Lingue e Linguaggi
Lingue Linguaggi 19 (2016), 61-75
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v19p61

FREQUENZA, LUNGHEZZA E OMONIMIA
Un'analisi degli omonimi nel vocabolario di base italiano

FEDERICA CASADEI
UNIVERSITÀ DELLA TUSCIA

Abstract – This paper aims to explore the relationship between word frequency, word length, and homonymy, through an analysis of the about 7,000 highest frequency lexemes that constitute the basic vocabulary in Italian (Vocabolario di Base, VDB). Data confirm that the development of homonymy is strongly related to word length: both in the overall lexicon and within VDB, word forms that are involved in homonymy are shorter than those that are not. At the same time, a strong correlation arises between word frequency and homonymy, since VDB lexemes are involved in homonymy to a greater extent than others. The percentage of lexemes whose forms have homonyms is much higher for the VDB (55%) than for less frequent lexemes (in the range of 10%-24%). Word length and word frequency seem to behave as two independent variables in favoring homonymy: the frequency being equal, shorter words have more homonyms; and the length being equal, more frequent words have more homonyms. This finding seems to support the hypothesis that the richness of homonymy in high frequency lexicon is due not only to the shortness of these words (i.e., the fact that the shorter the word, the more likely it is to find another word of accidentally the same form), but also to an organization principle of language. Given the disambiguating power of context, language might assign a greater amount of ambiguity to words that are easiest to process, i.e. shorter and more frequent words.

Keywords: homonymy; word frequency; word length; lexical semantics; statistical linguistics.

1. Omonimia e frequenza

1.1. Il ruolo della frequenza nei fenomeni linguistici

Il tema della frequenza è divenuto recentemente il punto di incontro di due approcci linguistici apparentemente tanto diversi da risultare inconciliabili, quello quantitativo-computazionale da un lato, quello funzionale-cognitivo dall'altro.

Che la frequenza d'uso di un'espressione linguistica abbia ricadute sia sulla sua forma sia sul suo contenuto è noto alla linguistica quantitativa sin dagli studi pionieristici di Zipf (1936, 1945, 1949), al quale si deve la prima formulazione sistematica di una serie di correlazioni tra la frequenza d'uso di una parola, la sua lunghezza (le parole più frequenti sono più brevi), la quantità di significati che può esprimere (le parole più frequenti sono più polisemiche), la sua versatilità semantica (le parole il cui significato è più generico sono più frequenti). Lo sviluppo delle analisi statistiche sulle lingue ha portato nei decenni successivi a formulare altre leggi, e lo studio delle leggi linguistiche è oggi al centro, in particolare, della cosiddetta linguistica sinergetica, un modello di linguistica quantitativa che tenta di spiegare le regolarità quantitative osservate nelle lingue alla luce di una teoria semiotica generale (v. Köhler 1986, 1990, 2005).

Parallelamente il tema della frequenza ha acquistato un ruolo centrale nella linguistica cognitiva e in particolare in quella che va sotto il nome di linguistica cognitivo-funzionale o *usage-based linguistics*. Questa espressione, coniata da Langacker (1987, p. 494) per indicare le teorie che rifiutano una netta distinzione tra competenza e uso, è

utilizzata oggi soprattutto in relazione all'approccio promosso da Bybee (1985), nel quale confluiscono studi diversi accomunati dall'assunto che le strutture linguistiche emergano dall'uso e che la frequenza svolga un ruolo decisivo nello sviluppo di tutti i fenomeni linguistici, nel comportamento comunicativo dei parlanti, nell'acquisizione del linguaggio, nel mutamento linguistico (v. Barlow, Kemmer 2000; Bybee 2001, 2007; Bybee, Hopper 2001; Tomasello 2003; Gries, Divjak 2012a, 2012b).

Non mancano, per la verità, dubbi sul ricorso alla frequenza come chiave interpretativa di ogni fenomeno linguistico, e non solo tra i critici dell'approccio *usage-based* come Newmeyer (1998, 2003). Anche tra coloro che riconoscono l'importanza delle informazioni statistiche che emergono dall'analisi dei corpora, ci si interroga se la frequenza sia di per sé un fattore esplicativo dei fenomeni o se non sia – per riprendere le parole di Greenberg (1966, p. 70) – un sintomo che a sua volta richiede di essere spiegato ("frequency is itself but a symptom and the consistent relative frequency relations which appear to hold for lexical items and grammatical categories are themselves in need of explanation").

Peraltro anche dal dettaglio delle analisi statistiche emerge che non sempre la frequenza è il *primus movens* delle relazioni quantitative osservabili nelle lingue. Ad esempio nella relazione tra la frequenza di un'unità linguistica e la sua politestualità, cioè la quantità di contesti in cui può occorrere, è quest'ultima, e non la frequenza, la variabile indipendente (Köhler 1986), sicché la relativa legge afferma che la frequenza di un costituente è funzione della sua politestualità, e non viceversa. E persino per quanto riguarda la relazione tra frequenza e lunghezza resta oggetto di discussione quale sia la direzione della relazione tra le due variabili, cioè se la prima dipenda dalla seconda o viceversa (v. Strauss *et al*. 2007, p. 277). Anche la direzione della relazione tra lunghezza e polisemia resta controversa, essendo anche qui possibili entrambe le opzioni (Grzybek 2015): poiché l'allungamento delle parole (per affissazione, composizione ecc.) nasce dalla necessità di renderne più specifico il significato, si potrebbe considerare la polisemia come variabile indipendente (cioè tanto meno una parola è polisemica, tanto più è lunga); se invece si assume che l'accorciamento dipenda dalla frequenza, si può considerare la polisemia come funzione della lunghezza (cioè la frequenza accorcia le parole, e le parole più corte sono più probabilmente polisemiche di quelle lunghe). Köhler (1999) ne conclude che aumento della lunghezza e riduzione della polisemia – o viceversa, accorciamento e aumento della polisemia – sono esiti simultanei di un unico processo. Fenk-Oczlon e Fenk (2010a, 2010b) ritengono invece che il primo e più forte impulso venga dalla frequenza, benché poi i due fattori finiscano per interagire nel senso che se da un lato la maggior frequenza d'uso è il fattore di innesco per lo sviluppo della polisemia, dall'altro l'elevata polisemia di un lessema contribuisce a favorirne una maggior frequenza, perché accresce la quantità di contesti in cui può essere usato e la quantità di espressioni formulaiche in cui può comparire.

1.2. Frequenza e ambiguità semantica

A prescindere dalle differenze sul ruolo attribuito alla frequenza nel funzionamento del linguaggio e nella competenza linguistica, è fuori discussione che vi sia una correlazione tra la frequenza di una forma e la quantità di significati e sensi di cui è portatrice, al punto che il grado di polisemia di un lessema può essere considerato un indicatore della sua frequenza. Così avviene, ad esempio, in WordNet, dove, in mancanza di dati affidabili sulla frequenza dei lessemi registrati, quest'ultima viene assegnata sulla base della loro polisemia ("the frequency data that are readily avilable in the technical literature and

semantically tagged corpora […] are inadequate for a database as extensive as WordNet", perciò "WordNet uses polysemy ad an index of familiarity", Fellbaum 1998, pp. 112-113).

