Pavel ŠTICHAUER
Università Carlo di Praga

APPROCCIO QUANTITATIVO ALLA PRODUTTIVITÀ MORFOLOGICA : ALCUNI SVILUPPI RECENTI

In questo articolo mi propongo di offrire un riassunto di alcuni sviluppi recenti della ricerca sulla nozione di produttività morfologica.[1] Mi limiterò a presentare uno specifico approccio quantitativo : quello sviluppato a partire dai lavori di Baayen (BAAYEN, 1992 ; 2001 ; 2008 ; 2009). Per farlo, mi avvarrò dei dati elaborati soprattutto da Davide Ricca (RICCA, 2005 ; 2008) e da Marco Baroni e Stefan Evert (BARONI, EVERT, 2006 ; EVERT, BARONI, 2006b ; 2007), e mi servirò del programma *zipfR* (http://zipfr.r-forge.r-project.org/), ideato dagli ultimi due studiosi. Si tratta di un *package* che si utilizza all'interno del noto software statistico *R* (http://www.r-project.org/).

1. Introduzione

La nozione di produttività rappresenta senz'altro uno di quei termini che fanno parte di ciò che Dal ha giustamente chiamato *bagaglio linguistico condiviso* (« bagage linguistique partagé », cfr. DAL, 2003 : 3), che fa sì che non ci sia bisogno, a volte, di definire la nozione con esattezza. Comunemente si ripropone la famosa formulazione di Schultink che definisce la produttività in termini di non-intenzionalità nella formazione di nuove parole, e in termini del numero potenzialmente infinito di tali parole derivate. È merito di Plag (PLAG, 1999 : 13-16) aver dimostrato l'inadeguatezza della definizione di Schultink e di aver richiamato l'attenzione a una nozione più interessante che è quella di Corbin (CORBIN, 1987 ; cfr. anche PLAG, 2006).

Corbin, infatti, osserva come la tradizionale nozione di produttività racchiuda in sé tre fenomeni diversi che vanno trattati separatamente : « ... la productivité désigne en fait à la fois la régularité des produits de la règle, la disponibilité de l'affixe, c'est-à-dire précisément la possibilité de construire des dérivés non attestés, de combler les lacunes du lexique attesté, et la rentabilité, c'est-à-dire la possibilité de s'appliquer à un grand nombre de bases et/ou de produire un grand nombre de dérivés attestés. » (CORBIN, 1987 : 177).

La disponibilità di un procedimento morfologico può dunque definirsi come semplice esistenza del dato procedimento all'interno di un sistema linguistico. Bauer, riprendendo il termine inglese *availability* (introdotto già da CARSTAIRS-MCCARTHY, 1992 : 37), aggiunge giustamente che « availability is a yes/no

[1] Il testo si basa, in gran parte, sul capitolo 2.4. della mia monografia *La produttività morfologica in diacronia : i suffissi* -mento, -zione *e* -gione *in italiano antico dal Duecento al Cinquecento*, Praha, Karolinum (in corso di stampa). L'articolo fa parte del progetto di ricerca n. 405/06/P009 finanziato dalla *Czech Science Foundation* (GAČR). Desidero ringraziare Davide Ricca per aver gentilmente messo a mia disposizione alcuni dei suoi dati sui composti verbonominali e per aver riletto una versione precedente di questo testo. Rivolgo un particolare ringraziamento a Marco Baroni per avermi avviato al programma *zipfR* e per aver discusso con me alcuni punti importanti. Infine, tengo a ringraziare Paolo Divizia che ha riletto una versione quasi definitiva dell'articolo e ha maggiormente contribuito a migliorarne la lingua.

question : either a morphological process is available or it is not. » (BAUER, 2001 : 205). Invece, la nozione di *rentabilité* (nella terminologia di Bauer *profitability*), cioè la « redditività » coglie un importante aspetto quantitativo dei processi della formazione di parole. Un procedimento formativo disponibile può, infatti, essere più o meno « redditizio » nella misura in cui porta alla formazione di un certo numero di parole[2].

Secondo tale concezione, ormai accettata, la discussione sulla produttività verte da un lato sugli aspetti qualitativi, associati alla disponibilità dei processi, dall'altro sugli aspetti quantitativi associati alla redditività dei processi stessi (cfr. DAL, 2003 ; PLAG, 2006).

