks? Speech and cognition in humans and machines» in *Technology and Language*, 2(4).

Chomsky, Noam (1957). *Syntactic structures,* Mouton & Co., The Hague Paris (*Le strutture della sintassi*, trad. Di, F. Antinucci, Laterza, Roma-Bari, 1974).

Chomsky, Noam (2011), «Keynote panel: the golden age — a look at the original roots of artificial intelligence, cognitive science, and neuroscience» http://languagelog.ldc.upenn.edu/myl/PinkerChomskyMIT.html

De Mauro, Tullio (1969), *Introduzione alla semantica*, Roma, Laterza, 1999.

De Mauro, Tullio (1982), *Minisemantica dei linguaggio non verbali e delle lingue*, Roma, Laterza, 2019.

Dupre, Gabe (2021), «(What) Can deep learning contribute to theoretical linguistics?» in *Minds and Machines*, 31(4), pp. 617-635.

Eco, Umberto (1974), *Trattato di semiotica generale*, Milano, La nave di Teseo, 2016.

Eco, Umberto (1997), *Kant e l'ornitorinco,* Milano, La nave di Teseo, 2016.

Ethayarajh, Kawin (2019), «How contextual are contextualized word representations? Comparing the geometry of bert, elmo, and gpt-2 embeddings» in *Proceedings of the 2019 EMNLP-IJCNLP*, pp. 55-65.

Fodor, Jerry (1987), *Psychosemantics. The problem of meaning in the philosophy of mind*, Cambridge, MIT Press, 1993.

Fodor, Jerry, e Pylyshyn, Zenon (1988), «Connectionism and cognitive architecture: a critical analysis», in *Cognition*, 28(1-2), pp. 3-71.

Friedman, Robert (2023), «Large Language Models and Logical Reasoning», in *Encyclopedia* 3, pp. 687-697.

Gastaldi, Juan Luis (2021), «Why can computers understand natural language? The structuralist image of language behind word embeddings», in *Philosophy & Technology*, 34(1), pp. 149-214.

Harnad, Stevan (1990), «The symbol grounding problem», in *Physica D: Nonlinear Phenomena*, 42(1-3), pp. 335-346.

Jakobson, Roman (2002). *Saggi di linguistica generale*, Milano, Feltrinelli.

Kovaleva, Olga et al. (2019), «Revealing the dark secrets of bert», in *arXiv*: 1908.08593.

Klumizev, Artur e Nivre, Joakim (2021), «Schrodinger's tree — On syntax and neural language models», in *arXiv*: 2110.08887.

Lenci, Alessandro (2023), «Understanding natural language understanding systems» in *Sistemi Intelligenti* 2, pp. 277-302.

Lo Piparo, Franco (2003), *Aristotele e il linguaggio. Cosa fa di una lingua una lingua*, Roma-Bari, Laterza.

Merrill, Williamo et al. (2021), «Provable limitations of acquiring meaning from ungrounded form: What will future language models understand?», in *Transactions of the ACL*, 9, pp.1047-1060.

Mikolov, Tomas et al. (2013). «Efficient Estimation of word Representations in Vector Space», in *arXiv*: 1301.3781.

Mu, Jesse e Andreas, Jacob (2020), «Compositional explanations of neurons», in *34th Conference on NeurIPS*.

Murty, Shikhar et al. (2022), «Characterizing intrinsic compositionality in transformers with tree projections», in *arXiv*: 2211.01288.

Pater, Joe (2019), « Generative linguistics and neural networks at 60: Foundation, friction, and fusion», in *Language* 95(1).

Piantadosi, Steven, e Hill, Felix (2022). «Meaning without reference in large language models», in *arXiv*: 2208.02957.

Pilehvar, Mohammad, e CamaCho-Collados, Jose (2021), *Embeddings in natural language processing: theory and advances in vector repreentations of meaning*. Cham, Springer.

Potts, Chris (2020). «Is it possible for language models to achieve language understanding?» https://chrisgpotts.medium.com/is-it-possible-for-language-models-to-achieve-language-understanding-81df45082ee2

Rogers, Anna et al. (2020). «A primer in bertology: what we know about how bert works», in *Transaction of the ACL* 8, pp. 842-866.

Sahlgren, Magnus e Carlsson, Fredrik (2021), «The singleton fallacy: why current critiques of language models miss the point», in *arXiv*: 2102.04310.

Saussure, Ferdinand, (1916), *Cours de linguistique Générale*, Losanna, Payot (*Corso di linguistica generale*, trad. di, T. De Mauro, Laterza, Roma-Bari, 2015).

Smilkov, Daniel et al. (2016). «Embedding projector: Interactive visualization and interpretation of embeddings», in *arXiv*: 1611.05469.

Søgaard, Anders (2022), « Understanding models understanding language», in *Synthese* 200(6).

Vaswani, Ashish et al. (2017). «Attention is all you need», in *arXiv*: 1706.03762.

Wittgenstein, Ludwig (1953), *Philosophische Untersuchungen* (*Ricerche filosofiche*, trad. di, M. Trinchero, Einaudi, Torino, 2014).

Wittgenstein, Ludwig (1921), «Tractatus Logico-Philosophicus», in *Ostwald's Annalen der Naturphilosophie* (Tractatus Logico-Philosophicus e quaderni 1914-1916, trad di. G. Conte, Einaudi, Torino, 2009).

Wolf, Thomas (2018). «Learning meaning in natural language processing – The semantics mega-thread», https://medium.com/huggingface/learning-meaning-in-natural-language-processing-the-semantics-mega-thread-9c0332dfe28e