La percezione della sinonimia : un'analisi statistica mediante modelli per ranghi

Carmela Cappelli, Angela D'Elia

Dipartimento di Scienze Statistiche – Università di Napoli Federico II
Via L. Rodinò, 22 – 80138 Napoli – Italia
{carmela.cappelli ; angela.delia}@unina.it

Abstract

In this paper we deal with the role of synonyms in the Italian language, focussing on the way they are perceived by the people. In particular, we propose to exploit statistical models for ranks data in order to analyse the rankings expressed by different raters towards the set of synonyms of some different words. Indeed, by means of these models, we can highlight the level of perceived synonymy with respect to a "target" word and the presence of uncertainty in the ranking process itself ; moreover, we can study the existence of a link between the raters' covariates (e.g. sex, age, education level, etc.) and the ranks they give among the synonyms of a given list.

Riassunto

L'articolo si colloca nell'ambito degli studi sull'uso dei sinonimi nella lingua italiana, focalizzando l'attenzione sulla percezione – da parte dei parlanti – della sinonimia tra parole. In particolare, si propone il ricorso a modelli statistici per variabili rango al fine di analizzare le graduatorie che diversi soggetti formulano con riguardo alle liste di sinonimi di alcune parole. In effetti, mediante tali modelli per ranghi è possibile quantificare sia il livello di sinonimia percepita rispetto ad una parola "obiettivo", sia il grado di incertezza presente durante l'elaborazione della graduatoria stessa. Inoltre, è possibile analizzare il legame esistente tra le principali caratteristiche dei soggetti (come sesso, età, livello di istruzione, ecc.) e le graduatorie che essi esprimono.[1]

Keywords : synonyms, rankings, word senses' identification, MUB model.

1. Introduzione

Nel corso degli ultimi decenni si è manifestato un crescente interesse verso l'uso di metodi e modelli statistici per lo studio di problemi di natura linguistica, come testimoniato dai numerosi testi che trattano dell'impiego della statistica per l'analisi di dati testuali (si vedano tra gli altri : Woods *et al.*, 1986 ; Lebart *et al.*, 1998 ; Bolasco, 1999).

La linguistica, tradizionalmente, opera una distinzione tra la morfo-sintassi, che attiene alle regole che presiedono alla formazione delle frasi o delle parole, e la semantica, che studia invece il significato delle parole o delle frasi e, dunque, attiene al contenuto di un testo. In quest'ultimo ambito, si è sviluppata in tempi recenti una notevole attenzione per le problematiche relative alla similarità semantica : cioè, la sinonimia (Lin *et al.*, 2003 ; Ploux e Ji, 2003).

[1] Il presente lavoro è frutto di una comune ricerca degli Autori. C. Cappelli ha scritto i paragrafi 1, 3, 4.1 ; A. D'Elia ha scritto i paragrafi 2, 4.2, 5.

In tale contesto, il presente lavoro affronta –mediante un approccio modellistico– un particolare aspetto del tema della similarità semantica : *il grado di sinonimia tra parole, così come risulta essere percepito da parte degli utilizzatori di una lingua*.

In effetti, la conoscenza e l'uso corretto dei sinonimi rappresenta un indicatore importante della padronanza di una lingua. Tuttavia tale uso non è univoco, legandosi al problema, noto in semantica, della polisemia. Il termine polisemia, introdotto dal linguista francese M. Bréal nel 1897, sta ad indicare la complessità semantica di una parola, ovvero la coesistenza di più significati in una stessa parola : si pensi, ad esempio, al sostantivo *albero*, cui corrispondono i diversi significati di pianta (botanica), grafico genealogico (araldica), organo di acciaio per reggere le vele (nautica) o per trasmettere movimento alle ruote (meccanica).

Al fine di meglio inquadrare il concetto di polisemia, è opportuno riflettere sulle distinte nozioni di *significato* e *senso*. Come discusso in De Mauro (2002), il significato rappresenta l'insieme di tutti i valori ed usi che una parola può assumere nella lingua ; il senso, invece, è il modo in cui la parola è sentita (percepita) dalla persona che la utilizza e/o la deve interpretare : esso, quindi, identifica valori ed usi determinati e particolari. Mentre il significato appartiene alla lingua ed alla comunità dei suoi utilizzatori ed ha quindi natura comune, il senso riguarda l'esprimersi individuale : è il modo in cui il significato si estrinseca da parte di chi parla o scrive. Appare, dunque, evidente che i dizionari – ed in particolare i dizionari dei sinonimi – diano conto del significato delle parole, individuando delle accezioni, ossia sensi consolidati in quanto ripetuti e ripresi dagli utilizzatori della lingua. Per contro, per i sensi, in quanto aventi natura occasionale e soggettiva, è lecito ipotizzare che siano strettamente legati alle caratteristiche personali e siano anche frutto di un meccanismo di percezione da parte dell'individuo (in base alla sua cultura, alle sue esperienze, ecc.).