Analogamente, in prospettiva diacronica, la polisemia può rappresentare un indicatore della trafila di diffusione di un lessema, secondo l'ipotesi di Alinei che le aree di prima diffusione di un lessema siano quelle in cui esso presenta maggiore densità semantica, essendo a sua volta quest'ultima dipendente dalla frequenza ("semantic density is first of all the product of the familiarity that a given word has locally", Alinei 1974, p. 17). Questa ipotesi sembra coerente con gli studi sulla correlazione tra polisemia ed età dei lessemi, dai quali emerge che le parole più antiche sono più polisemiche (v. Polikarpov 1999; Poddubbny, Polikarpov 2015).

Meno indagata, invece, è la relazione tra frequenza e omonimia. Quest'ultima, anzi, sembra essere l'unico grande fenomeno lessicale per cui non è stata formulata nessuna legge quantitativa (non è mai menzionata, ad esempio, sul sito dedicato alle *Laws in Quantitative Linguistics*, http://lql.uni-trier.de/index.php/Main_Page).

Certamente vi è una stretta relazione tra omonimia e lunghezza, visto che l'omonimia è spesso il prodotto di fenomeni di accorciamento dovuti alla perdita di materiale fonico e/o grafico in forme originariamente diverse o all'abbreviazione ([1]*metro* 'unità di misura' ~ [2]*metro* abbr. di *metropolitana*). E in generale delle tre maggiori cause di omonimia – la convergenza fonetica di forme etimologicamente diverse ([1]*acconto* 'caparra' der. di *conto* ~ [2]*acconto* 'amico' dal lat. tardo *accŏgnĭtum*), il prestito ([1]*braga* 'calzone' ~ [2]*braga* rumeno 'tipo di birra'), la divergenza semantica tra accezioni di un lessema polisemico ([1]*vite* 'pianta' ~ [2]*vite* 'asticella filettata'), le prime due, di gran lunga più frequenti della terza, agiscono più facilmente nel caso di forme brevi.

Perciò Jespersen, partendo dall'osservazione che l'omonimia è tanto più probabile quanto più una parola è breve ("the shorter the word, the more likely is it to find another word of accidentally the same sound", 2010, p. 334), calcola approssimativamente che in inglese gli omonimi monosillabici siano circa quattro volte più numerosi di quelli polisillabici. Sulla base dello stesso assunto, Polikarpov (1997) ritiene che il diverso numero di gruppi omonimici presenti in due repertori di omonimi russi e inglesi, rispettivamente circa 500 e oltre 2.000 casi, sia conseguenza del fatto che in russo le parole sono in media 1.4 volte più lunghe che in inglese. Dati più precisi sono forniti dall'analisi di Ke (2006), uno dei pochissimi lavori sull'omonimia che prenda in considerazione il lessico di maggior frequenza: considerando le 5.000 parole più frequenti in inglese, tedesco e olandese, emerge che in tutte e tre le lingue la percentuale di omofoni è molto più alta nelle 100 parole di maggior frequenza (35% in inglese, 16% in tedesco, 11% in olandese) e che nella maggior parte dei casi – anzi nella totalità dei casi per l'inglese – si tratta di monosillabi; la quantità di monosillabi nel lessico risulta essere il miglior indicatore della quantità di omonimi nelle lingue considerate.

Da un lato, dunque, sembra esserci un'ovvia correlazione tra omonimia, lunghezza e frequenza. Poiché l'omonimia si verifica più facilmente se le forme sono brevi, e poiché le parole più frequenti sono più brevi (e, in generale, la frequenza determina una maggiore incidenza di fenomeni di accorciamento), sembra logico aspettarsi che vi siano più omonimi tra i lessemi di maggior frequenza.

D'altro canto omonimia e frequenza possono collidere, perché l'esistenza di omonimi di pari frequenza può comportare la scomparsa di uno dei due, per evitare fraintendimenti, secondo la classica tesi del conflitto omonimico di Gilliéron. E la probabilità di scomparire sarà maggiore nel caso di omonimi monosillabici, che per la loro inconsistenza formale sono più soggetti a decadere; così, ad esempio, in francese antico il conflitto tra i due omofoni *pis* 'petto' (dal lat. *pectus*) e *pis* 'peggio' (dal lat. *peius*) causa

la perdita del primo e la sua sostituzione con *poitrine* 'pettorale'. Perciò, nel prendere in esame i fattori che influiscono sulla frequenza delle forme, Bloomfield (1933, p. 396) cita in particolare l'omonimia ("homonymy in general may injure the frequency of a form"). Analogamente Lyons (1968, p. 90) ritiene che la risoluzione dei conflitti omonimici illustrata da Gilliéron rappresenti l'esempio ideale dell'equilibrio omeostatico tra il principio zipfiano del minimo sforzo, che ha come esito l'accorciamento delle forme più frequenti, e l'esigenza di comprensione, che inibisce l'effetto di accorciamento causato dalla frequenza.

In sostanza, la relazione tra frequenza e omonimia risponderebbe a una sequenza del tipo: frequenza > accorciamento > omonimia > perdita di frequenza, per cui la maggior frequenza determina accorciamenti che causano omonimie, a loro volta causa di riduzione della frequenza.

Questa ipotesi, tuttavia, si basa sull'assunto che l'omonimia non sia altro che un ostacolo al buon funzionamento della lingua – una patologia linguistica, nella terminologia di Gilliéron (v. Gilliéron 1921), o una fonte di ambiguità, come viene solitamente indicata soprattutto nella linguistica formale la mancata relazione 1:1 tra forme e significati, si tratti di polisemia o di omonimia. Con l'aggravante, nel caso dell'omonimia, di essere l'esito di eventi accidentali che si verificano (e potrebbero non verificarsi) nell'evoluzione delle lingue, privo, a differenza della polisemia, di qualunque valore semiotico o cognitivo. Illustra bene questo punto Ullmann (1966) quando afferma che se è impensabile una lingua senza polisemia, una lingua senza omonimia è invece facilmente immaginabile e risulterebbe anzi più efficiente: "polysemy is in all probability a semantic universal inherent in the fundamental structure of language" (p. 194), viceversa "homonymy is not necessarily an unrestricted universal. […] one could easily imagine an idiom without any homonyms; it would be, in fact, a more efficient medium" (p. 197).

