Categorie grammaticali e classi di parole. Statuto e riflessi metalinguistici

a cura di Francesco Dedè

Roma 2016

Volume pubblicato con il contributo del Dipartimento di Studi letterari, filologici e linguistici dell'Università degli Studi di Milano.

ISBN 9788898640188

INTERNET http://www.ilcalamo.it
E-mail: info@ilcalamo.it

I volumi pubblicati nella Collana sono sottoposti a un processo di peer review che ne attesta la validità scientifica

Per ordinazioni / Orders to be sent to:

Editrice "Il Calamo" s.n.c.
Tel. 06.98968058 - Fax 06.98968062

INDICE

Marianna D'Errico, Nicola Grandi,
Serena Paternesi Meloni, Fabio Tamburini

INDUZIONE DI CATEGORIE GRAMMATICALI E LESSICALI

1. Introduzione

La classificazione delle parole in parti del discorso è uno dei temi più antichi e insieme più controversi della teoria linguistica. Esprime in modo esemplare il paradosso Baker (2003: 1) in apertura del suo volume *Lexical Categories*: ‹‹It's ironic that the first thing one learns can be the last thing one understands››.
Fin dai primi studi sul linguaggio umano, infatti, a partire dalla distinzione platonica in *onoma* e *rhema*, è stato riconosciuto che gli elementi che compongono il discorso possono essere raggruppati in diverse classi, sulla base di proprietà semantiche e/o formali condivise.
Scrive a tal proposito Prandi (2006: 273):

> Le parti del discorso sono classi di parole che occupano posizioni simili nella struttura delle frasi, presentano proprietà grammaticali interne simili se sono variabili, e hanno funzioni e contenuti concettuali caratteristici.

La tradizione grammaticale occidentale ha unanimemente considerato le parti del discorso imprescindibili strumenti di descrizione e analisi delle lingue naturali, tanto che non c'è grammatica tradizionale che non includa una loro rassegna: esse ‹‹possono essere definite come la parte essenziale di ogni sistema grammaticale e in esse si pongono, in ultima istanza, tutti i problemi di ordine grammaticale›› (Hjelmslev 1998:233).

Tuttavia, nonostante l'evidenza intuitiva alla base almeno della identificazione delle parti del discorso maggiori (con la parziale eccezione dell'aggettivo, di cui, secondo alcuni studi tra cui Chafe 2012, alcuni gruppi di lingue sarebbero prive), gli approcci alla suddivisione delle parole in parti del discorso (in questa sede, per motivi di spazio, non si potrà dare conto della variazione terminologica presente nell'ambito: classi di parole, parti del discorso, categorie sintattiche e/o grammaticali, ecc.) sono molto difformi e si registrano differenze sostanziali tra quadri teorici diversi. Spesso queste differenze sono determinate dalla tendenza a stabilire, quasi *a priori*, una serie di parametri per definire l'appartenenza di una parola ad una classe e a far poi seguire la verifica sui dati. In un qua-

dro di questo tipo, emerge in tutta la sua evidenza la situazione complessa ad esempio di classi come quella dei participi o degli infiniti, la cui morfologia è spesso aggettivale e nominale rispettivamente, ma la cui configurazione semantica sembra invece verbale. Inoltre, anche all'interno di singole classi le parole paiono assumere comportamenti non sempre del tutto omogenei: vi sono nomi che hanno il plurale e nomi che non hanno plurale, verbi che selezionano un ausiliare e verbi che ne selezionano un altro, ecc.

Lo scopo di questo contributo è quello di tentare una definizione il più possibile a-teorica delle parti del discorso, derivando il set di categorie dall'evidenza distribuzionale dei dati, ovvero dalle informazioni sui contesti linguistici nei quali una parola occorre. Lo slancio per un'indagine di questo tipo coincide infatti con la disponibilità di corpora di grandi dimensioni in grado, quindi, di fornire un'adeguata base empirica all'analisi. La definizione di categoria così ottenuta dipende unicamente dalla rappresentazione del contesto e dall'analisi della similarità distribuzionale, senza che nessuna "teoria" preceda la classificazione. L'ipotesi fondamentale di lavoro è che due parole formalmente e semanticamente simili e dall'analogo comportamento sintattico appariranno in contesti affini; la formalizzazione di tale assunto è racchiusa nell'ipotesi di distribuzionalità di Harris (1968) e più in generale all'interno della metodologia dello strutturalismo americano, per cui soltanto un metodo empirico legato alle regolarità distribuzionali e statistiche avrebbe potuto assicurare alla riflessione linguistica una fondatezza scientifica. In senso stretto, dunque, un punto di vista distribuzionale è teso alla descrizione delle relazioni strutturali tra gli elementi linguistici e i loro contesti: da ciò deriva la rilevanza determinante della rappresentazione del contesto e del modo in cui viene raccolta l'informazione contestuale, e soprattutto la possibilità di classificare nello stesso modo elementi che presentano gli stessi contorni linguistici attraverso la definizione di opportune misure di similarità tra i contesti stessi.

2. La derivazione empirica delle parti del discorso

Gli studi nella direzione della derivazione empirica di una classificazione per parti del discorso (ad esempio, Kiss 1973; Brill, Marcus 1992; Finch, Chater 1992; Schütze 1993; 1995; Pereira *et al*. 1993; Redington *et al*. 1998; Clark 2000; Tamburini *et al*. 2002; 2008) condividono gli stessi presupposti: (a) la nozione di contesto come *distribuzione delle parole* che occorrono nell'immediato intorno della parola target, tipicamente una o due parole a destra e a sinistra; (b) la definizione di opportune metriche, su base statistico-probabilistica, che consentano la comparazione delle

distribuzioni dei contesti delle parole e (c) la creazione di gruppi di parole sulla base della condivisione delle informazioni contenute nelle distribuzioni contestuali attraverso opportune metodologie di *clustering* (Aggarwal e Reddy 2014).

Questo approccio, assolutamente promettente, si è tuttavia rivelato non privo di difetti: la frequenza di co-occorrenza multipla fra i termini all'interno di un testo tende ad assumere rapidamente valori molto bassi, spesso nulli, man mano che ci si sposta dalle parole lessicalmente "vuote" alle parole lessicalmente "piene" (Zipf, 1949). Per ridurre l'interferenza di questo aspetto, noto come problema della *sparsità dei dati,* è necessario considerare orizzonti contestuali molto ridotti vincolando l'analisi necessariamente su un piano prevalentemente morfosintattico ed escludendo quasi completamente la possibilità di induzione di categorie a livello semantico.

A questi problemi ha cercato di dare risposta una modalità parzialmente diversa di rappresentazione della parola e del contesto nata all'interno del quadro connessionista e definita *word embedding*. Il successo degli esperimenti effettuati da Elman (1990) nel predire le successioni di parole utilizzando reti neurali ricorsive e i pionieristici esperimenti di McClelland e Rogers (2003a; 2003b) sull'emersione delle categorie semantico-lessicali, utilizzando reti neurali multistrato in grado di cogliere i diversi aspetti della cognizione semantica e di dare una spiegazione a molti fenomeni ad essa legati, che non potevano essere efficacemente spiegati utilizzando le teorie cognitiviste e linguistiche tradizionali, hanno aperto la strada allo sviluppo di nuove rappresentazioni dei termini in grado di coglierne appieno le proprietà distribuzionali e contestuali.