Oggetto del presente lavoro è lo studio dei sinonimi come possibili sensi di una parola "obiettivo" : tale studio, alla luce della problematica esposta, può essere condotto mediante l'analisi delle scelte individuali espresse in termini di graduatorie di similarità percepita. In particolare, se si assume che ciascuna parola sia caratterizzata da uno spazio semantico definibile attraverso l'insieme dei suoi sinonimi[2] (Cappelli, 2003 ; Cappelli e Corduas, 2003), l'analisi delle graduatorie elaborate dagli utenti sugli elementi facenti parte di tale spazio (ordinandoli in base al grado di sinonimia percepita) può essere efficacemente condotta mediante l'impiego di modelli statistici per ranghi. La proposta di utilizzo di tali modelli, in effetti, deriva dalla considerazione che essi consentono di quantificare il livello di sinonimia percepita e di individuare l'esistenza di sensi in base al legame tra le principali caratteristiche dei soggetti e le graduatorie da essi stessi espresse.

Il presente lavoro è così organizzato : nel paragrafo 2 vengono descritte, nel dettaglio, le motivazioni dell'approccio proposto e la metodologia sviluppata per l'analisi della percezione della sinonimia, mediante l'impiego di un modello mistura per variabili rango ; nel paragrafo 3 si illustrano le caratteristiche della indagine condotta per la valutazione della sinonimia percepita ; la presentazione ed il commento dei risultati della indagine costituiranno oggetto dei paragrafi 4.1 e 4.2. Alcune considerazioni finali concludono il lavoro.

[2] E' evidente che si tratta di una ipotesi di lavoro ; nulla vieta che se ne adottino altre. Ad esempio, che si definisca lo spazio semantico di una parola a partire dall'insieme dei suoi antonimi, oppure dei sinonimi e degli antonimi.

2. La metodologia

Si consideri una parola obiettivo w e sia $S_w = [s_1, \ldots, s_j, \ldots, s_m]$ l'insieme di tutti i suoi possibili sinonimi, individuati sia mediante ricerca manuale che elettronica (come esplicitato in Cappelli, 2003 ; Cappelli e Corduas, 2003).

Si assuma, inoltre, che n soggetti elaborino una graduatoria degli m elementi di S_w secondo un criterio di sinonimia rispetto a w : ogni "giudice", cioè, assegna rango R=1 al sinonimo (tra gli m a disposizione) che percepisce come più *simile* alla parola w, rango R=2 a quello che percepisce come successivo, e così via, fino ad arrivare al vocabolo che viene percepito come il più lontano in termini di sinonimia e che, quindi, riceve rango R=m.

In tal modo, ad ogni prefissato elemento s_j (j = 1, 2, …, m) di S_w è associato un vettore di ranghi osservati $\mathbf{r} = (r_1, r_2, \ldots, r_n)'$ che rappresentano una misura del grado di sinonimia tra s_j e la parola obiettivo w così come viene percepito dagli n giudici. Tale tipologia di dati può essere efficacemente analizzata mediante modelli per variabili rango (per una rassegna : Marden, 1995 ; D'Elia e Piccolo, 2002).

Al fine di proporre un modello statistico adeguato a rappresentare il meccanismo generatore dei ranghi osservati $\mathbf{r}$ per ogni s_j, è utile riflettere sulla procedura psico-linguistica che, presumibilmente, presiede all'elaborazione di una graduatoria di sinonimia, da parte di soggetti non-esperti. In effetti, l'assegnazione di ranghi ad un insieme di elementi (parole, concetti, oggetti, ecc.) richiede un procedura di ordinamento cui fare riferimento : a tal proposito, la letteratura psicometrica ha dato grande rilievo al criterio dei confronti appaiati (*paired comparisons*), in base al quale la formulazione di graduatorie di m elementi deriva dagli $m(m-1)/2$ confronti tra tutte le possibili coppie di *items* (Bradley e Terry, 1952). E' risultato, altresì, vero che tale procedura è accompagnata da una componente di incertezza che, generalmente, aumenta con m e caratterizza soprattutto i ranghi assegnati a quegli elementi verso i quali non esiste un giudizio netto da parte del soggetto.

Con riguardo alle graduatorie di sinonimia, possiamo assumere che il rango assegnato ad un sinonimo s_j di una parola w sia il risultato di un processo articolato in due componenti, che intervengono con ruoli distinti. La prima componente è relativa alla valutazione (di ciascun soggetto) concernente la sinonimia di s_j rispetto a w : ci sembra lecito ipotizzare che tale valutazione avvenga secondo uno schema di confronti appaiati, mediante il quale confrontando tutte le possibili coppie di sinonimi a disposizione si individua il vocabolo maggiormente prossimo a w (quello "vincente" in tutti i confronti), e così via[3]. La seconda componente, invece, esprime l'incertezza che è presente nell'assegnazione del rango, a causa della natura intrinsecamente "sfocata" dei possibili sensi di w nella percezione linguistica.