Alcuni studi recenti (Wasow *et al*. 2003; Wasow 2015; Piantadosi *et al*. 2011, 2015) suggeriscono però una prospettiva diversa, cioè che l'ambiguità sia una proprietà *desiderabile* dei sistemi linguistici, perché ne aumenta l'efficienza comunicativa. In particolare Piantadosi *et al*. (2015) ritengono che l'ambiguità nelle lingue comporti due vantaggi: anzitutto evitare l'eccessiva ridondanza e la trasmissione di informazioni inutili (come si avrebbe invece usando forme non ambigue anche laddove il contesto sia sufficiente alla disambiguazione), in secondo luogo consentire il riutilizzo di parole la cui produzione e comprensione è più facile. A conferma di questa seconda ipotesi, l'analisi degli omonimi in inglese, tedesco e olandese mostra che in tutte e tre le lingue la quantità di significati associati a una stessa forma è tanto maggiore quanto più questa è breve, frequente e fonotatticamente probabile. Che il lessico di alta frequenza sia più ricco di omonimi, dunque, si spiegherebbe non solo con la maggiore brevità delle forme in questione (cioè con il fatto, del tutto casuale, che queste trovano più facilmente degli omofoni e/o omografi), ma con un principio semiotico generale volto a ottimizzare la capacità comunicativa del codice linguistico.

In questo lavoro si propone un'analisi dell'omonimia nel lessico italiano di maggior frequenza con l'obiettivo di verificare se, e in quale misura, anche in italiano vi sia una correlazione tra frequenza, lunghezza e sviluppo di omonimie. Data la lunghezza media delle parole in italiano, è prevedibile che la quantità assoluta di omonimi risulti inferiore rispetto a quella osservata in lingue in cui le parole sono più brevi, e infatti l'italiano è tradizionalmente citato (v. ad esempio Ullmann 1966, p. 186) come una lingua a basso grado di omonimia rispetto, tipicamente, all'inglese. Tuttavia, come mostrano le analisi sia di Ke (2006) che di Piantadosi *et al*. (2015) sulle tre lingue rappresentate nel database lessicale CELEX (inglese, tedesco e olandese), la relazione tra frequenza,

lunghezza e omonimia vale crosslinguisticamente, dunque ci si può aspettare che valga anche per l'italiano.

Le analisi presentate si basano sui dati ricavabili dal maggiore dizionario italiano, il Gradit di De Mauro, e da HOMO, un database di circa 113.000 omonimi italiani compilato prendendo in considerazione tutte le forme dei paradigmi dei lessemi registrati dal Gradit (per una descrizione dei criteri di costruzione di HOMO v. Casadei 2016). Entrambe le fonti utilizzate includono solo forme omografe e, per la maggior parte, anche omofone (in HOMO i non omofoni sono il 4,5% del totale); diversamente quindi dai lavori di Ke (2006) e di Piantadosi *et al*. (2015), che prendono in considerazione gli omofoni anche non omografi (del tipo *wait* e *weight*), i dati qui presentati si riferiscono per lo più a omonimi perfetti, uguali sia nella grafia che nella pronuncia.

2. Frequenza d'uso e produttività omonimica

HOMO contiene 112.344 forme omonime riconducibili a 35.557 lessemi. Per ciascuna forma è indicata nel database, tra le altre informazioni, la fascia d'uso del lessema corrispondente, assegnata sulla base delle indicazioni del Gradit. Il dizionario, infatti, riporta per ciascun lemma una marca che ne segnala l'ambito di diffusione. Le marche d'uso usate dal Gradit possono essere ripartite in quattro classi:

1) vocabolario di base (VDB), costituito da circa 7.000 lessemi di massima frequenza, ripartiti in tre fasce di frequenza: il vocabolario fondamentale (FO), formato dai circa 2.000 lessemi più usati in assoluto in italiano; il vocabolario di alto uso (AU), formato da circa 2.700 lessemi di alta, benché minore, frequenza; e il vocabolario di alta disponibilità (AD), formato da circa 2.000 lessemi relativamente infrequenti nello scritto e nel parlato ma comunque noti ai parlanti perché legati a oggetti e azioni di grande rilevanza nella vita quotidiana;
2) vocabolario comune (marca CO), costituito da circa 50.000 lessemi estranei al VDB ma che si ritengono noti a chiunque abbia un livello di istruzione medio-superiore;
3) vocabolario tecnico-specialistico (TS), costituito da lessemi usati prevalentemente o solo in ambito scientifico, tecnologico o professionalmente settoriale;
4) ambito d'uso non tecnico-scientifico ma non comune: marche BU (basso uso), OB (obsoleto), LE (uso solo letterario), DI (uso dialettale), RE (uso regionale).

La Tabella 1 mostra come si distribuiscono le forme inventariate in HOMO, e i lessemi ai quali esse sono riconducibili, nelle varie fasce d'uso individuate dal Gradit. Come si vede, l'ambito d'uso nel quale si colloca la maggior parte degli omonimi è quello del vocabolario comune, seguito a poca distanza da quello tecnico-specialistico. Il vocabolario di base, invece, dà luogo solo all'11% delle omonimie.

fascia d'uso	n° di forme	n° di lessemi
CO	33.400 (30%)	11.979 (34%)
TS	28.219 (25%)	10.895 (31%)
BU	14.480 (13%)	3.542 (10%)
OB	12.522 (11%)	2.481 (7%)
VDB	11.532 (10%)	3.732 (11%)
FO	4.371 (4%)	1.321 (4%)
AU	4.306 (4%)	1.434 (4%)
AD	2.855 (2%)	977 (3%)
RE	6.722 (6%)	1.563 (4%)
LE	4.917 (4%)	1.110 (3%)
DI	373 (0,3%)	83 (0,2%)
ES	179 (0,1%)	172 (0,5%)
	112.344	35.557

Tabella 1
Fasce d'uso degli omonimi in HOMO (AD = alta disponibilità, AU = alto uso, BU = basso uso, DI = dialettale, ES = esotismo, FO = fondamentale, CO = comune, LE = letterario, OB = obsoleto, RE = regionale, TS = tecnico-specialistico, VDB = vocabolario di base).

Se si considera il totale degli omonimi italiani, dunque, il contributo del vocabolario di base appare piuttosto esiguo. Tuttavia si tratta di un dato che va letto alla luce di due considerazioni.

Anzitutto si può osservare che l'11% è comunque di una quota non irrilevante, posto che i lessemi del vocabolario di base rappresentano una porzione minima del lessico (meno del 3%, se si considera che il Gradit include 6.728 lessemi di base su un lemmario di oltre 260.000 entrate); per fare un paragone, i lessemi di fascia OB, che generano una quota di forme omonime pressoché uguale a quella prodotta dal vocabolario di base, sono però più del doppio di questi ultimi (circa 15.000 nel Gradit, pari a circa l'8% del lemmario).