2. La nozione di disponibilità : un approccio qualitativo

Il suffisso *-bile*, in formazioni quali, ad esempio, *consultabile*, *riparabile*, ecc. può essere considerato un procedimento disponibile e anche altamente redditizio. Infatti, attenendoci ai dati tratti dal *DISC*, abbiamo all'incirca 1000 aggettivi (tra cui anche quelli che non possono naturalmente dirsi parole complesse costruite – nel senso di CORBIN, 1987 – come ad es. *mirabile*), la cui distribuzione diacronica ci fa capire che siamo di fronte a un suffisso che è sempre stato produttivo nella storia della lingua italiana (128 aggettivi risalgono agli anni 1950-1974, ossia il 12,7% del totale). Per il suffisso *evole*, invece, i dati lessicografici ci dicono che l'apice della sua produttività risale al quattordicesimo secolo ; inoltre, per gli ultimi cinquant'anni, il *DISC* registra solo quattro formazioni. Il suffisso è quindi da considerarsi ormai indisponibile nell'italiano contemporaneo (cfr. ŠTICHAUER, 2007 : 25). Per quel che riguarda, invece, i composti verbonominali, la loro produttività – nel senso di disponibilità – sembra essere costante : questi composti, infatti, sono sempre stati sistematicamente possibili, ma solo negli ultimi decenni la loro redditività, per diverse ragioni (per lo più extralinguistiche), ha subito un incremento, come si ricava da una breve consultazione della distribuzione diacronica di tali composti (cfr. ŠTICHAUER, 2007 : 109-113).

Da un punto di vista qualitativo, che prescinda dalle considerazioni quantitative, il suffisso *-bile* e la composizione verbonominale sono quindi due processi disponibili nella formazione delle parole in italiano, che possono essere descritti in rapporto alle restrizioni che ne delimitano il dominio di applicazione. Inoltre, esistono pure le regole di formazione di parole (d'ora in poi RFP) realizzate da più di un mezzo morfologico, come ad esempio i suffissi *-ità/-ezza* per i nomi di qualità (ad esempio *riparabilità / confortevolezza*) o *-zione/-mento* (ad esempio *banalizzazione / impallidimento*) per i nomi di azione che si differenziano tra di loro per le diverse restrizioni che impongono alle loro basi. In tal modo, « la produttività di una RFP non è (...) identificabile con la frequenza con cui essa si applica né, di conseguenza, con il numero di parole che essa forma : è necessario invece tener conto delle restrizioni (morfologiche, fonologiche, ecc.). » (SCALISE, 1994 : 106).

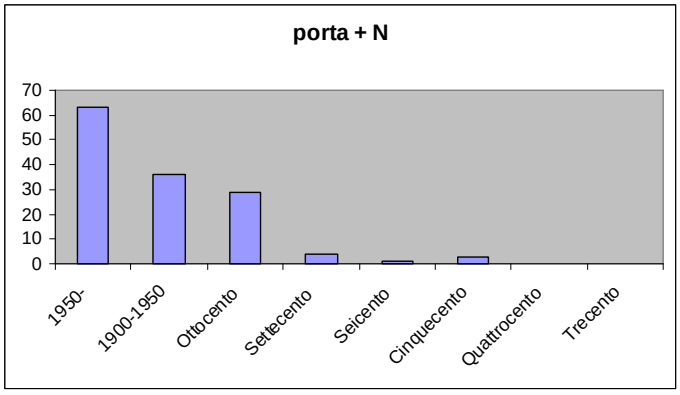
[2] Mi attengo all'interpretazione di Bauer discostandomi, dunque, da GAETA, RICCA (2002 : 231 e n. 5), ove dicono che tale aspetto quantitativo sia appunto prossimo alla *disponibilité* corbiniana ; a mio avviso si tratta piuttosto di *redditività*.

Purtuttavia, l'aspetto quantitativo, che riguarda sia la frequenza che il numero di parole formate, è un fattore importante che negli ultimi anni, soprattutto grazie al rapido sviluppo di diversi mezzi elettronici (dizionari e corpora), è diventato senz'alcun dubbio centrale nella trattazione della produttività.