Tali due componenti possono essere rappresentate dalle variabili casuali Binomiale traslata e Uniforme discreta, rispettivamente, mediante un modello mistura definito MUB (D'Elia e Piccolo, 2003). Sia, infatti, r il rango assegnato ad un sinonimo s_j della parola w ; allora, è possibile considerare r come una realizzazione della variabile casuale $R \sim \text{MUB}(m, \pi, \xi)$ se :

$$\Pr(R = r) = \pi P_B(r) + (1-\pi) P_U(r), \qquad r = 1, 2, \ldots, m,$$

[3] Evidentemente, è anche possibile ipotizzare che tale processo si svolga in modo gerarchico, quando si è in presenza di una parola w che ammette una lista di sinonimi S_w i cui significati sono riconducibili a k concetti ($C_1, \ldots, C_h, \ldots, C_k$) ben distinti. In tal caso, quindi, i confronti appaiati avvengono a due livelli : prima tra i concetti, e successivamente tra le parole all'interno di uno stesso concetto C_h (h=1, 2, …, k).

$$P_B(r) = \binom{m-1}{r-1}(1-\xi)^{r-1}\xi^{m-r}; \qquad P_U(r) = \frac{1}{m}; \qquad r = 1,2,...,m;$$

dove :

e $\pi \in [0, 1]$; $\xi \in [0, 1]$.

Con riferimento alle graduatorie di sinonimia percepita, i parametri del modello MUB possono così essere interpretati : m è il numero dei sinonimi di w, e come tale è fisso e noto a priori. Il parametro π è inversamente legato all'incertezza presente nel processo di formulazione di una graduatoria di sinonimia : difatti, il fattore $(1-\pi)$ è una misura dell'incertezza che compete al rango assegnato a s_j.

Per quanto riguarda, invece, il significato del parametro ξ, può essere utile considerare le due seguenti espressioni :

$$\Pr(R=1) = \pi\xi^{m-1} + (1-\pi)/m; \qquad E(R) = \pi(m-1)(1/2-\xi) + (m+1)/2.$$

Da esse, infatti, emerge che – a parità di π – al crescere di ξ aumenta Pr(R = 1), cioè la probabilità che s_j sia considerato il sinonimo più vicino alla parola w, e viceversa diminuisce $E(R)$, conducendo così ad una distribuzione di probabilità con un valore atteso del rango assegnato a s_j più basso. Ne deriva che il parametro ξ può essere interpretato come una misura della forza della sinonimia percepita : in particolare, poiché per ξ = 1/2 si ottiene una distribuzione simmetrica intorno a $E(R) = (m+1)/2$, tale valore del parametro (1/2) può essere considerato come una soglia tra sinonimia sentita in modo debole e forte. In altri termini, il parametro ξ può essere altresì considerato come una *misura della scambiabilità* tra la parola w e il sinonimo s_j, così come risulta dalle evidenze empiriche.

Per quanto concerne, poi, le stime dei due parametri (π, ξ) caratterizzanti la distribuzione, esse sono derivabili con il metodo della massima verosimiglianza, la cui complessità computazionale richiede il ricorso all'algoritmo E-M (*Expectation – Maximisation*). L'efficacia di tale procedura per la stima di modelli mistura è ampiamente documentata nella letteratura statistica (McLachlan e Krishnan, 1997 ; McLachlan e Peel, 2000), ed è stata riscontrata anche per il modello MUB (D'Elia e Piccolo, 2003).

Il modello MUB può essere esteso al fine di contemplare la presenza di covariate relative ai soggetti che esprimono le graduatorie di sinonimia. In particolare, seguendo una logica analoga a quella dei Modelli Lineari Generalizzati (McCullagh e Nelder, 1989), è possibile introdurre un legame tra il vettore di variabili esplicative $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})'$ specifico di ciascun i-esimo soggetto (i = 1, 2, …, n) e i parametri π e ξ, considerati separatamente (D'Elia, 2003) oppure congiuntamente (Piccolo, 2003).

La specificazione del modello MUB, quindi, avviene mutuando la funzione logistica dei modelli logit, mediante la quale si crea una corrispondenza tra l'insieme reale e l'intervallo [0, 1] su cui sono definiti sia π che ξ :

$$(\pi \mid \mathbf{X} = \mathbf{x}_i) = \frac{1}{1+\exp(-\mathbf{x}_i\boldsymbol{\beta})}; \qquad (\xi \mid \mathbf{X} = \mathbf{x}_i) = \frac{1}{1+\exp(-\mathbf{x}_i\boldsymbol{\gamma})}.$$

Ivi $\mathbf{X}$ rappresenta la matrice del disegno di dimensioni (n x p+1), contenente i valori delle covariate degli n soggetti, mentre $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ sono i vettori dei coefficienti relativi a tali covariate.

In particolare, se si adotta la specificazione relativa al solo parametro π, si modella una situazione nella quale le caratteristiche dei soggetti che elaborano le graduatorie di sinonimia hanno effetto solo sull'incertezza, ma non sul grado di sinonimia percepita ; viceversa, se si adopera solo la specificazione relativa al parametro ξ, ciò equivale ad assumere che il grado di sinonimia avvertita dipenda dalle covariate, ma non l'incertezza nell'assegnazione del rango. Chiaramente, l'utilizzo contemporaneo di entrambe le specificazioni conduce ad un modello nel quale sia il grado di sinonimia percepita che l'incertezza nella scelta sono interpretabili attraverso le caratteristiche dei soggetti intervistati.