Le parole del lessico di riferimento, anziché essere trattate come semplici stringhe di caratteri, vengono trasformate, tramite metodi matematici complessi e/o reti neurali artificiali di vario tipo, in vettori "densi" di numeri reali che contengono e rappresentano efficacemente tutte le informazioni contestuali di un termine all'interno di uno specifico corpus. Non abbiamo la possibilità in questa sede di approfondire gli aspetti tecnici della creazione di un *word embedding*, ma possiamo pensare a una sorta di procedura capace di fondere tutte le informazioni contestuali di un termine, prevalentemente di co-occorrenza posizionale, e di estrarne la porzione più rilevante trasformandole in un insieme ordinato di numeri reali. La potenza di questi metodi di rappresentazione risiede nel fatto che a parole simili dal punto di vista distribuzionale, sono associati con alta probabilità vettori molto simili dal punto di vista matematico: in questo contesto è semplice definire una distanza su base geometrica riferita ai vettori alla quale corrisponde, implicitamente, una nozione di similarità sintattico/semantica distribuzionale dei termini.

A partire da queste nozioni quantitative di similarità sui vettori è possibile costruire visualizzazioni bidimensionali delle proprietà distribuzionali dei termini che sono in grado di mantenere le proprietà di vicinanza dei vettori originali, e quindi dei relativi termini, utilizzando ad esempio il metodo proposto da van der Maaten e Hinton (2008).

Come notano vari studiosi (es. Turian *et al*. 2010; Lai *et al*. 2015) questi vettori sono in grado di catturare sia le informazioni sintattiche che semantiche di una parola a partire da un *corpus* non annotato di grandi dimensioni. Alterando i parametri relativi alla costruzione del modello è possibile dare predominanza all'uno o all'altro aspetto e quindi costruire tipologie differenti di vettori idonee a supportare tentativi di identificazione induttiva di categorie linguistiche a vari livelli.

In linguistica computazionale i modelli di rappresentazione basati su *word embedding* stanno assumendo un ruolo centrale soprattutto in task di categorizzazione e come input per un'analisi di *clustering* non supervisionata, capace di rivelare le similarità e dunque indurre classi di parole. La storia di queste rappresentazioni, basate su reti neurali artificiali, si fonda, come riconosciuto da Mikolov *et al*. (2013), sui lavori di Hinton, McClelland e Rumelhart (1986) e di Elman (1990). Tra i diversi modelli neurali di linguaggio sviluppati invece negli ultimi anni è doveroso citare Bengio *et al*. (2003), Mnih e Hinton (2007), Collobert e Weston (2008), Mikolov (2010); in questo lavoro verranno utilizzati, nello specifico, due algoritmi per l'apprendimento di rappresentazioni distribuite di parole basati su reti neurali multistrato implementati nel toolkit *word2vec*[1] (Mikolov *et al*. 2013).

Lo scopo di questa indagine, in sintesi, è quello di rispondere ad una domanda come la seguente: le parole sono state classificate in parti del discorso sulla base, ad esempio, di parametri di ordine funzionale o sulla base del loro significato inerente; è possibile invece raggruppare assieme le parole in virtù del contesto di occorrenza, senza tener conto, almeno in una prima fase, di altri parametri? In altri termini, una indagine fondata in via esclusiva sul reale contesto di occorrenza delle parole consente di raggruppare le medesime in classi omogenee? In questo caso, dunque, sarebbe l'uso concreto che si fa delle parole a definire il perimetro delle varie parti del discorso, collocando nella medesima classe parole caratterizzate dagli stessi ambiti di impiego.

[1] https://code.google.com/archive/p/word2vec/

3. Il caso dell'infinito

Per mostrare la validità di questo approccio prima di presentare, in modo sommario, il quadro complessivo delle parti del discorso in italiano, ci soffermeremo brevemente sulla situazione dell'infinito. Come è noto, la posizione dell'infinito nel quadro delle parti del discorso è complessa. Di norma l'infinito viene collocato nel paradigma del verbo; esso, dunque, sarebbe una voce verbale a tutti gli effetti. Tuttavia, l'infinito, almeno in italiano e nelle principali lingue indoeuropee, è invariabile: non si coniuga per tempo, modo, persona, ecc. Ad esempio, nel quadro elaborato da Croft (2000), basato su una classificazione delle parole in base alla combinazione di tre classi di significato (azione, oggetto e proprietà) e di tre funzioni pragmatiche (predicazione, referenza, modificazione), nella quale sono tipicamente verbi le parole che indicano azioni quando usate in funzione predicativa, l'infinito viene collocato ai margini della classe dei verbi in quanto, pur indicando una azione, esso ha prevalentemente una funzione simile a quella referenziale.

Nella nostra indagine basata sul corpus CORIS (Rossini Favretti *et al.* 2002), abbiamo condotto un duplice esperimento sull'infinito, valutandone il contesto sia in prospettiva sintattica, sia in prospettiva semantica. Dal punto di vista sintattico gli infiniti analizzati (*vedere*, *andare*, *parlare*, *trovare*, *capire*, *pensare*, *prendere*, *mettere*, *evitare*, *passare*, *arrivare*, *diventare*, *entrare*, *ottenere*, *tornare*) sembrano isolarsi nel piano, più lontani da tutte le altre classi, ma comunque più vicini ai nomi che ai verbi di modo finito (Figura 1).