3. La nozione di redditività : un approccio quantitativo

Per un approccio quantitativo si offrono principalmente due possibilità di misurare la redditività, una basata sui dizionari (*dictionary-based*), l'altra su corpora elettronici (*corpus-based*). Se riprendiamo l'esempio dei composti verbonominali visto sopra, si possono fare una serie di osservazioni. Basandoci sul *DISC*, possiamo estrarre ad esempio tutti i composti con struttura *porta + NOME* (del tipo *portalettere* che, tra i composti verbonominali, è la struttura più frequente ; cfr. RICCA, 2005 : 471-2 ; 2008 ; RADIMSKÝ, 2006 : 100). Si otterranno 139 composti (indipendentemente dal fatto che si tratti di sostantivi o di aggettivi). La loro distribuzione diacronica indica che la maggior parte dei composti, ossia 99 formazioni, risalgono al Novecento (63 formazioni a partire dal 1950, 36 composti tra gli anni 1900-1950), mentre solo 29 datano all'Ottocento, 4 formazioni appartengono al Settecento, uno solo risale al Seicento, e infine 3 formazioni sono fatte risalire, dallo stesso DISC, al Cinquecento. Riassumo graficamente i dati nella figura 1 :

Figura 1. Composti *porta + NOME* e loro attestazioni secondo il *DISC*



Il quadro che si ottiene in questo modo si basa sulla nozione di redditività intesa come *type frequency*, cioè come frequenza di lemma (cfr. BAUER, 2001 : 47). Il processo produttivo (= redditizio) può essere definito in base al numero di parole nuove create in un dato periodo : « The more productive a morphological process is, the more [are the] coinages that occur created by that morphological process in a given time period. (...) the productivity of the process can be equated with the average number of new words (...) that are used in the language within a specified time period. » (BAUER, 2001 : 156).

Come è risaputo, tale indagine, limitata a quanto registrato dai dizionari e applicata a processi intuitivamente redditizi (com'è appunto la composizione verbonominale), perde un importante aspetto della produttività, cioè la potenzialità

o, meglio, la *probabilità* di trovare o di creare una parola completamente nuova, ancora non registrata dai dizionari, eppure del tutto trasparente e correttamente formata dal punto di vista delle regole di formazione di parole (sulle limitazioni e sui vantaggi della ricerca basata sui dizionari cfr. BAUER, 2001 : 156-161 ; GAETA, RICCA, 2003 : 63-65).

L'approccio basato su corpora elettronici, senza i quali la ricerca linguistica è ormai impensabile, richiede l'introduzione di una seconda dimensione che è quella associata alla nozione di *token frequency*. Infatti, in un corpus una data parola (definita come *lemma* ovvero *tipo*, « type ») può comparire varie volte e la frequenza con cui occorre (« tokens ») diventa un fattore importante per due motivi.

Il primo, di ordine psicolinguistico, è che le parole ad alta frequenza sono tendenzialmente parole lessicalizzate, immagazzinate nel lessico mentale, molto spesso anche poco trasparenti che devono essere apprese come tali senza ricorrere a una regola di formazione di parole. Invece, le parole a bassa frequenza sono di solito morfologicamente e semanticamente trasparenti e possono, di conseguenza, essere analizzate per mezzo di una regola di formazione di parole, in base all'istruzione semantica della regola stessa (cfr. BAAYEN, 1992).

Il secondo motivo è di ordine matematico : il numero di occorrenze è, ovviamente, dato dalla dimensione del corpus in questione: con la progressiva crescita della dimensione del corpus, il numero di occorrenze tenderà necessariamente ad aumentare (cfr. BAAYEN, 1992 : 113).

In un'indagine basata su un corpus, dobbiamo dunque tener conto di almeno tre variabili : il numero di lemmi / tipi (*type frequency*, d'ora in poi *V*) e il loro numero di occorrenze (*token frequency*, d'ora in poi *N*) che, se calcolati per un dato affisso ad esempio, corrispondono a quello che BAAYEN (1992 : 113) chiama *item sample* ; e la dimensione del corpus che rappresenta il campione di base (*frame sample*, d'ora in poi *F*). Il rapporto tra queste variabili è il seguente : « ... the values of *N* and *V*, as calculated from the item sample, depend on the size of the frame sample. For larger frame samples, larger values of *N* and *V* are to be expected for the item sample. » (BAAYEN, 1992 : 113).