Anche nel caso del modello MUB esteso con la presenza delle covariate, la stima dei parametri e/o dei coefficienti delle variabili esplicative avviene mediante il ricorso all'algoritmo E-M, opportunamente modificato per tener conto delle covariate.

Le stime ottenute possono, quindi, essere adoperate per quantificare non solo la percezione della sinonimia e la connessa incertezza, ma anche per valutare l'impatto che le caratteristiche individuali hanno in tali componenti. In particolare, mediante il modello MUB con l'inclusione di covariate è possibile individuare profili distinti di soggetti (ad esempio, in base al livello di istruzione) e misurare come i diversi livelli di una o più variabili conducano a diverse graduatorie di sinonimia percepita.

3. L'indagine

L'indagine si è svolta nell'arco di un mese mediante somministrazione di un questionario a 654 soggetti non esperti ed appartenenti ad ambiti sociali e culturali sufficientemente differenziati. Esso è articolato in due sezioni : la prima è relativa alle informazioni sul soggetto e sul contesto in cui vive e opera ; la seconda parte, invece, prevede l'elaborazione di graduatorie di sinonimia.

A tal fine sono stati prescelti vocaboli appartenenti a categorie grammaticali distinte : cioè sostantivi, aggettivi, verbi. I vocaboli scelti sono : *scolaro*, *solare*, *piantare*. Per ciascun vocabolo la lista estesa dei sinonimi è stata ottenuta a partire dai cinque maggiori dizionari italiani dei sinonimi[4], e mediante consultazione del sito web :
http ://parole.virgilio.it/parole/sinonimi_e_contrari/.

Nel questionario le liste dei sinonimi di ciascun vocabolo sono state presentate in ordine alfabetico. Ad ogni soggetto partecipante all'indagine è stato chiesto di elaborare una graduatoria dei sinonimi, ordinandoli a partire da quello che lui/lei riteneva maggiormente simile (e, quindi, scambiabile) rispetto alla parola *w* di riferimento. E'stato, inoltre, chiesto di non assegnare lo stesso rango a sinonimi distinti all'interno di una stessa lista, e di non consultare vocabolari, dizionari, grammatiche, ecc.

4. Principali evidenze empiriche

La nostra proposta per un nuovo approccio all'analisi della sinonimia consente di misurare e valutare in modo oggettivo il grado di similarità percepito tra una parola ed i suoi possibili sinonimi. Inoltre, nel caso del modello con covariate, è possibile individuare profili di utenti in base ai sensi della parola che vengono da quest'ultimi privilegiati e porre, quindi, in corrispondenza le caratteristiche individuali e il grado di sinonimia assegnato.

[4] Gabrielli (1967, Loescher) ; Pittano (1987, Zanichelli) ; Quartu (1994, Rizzoli) ; De Mauro (2002, Mondadori) ; Stoppelli (2002, Garzanti).

In questo paragrafo illustriamo alcuni risultati (i principali, per motivi di spazio) emersi dall'indagine, al fine di mettere il luce le potenzialità della metodologia proposta e descritta nel paragrafo 2.

4.1. *La stima del grado di sinonimia percepita*

Nelle tabelle 1, 2 e 3 sono illustrati i risultati della applicazione del modello MUB (senza inclusione di covariate) ai tre vocaboli considerati. Per ciascun vocabolo è riportato l'elenco dei relativi sinonimi con l'indicazione delle stime $\hat{\xi}$ e $\hat{\pi}$ dei corrispondenti parametri del modello (e i rispettivi errori standard), il rango medio $\bar{r}_n$ e la varianza dei ranghi osservati. Si è ritenuto di ordinare le graduatorie in base al valore stimato del parametro ξ, poiché, come già detto nel paragrafo 2, esso può essere interpretato come una *misura di scambiabilità* tra i vocaboli e quindi fornisce una indicazione immediata del grado di sinonimia percepito tra s_j e w. Si noti inoltre che le stime del parametro ξ dipendono dal valore $\bar{r}_n$ (media dei ranghi osservati) ma in modo non lineare : pertanto, le graduatorie basate sull'uno o sull'altro criterio possono coincidere (come nel caso del vocabolo *scolaro*), ma non necessariamente. Per quanto concerne il parametro π, si è detto nel paragrafo 2 che esso è inversamente legato all'incertezza nella formulazione della graduatoria : pertanto, per ciascun sinonimo, il complemento ad uno della stima di π, (1-$\hat{\pi}$), fornisce una misura della incertezza nell' assegnazione del relativo rango a s_j rispetto a w.