Ma soprattutto, se si guarda quale sia nelle varie fasce d'uso la percentuale di lessemi coinvolti in omonimie, emerge che il vocabolario di base, in proporzione al numero di lessemi che lo costituiscono, è in realtà la fascia d'uso più coinvolta nell'omonimia. Come mostra la Tabella 2, infatti, risulta coinvolto in omonimie ben il 55% dei lessemi del vocabolario di base contro il 24% dei lessemi di uso comune, il 15-17% di quelli di basso uso e obsoleti, il 10% dei tecnico-specialistici.[1]

[1] Calcolare la distribuzione dei lessemi del Gradit nelle varie fasce d'uso è complicato per motivi legati sia ai criteri di lemmatizzazione del dizionario, sia al funzionamento del modulo di ricerca della versione elettronica. Quest'ultimo non distingue il caso in cui una marca d'uso sia presente nell'intestazione del lemma dal caso in cui compaia solo in una delle accezioni (sicché ad esempio conteggia tra i lemmi OB anche *perverso*, che invece è CO in intestazione); vi sono poi lemmi che hanno in intestazione più di una marca e viceversa altri che non ne hanno nessuna perché sono trattati con rinvio secco ad altro lemma. Pur cercando di tenere conto di questi elementi, il conteggio ha un certo margine di approssimazione. Il totale 223.649 indicato nella Tabella 2 non corrisponde al lemmario del Gradit (260.709) perché esclude circa 37.000 lemmi privi di marca d'uso.

fascia d'uso	n° lessemi nel Gradit	n° lessemi in HOMO	%
VDB	6.728	3.732	55
FO	2.077	1.321	64
AU	2.663	1.434	54
AD	1.988	977	49
CO	49.845	11.979	24
RE	7.124	1.563	22
LE	5.551	1.110	20
OB	14.488	2.481	17
BU	24.320	3.542	15
DI	558	83	15
TS	111.301	10.895	10
ES	3.734	172	5
	223.649*	35.557	

Tabella 2
Quantità di lessemi coinvolti in omonimie nelle varie fasce d'uso (*v. nota 1).

Quest'ultimo dato sembra indicare una forte correlazione tra frequenza e sviluppo di omonimie. Lasciando da parte gli esotismi, la cui scarsa partecipazione all'omonimia si spiega col fatto che hanno spesso una forma grafica tale da rendere improbabile l'esistenza di un omonimo italiano,[2] e considerando solo gli altri casi, la quota di lessemi le cui forme hanno omonimi risulta maggiore tanto più è ampio l'ambito d'uso, con uno scarto nettissimo tra la quantità di omonimie che coinvolgono il lessico di maggior frequenza rispetto a tutte le altre fasce. La stessa correlazione, inoltre, si manifesta all'interno del vocabolario di base, poiché la quantità di lessemi coinvolti in omonimie è massima nel vocabolario fondamentale (64%) e decresce progressivamente nel vocabolario di alto uso (54%) e in quello di alta disponibilità (49%).

Un altro segnale della relazione tra frequenza d'uso e sviluppo di omonimie proviene dall'analisi delle omonimie che coinvolgono più di due forme. Nella stragrande maggioranza dei casi l'omonimia in italiano si verifica tra due forme, mentre sono molto meno frequenti le omonimie di tre e più forme ([1]*brocca* 'caraffa' ~ [2]*brocca* 'bulletta' ~ [3]*brocca* vc. del verbo *broccare*; [1]*indi* avv. ~ [2]*indi* pl. di *indio* 'degli Indi' ~ [3]*indi* pl. di *indio* 'elemento chimico' ~ [4]*indi* pl. di *indo* 'indiano') e sono decisamente rare quelle tra sei e più forme: nell'intero database HOMO sono solo una decina i gruppi formati da otto omonimi, e solo in quattro casi il gruppo arriva a contarne nove. Tuttavia, come si vede dalla Tabella 3, la percentuale di gruppi composti da più di due forme risulta più alta tra quelli che coinvolgono il vocabolario di base rispetto al totale generale: nell'insieme di HOMO i gruppi composti da solo due omonimi sono il 79%, mentre nel sottoinsieme che coinvolge il vocabolario di base la percentuale scende al 65% e aumenta il numero di gruppi composti da tre o più forme. E anche se risulta comunque eccezionale che il gruppo arrivi a contare otto o nove omonimi, questi casi vedono sempre la presenza di uno o più

[2] I pochi esotisimi registrati in HOMO sono lessemi di altre lingue omografi di forme italiane, come l'inglese *file*, il tedesco *ossi*, il cinese *pipa*, il polacco *rada*; quasi sempre si tratta di omonimie parziali, perché le forme in questione sono omografe ma non omofone.

lessemi del vocabolario di base; così avviene, ad esempio, nel gruppo formato dalle otto forme *piano*[3] e in quello formato dalle nove forme *pari*.[4]

gruppi omonimici che includono una forma VDB	forme nel gruppo	forme coinvolte
7.578	2	15.156 (65%)
1.694	3	5.082 (22%)
452	4	1.808 (8%)
151	5	755 (3%)
46	6	276 (1%)
23	7	161 (0,7%)
7	8	56 (0,2%)
2	9	18 (0,1%)
9.953		23.312

gruppi omonimici in HOMO	forme nel gruppo	forme coinvolte
44.306	2	88.612 (79%)
5.475	3	16.425 (15%)
1.217	4	4.868 (4%)
294	5	1.470 (1%)
99	6	594 (0,5%)
37	7	259 (0,2%)
10	8	80 (0,1%)
4	9	36
51.442		112.344

Tabella 3
Omonimie che coinvolgono il VDB (a sinistra) rispetto al totale di HOMO (a destra).

Nel complesso i dati esaminati fin qui confermano l'esistenza di una relazione tra frequenza e omonimia, tuttavia non dicono se sia la maggior frequenza di per sé a causare sviluppo di omonimie o se ciò non dipenda da altre variabili e in particolare dalla lunghezza. Essendo l'omonimia più probabile se le forme sono brevi, e data la relazione tra frequenza e brevità, la variabile cruciale potrebbe essere la lunghezza a prescindere dalla frequenza.

3. Lunghezza delle parole e sviluppo di omonimie

La scelta di quale unità di misura usare per valutare la lunghezza delle parole – sillabe, lettere, fonemi o altro – è una delle questioni ancora aperte nella letteratura sull'argomento (v. Strauss *et al*. 2007) e le varie opzioni possono presentare pro e contro sia teorici che pratici (v. Grzybek 2015, pp. 90-93). Ad esempio utilizzare come unità di misura le lettere esclude qualunque confronto con lingue che hanno sistemi di scrittura non alfabetici e crea problemi per l'analisi degli omofoni non omografi (del tipo *anno* e *hanno* in italiano, *time* e *thyme* in inglese), che spesso differiscono per numero di lettere pur essendo identici nella pronuncia;[5] d'altro canto l'analisi in lettere è la più facile da eseguire in modo automatico, ed è infatti molto usata nei lavori di ambito computazionale fin dagli esordi (v. ad esempio Miller *et al*. 1958). Seguendo la prassi più diffusa, è utilizzata qui come unità di misura prevalente la sillaba; comunque i conteggi effettuati anche sul numero di lettere danno esiti sostanzialmente sovrapponibili.