A titolo d'esempio, possiamo riprendere il caso dei composti verbonominali con struttura *porta + NOME*. I dati, elaborati da Davide Ricca[3] (cfr. RICCA, 2008), sono riassunti nella tabella :

Tabella 1. Composti *porta + NOME* presenti nel corpus *La Repubblica* (secondo RICCA, 2008)

composti verbonominali	lemmi (*V*)	occorrenze (*N*)	la dimensione del corpus (*F*)
porta + NOME	214	46289	~ 330 milioni di occorrenze[4]

Questa tabella riassuntiva non è molto informativa, tranne che per il fatto che il numero dei lemmi risulta più alto rispetto a quanto registrato da un dizionario[5], il

[3] Ringrazio Davide Ricca per aver messo gentilmente a mia disposizione la sua tabella definitiva dei composti *porta + N* con le relative frequenze. (I composti VN in cui N compare sia al singolare che al plurale sono sempre stati considerati come un unico lemma, anche quando tutte e due le varianti compaiono nel composto al singolare.)

[4] Di solito, viene riportata la dimensione 380 mil. di occorrenze, ma come fa notare Davide Ricca (cfr. RICCA, 2008), tale dato include anche la punteggiatura che è ovviamente irrilevante.

che è una situazione normale quando si ha a che fare con un processo produttivo. Tuttavia, una volta ottenuta in questo modo la lista di frequenze, si può procedere a creare tre strutture molto importanti per una valutazione quantitativa della produttività. Queste strutture, di cui ci occuperemo nei prossimi sottocapitoli, sono, rispettivamente, lo *spettro di frequenza*, la *curva di accrescimento del lessico* e il *ritmo di accrescimento del lessico*. Per creare questi oggetti ci serviremo del programma *zipfR*, sviluppato da Stefan Evert e Marco Baroni (cfr. EVERT, BARONI, 2006b ; 2007 ; BARONI, EVERT, 2006). Si tratta di un *package* utilizzabile all'interno del noto programma statistico *R* (cfr. il tutorial dello *zipfR* disponibile all'indirizzo web : http://zipfr.r-forge.r-project.org/ ; per un'introduzione generale alla statistica lessicale in *R*, cfr. BAAYEN, 2008).

3.1. Lo spettro di frequenza

Lo spettro di frequenza (*frequency spectrum*) è una lista in cui i lemmi sono ordinati a seconda della loro frequenza, o più precisamente, a seconda del *rango di frequenza* (d'ora in poi *m*) assegnato a un dato lemma (cfr. BARONI, 2009 ; BAAYEN, 2008 : 222-236). Tutti i lemmi con la frequenza 1 (cioè i cosiddetti *hapax legomena* la cui importanza verrà discussa più avanti) verranno assegnati al rango 1 (*m = 1*, ovvero V_1), tutti i lemmi che occorrono nel corpus due volte apparterranno al rango 2 (*m = 2*, ovvero V_2), e così via di seguito fino all'ultimo rango rappresentato dalla parola caratterizzata dalla più alta frequenza. Lo spettro di frequenza è particolarmente utile perché ci permette di vedere il numero di lemmi come funzione del loro rango *V(m)*. Nel caso dei composti *porta + NOME*, qui in discussione, si ottiene il quadro riportato nella tabella 2 che, oltre ai valori *V* e *N*, contiene anche quelli che corrispondono ai primi ranghi di frequenza :

Tabella 2. I composti *porta + NOME* presenti nel corpus *La Repubblica* : distribuzione delle loro frequenze

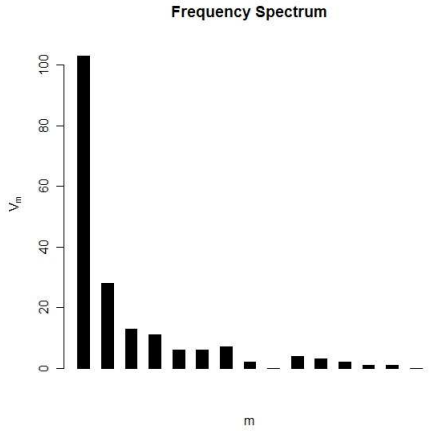
composti verbonominali	*V*	*N*	V_1 *(m = 1)*	V_2 *(m = 2)*	V_3 *(m = 3)*	V_n...	V_{27576}
porta + NOME	214	46289	103	28	13	...	1