Sinonimi	$\hat{\xi}$	*e.s.*	$\hat{\pi}$	*e.s.*	$\bar{r}_n$	Var(r)
Alunno	0.924	*0.005*	0.955	*0.012*	1.561	0.802
Studente	0.817	*0.007*	0.917	*0.019*	2.249	1.349
Allievo	0.712	*0.007*	0.990	*0.011*	2.745	0.823
Educando	0.376	*0.008*	0.963	*0.022*	4.720	1.409
Discepolo	0.286	*0.008*	0.945	*0.020*	5.223	1.397
Discente	0.237	*0.009*	0.824	*0.028*	5.372	1.995
Seguace	0.125	*0.006*	0.941	*0.015*	6.130	1.202

Tabella 1. Sinonimia percepita rispetto a "scolaro".

Nel caso del vocabolo *scolaro*, il sinonimo percepito come più prossimo, e quindi dotato del maggior grado di scambiabilità, è *alunno* ; si ricordi, infatti, che essendo $\xi \in [0, 1]$ il valore 0.924 sta ad indicare una grado di scambiabilità quasi perfetto. Meno prossimi, ma comunque percepiti come altamente scambiabili sono i sinonimi *studente* ed *allievo*. I rimanenti sinonimi sono invece caratterizzati da un valore stimato di ξ di molto inferiore. Quindi, si può dire che nel caso della parola *scolaro*, vi è un nucleo forte di sinonimi formato dai vocaboli *alunno*, *studente* ed *allievo* che sono legati al senso di *scolaro* come colui che apprende delle nozioni nell'ambito di un programma educativo. Il termine *educando* che si colloca al centro della graduatoria sembra invece identificare un senso separato, mentre i tre i rimanenti sinonimi che individuano un senso della parola in oggetto come colui che apprende un credo o una dottrina, sono percepiti come meno scambiabili rispetto al vocabolo *scolaro* e quindi sono sinonimi deboli nella percezione degli utilizzatori. Come si è detto in precedenza, le stime dei parametri sono legate a $\bar{r}_n$ che in tale caso fornisce una graduatoria coincidente con quella ottenuta mediante le stime di ξ ed i cui valori, se si guarda al loro campo di variazione, rispecchiano la distinzione effettuata tra sinonimi forti e deboli.

Per quanto riguarda invece l'aspetto della incertezza nella formulazione della graduatoria, i sinonimi sono caratterizzati tutti da una bassissima incertezza, che è presumibilmente riconducibile alla ridotta numerosità dei sinonimi in questione (m = 7).

Nel caso del vocabolo *solare* l'esame della graduatoria consente di identificare una sorta di percorso che conduce dal senso di *solare* inteso come capacità (figurata) di irradiare luce (*radioso*, *raggiante*, *luminoso*, *splendente*, *brillante*, *scintillante*, *sfolgorante*) a quello di *solare* come comprensibile (*lampante*, *visibile*, *evidente*, *palese*, *indiscutibile*, *indubitabile*, *innegabile*, *lapalissiano*). Si noti come il termine *chiaro*, associabile ad entrambi i sensi, occupi una sorta di posizione di passaggio nell'ambito della graduatoria giocando un ruolo di termine "di transizione".

Sinonimi	$\hat{\xi}$	*e.s.*	$\hat{\pi}$	*e.s.*	$\bar{r}_n$	Var(r)
Radioso	0.914	*0.004*	0.648	*0.024*	3.841	9.468
Raggiante	0.913	*0.004*	0.634	*0.025*	3.940	10.185
Luminoso	0.904	*0.005*	0.602	*0.026*	4.017	9.549
Splendente	0.832	*0.005*	0.774	*0.023*	4.320	7.954
Brillante	0.763	*0.006*	0.733	*0.026*	5.115	8.019
Scintillante	0.675	*0.007*	0.719	*0.028*	6.746	9.287
Sfolgorante	0.625	*0.011*	0.437	*0.034*	7.879	13.861
Chiaro	0.591	*0.008*	0.666	*0.031*	6.913	9.083
Lampante	0.440	*0.009*	0.604	*0.032*	9.125	10.107
Visibile	0.422	*0.009*	0.575	*0.033*	10.003	10.508
Evidente	0.406	*0.006*	0.799	*0.026*	9.599	7.160
Palese	0.266	*0.008*	0.608	*0.030*	11.436	10.450
Indiscutibile	0.175	*0.005*	0.823	*0.021*	12.783	6.467
Indubitabile	0.146	*0.004*	0.884	*0.017*	13.318	5.416
Innegabile	0.123	*0.004*	0.825	*0.020*	13.416	6.971
Lapalissiano	0.023	*0.003*	0.395	*0.024*	13.182	9.366

Tabella 2. Sinonimia percepita rispetto a "solare".

Un ulteriore aspetto da sottolineare riguarda la incertezza nella formulazione dei ranghi che appare maggiore rispetto al caso del vocabolo *scolaro* precedentemente esaminato. I più elevati valori di (1-$\hat{\pi}$) sono da ascrivere alla maggiore lunghezza della lista di sinonimi che tende ad accrescere l'incertezza nella assegnazione del rango e anche la variabilità dei ranghi osservati. E' opportuno notare che, nonostante i differenti valori di m (7 e 16), una comparazione tra i vocaboli, in termini di incertezza, è comunque possibile, anche se non è di immediato interesse per i fini di questo lavoro.