[3] Il gruppo include il singolare dei sostantivi *piano* 'superficie piana' (FO), *piano* 'progetto' (FO), *piano* 'pianoforte' (CO), il singolare maschile degli aggettivi *piano* 'piatto' (FO) e *piano* 'relativo a un papa di nome Pio' (TS), le voci dei verbi *pianare* 'spianare' (TS), *piare* 'pigolare' (BU) e *piare* 'germogliare' (RE).

[4] Il gruppo include le voci dei verbi *parare* 'fermare' (AD), *parere* 'sembrare' (FO) e *pariare* 'divertirsi' (RE), la forma invariabile degli aggettivi *pari* 'uguale' (FO) e *pari* 'dei Pari' (TS), il plurale degli aggettivi *pare* 'uguale' (OB), *paro* 'uguale' (RE) e *pario* 'relativo all'isola di Paro' (CO), il plurale del sostantivo *paro* 'genere di uccello' (TS).

[5] O, per meglio dire, *quasi* identici, se è vero come emerge dallo studio di Gahl (2008) che nel parlato spontaneo il membro di maggior frequenza di una coppia di omofoni risulta più breve del suo corrispettivo di minor frequenza (ad esempio *time* è più breve di *thyme*).

Prendendo come riferimento il lemmario del Gradit, la lunghezza media dei lessemi italiani risulta essere di 4,3 sillabe e di circa 10 (9,96) lettere. È piuttosto difficile il confronto con i dati reperibili in altri lavori sia sull'italiano che su altre lingue, perché i conteggi variano molto a seconda che siano eseguiti su corpora o su dizionari e, in questo secondo caso, a seconda che si tratti di dizionari di piccole o di grandi dimensioni (quest'ultima variabile è molto significativa, poiché tanto più è ampio il lessico, tanto più lunghe sono in media le parole che esso contiene). Comunque il dato ricavato dal Gradit sul numero di lettere è identico a quello fornito da Smith (2012) e molto vicino a quello di Parikh (2015). Per fare un confronto con lingue in cui i lessemi sono rispettivamente più corti e più lunghi che in italiano, dall'analisi di corpora e dizionari la lunghezza media dei lessemi in inglese risulta essere di 7,6-8 lettere con un massimo di 9,6 nei dizionari maggiori, ricchi di termini tecnico-scientifici (v. Németh, Zainkó 2001; Henrich 2008; Smith 2012; Norvig 2013) mentre in tedesco oscilla tra 10,5 e 11,6 (Németh, Zainkó 2001; Parikh 2015).[6]

A conferma della relazione zipfiana tra frequenza e lunghezza, i lessemi del vocabolario di base sono più brevi della media generale e tanto più brevi nella fascia di maggior frequenza, cioè quella del vocabolario fondamentale. Come si vede dalla Tabella 4, infatti, la lunghezza media nel vocabolario di base è di 3,3 sillabe, che scendono a 3 nel vocabolario fondamentale. Esito analogo se si considerano le lettere: la lunghezza media nel vocabolario di base è di 7,7 lettere, che scendono a 7 nel vocabolario fondamentale e salgono a 8 nei lessemi di alto uso e a 8,2 in quelli di alta disponibilità.

n° sillabe	tutti i lessemi	VDB	FO	AU	AD
1	2.012 (0,8%)	112 (2%)	76 (4%)	17 (0,6%)	19 (1%)
2	16.323 (6%)	1.336 (20%)	567 (27%)	441 (17%)	328 (16%)
3	45.429 (18%)	2.493 (37%)	797 (38%)	986 (37%)	710 (36%)
4	84.897 (33%)	2.077 (31%)	532 (26%)	869 (33%)	676 (34%)
5	63.186 (25%)	611 (9%)	99 (5%)	299 (11%)	213 (11%)
6	28.121 (11%)	94 (1%)	6 (0,3%)	49 (2%)	39 (2%)
7	10.332 (4%)	5 (0,1%)		2 (0,1%)	3 (0,1%)
8	2.854 (1%)				
9	811 (0,3%)				
10	228 (0,1%)				
11	87				
12	24				
13	6				
14	1				
15	2*				
	254.313**	6.728	2.077	2.663	1.988
M sillabe	4,3	3,3	3	3,4	3,4

Tabella 4
Numero di sillabe dei lessemi registrati nel Gradit (*con trattino; **il totale non corrisponde al numero di lemmi del Gradit, 260.709, perché sono esclusi dal conteggio sia simboli/sigle privi di sillabazione sia le locuzioni).

[6] Che siano ricavati da corpora o da dizionari, questi dati riguardano la lunghezza dei *types*/lessemi, non delle loro occorrenze. Se si considerano invece le singole occorrenze nei corpora, la lunghezza media si abbassa drasticamente, perché le parole brevi hanno più occorrenze; ad esempio per l'inglese è stimata in 4,5 lettere da Németh e Zainkó (2001) e in 4,79 lettere da Norvig (2013) (il cui conteggio, effettuato sui circa 743 miliardi di occorrenze dei Google Books, esclude le parole con meno di 100.000 occorrenze).

In particolare mentre nel lessico generale esistono lessemi che arrivano a 14-15 sillabe e 31-32 lettere,[7] nel vocabolario di base nessun lessema supera le 7 sillabe (*indimenticabile*, *irriconoscibile*), che scendono a 6 nel vocabolario fondamentale (*comunicazione*, *internazionale*, *responsabilità*). Viceversa il numero di monosillabi e bisillabi nel vocabolario fondamentale è più del quadruplo che nel lessico generale, benché i conteggi relativi a quest'ultimo sovrastimino, e non di poco, la quota di monosillabi.[8]

La Tabella 5 mostra invece come si distribuiscono, sempre in base al numero di sillabe, i lessemi le cui forme sono coinvolte in omonimie, cioè quelli del Gradit che hanno almeno un omonimo e quelli registrati in HOMO (questi ultimi tutti caratterizzati dal fatto di avere almeno una forma che ha almeno un omonimo):

n° sillabe	lessemi Gradit con omonimi	lessemi HOMO	forme HOMO
1	448 (3%)	294 (1%)	290
2	3.207 (22%)	4.958 (14%)	14.465
3	5.170 (36%)	12.155 (34%)	40.018
4	4.296 (30%)	12.426 (35%)	38.433
5	1.084 (7%)	4.558 (13%)	15.781
6	197 (1%)	957 (3%)	2.890
7	36 (0,2%)	182 (0,5%)	431
8	/	8	20
9	4	10	16
	14.442	35.548*	112.344
M sillabe	3,2	3,5	3,5

Tabella 5
Numero di sillabe dei lessemi/forme che hanno almeno un omonimo (* il totale non corrisponde a quello indicato sopra, 35.557, perché sono esclusi dal conteggio sia simboli/sigle privi di sillabazione sia le locuzioni).