Come si evince dalla tabella, nel campione dei composti *porta + NOME* ci sono 214 lemmi dei quali 103 occorrono nel corpus precisamente una sola volta ; ci sono poi 28 lemmi la cui frequenza equivale a 2 ; 13 lemmi con la frequenza 3 ; e infine la parola con più alta frequenza (*portavoce*) si trova nel testo ben 27576 volte. Per vedere meglio la distribuzione delle frequenze, si può creare un istrogramma (nella figura 2) in cui sull'asse *x* abbiamo il rango *m* e sull'asse *y* il numero di lemmi come funzione del rango *V(m)*. Come si vede chiaramente, di gran lunga più numeroso è il gruppo di *hapax legomena*. In generale, si può anche affermare che le parole a bassa frequenza tendono ad essere una netta maggioranza[6]:

[5] Tuttavia, come osserva Davide Ricca, un dizionario di dimensione maggiore come GRADIT riporta un maggior numero di composti VN (per l'esattezza 242 ; cfr. RICCA, 2005 : 471) che però non sono gli stessi che si ricavano dal corpus.

[6] È ben noto che tale distribuzione è consueta indipendentemente dal tipo e dalla dimensione del corpus e anche dal tipo di lemmi in questione, siano essi parole semplici o complesse (cfr. BARONI, 2007 ; BARONI, 2009).

Figura 2. Spettro di frequenza dei composti *porta + NOME*



Per verificare l'ipotesi sulle parole ad alta frequenza rispetto a quelle a bassa frequenza, si veda anche la tabella 3, dove vengono riportati gli esempi concreti dei composti *porta + NOME* (le parole più frequenti sono nella colonna sinistra, qualche esempio di quelle a bassa frequenza si trova nella colonna a destra) :

Tabella 3. Esempi dei composti *porta + NOME* più frequenti e quelli meno frequenti (secondo RICCA, 2008)

Frequenza	Lemma	Frequenza	Lemma
27576	portavoce	4	portaritratti
12488	portafogli(o)	4	portagioielli
1725	portaerei	3	portacomputer
1127	portaborse/a	3	portacravatte
503	portabandiera/e	2	portacassette
367	portabagagli(o)	2	portafoto
342	portafortuna	1	portabiglietti
252	portavalori	1	portabollo
230	portacenere	1	portacaffè
207	portalettere	1	portacoltelli
70	portapacchi	1	portagettoni
48	portacolori	1	portaguanti

Come prevedibile, i composti più frequenti sono quelli lessicalizzati, come ad es. *portavoce*, *portabagagli*, *portacenere* e *portalettere*. Invece, tra i composti poco frequenti troviamo anche neologismi o formazioni recenti e (spesso) effimere.

L'idea, appunto, che i neologismi debbano essere cercati tra le formazioni a bassa frequenza, verrà discussa più avanti ; prima, però, introdurremo ancora la seconda struttura che è quella della curva di accrescimento del lessico.

3.2. La curva di accrescimento del lessico

Come abbiamo già avuto modo di vedere, il numero di occorrenze è dato dalla dimensione del corpus ; così il numero di lemmi dipende indirettamente dal numero di occorrenze, come dice BAAYEN (1992 : 113): « ... for some fixed

morphological process, *V* can be viewed as a function of *N* : for increasing numbers of tokens in the item sample, obtained by increasing the frame sample, *V* will also increase. » Si può rappresentare questa funzione per mezzo di un grafico in cui sull'asse *x* si colloca il numero di occorrenze *N*, e sull'asse *y* il numero di lemmi *V*. La curva che così viene creata si chiama curva di accrescimento del lessico (*vocabulary growth curve* ; cfr. BAAYEN, 1992 : 113 ; 2008 : 222 ; BARONI, EVERT, 2006).

Si possono avere tre tipi di curva. La prima, empirica, si basa sul progressivo spoglio in cui a ogni tappa, rappresentata da un certo numero di occorrenze *N*, corrisponde un dato numero di lemmi *V*. All'interno del package *zipfR*, rappresentiamo, come esempio, la curva empirica del prefisso *ri-*. I dati, estratti dal corpus *La Repubblica* ed elaborati da Marco Baroni (cfr. BARONI, 2007), sono riassunti nella tabella 4 :

Tabella 4. I verbi prefissati con *ri-* presenti nel corpus *La Repubblica* (secondo BARONI, 2007 ; BARONI, EVERT 2006)

RFP	*V*	*N*	V_1 *(m = 1)*	V_2 *(m = 2)*	V_3 *(m = 3)*	V_4 *(m = 4)*	V_n *...*
ri+VERBO	1098	1399898	346	105	74	43	...