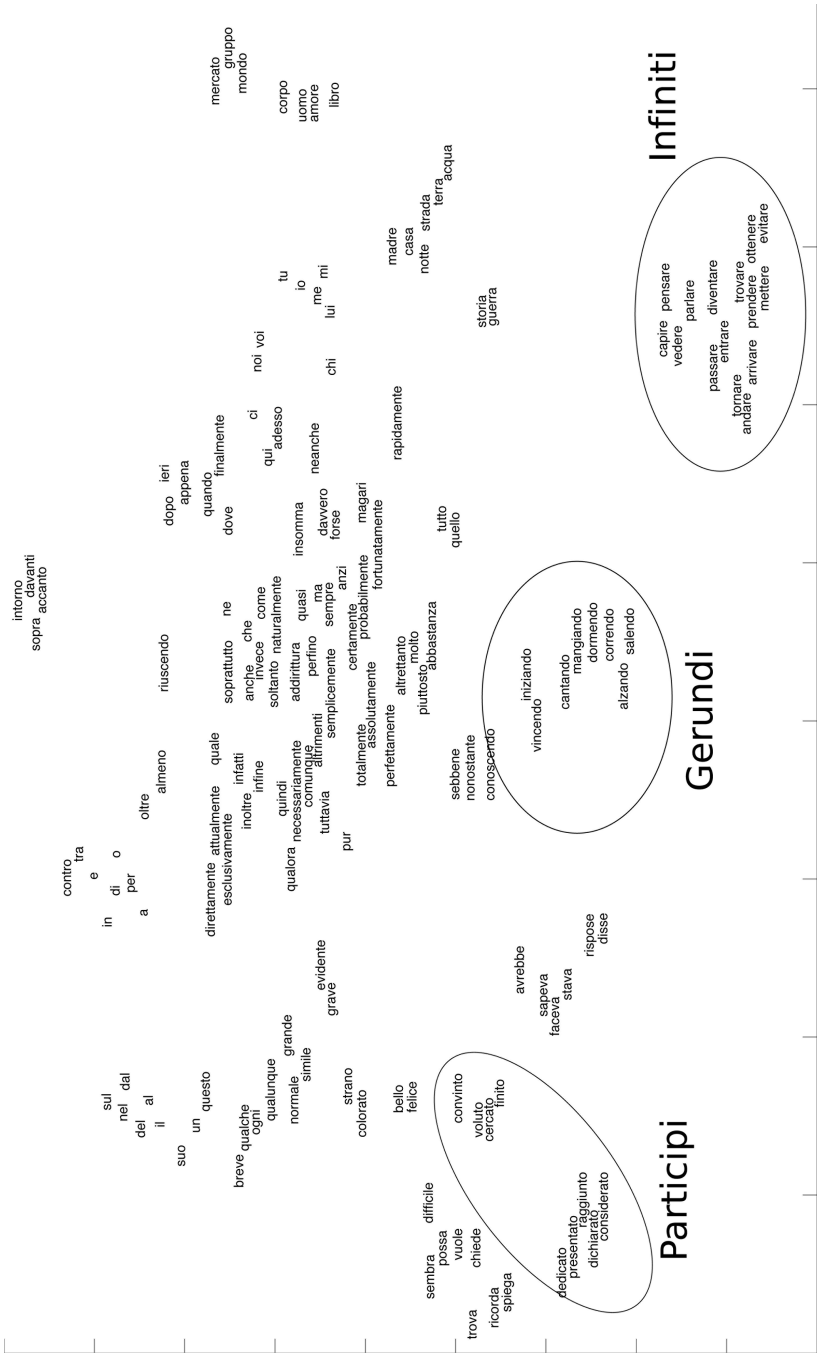

Figura 1: la posizione dell'infinito in base al contesto sintattico

La prossimità degli infiniti alla categoria dei nomi, piuttosto che a quella dei verbi sembrerebbe avvalorare le caratteristiche nominali che l'infinito è in grado di assumere, vale a dire la capacità di costituire il nucleo di un SN e avere gli stessi contorni sintattici di una testa nominale, come determinanti e modificatori (incidentalmente: si noti la posizione dei participi, adiacente a quella degli aggettivi).

Dal punto di vista del contesto semantico (Figura 2), gli infiniti hanno mostrato una (logica e prevedibile) somiglianza con i nomi di azione. Entrambi, infatti, paiono caratterizzati da una atipicità di fondo nel quadro elaborato da Croft citato in precedenza. Se gli infiniti sono verbi con una funzione pragmatica simile a quella dei nomi, i nomi d'azione sono, appunto, nomi, ma con un significato più tipicamente verbale. Nei modelli linguistici tradizionali, che separano nettamente la componente sintattica da quella semantica, queste due classi di parole vengono rigidamente mantenute distinte dal punto di vista delle loro manifestazioni sintattiche, nonostante una certa similiarità dal punto di vista semantico. I modelli tradizionali faticano a risolvere questa ambiguità, che invece può essere facilmente superata con una categorizzazione del lessico che registri le effettive regolarità d'uso, cioè le reali proprietà distribuzionali delle parole. Gli infiniti già citati sopra si collocano, nel piano, accanto ai nomi d'azione (*donazione, restituzione, risanamento, invecchiamento, sparizione, esplosione, montaggio, atterraggio, masterizzazione, partenza*). Questi ultimi, invece, occupano una posizione intermedia tra gli infiniti stessi e i nomi più 'tipici', come quelli che designano animali, vegetali, oggetti e umani:

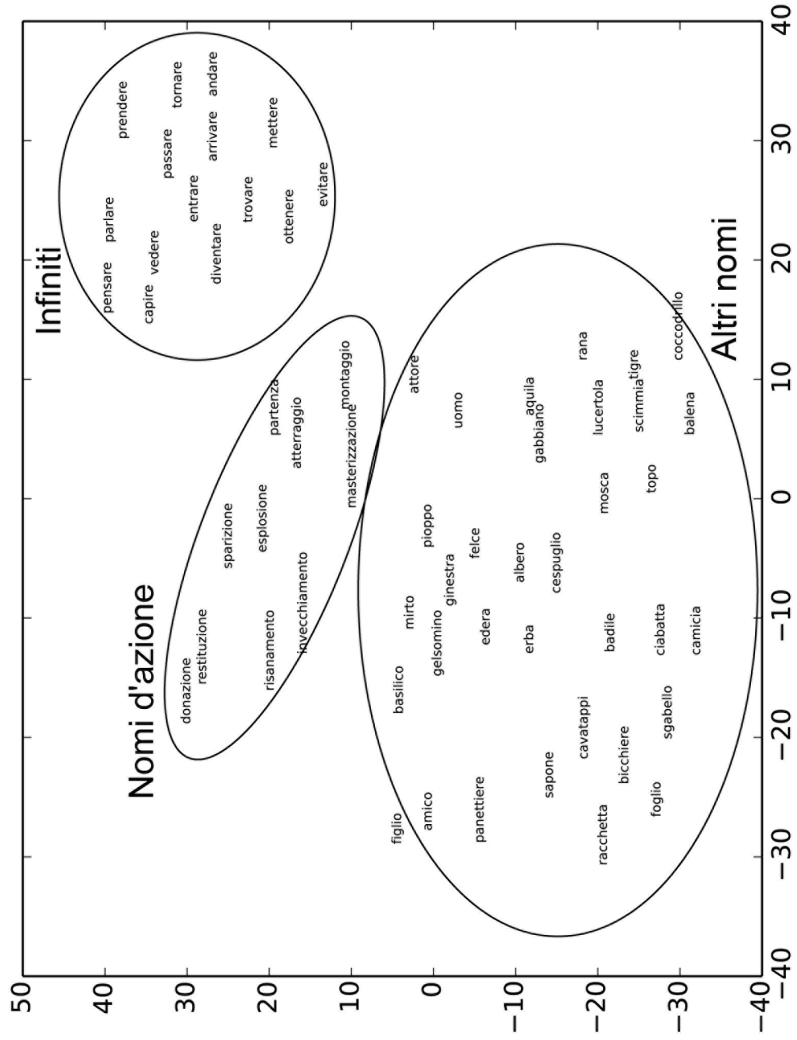

Figura 2: la posizione degli infiniti rispetto al contesto semantico

Come si può notare, se estraessimo in automatico i contesti di occorrenza degli infiniti ed usassimo questo parametro per definirne la collocazione nel quadro delle parti del discorso, otterremmo una loro collocazione a mezza via tra nomi e verbi di modo finito. Il modello appena descritto ha poi il vantaggio di registrare in modo evidente le ampie aree di sovrapposizione tra infiniti e nomi d'azione che, nella maggior parte delle classificazioni tradizionali, vengono sacrificate sull'altare della inviolabilità del confine tra la classe del nome e quella del verbo.