L'ultimo vocabolo considerato è quello del verbo *piantare* che presenta una lista di sinonimi piuttosto lunga (m = 20) nell'ambito della quale è possibile individuare vari sensi.

Sinonimi	$\hat{\xi}$	*e.s.*	$\hat{\pi}$	*e.s.*	$\bar{r}_n$	Var(*r*)
Seminare	0.975	*0.003*	0.487	*0.023*	5.630	31.646
Coltivare	0.949	*0.004*	0.467	*0.025*	6.219	32.755
Interrare	0.932	*0.005*	0.350	*0.026*	6.844	28.743
Innestare	0.860	*0.013*	0.203	*0.028*	9.257	35.291
Conficcare	0.744	*0.018*	0.191	*0.031*	9.183	27.064
Ficcare	0.721	*0.013*	0.279	*0.032*	9.353	23.571
Infilare	0.658	*0.013*	0.327	*0.033*	9.671	21.900
Inserire	0.616	*0.012*	0.348	*0.033*	9.587	20.105
Introdurre	0.601	*0.015*	0.281	*0.033*	10.783	24.347
Mettere	0.552	*0.015*	0.281	*0.034*	10.641	24.447
Collocare	0.511	*0.015*	0.280	*0.034*	10.392	24.515
Porre	0.479	*0.013*	0.348	*0.034*	11.477	21.720
Sistemare	0.428	*0.016*	0.268	*0.023*	12.838	23.643
Mollare	0.216	*0.012*	0.273	*0.030*	12.275	32.019
Abbandonare	0.214	*0.026*	0.118	*0.029*	11.618	40.371
Lasciare	0.207	*0.011*	0.293	*0.030*	11.963	33.133
Cessare	0.174	*0.011*	0.267	*0.029*	12.635	31.376
Smettere	0.170	*0.010*	0.318	*0.029*	12.446	33.645
Interrompere	0.155	*0.007*	0.423	*0.029*	13.765	26.608
Troncare	0.038	*0.008*	0.154	*0.022*	12.593	35.330

Tabella 3. Sinonimia percepita rispetto a "piantare".

Innanzitutto, si osservi che tale verbo viene percepito come avente un senso legato alla attività agricola con tre sinonimi (*seminare*, *coltivare* ed *interrare*) caratterizzati da un grado di scambiabilità molto elevato e un quarto (*innestare*) meno marcato, ma comunque percepito come altamente scambiabile. Un secondo senso è individuato dai sinonimi *conficcare*, *ficcare*, *infilare*, *inserire*, *introdurre*, caratterizzati da un grado percepito di scambiabilità medio. Gli ultimi due sensi individuabili sono costituiti da sinonimi deboli : *piantare* come collocare in un posto o ordine (*mettere*, *collocare*, *porre*, *sistemare*) e *piantare* come l'atto di porre fine. Si noti come in questo caso, per tutti i sinonimi il grado di incertezza sia particolarmente elevato.

4.2. *L'effetto delle caratteristiche individuali sulla percezione della sinonimia*

Il modello MUB per l'analisi delle graduatorie di sinonimia consente anche di includere nella sua specificazione la presenza di covariate, vale a dire di variabili esplicative relative alle caratteristiche individuali, mediante le quali è possibile interpretare le graduatorie espresse.

In particolare, sembra interessante poter individuare quali variabili condizionano la posizione in graduatoria di un fissato sinonimo s_j (j=1, 2,…, m) rispetto alla parola obiettivo w, e quantificare direzione e forza di tale impatto. Inoltre, l'opportuna combinazione di valori delle variabili esplicative, risultate rilevanti, permette la definizione di profili di utenti della

lingua, la cui percezione di sinonimia del termine s_j rispetto a *w* appare significativamente differenziata.

Al fine di evidenziare tali potenzialità, illustriamo di seguito solo alcuni risultati emersi dall'indagine condotta rispetto alle parole obiettivo, le cui stime delle graduatorie di sinonimia percepita sono state commentate nel precedente paragrafo 4.1.

• Con riferimento al sostantivo *scolaro*, discutiamo le evidenze emerse per il sinonimo *discente* (che presenta un grado di scambiabilità molto modesto, cioè $\hat{\xi}$ = 0.237). Per tale sinonimo sono risultate significative rispetto al parametro ξ (misura di sinonimia) le variabili esplicative "numero di componenti della famiglia", "possesso della laurea", "lettore assiduo (di libri)", come si evince dalla Tabella 4. La misura dell'incertezza, espressa da $(1 - \hat{\pi})$, invece non è risultata dipendere da alcuna covariata.

Covariate	**Stime ($\hat{\gamma}$)**	**Errori standard**
Costante	-0.831	*0.167*
Numero componenti famiglia	-0.115	*0.038*
Laurea (NO=0, SI=1)	0.896	*0.110*
Lettore assiduo (NO=0,SI=1)	0.290	*0.121*
	($\hat{\pi}$)	**Errore standard**
	0.867	*0.013*

Tabella 4. Modello MUB con covariate per il sinonimo "discente".