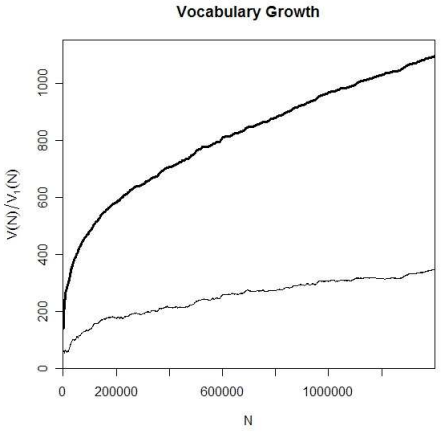
La crescita dei lemmi *V*, come funzione del progressivo incremento di *N*, si evince dalla tabella 5, in cui ai singoli campioni di *N* (*N* = 200.000, 400.000, 600.000, ecc.) corrisponde il numero di lemmi *V* e il numero di *hapax legomena* V_1.

Tabella 5. Numero di occorrenze *N*, di lemmi *V* e di *hapax legomena* V_1 dei verbi prefissati con *ri-* (secondo BARONI, 2007 ; BARONI, EVERT, 2006)

N	*V*	V_1
200000	583	176
400000	707	213
600000	810	259
800000	881	274
1000000	969	306
1200000	1029	314

La curva di accrescimento corrispondente si vede nel seguente grafico, in figura 3, in cui si riporta l'incremento sia dei lemmi *V* che quello degli *hapax legomena* V_1 (cfr. BARONI, EVERT, 2006). La prima curva, in grassetto, riporta il numero di lemmi V, la seconda, più sottile sotto, il numero di *hapax legomena* V_1.

Figura 3. Curve di accrescimento empiriche dei verbi prefissati con *ri-* (secondo BARONI, EVERT, 2006 : sezione 3.2)



Come osservano BARONI, EVERT (2006), l'andamento di tali curve « ... is typically not very smooth, as it reflects all the quirks due to the non-random distribution of words and texts in a corpus. » Si possono invece ottenere delle curve con un andamento più regolare mediante la tecnica della interpolazione ed estrapolazione binomiale (cfr. BARONI, EVERT, 2006 ; BAAYEN, 2001 : cap. 2.6).[7]

L'interpolazione binomiale permette, dato un certo spettro di frequenza, di calcolare i valori stimati (*expected values*) di *V* e di *V(m)*, che vengono rappresentati come $E(V)$, $E(V_m)$, per diversi valori di *N* fino al valore reale del campione. In tal modo, si può ottenere la seguente tabella 6 con i valori stimati per gli *N* che corrispondono a quelli indicati già sopra.

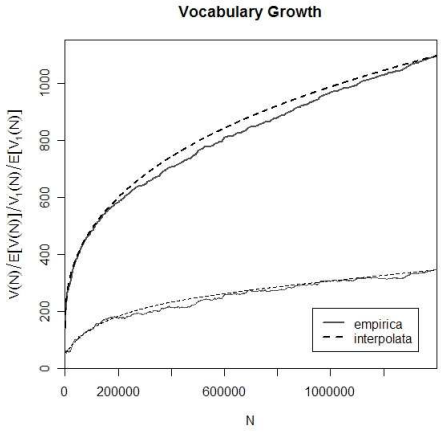
Tabella 6. Valori di *V* e V_1 stimati mediante l'interpolazione binomiale (BAAYEN, 2001)

N	$E(V)$	$E(V_1)$
200000	599,6121	182,7075
400000	743,247	232,2076
600000	843,4894	262,5607
800000	922,3546	286,1815
1000000	988,4883	307,0424
1200000	1046,216	326,6862

Per valutare il grado di corrispondenza tra la curva empirica e quella interpolata, si può ricorrere al confronto, illustrato nella figura 4, delle due curve.

[7] Per i dettagli matematici cfr. BAAYEN, 2001 : 64-69 (interpolazione), 69-76 (estrapolazione). Si veda anche la documentazione allo *zipfR* (s.v. *Binomial Interpolation*) e i riferimenti ivi citati.