4. Le parti del discorso dell'italiano

Nel seguito della nostra ricerca, il modello appena descritto è stato dunque applicato, su più larga scala, per una derivazione induttiva delle parti del discorso dell'italiano. Per ragioni di spazio, non è possibile descrivere, nel dettaglio, l'esperimento condotto, per il quale si rinvia a Paternesi Meloni (2016). Esso ha generato, sulla base di regolarità distribuzionali, tredici parti del discorso: nomi, verbi (forme finite + gerundi, infiniti e participi passati), aggettivi, determinanti, pronomi, preposizioni, operatori relazionali, avverbi (deittici, specificatori e frasali) e connettivi (Figura 3):

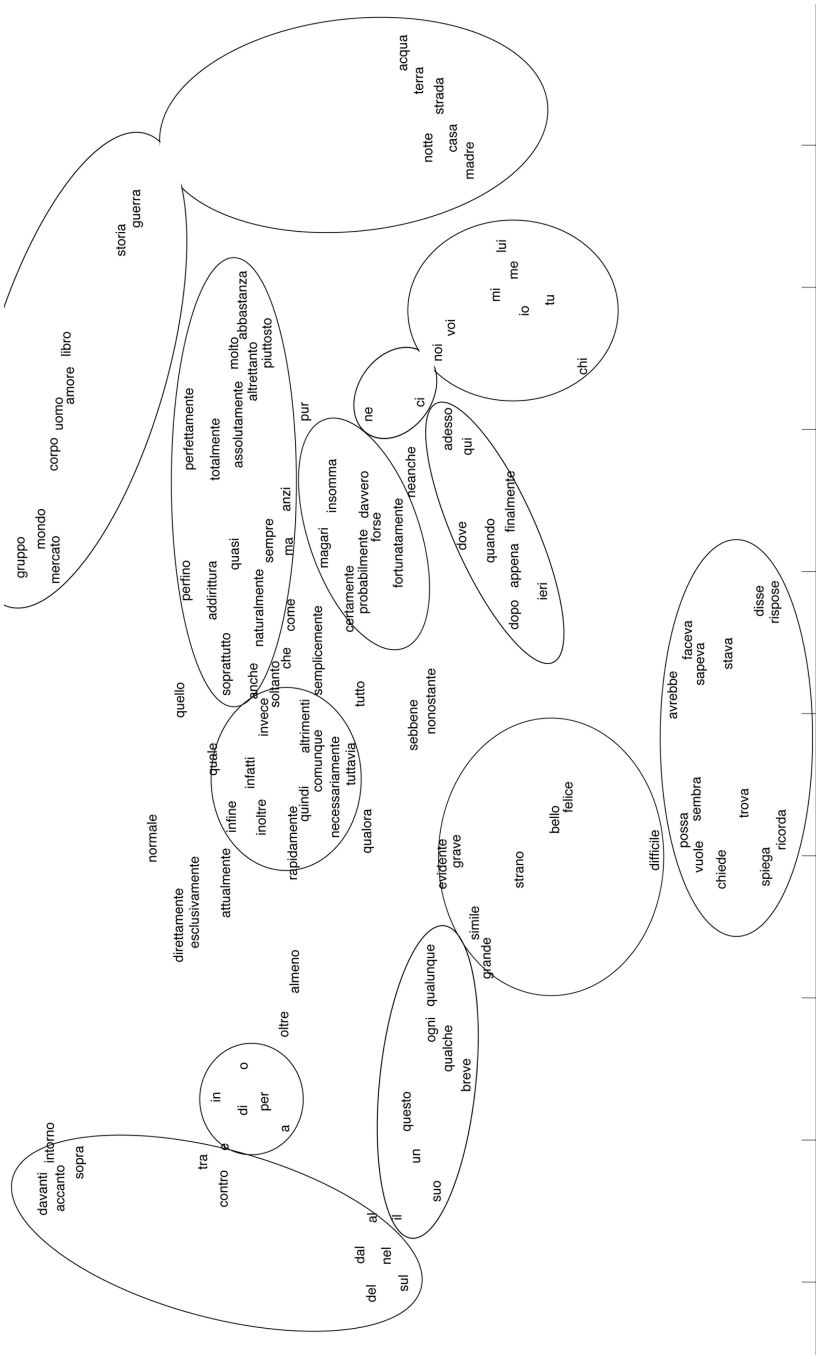

Figura 3: la rappresentazione sul piano delle parti del discorso in italiano in base al contesto sintattico

Benché a livello numerico e terminologico l'inventario sembrerebbe non discostarsi molto da quello delle grammatiche tradizionali, l'osservazione della composizione interna e dei rapporti spaziali delle categorie suggerisce, invece, diverse operazioni di riassegnazione e sottocategorizzazione di rilevanza sintattica, ignorate dagli impianti descrittivi tradizionali prevalentemente morfo-semantici.

Le prime considerazioni coinvolgono le classi dei nomi e dei verbi, disposte in due *cluster* distinti e ben definiti, posti agli antipodi del piano; d'altronde, fin dalle prime riflessioni sulle parti del discorso, è stato riconosciuto il valore universale della distinzione nome-verbo, riconducibile alle due funzioni pragmatiche primarie di riferimento e predicazione.

Dato un supporto anche distribuzionale della categorizzazione del nome e del verbo, è stata indagata del primo la struttura interna, mentre del secondo i rapporti con i cosiddetti modi non finiti, ovvero infinito, participio e gerundio. Per quello che concerne il nome emergono come primarie la sottocategorizzazione in nome massa vs. numerabile e le diverse strutture argomentali dei nomi deverbali (Renzi *et al.* 1991a:326; Graffi 1994:58), che sembrano scalzare i classici tratti semantici della concretezza e dell'astrattezza. Per quanto riguarda i modi verbali non finiti, le analisi hanno segnalato il comportamento pienamente verbale del gerundio, rispetto a quelli più ibridi dell'infinito e del participio: del primo, oggetto di un esperimento *ad hoc,* si è detto sopra; il secondo si dispone in una porzione di piano compresa tra verbi e aggettivi (si veda la Figura 4 in 3D che illustra al meglio il fenomeno rispetto alla Figura 3 in 2D).