Che i lessemi inventariati in HOMO risultino mediamente più lunghi, sia pure di poco, rispetto a quelli del Gradit si spiega col fatto che i criteri di lemmatizzazione in HOMO penalizzano le forme più brevi; ad esempio non compaiono nel database gli affissi e altre forme tipicamente brevi come le sigle, e ne sono esclusi i simboli e le abbreviazioni quando siano omonimi di forme non pienamente lessicali quali altri simboli o sigle (non vi compare quindi, ad esempio, l'omonimia tra *cc* simbolo di centimetro cubico e *cc* sigla di *courtesy copy*, mentre compare quella tra *ala* simbolo di alanina e *ala* sostantivo). Inoltre HOMO registra anche le omonimie non nella forma di citazione, che coinvolgono forme di lessemi i quali considerando la sola forma di citazione non avrebbero omonimi e che sono per lo più tecnicismi lunghi; ad esempio nel lemmario di HOMO compaiono il sostantivo *magnetofluidodinamica* e l'aggettivo *magnetofluidodinamica* (il cui femm.

[7] Si tratta, prevedibilmente, di termini scientifici formati dal cumulo di vari confissi. I più lunghi risultano essere *celioisterosalpingo-ooforectomia* e *pentagonododecaedrico-tetraedrico*, registrati però dal Gradit con il trattino, seguiti da *colangiocolecistocoledocetomia* (14 sillabe, 31 lettere) e *pleuroepicheliognatouranoschisi* (13 sillabe, 31 lettere).

[8] Dei 2.012 monosillabi indicati nella Tabella 4 almeno la metà sono prestiti (*bang*, *chef*, *film*, *hot*, *mix*) e molti dei restanti sono affissi, interiezioni, fonosimboli o anche nomi propri messi a lemma nel Gradit perché compaiono all'interno di locuzioni polirematiche. Escludendo questi casi, i monosillabi ammontano a poche centinaia, a conferma della bassa monosillabicità dell'italiano; per fare un paragone, in inglese i monosillabi sono stimati intorno a 10.000 e fino a 14.000 se si includono quelli che ammettono anche una pronuncia bisillabica, v. Moser (1969).

sing. è omonimo del sostantivo), che mancano invece, ovviamente, tra gli omonimi del Gradit.

Sia nel Gradit che in HOMO, comunque, la lunghezza media dei lessemi coinvolti in omonimie risulta inferiore a quella generale (3,2-3,5 sillabe contro il 4,3 del lessico generale) e pressoché identica a quella del lessico di maggiore frequenza. Questo dato sembra indicare che lo sviluppo di omonimie correla con la lunghezza dei lessemi a prescindere dalla loro frequenza, ovvero che la maggiore brevità delle forme è di per sé un fattore che favorisce lo sviluppo di omonimie. Indipendentemente dalla fascia d'uso dei lessemi, cioè, quelli le cui forme hanno omonimi sono nettamente più corti; l'84% delle omonimie in HOMO e ben il 91% di quelle nel Gradit riguarda lessemi lunghi al massimo 4 sillabe. Ciò sembra suggerire che la lunghezza sia un fattore indipendente per lo sviluppo di omonimie.

Una conferma di questa ipotesi proviene dall'analisi dei gruppi omonimici formati da cinque o più forme. Considerando la lunghezza media delle forme (non lessemi) che rientrano nei gruppi omonimici più ampi, cioè quelli costituiti da quattro o più omonimi, si osserva che la percentuale di bisillabi aumenta in proporzione all'ampiezza del gruppo, per arrivare al 100% nei gruppi che coinvolgono nove forme; come mostra la Tabella 6, infatti, questi ultimi sono tutti costituiti da bisillabi. E questo del tutto a prescindere dalla frequenza d'uso delle forme interessate, molte delle quali infatti sono di bassa frequenza: ad esempio delle otto forme *radi* solo due appartengono al vocabolario di base (*radi* voce del vb. AD *radere* 'rasare' e *radi* masch. pl. dell'agg. AU *rado* 'raro'); delle nove forme *matta* una sola è FO (il femminile dell'agg./sost. *matto* 'pazzo') mentre le altre sono RE, BU, OB o TS; e nessuna delle nove forme *ronchi* appartiene al vocabolario di base o comune, sono tutte BU, RE, LE o TS.

n° forme nel gruppo	bisillabi
4	35%
5	54%
6	65%
7	57%
8	90%
9	100%

Tabella 6
Percentuale di bisillabi nei gruppi omonimici che comprendono quattro e più omonimi.

Che la lunghezza sia una variabile cruciale sembra emergere anche dall'analisi delle omonimie che coinvolgono forme di lessemi del vocabolario di base. Se si considerano, di tutti i lessemi del vocabolario di base, solo quelli le cui forme sono coinvolte in omonimie, si osserva che la loro lunghezza media è inferiore a quella del vocabolario di base generale. In quest'ultimo infatti il numero medio di sillabe è 3,3 (cfr. Tabella 4), mentre, come mostra la Tabella 7, la lunghezza media dei lessemi VDB coinvolti in omonimie è di 3-3,1 sillabe (a seconda che si consideri il Gradit o HOMO). La stessa differenza si osserva all'interno delle varie fasce del vocabolario di base: ad esempio la lunghezza media dei lessemi del vocabolario fondamentale risulta essere di 3 sillabe nel Gradit (v. Tabella 4), mentre scende a 2,5 sillabe in quelli, sempre di fascia FO, che hanno omonimi (v. Tabella 7) e aumenta a 3,3 sillabe in quelli non coinvolti in omonimie. A parità di frequenza, dunque, l'esistenza di omonimie correla con la maggiore brevità delle forme coinvolte.

n° sillabe	lessemi Gradit con omonimi				lessemi HOMO			
	VDB	FO	AU	AD	VDB	FO	AU	AD
1	54	36	5	13	58 (1,5)	38 (3)	7 (0,5)	13 (1)
2	370	148	131	91	955 (26)	412 (31)	314 (22)	229 (23)
3	362	126	140	96	1.513 (40)	532 (40)	581 (40)	400 (41)
4	167	45	62	60	1.021 (27)	302 (23)	433 (30)	286 (29)
5	17	2	10	5	176 (5)	36 (3)	96 (7)	44 (4)
6	6			6	9 (0,2)	1 (0,1)	3 (0,2)	5 (0,5)
M sillabe	3	2,5	3	3	3,1	3	3,2	3,1

Tabella 7
Numero di sillabe dei lessemi VDB che hanno almeno un omonimo.

Infine, un'ulteriore indicazione dell'importanza della lunghezza per lo sviluppo di omonimie viene dall'analisi delle omonimie nell'ambito dei termini tecnico-specialistici (TS). Di tutte le fasce d'uso quella del lessico specialistico è la meno 'omonimogena', poiché, come si è visto sopra (v. Tabella 2), solo il 10% dei lessemi TS presenti nel Gradit risulta coinvolto in omonimie. Analizzando la lunghezza dei lessemi in questa fascia d'uso, risulta che mentre lunghezza media dei lessemi TS nel loro insieme è di 4,5 sillabe, la lunghezza media dei lessemi TS coinvolti in omonimie è di 3,6 sillabe, dunque notevolmente inferiore. Di tutti i lessemi TS coinvolti in omonimie circa la metà sono al massimo trisillabi e l'80% sono al massimo quadrisillabi, e al sopra delle 7 sillabe i casi di omonimia sono del tutto sporadici (una decina in tutto, ad esempio [1]*telecinematografia* comp. da [1]*tele-* 'trasmissione televisiva di pellicole cinematografiche' ~ [2]*telecinematografia* comp. da [2]*tele-* 'ripresa cinematografica mediante teleobiettivi').