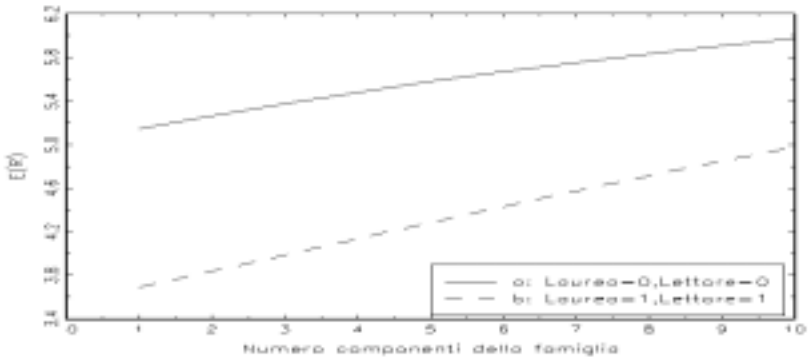
Le stime ottenute evidenziano che il possesso della laurea e il fatto di essere lettori assidui esercitano un impatto positivo sul grado di sinonimia percepita, in quanto determinano un aumento del parametro ξ, che può essere considerato come una misura di scambiabilità del sinonimo *discente* rispetto alla parola obiettivo *scolaro*. Un impatto di segno opposto è, invece, esercitato dalla variabile "numero di componenti della famiglia".

Nella successiva Tabella 5 sono riportati, per alcuni possibili profili di parlanti, i corrispondenti valori attesi del rango assegnato a *discente* nella graduatoria di sinonimia e le corrispondenti stime del parametro ξ.

Numero componenti famiglia	Laurea (NO=0, SI=1)	Lettore assiduo (NO=0, SI=1)	$\hat{\xi}$	E(*R*)
3	0	0	0.236	5.376
6	0	0	0.180	5.670
3	1	1	0.503	3.985
6	1	1	0.418	4.429

Tabella 5. Profili e valori attesi del rango per il sinonimo "discente"

In tal modo, emerge che un profilo culturale elevato (laureato e lettore assiduo) insieme con l'appartenenza ad una famiglia poco numerosa determinano il più alto grado ($\hat{\xi}$ = 0.503) di sinonimia percepita della parola *discente* rispetto a *scolaro* ; all'opposto si colloca il profilo



culturale modesto congiunto alla provenienza da famiglie abbastanza numerose ($\hat{\xi}$ = 0.180). Ciò appare giustificabile, se si considera che nella lingua italiana il termine *discente* non risulta essere di uso comune, ed è da considerarsi una parola cosiddetta "colta", per la quale è lecito ipotizzare che esistano difficoltà di attribuzione di senso da parte di persone di media/modesta cultura.

Il ruolo svolto dal "numero di componenti della famiglia" in rapporto al valore atteso del rango assegnato a *discente* è evidenziato anche nella Figura 1, che conferma come al crescere del "numero di componenti" diminuisca la percezione di sinonimia rispetto a *scolaro* (in quanto aumenta il rango atteso).

• Per quanto riguarda il verbo *piantare*, illustriamo qui i risultati ottenuti per il sinonimo *seminare*, che presenta il maggior grado di scambiabilità stimato ($\hat{\xi}$ = 0.975), come si evinceva dalla precedente Tabella 3.

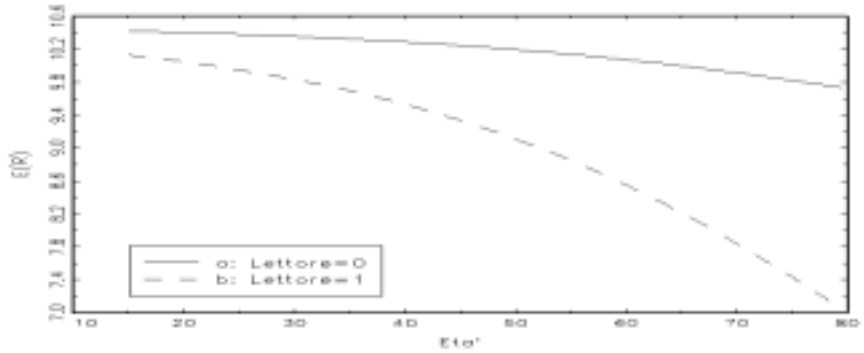
Per tale sinonimo sono risultate significative rispetto al parametro ξ (misura di sinonimia) la variabile esplicativa "lettore assiduo", e rispetto al parametro π la variabile "età" (Tabella 6).

Covariate	**Stime ($\hat{\gamma}$)**	**Errori standard**
Costante	0.446	*0.088*
Lettore assiduo (NO=0,SI=1)	5.049	*0.797*
	($\hat{\beta}$)	**Errori standard**
Costante	-3.798	*0.380*
Età	0.041	*0.010*

Tabella 6. Modello MUB con covariate per il sinonimo "seminare".