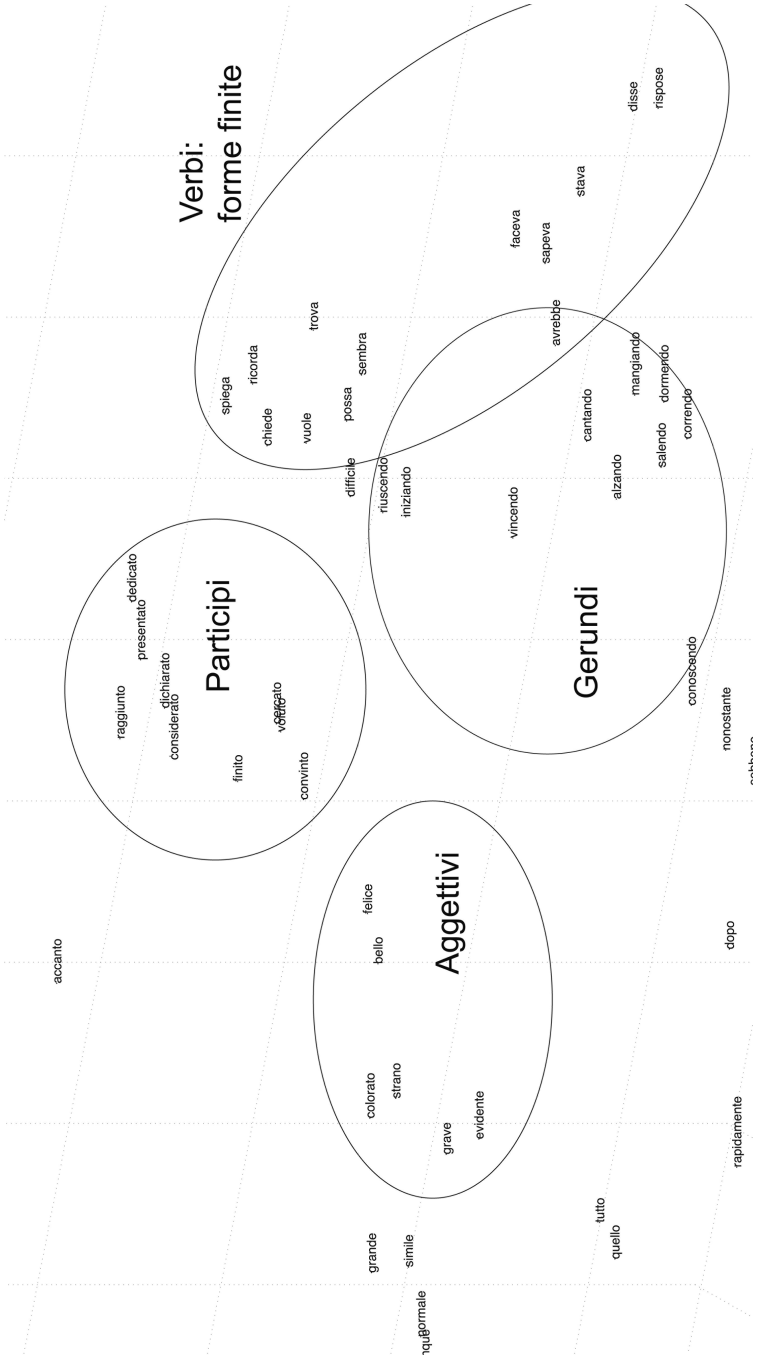

Figura 4: Disposizione nello spazio delle forme verbali non finite.

Continuando con le parti variabili del discorso, la categorizzazione empirica ha permesso di riclassificare gli aggettivi, gli articoli e i pronomi, attraverso riassegnazioni di elementi a categorie diverse e rielaborazioni delle definizioni tradizionali.

L'estensione e la varietà della classe degli aggettivi sono state ridotte, includendo i soli aggettivi definiti dalla tradizione come 'qualificativi', elementi dotati cioè di significato lessicale, specializzati nella modificazione a vario livello del nome all'interno del sintagma nominale e caratterizzati da una posizione non marcata post-nominale, che non concorrono, a differenza dei determinativi, alla specificazione pre-nominale. Queste diverse proprietà distribuzionali si sono infatti manifestate in una collocazione spaziale ben precisa, per cui il *cluster* degli aggettivi rivelava due addensamenti interni, ciascuno corrispondente alle due tradizionali sottocategorie; il polo dei determinativi era localizzato in prossimità dell'articolo, a indicare un comportamento distribuzionale simile, che ha suggerito di non conservare una classe indipendente per l'elemento grammaticale, ma di farlo confluire nella classe dei determinanti.

L'individuazione della classe dei determinanti non è stata certo un contributo innovativo di questo lavoro: già Renzi (1988) aveva segnalato l'articolo tra i determinanti e più in generale tra i modificatori del SN, motivando a livello teorico la condivisione di specifiche proprietà sintattiche; l'aspetto forse più originale è invece determinato dall'osservazione della sua composizione e strutturazione. La categoria dei determinanti sembra infatti un perfetto esempio di prototipo funzionalista: usando la terminologia di Bertinetto (2003), mentre l'articolo definito e indefinito insieme ai dimostrativi e ai quantificatori indefiniti, ne costituisce il nucleo più rappresentativo, possessivi, interrogativi e relativi popolano, a distanze differenti dal centro, la periferia, dotata di contorni sfumati così da consentire a questi elementi linguistici di transitare verso proprietà sintattiche tipiche delle categorie attigue. Grazie a questa specifica configurazione che si può rilevare alla periferia di questa classe, la classificazione induttiva ha provato a motivare il comportamento ibrido (a) di alcuni dimostrativi in relazione agli aggettivi cosiddetti anaforici; (b) dei possessivi, non considerati da molte grammatiche come determinanti ma aggettivi veri e propri, con la particolarità del tratto della persona; e soprattutto (c) dei relativi, in rapporto alla distribuzione dei quali è stata confermata la natura di complementatore di *che*. A proposito di quest'ultimo punto, mentre parole come *cui* e *quale* sono identificabili a livello distribuzionale come determinanti, seppur in maniera non prototipica, un elemento come *che*, tradizionalmente classificato nel gruppo dei relativi, non mostra prossimità con esse, ma si dispone nel gruppo delle parole grammaticali, non evi-

denziando alcun comportamento pronominale. Alcuni studiosi come Salvi e Vanelli (2004:289-290), Vanelli (2010:49-53) e Salvi (2013:73-75) hanno quindi ipotizzato che non esista un *che* relativo: ‹‹il *che* che si trova all'inizio delle frasi relative, nel sistema grammaticale dell'italiano, non è un relativo, ma è la stessa congiunzione subordinante *che* che troviamo all'inizio delle subordinate complemento di un verbo›› (Salvi 2013:73); nella terminologia moderna questo elemento grammaticale, non più considerato omofono del pronome relativo, prende il nome di complementatore, ‹‹cioè di elemento che serve a trasformare una struttura frasale in un complemento›› (Salvi 2003:127).