4. Osservazioni conclusive

Dall'analisi dei dati sull'omonimia nel lessico italiano ricavabili dal Gradit e dal repertorio di omonimi HOMO (v. Casadei 2016) emerge una forte relazione tra frequenza e sviluppo di omonimie. Infatti i lessemi di maggiore frequenza – quelli del vocabolario di base – risultano coinvolti in omonimie in misura notevolmente maggiore rispetto ai lessemi di tutte le altre fasce d'uso: la percentuale di lessemi le cui forme hanno omonimi è del 55% nel VDB, mentre nelle altre fasce d'uso non supera il 24% (v. Tabella 2). Inoltre la correlazione tra frequenza e omonimia si manifesta anche all'interno del vocabolario di base, poiché la percentuale di lessemi coinvolti in omonimie risulta massima nel vocabolario fondamentale (64%) e decresce progressivamente nel vocabolario di alto uso (54%) e in quello di alta disponibilità (49%).

Allo stesso tempo i dati confermano che la lunghezza delle forme è una variabile cruciale per lo sviluppo di omonimie. Risulta infatti che

1) nel lessico nel suo complesso, quindi considerando tutte le fasce d'uso comprese quelle di bassa frequenza, la lunghezza media dei lessemi coinvolti in omonimie è inferiore a quella generale (3,2-3,5 sillabe contro il 4,3 del lessico nel suo insieme, v. Tabella 5);
2) anche all'interno del lessico di alta frequenza, le forme coinvolte in omonimie sono mediamente più brevi di quelle che non hanno omonimi (v. Tabella 7).

Soprattutto, ciò che sembra emergere è che frequenza e lunghezza agiscono come due variabili indipendenti nel favorire l'omonimia. A parità di frequenza le parole più brevi hanno più omonimi (ad esempio tra i lessemi di massima frequenza, cioè di fascia FO, la

lunghezza media di quelli che hanno omonimi è di 2,5 sillabe mentre la lunghezza media di quelli che non ne hanno è di 3 sillabe). E a parità di lunghezza le parole più frequenti hanno più omonimi (ad esempio i lessemi di alto uso e quelli di alta disponibilità hanno identica lunghezza media di 3,4 sillabe ma tra i primi la quota di omonimie è del 54% mentre tra i secondi è del 49%).

In sostanza, lo sviluppo di omonimie risulta favorito non solo dalla maggior brevità delle forme, ma anche dalla loro maggior frequenza.

Questo risultato appare coerente con l'ipotesi di Piantadosi *et al*. (2015) accennata nel Paragrafo 1.2., secondo cui la ricchezza di omonimie nel lessico di alta frequenza delle lingue si spiegherebbe non solo con la minore lunghezza delle forme in questione (cioè con il fatto, del tutto accidentale, che queste trovano più facilmente degli omofoni e/o omografi), ma con un principio generale di organizzazione del codice linguistico. Data la capacità disambiguante del contesto, le lingue sfrutterebbero la possibilità di assegnare un maggior carico di ambiguità – sia per polisemia, sia per omonimia – alle forme lessicali le cui caratteristiche ne facilitano l'elaborazione, ovvero quelle più brevi e/o di maggior frequenza.

Certo, resta una profonda differenza di natura tra omonimia e polisemia, essendo la prima un fenomeno più 'accessorio' e la seconda invece il frutto di un meccanismo di estensione del significato intrinseco a tutte le lingue e legato a quella che è forse la loro proprietà semiotica più importante, cioè l'indeterminatezza semantica dei segni. E questa differenza è testimoniata anche dal diverso peso quantitativo dei due fenomeni nel lessico di alta frequenza, nel quale l'omonimia coinvolge, come si è visto, circa la metà dei lessemi, laddove la polisemia coinvolge pressoché la totalità delle forme (sono infatti polisemici l'89% dei lessemi del vocabolario di base, con un massimo del 96% nel vocabolario fondamentale, v. Casadei 2014). Tuttavia il fatto che anche nel caso dell'omonimia sembri verificarsi un 'effetto frequenza' suggerisce una diversa lettura di questo fenomeno, che non appare riducibile a una mera casualità o a un difetto il cui unico esito sia rendere meno efficiente il sistema linguistico.

Bionota: Federica Casadei è professore associato di Didattica delle lingue moderne all'Università della Tuscia (Viterbo). La sua area di interesse principale è la semantica lessicale, con particolare attenzione per il linguaggio figurato e la metafora, le espressioni idiomatiche, i lessici settoriali e scientifici. Su questi temi, oltre a numerosi articoli, ha pubblicato i volumi *Metafore ed espressioni idiomatiche* (Bulzoni 1996) e *Lessico e semantica* (Carocci 2003). Di recente si è occupata dei fenomeni di polisemia e omonimia nel lessico italiano, v. gli articoli *La polisemia nel vocabolario di base dell'italiano* (in "Lingue e Linguaggi" 12, 2014) e *L'omonimia nel lessico italiano* (in corso di stampa in "Studi di Lessicografia Italiana" 33, 2016).

Recapito autore: f.casadei@unitus.it

Riferimenti bibliografici

Alinei M. 1974, *Semantic density in linguistic geography*, in Weijnen A.A. and Alinei M. (eds.), *The wheel in the Atlas Linguarum Europae: heteronyms and semantic density*, North-Holland, Amsterdam, pp. 16-28.

Barlow M. and Kemmer S. (eds.) 2000, *Usage Based Models of Language*, University of Chicago Press, Chicago.

Bloomfield L. 1933, *Language*, Allen & Unwin, London; trad. it. di Antinucci F., Cardona G. 1974, *Il linguaggio*, Il Saggiatore, Milano.

Bybee J. 1985, *Morphology: A study on the relation between meaning and form*, Benjamins, Amsterdam.

Bybee J. 2001, *Phonology and Language Use*, Cambridge University Press, Cambridge.

Bybee J. 2007, *Frequency of Use and the Organization of Language*, Oxford University Press, Oxford.

Bybee J., Hopper P. (eds.) 2001, *Frequency and the Emergence of Linguistic Structure*, Benjamins, Amsterdam.

Casadei F. 2014, *La polisemia nel vocabolario di base dell'italiano*, in "Lingue e Linguaggi" 12, pp. 35-52.

Casadei F. 2016, *L'omonimia nel lessico italiano*, in "Studi di Lessicografia Italiana" 33, pp. 187-228.