Le stime ottenute evidenziano, anche in questo caso, che il fatto di essere lettori assidui esercita un impatto positivo sul grado di sinonimia percepita, in quanto determina un aumento del parametro ξ, che può essere considerato come una misura di scambiabilità del sinonimo *seminare* rispetto alla parola obiettivo *piantare*. Per quanto concerne, invece, la variabile "età", essa influisce sul grado di incertezza : infatti, il segno positivo della stima del rispettivo coefficiente evidenzia che all'aumentare dell'"età" cresce anche $\hat{\pi}$, diminuendo quindi l'incertezza.

In effetti, poiché il valore atteso del rango di un sinonimo s_j rispetto a w dipende sia dal grado di scambiabilità percepita che dalla misura di incertezza, entrambe le variabili individuate ("lettore assiduo" ed "età") esercitano un ruolo nel determinare la posizione attesa di *semi-*



nare nella graduatoria, condizionatamente a determinati profili di utenti. Ciò è evidenziato nella successiva Figura 2.

Dalla Figura emerge che la variabile "lettore assiduo" diventa via via più rilevante al crescere dell'"età", con la quale, evidentemente, interagisce. In particolare, appare che il senso di *piantare* inteso come *seminare* è privilegiato da persone più adulte e lettori assidui, che quindi prediligono l'accezione più tradizionale del verbo in questione.

5. Considerazioni finali

L'articolo ha evidenziato come nello studio della sinonimia percepita possa essere utile il ricorso a modelli per variabili rango, che permettono da un lato di stimare in modo oggettivo sia il grado di sinonimia che l'eventuale misura di incertezza presente nel processo, e d'altro canto consentono l'individuazione e la quantificazione dell'impatto esercitato su tale percezione dalle caratteristiche personali dei parlanti. Tali potenzialità possono essere utilmente sfruttate in numerosi ambiti, nei quali è importante associare ad un termine il senso privilegiato dalla maggior parte degli utenti, o da una parte di essa con particolari caratteristiche (si pensi, ad esempio, al problema della cosiddetta *information retrivial*).

Ulteriori sviluppi sono possibili sia dal punto di vista metodologico che in termini di campi di applicazione. Sul primo versante, infatti, sarebbe utile specificare un modello per ranghi di tipo gerarchico, in modo da tener conto del fatto che l'elaborazione di una graduatoria di sinonimia può avvenire mediante due stadi : graduatoria tra i differenti sensi e, poi, assegnazione dei ranghi all'interno di un singolo senso. Dal punto di vista delle applicazioni, invece, sarebbe interessante investigare il ruolo della sinonimia percepita con riferimento a lingue diverse, per cogliere l'eventuale presenza di strutture semantiche analoghe.

Ringraziamenti : Il presente lavoro è stato svolto nell'ambito dei progetti di ricerca afferenti al Dipartimento di Scienze Statistiche, Università di Napoli Federico II ; ci si è inoltre avvalsi dei fondi della L.R. 5/2002.

Bibliografia

Bolasco S. (1999). *Analisi Multidimensionale dei Dati*. Carocci.

Bradley R.A. e Terry M.A. (1952). Rank analysis of incomplete block designs I. *Biometrika*, vol. (39) : 324-345.

Cappelli C. (2003). Identifying word senses from synonyms : a cluster analysis approach. *Quaderni di Statistica*, vol. (5) : 105-117.

Cappelli C. e Corduas M. (2003). Assessing synonymy links : a cluster analysis approach. In *Book of Short Papers CLADAG 2003* : 91-94.

D'Elia A. (2003). A mixture model with covariates for ranks data : some inferential developments. *Quaderni di Statistica*, vol. (5) : 1-25.

D'Elia A. e Piccolo D. (2002). Analisi statistica delle preferenze : metodi e modelli a confronto. In Frosoni B., Magagnali U. e Boari G. (Eds), *Studi in onore di Angelo Zanella*. Vita e Pensiero : 167-187.

D'Elia A. e Piccolo D. (2003). A mixture model for preferences data analysis. *Submitted*.

De Mauro T. (2002). *Dizionario della lingua italiana*. Mondadori.

Lebart L., Salem A. e Berry L. (1998). *Exploring Textual Data*. Kluwer Academic Publishers.

Lin D., Zhao S., Qin L. e Zhou M. (2003). Identifying Synonyms among Distributionally Similar Words. In *Proceedings of IJCAI-03* : 1492-1493.

Marden J.I. (1995). *Analyzing and Modeling Rank Data*. Chapman & Hall.

McCullagh P. e Nelder J. (1989). *Generalized Linear Models (2nd edition)*. Chapman & Hall.

McLachlan G. e Krishnan T. (1997). *The E-M Algorithm and Extensions*. J. Wiley & Sons.

McLachlan G. e Peel D. (2000). *Finite Mixture Models*. J. Wiley & Sons.

Piccolo D. (2003). Computational issues in the E-M algorithm for ranks model estimation with covariates. *Quaderni di Statistica*, vol. (5) : 27-48.

Ploux S. e Ji H. (2003). A model for matching semantic maps between languages (French/English, Englissh/French). *Computational Linguistics*, vol. (29) : 155-178.

Woods A., Fletcher P. e Hughes A. (1986). *Statistics in Language Studies*. Cambridge University Press.