La definizione empirica della categoria dei pronomi provoca una decisiva frattura rispetto all'omonima categoria tradizionale, sia perché esclude le forme pronominali dei determinanti, sia perché comporta una reinterpretazione etimologica dell'etichetta in pro-SN, in linea con le proposte teoriche più recenti (Vanelli 2010:41). L'evidente prossimità spaziale dei pronomi personali e delle forme pro-SN degli aggettivi determinativi che sostituiscono la testa nominale – a differenza delle forme cosiddette pronominali dei determinanti, coinvolte in realtà in un processo di ellissi del nome testa – alla classe dei nomi, consente di pensare ad una classe di pronomi più ristretta di quella tradizionale, definibile secondo la nuova interpretazione distribuzionale; quest'ultima garantisce inoltre una copertura descrittiva per la distribuzione periferica delle forme clitiche che sono considerate più genericamente dei prosintagmi con proprietà non solo pronominali, ma anche circostanziali, rivelate dalla vicinanza con gli avverbi deittici di luogo.

È nella classificazione delle parole invariabili, però, che si concentrano le maggiori debolezze e difficoltà descrittive della teoria tradizionale: la scelta di tratti categoriali esclusivi e la conseguente individuazione di confini precisi per le classi delle preposizioni, degli avverbi e delle congiunzioni sono infatti i punti principali di differenziazione tra i vari sistemi di parti del discorso proposti dalle grammatiche; gli stessi approcci moderni revisionisti, d'altro canto, propongono sottoclassificazioni e sistemazioni non sempre compatibili tra loro.

L'osservazione dei rapporti distribuzionali degli elementi in questione fornisce alcuni spunti per avanzare una loro possibile classificazione in preposizioni, operatori relazionali, avverbi deittici e frasali, specificatori e connettivi.

Per quanto riguarda il *cluster* delle preposizioni, sono emersi due aspetti importanti; il primo legato alla distribuzione delle preposizioni cosiddette improprie, spesso riclassificate tra gli avverbi di luogo e di tempo in virtù di un loro possibile uso intransitivo: la distanza presente tra

le forme avverbiali e le preposizioni lessicali ha tuttavia evidenziato un uso prevalente di queste ultime come teste preposizionali che ricorrono generalmente con un complemento, all'interno dunque di una distribuzione simile a quella delle preposizioni proprie; il secondo punto riguarda la medesima localizzazione delle preposizioni semplici e degli operatori logici di congiunzione e disgiunzione, motivata attraverso varie osservazioni che hanno condotto anche ad una riflessione sul comportamento dei diversi operatori di coordinazione e sul loro valore connettivale.

La configurazione nel piano/spazio pertanto suggerisce una classificazione che raccoglie sotto la tradizionale etichetta di 'preposizione' le preposizioni polisillabiche, mentre sotto quella di operatori di relazione le preposizioni grammaticali e gli operatori logici: la scelta di questa denominazione è sembrata appropriata per dare coerenza esplicativa al comune comportamento relazionale di questi elementi linguistici, includendo anche la funzione di complementatore che le preposizioni possono assumere nell'introduzione di subordinate infinitive.

Le restanti parole invariabili, variamente classificate tra avverbi, congiunzioni e connettivi, presentano una distribuzione piuttosto compatta al centro del piano in Figura 3, quasi a supportare, ad una semplice osservazione preliminare, la criticità di una loro categorizzazione, così come è stato segnalato dalla tradizione grammaticale; attraverso l'interpretazione delle posizioni dei punti associati alle parole *target*, l'indagine ha tuttavia portato all'individuazione di una serie di *cluster*, a partire dai quali è stata avanzata l'ipotesi di una sottoclassificazione distribuzionalmente motivata degli avverbi in deittici (*dopo*, *ieri*, *qui*), specificatori (*soprattutto*, *totalmente*, *molto*) e frasali (*fortunatamente*, *probabilmente*, *forse*), in aggiunta alla classe indotta dei connettivi, riconosciuta dalla maggior parte delle grammatiche e definita in termini di pianificazione testuale.

Gli avverbi deittici sono stati classificati in lavori recenti (Salvi, Vanelli 2004:180) come forme pronominali, sia per il comune rinvio extratestuale che per la proprietà di questi avverbi di comportarsi come un SN: la prossimità spaziale delle due categorie non solo ha permesso di notare questa affinità a livello distribuzionale, ma ha anche indirettamente confermato la categorizzazione di altre forme invariabili, come i clitici, nel *cluster* dei pronomi; la proposta dell'etichetta di specificatore per la classe dei focalizzatori e dei quantificatori ha voluto mettere in evidenza la loro funzione e posizione rispetto alla testa dei costituenti di vario tipo che modificano; il raggruppamento empiricamente emerso in prossimità dei connettivi viene identificato come costituito da elementi avverbiali frasali la portata della cui modificazione si estende invece all'intero atto linguistico.

Alla luce della distribuzione dei dati nel piano/spazio, la proposta avanzata per una possibile classificazione empirica delle categorie grammaticali dell'italiano può essere formalizzata come segue:

<table>
<tr><th colspan="3">CATEGORIE VARIABILI</th></tr>
<tr><td>Verbo: Modi finiti + gerundio</td><td>Verbo: Infiniti</td><td>Verbo: Participi passati</td></tr>
<tr><td>Nome
sottoclassificabile in:
• Massa (± astratti)
• Numerabili
• Argomentali</td><td colspan="2">Pronome
• Pronomi personali
• Forme determinative pro-SN
• (Clitici pro-sintagma)</td></tr>
<tr><td>Aggettivo
• Aggettivi qualificativi (funzione attributiva, posizione non marcata post-nominale)</td><td colspan="2">Determinante
(posizione prenominale)
• Centro
▫ Articoli definiti e indefiniti
▫ Dimostrativi locali (e forme con ellissi testa SN)
▫ Quantificatori indefiniti (e forme con ellissi testa SN)
• Periferia
▫ Possessivi
▫ Relativi
▫ Interrogativi</td></tr>
<tr><th colspan="3">CATEGORIE INVARIABILI</th></tr>
<tr><td colspan="3">Preposizione
• Preposizioni polisillabiche (teste di SP)
• Preposizioni articolate</td></tr>
<tr><td colspan="3">Operatore relazionale
• Preposizioni semplici
• Operatori logici di congiunzione e disgiunzione</td></tr>
<tr><td colspan="3">Avverbi
• Deittici
• Specificatori (focalizzatori e quantificatori)
• Frasali</td></tr>
<tr><td colspan="3">Connettivi</td></tr>
</table>

Tabella 1: le parti del discorso in italiano su base distribuzionale

5. Conclusioni e ulteriori prospettive di ricerca

Quanto riportato nei paragrafi precedenti dimostra quanto sia promettente il metodo di induzione delle parti del discorso a partire dalla distribuzione reale delle parole. Somiglianze e regolarità nel contesto d'uso delle parole, consentono cioè di raggruppare le parole stesse in classi che rivelano poi anche una omogeneità dal punto di vista funzionale. La prospettiva è dunque ribaltata rispetto ad altri approcci alla questione: non si parte più dalla funzione per cercare poi le sue realizzazioni formali; sono invece le analogie sul piano strutturale e distribuzionale a indicare una plausibile omogeneità sul piano funzionale. Dal momento che una analoga funzione tende ad essere realizzata di norma da strategie simili, la possibilità di estrarre in automatico da grandi *corpora* costrutti con identica struttura permette di ottenere in modo rapido ed efficace anche informazioni sul piano funzionale. Ma le potenzialità di questo approccio non si esauriscono in questa considerazione. I *cluster* che coincidono con le parti del discorso elencate sopra non hanno infatti una struttura interna omogenea. In altri termini, se passiamo all'analisi del contesto semantico, adattando i parametri di calcolo dei *word embedding* a questo scopo, la posizione di ogni singolo elemento nelle riproduzioni sul piano è indicativa della sua rappresentatività rispetto alla categoria: la vicinanza al centro del *cluster* è dunque direttamente proporzionale alla prototipicità di una parola. Come dimostra D'Errico (2016), le parole più prototipiche di una categoria si presentano collocate nel grafico 2D in posizioni centrali rispetto al resto del gruppo, mentre le meno prototipiche stanno più al margine (Figura 5).