Figura 4. Curve di accrescimento : confronto tra curva empirica e curva interpolata (secondo BARONI, EVERT, 2006 : sezione 3.3)



L'ultimo tipo di curva è quello estrapolato. L'estrapolazione binomiale consente di oltrepassare il numero reale di *N* dato dal campione. L'interesse di tale procedimento sta nel fatto che avendo un campione limitato non si possono azzardare delle ipotesi sul totale della popolazione, ovvero sul numero possibile di tutti i lemmi a cui una data regola può dar luogo (cfr. BARONI, EVERT, 2006). Inoltre, avendo un campione limitato e, in molti casi, di dimensioni diverse per ogni processo morfologico, non si può procedere al mutuo confronto dei processi, perché ovviamente, il confronto non sarebbe completo dal momento che le due curve non sarebbero quantitativamente paragonabili.

Per poter procedere all'estrapolazione, si deve ricorrere ai modelli statistici elaborati per la distribuzione delle frequenze. La distribuzione delle frequenze in un corpus esibisce un *pattern* tipico (cfr. BARONI, 2009) che permette di farla rientrare, dal punto di vista della tipologia delle distribuzioni, nel gruppo di ciò che si chiama *Large-Number-of-Rare-Events* (cfr. BAAYEN, 2001 : cap. 4, BARONI, EVERT, 2006). Attualmente, il programma *zipfR* offre tre modelli (*Generalized Inverse Gauss Poisson*, GIGP ; *Zipf-Mandelbrot*, ZM ; *finite Zipf-Mandelbrot*, fZM), le cui caratteristiche matematiche, sulle quali siamo costretti a sorvolare, sono descritte in BAAYEN, 2001 (GIGP) e in EVERT, 2004 (ZM e fZM).

3.3. Il ritmo di accrescimento del lessico

L'ultimo tipo di struttura che permette di cogliere la redditività è il famoso *P* che, come indice della produttività, è diventato, fin dal primo lavoro di BAAYEN, 1992, il mezzo più utilizzato. Tuttavia, tale indice, senza la descrizione delle due strutture che abbiamo introdotto sopra, rimane una formula priva di senso. Infatti, il valore *P*, calcolato nel modo seguente $P = V_1 / N$ (cioè il numero dei lemmi con la frequenza 1, ossia gli *hapax legomena*, diviso per il totale di occorrenze del tipo morfologico in questione), indica il ritmo con cui il lessico cresce (*vocabulary growth rate*, cfr. BAAYEN, 1992 : 115 ; 2008 : 222). Tale ritmo viene inteso come la probabilità di trovare un lemma nuovo in relazione al progressivo aumento della dimensione del corpus : « ... it expresses in a very real sense the probability that

new types will be encountered when the item sample is increased. » (BAAYEN, 1992 : 115)[8].

In tal modo, possiamo aggiungere ai dati già presentati sopra il valore *P* che indicherà il ritmo con cui il lessico, inteso in senso stretto come costituito dai lemmi formati da una data RFP, cresce. I valori di *P* per i composti verbonominali *porta + NOME* e per i verbi prefissati con *ri-* si riportano nella tabella 7.

Tabella 7. Valori di *V, N, V_1* e corrispondente indice di produttività *P* (BAAYEN, 1992 : 115) per i verbi in *ri-* e i composti *porta + NOME*

RFP	*V*	*N*	V_1	*P* (V_1 / *N*)
ri + VERBO	1098	1399898	346	0,00024
porta + NOME	214	46289	103	0,00222

I due procedimenti, in tale prospettiva (e a prescindere dallo scarso interesse di paragonare il prefisso *ri-* e la composizione verbonominale) si differenziano per il diverso ritmo di crescita. Ciò vuol dire semplicemente, in vista della definizione di cui sopra, che se la dimensione del corpus aumenterà, la probabilità di trovare un nuovo composto *porta + NOME* sarà più alta di quanto non lo sia per un nuovo verbo prefissato con *ri-*. Naturalmente, tale confronto, per stabilire quale dei due è più redditizio, come già accennato, avrà maggior senso se compiuto tra due mezzi morfologici (per esempio, suffissi) che appartengono alla stessa RFP, e quindi sono teoricamente in concorrenza (com'è, tanto per dare un esempio, la coppia *-zione / -mento*). Inoltre, tale confronto, come si vedrà nella sezione seguente, sarà corretto solo a parità di *N*, cioè con il valore di N identico nei due procedimenti.

La caratterizzazione quantitativa della produttività come formulata finora suscita diversi punti di discussione (alcuni dei quali già noti) che cercheremo di riassumere nella sezione seguente.