Fellbaum C. (ed.) 1998, *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge.

Fenk-Oczlon G. and Fenk A. 2010a, *The association between word frequency and polysemy: a chicken and egg problem?*, in Solovyev V. and Polyakov V. (eds.), *Proceedings of the XIIth International Conference "Cognitive Modeling in Linguistics"*, Kazan State University Press, Kazan, pp. 167-170.

Fenk-Oczlon G. and Fenk A. 2010b, *Frequency effects on the emergence of polysemy and homophony*, in "International Journal of Information Technologies and Knowledge" 4 [2], pp. 103-109.

Gahl S. 2008, Time *and* thyme *are not homophones: the effect of lemma frequency on word durations in spontaneous speech*, in "Language" 84 [3], pp. 474-496.

Gilliéron J. 1921, *Pathologie et thérapeutique verbale*, Champion, Paris.

Gradit = *Grande Dizionario Italiano dell'Uso* ideato e diretto da Tullio De Mauro, 6 voll., UTET, Torino, 1999 (2a ed. 8 voll., *ivi*, 2007).

Greenberg J.H. 1966, *Language Universals: With Special Reference to Feature Hierarchies*, De Gruyter, Berlin/New York.

Gries S.Th. and Divjak D. (eds.) 2012a, *Frequency Effects in Language Learning and Processing*, De Gruyter, Berlin/New York.

Gries S.Th. and Divjak D. (eds.) 2012b, *Frequency Effects in Language Representation*, De Gruyter, Berlin/New York.

Grzybek P. 2015, *Word Length*, in Taylor J.R. (ed.), *The Oxford Handbook of the Word*, Oxford University Press, Oxford, pp. 89-119.

Henrick J. 2008, *On word-length and dictionary size*. http://www.thefreelibrary.com/On+word-length+and+dictionary+size.-a0189832222 (23.6.2016).

Jespersen O. 2010, *Monosyllabism in English*, in Jespersen O., *Selected Writings of Otto Jespersen*, Routledge, New York, pp. 325-341 (1ª ed. 1929, in *Proceedings of the British Academy*, vol. 14, Milford, London).

Ke J. 2006, *A cross-linguistic quantitative study of homophony*, in "Journal of Quantitative Linguistics" 13, pp. 129-159.

Köhler R. 1986, *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*, Universitätsverlag Brockmeyer, Bochum.

Köhler R. 1990, *Elemente der synergetischen Linguistik*, in Hammerl R. (Hrsg.), *Glottometrika 12*, Universitätsverlag Brockmeyer, Bochum, pp. 179-188.

Köhler R. 2005, *Synergetic Linguistics*, in Köhler R., Altmann G. and Piotrowski R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook* , De Gruyter, Berlin/New York, pp. 760-775.

Langacker R.W. 1987, *Foundations of Cognitive Grammar*, vol. I, *Theoretical Prerequisites*, Stanford University Press, Stanford.

Lyons J. 1968, *Introduction to theoretical linguistics*, Cambridge University Press, Cambridge; trad. it. Antinucci F., Mannucci E. 1971, *Introduzione alla linguistica teorica*, Laterza, Bari.

Miller G.A., Newman E.B. and Friedman E.A. 1958, *Length-Frequency Statistics for Written English*, in "Information and Control" 1, pp. 370-389.

Németh G. and Zainkó C. 2001, *Word unit based multilingual comparative analysis of text corpora*, in *INTERSPEECH 2001, 7th European Conference on Speech Communication and Technology*, pp. 2035-2038 (electronic edition http://www.isca-speech.org/archive/archive_papers/eurospeech_2001/e01_2035.pdf).

Newmeyer F.J. 1998, *Language Form and Language Function*, The MIT Press, Cambridge.
Newmeyer F.J. 2003, *Grammar is grammar and usage is usage*, in "Language" 79 [4], pp. 682-707.
Norvig P. 2013, *English Letter Frequency Counts: Mayzner Revisited*. http://norvig.com/mayzner.html (23.06.2016).
Parick R. 2015, *Distribution of Word Lengths in Various Languages*. http://www.ravi.io/language-word-lengths (23.06.2016).
Piantadosi S.T., Tily H. and Gibson E. 2011, *Word lengths are optimized for efficient communication*, in "Proceedings of the National Academy of Sciences" 108 [9], pp. 3526-3529.
Piantadosi S.T., Tily H. and Gibson E. 2015, *The communicative function of ambiguity in language*, in "Cognition" 122 [3], pp. 280-291.
Poddubbny V. and Polikarpov A.A. 2015, *Evolutionary derivation of laws for polysemic and age-polysemic distributions of language signs ensembles*, in Tuzzi A., Benešová M. and Macutek J. (eds.), *Recent Contributions to Quantitative Linguistics*, De Gruyter, Berlin/New York, pp. 115-124.
Polikarpov A.A. 1997, *Some factors and regularities of analytic/synthetic development of language systems*, paper presented at the XIII International Conference on Historical Linguistcs, 10-17 August 1997, Duesseldorf, Heinrich Heine Universitaet. http://www.philol.msu.ru/~lex/articles/fact_reg.htm (23.06.2016).
Polikarpov A.A. 1999, *Cognitive Model of Lexical System Evolution and its Verification*, in *HumLang. A Site on General Linguistics*, Laboratory for General and Computational Lexicology and Lexicography, Moscow Lomonosov State University. http://www.philol.msu.ru/~humlang/articles/h_cyc_n.htm (23.06.2016).
Smith R. 2012, *Distinct word length frequencies: distributions and symbol entropies*, in "Glottometrics" 23, pp. 7-22.
Strauss U., Grzybek P. and Altmann G. 2007, *Word Length and Word Frequency*, in Grzybek P. (ed.), *Contributions to the Science of Text and Language: Word Length Studies and Other Issues*, Springer, Dordrecht, pp. 277-294.
Tomasello M. 2003, *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Harvard University Press, Cambridge (MA).
Ullmann S. 1966, *Semantic universals*, in Greenberg J.H. (ed.), *Universals of language*, The MIT Press, Cambridge, pp. 172-207.
Wasow Th., Perfors A. and Beaver D. 2003, *The Puzzle of ambiguity*, in Orgun O. and Sells P. (eds.), *Morphology and The Web of Grammar: Essays in Memory of Steven G. Lapointe*, CSLI Publications, Stanford University, Stanford.
Wasow Th. 2015, *Ambiguity avoidance is overrated*, in Winkler S. (ed.,) *Ambiguity: Language and Communication*, De Gruyter, Berlin/New York, pp. 29-47.
Zipf G.K. 1936, *The Psycho-Biology of Language*, Routledge & Sons, London.
Zipf G.K. 1945, *The meaning-frequency relationship of words*, in "Journal of General Psychology" 33, pp. 251-256.
Zipf G.K. 1949, *Human behaviour and the principle of least effort. An introduction to human ecology*, Addison-Wesley Press, Cambridge.