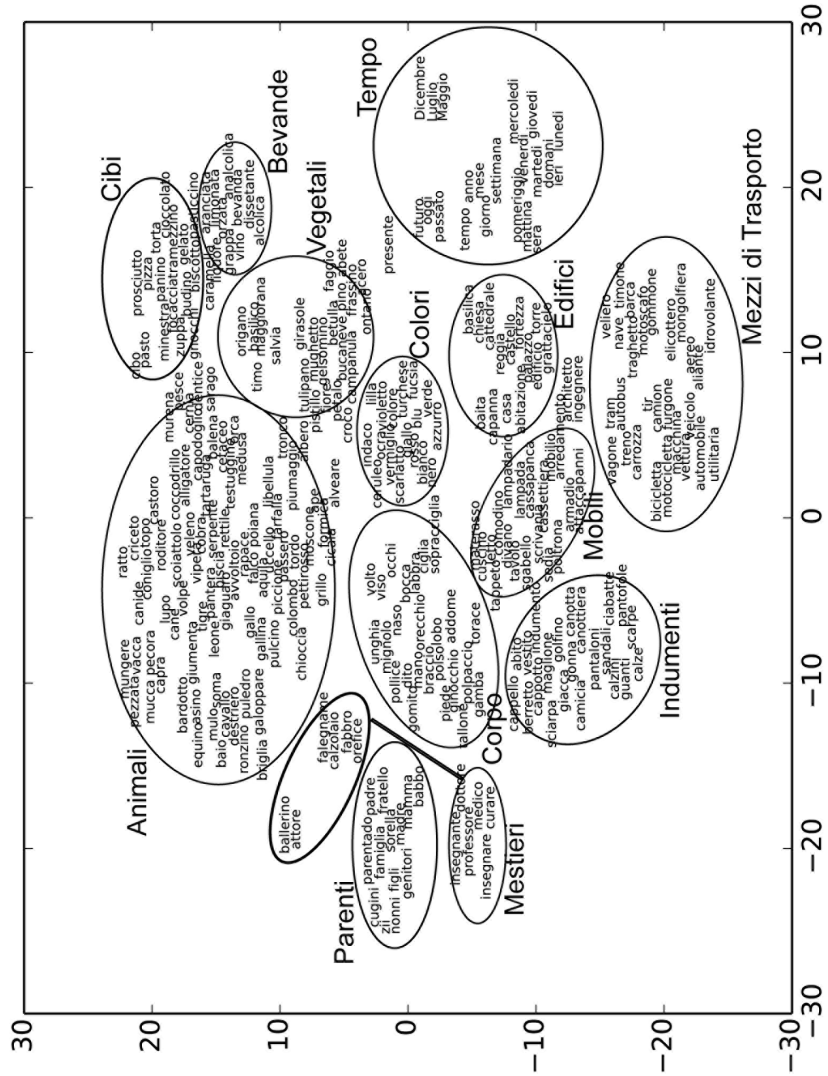

Figura 5: La struttura interna della parti del discorso: il caso dei nomi

Si veda ad esempio come i colori primari siano al centro del gruppo dei colori, come *macchina* sia in posizione centrale rispetto ai mezzi di trasporto di terra, come *piccione* sia centrale negli uccelli, *cavallo* negli equini, *topo* tra i roditori, *vipera* tra i serpenti, *barca* e *nave* nei mezzi di acqua, *aereo* ed *elicottero* tra i velivoli, come *mano* e *dita* siano centrali nel gruppo delle parti del corpo. Parole poco prototipiche come *medusa* sono invece al margine del gruppo degli animali marini (la medusa è l'unica del gruppo a non avere branchie, coda e pinne); *girasole* occupa una posizione marginale tra i fiori e propende verso gli alberi, probabilmente in virtù delle sue grandi dimensioni e del fatto che è anche coltivato a scopi alimentari; le calzature sono le più esterne tra gli indumenti, l'idrovolante è marginale rispetto ai velivoli e prossimo ai mezzi di acqua. Insomma, il sistema utilizzato per indurre automaticamente una classificazione delle parole non si limita a fornirci informazioni 'grossolane' sulla loro suddivisione in parti del discorso, ma ci fornisce anche indicazioni di granularità più fine sulla suddivisione interna delle classi e, indirettamente, sul retroterra culturale della comunità dei parlanti. Ad esempio, nell'esperimento che abbiamo condotto (per i dettagli si rinvia di nuovo a D'Errico 2016), la suddivisione interna degli animali non avviene secondo la tipica tassonomia linneiana che ci saremmo aspettati, ma secondo criteri riconducibili alle loro caratteristiche fisiche più evidenti, all'ambiente, allo stile di vita e al rapporto con l'uomo. I cetacei, pure essendo mammiferi, paiono inseriti nel gruppo che potremmo denominare "degli animali marini", i mammiferi terrestri risultano suddivisi in domestici e selvatici, ecc.

Insomma, un approccio basato sull'uso reale e concreto delle parole sembra in grado di fornire informazioni che vanno ben al di là del mero comportamento formale o funzionale delle stesse.

Riferimenti bibliografici

Aggarwal - Reddy 2014 = C. Aggarwal Charu, K. Reddy Chandan, *Data Clustering: Algorithms and Applications*, CRC Press, 2014.

Baker 2003 = M. Baker, *Lexical categories. Verbs, Nouns, and Adjectives*, Cambridge, Cambridge University Press, 2003.

Bengio *et al*. 2003 = Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, *A Neural Probabilistic Language Model*, ‹‹Journal of Machine Learning Research›› 3 (2003), pp. 1137-1155.

Bertinetto 2003 = P. M. Bertinetto, *'Centro' e 'periferia': una mappa per orientarsi*, in *Modelli recenti in linguistica. Atti del Convegno della Società Italiana di Glottologia*, a cura di D. Maggi, D. Poli, Roma, Il Calamo, 2003, pp. 157-211.

Brill - Marcus 1992 = E. Brill, M. Marcus, *Tagging an unfamiliar text with minimal human supervision*, in *Proceedings of the Fall Symposium on Probabilistic Approaches to Natural Language (Working Notes)*, Cambridge, MA, 1992, pp. 10-16.

Chafe 2012 = W. Chafe, *Are adjectives universal? The case of Northern Iroquoian,* ‹‹Linguistic Typology›› 16/1 (2012), pp. 1-39.

Clark 2000 = A. Clark, *Inducing syntactic categories by context distribution clustering*, in *Proceedings of the workshop on Learning Language in Logic and the conference on Computational Natural Language Learning*, 2000, pp. 91-94.

Collobert - Weston 2008 = R. Colloert, J. Weston, *A unified architecture for natural language processing: deep neural networks with multitask learning*, in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 160-167.

Croft 2000 = W. Croft, *Parts of speech as language universals and as language-particular categories* in *Approaches to the Typology of Word Classes*, a cura di P. Vogel e B. Comrie, Berlin – New York, Mouton de Gruyter, 2000, pp. 65-102.

D'Errico 2016 = M. D'Errico, *Definizione induttiva di categorie semantiche a livello lessicale*. Tesi Magistrale, 2016, Università di Bologna.

Elman 1990 = J. L. Elman, *Finding structure in time*, ‹‹Cognitive Science›› 14 (1990), pp. 179-211.

Finch - Chater 1992 = S. Finch, N. Chater, *Bootstrapping syntactic categories using statistical methods*, in *Background and Experiments* in *Machine Learning of Natural Language*, a cura di W. Daelemans, D. Powers, Tilburg, NL, 1992, pp. 229-236.

Graffi 1994 = G. Graffi*, Le strutture del linguaggio. Sintassi*, Bologna, Il Mulino, 1994.

Harris 1968 = Z. Harris, *Mathematical Structures of Language*, New York, Wiley, 1968.

Hjelmslev 1998 = L. Hjelmslev, *Principes de grammaire generale* 1928, trad. it. *Principi di Grammatica Generale* 1998, Bari, Levante, 1998.

Kiss 1973 = G.R. KISS, *Grammatical word classes: A learning process and its simulation*, ‹‹Psychology of Learning and Motivation›› 7 (1973), pp. 1-41.

Lai *et al.* 2015 = S. LAI, L. KANG; L. XU, J. ZHAO, *How to generate a good word embedding*, 2015, arXiv:1507.05523.

Maaten - Hinton 2008 = L. VAN DER MAATEN, G. HINTON, *Visualizing High-Dimensional Data Using t-SNE*, ‹‹Journal of Machine Learning Research›› 9 (2008), pp. 2579-2605.

McClelland - Rogers 2003a = J.L. MCCLELLAND, T.T. ROGERS, *The parallel distributed processing approach to semantic cognition*, ‹‹Nature›› 4 (2003), pp. 310-322.

McClelland - Rogers 2003b = J.L. MCCLELLAND, T.T. ROGERS, *Semantic Cognition*, MIT Press, 2003.

Mikolov *et al.* 2010 = T. MIKOLOV, M. KARAFIAT, L. BURGET, J. CERNOCKY, S. KHUDANPUR, *Recurrent neural network based language model*, in *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, JP, 2010, pp. 1045-1048.

Mikolov *et al.* 2013 = T. MIKOLOV, K. CHEN, G. CORRADO, J. DEAN, *Efficient estimation of word representations in vector space*, in I*nternational Conference on Learning Representations Workshop Track*, 2013, arXiv:1301.3781.

Mnih - Hinton 2007 = A. MNIH, G. HINTON, *Three new graphical models for statistical language modeling*, in *Proceedings of the 24th International Conference on Machine Learning*, 2007, pp. 641-648.

Paternesi Meloni 2016 = S. PATERNESI MELONI, *Induzione di categorie grammaticali a partire dai dati*. Tesi Magistrale, 2016, Università di Bologna.

Pereira *et al.* 1993 = F. PEREIRA, N. TISHBY, L. LEE, *Distributional clustering of English words,*in *Proceedings of the 31st Association Computational Linguistics*, Columbus, Ohio, 1993, pp. 183-190.

Prandi 2006 = M. PRANDI, *Le regole e le scelte: introduzione alla grammatica italiana*, Torino, UTET, 2006.

Redington *et al.* 1998 = M. REDINGTON, N. CHATER, S. FINCH, *Distributional information: a powerful cue for acquiring syntactic categories*, ‹‹Cognitive Science›› 22/4 (1998), pp. 425-469.

Renzi *et al.* 1991a = L. RENZI, G. SALVI, A. CARDINALETTI (a cura di), *Grande grammatica italiana di consultazione*, Bologna, Il Mulino, 3 Voll. 1991a [1988], 1991b, 1995.

Rossini Favretti *et al.* 2002 = R. ROSSINI FAVRETTI R., F. TAMBURINI F., C. DE SANTIS. *CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model* in *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, a cura di A. Wilson, P. Rayson, P., T. McEnery, Lincom-Europa, Munich, 2002, pp. 27-38.

Salvi 2013 = G. SALVI, *Le parti del discorso*, Roma, Carocci, 2013.

Salvi - Vanelli 2004 = G. SALVI, L. VANELLI, *Nuova grammatica italiana*, Bologna, Il Mulino, 2004.

Schütze 1993 = H. SCHÜTZE, *Part-of-speech induction from scratch*, in *Proceedings of the 31st Association Computational Linguistics*, Columbus, Ohio, 1993, pp. 251–258.

Tamburini *et al.* 2002 = F. TAMBURINI, C. DE SANTIS, E. ZAMUNER, *Identifying phrasal connectives in Italian using quantitative methods*, in *Phrases and Phraseology Data and Description*, a cura di S. Nuccorini, Berlin, Peter Lang, 2002, pp. 45-64.

Tamburini *et al.* 2008 = F. TAMBURINI, C. SEIDENARI, A. BOLOGNESI, R. BERNARDI, *Italian lexical-classes Definition using automatic methods*, in *Frames, Corpora and Knowledge Representation*, Bologna, Bononia University Press, 2008, pp. 95-120.

Turian *et al.* 2010 = J. TURIAN, L. RATINOV, Y. BENGIO, *Word representations: a simple and general method for semi-supervised learning*, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010, pp. 384-394.

Vanelli 2010 = L. VANELLI, *Grammatiche dell'italiano e linguistica moderna*, Padova, Unipress, 2010.

Zipf 1949 = G.K. ZIPF, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Cambridge, MA, Addison-Wesley Press, 1949.

ABSTRACT

The aim of this paper is to give an 'a-theoretical' definition of the main parts of speech, extracting the set of categories from the actual distribution of data, or, in other words, from the contexts of occurrence of words. The definitions of the parts of speech obtained in this way depend uniquely on contextual information and on the analysis of distributional similarities among words, and are not conditioned by any theoretical framework. The research hypothesis is that two words which are formally and semantically similar and which share the same syntactic behavior will occur in similar contexts. As a consequence, if we classify words according to their contexts of occurrence, we should expect that formally and semantically similar words will turn up in the same class.

So, if we investigate a huge, representative corpus of a language, we should be able to automatically extract all the parts of speech by means of a survey of the contexts of occurrences. In this article we will test this approach on Italian, basing our analysis on CORIS, a representative corpus of written Italian.