4. Problemi e soluzioni

Per poter adottare un approccio quantitativo di questo tipo, bisogna risolvere alcuni problemi di ordine metodologico. Dobbiamo infatti affrontare due problemi. Il primo riguarda lo statuto degli *hapax legomena* e il loro rapporto con i neologismi. Il secondo riguarda la dimensione del corpus e la dimensione del campione *N* che ha un peso molto importante nel computo della produttività (per una rassegna di questi problemi, cfr. PLAG, 1999 ; DAL, 2003 ; GAETA, RICCA, 2002 ; 2003 ; 2006).

4.1. *Hapax legomenon* = neologismo?

Lo statuto di *hapax legomenon* è doppio. Da una parte, dal punto di vista linguistico (tradizionale), ci si deve chiedere se le parole con frequenza 1 siano davvero delle parole completamente nuove che testimonino della produttività di una data RFP. L'equivalenza tra *hapax* e neologismo è stata messa in discussione varie volte.

[8] Cfr. anche BAAYEN, 2008 : 222: « The growth rate is a probability, the probability that, after having read *N* tokens, the next token sampled represents an unseen type, a word that did not occur among the preceding *N* tokens. »

Innanzitutto, si può facilmente dimostrare che alcuni degli *hapax* hanno tale frequenza soltanto all'interno del corpus e che, al di fuori dello stesso, se ne possono trovare molte occorrenze (cfr. DAL, 2003 : 17-18 ; PLAG, 1999 : 26-28 ; BAUER, 2001 : 152-153). Tuttavia, è certamente tra gli *hapax* che i neologismi vanno cercati, per quanto la definizione di neologismo sia vaga e sfuggente (cfr. PLAG, 2006). Una delle soluzioni pratiche, proposta da GAETA, RICCA (2002 : 227-229 ; cfr. anche BAAYEN, RENOUF, 1996 ; PLAG, 1999 : 26-27), consiste in un confronto tra dati lessicografici e dati basati su corpora : « ... possiamo infatti stipulare di considerare neologismi le parole non riportate in un dizionario molto ricco ed aggiornato, e per di più uscito posteriormente rispetto al materiale raccolto nel (...) corpus. » (GAETA, RICCA, 2002 : 227). Gaeta e Ricca, in seguito a tale scelta metodologica, mostrano molto bene come i neologismi così definiti siano presenti tra gli *hapax legomena* e come, invece, il loro numero decresca con l'aumento progressivo della frequenza (cfr. GAETA, RICCA, 2002 : 228).

Dal punto di vista psicolinguistico, invece, lo statuto di *hapax* è molto importante perché, come abbiamo già accennato, le parole a bassa frequenza richiedono l'attuazione di un accesso lessicale basato sulle regole. Come osserva PLAG (2006 : 542), « [such] words are crucial for the determination of the productivity of a morphological process because in very large corpora *hapax*es tend to be words that are unlikely to be familiar to the hearer or reader. Complex unknown words can be understood at least in those cases where an available word-formation rule allows the decomposition of the newly encountered word into its constituent morphemes and thus the computation of the meaning on the basis of the meaning of the parts. The word-formation rule in the mental lexicon guarantees that even complex words with extremely low frequency can be understood. Thus, with regard to productive processes, we expect large numbers of low frequency words and small numbers of high frequency words, with the former keeping the rule alive. In contrast, unproductive morphological categories will be characterized by a preponderance of words with rather high frequencies and by a small number of words with low frequencies. »

Di conseguenza, non c'è bisogno di controllare una per una le parole con la frequenza 1 per stabilire se si tratti di neologismi o meno. Semplicemente, la presenza di una larga classe di *hapax* segnala un procedimento morfologico più o meno redditizio. Anzi, la classe di *hapax*, come abbiamo già avuto modo di vedere, è di gran lunga più numerosa in tutti i tipi di RFP e in tutti i tipi testuali ; il numero di lemmi, in tale distribuzione, « ... decresce in maniera molto netta man mano che cresce la frequenza. » (GAETA, RICCA, 2002 : 229). Questa tendenza si vede molto bene nella prima struttura che abbiamo introdotto sopra : lo spettro di frequenza. Lo spettro di frequenza dei composti *porta + NOME*, che abbiamo già visto e che ripetiamo qui nella figura 5, è sostanzialmente identico a quello del prefisso *ri-* (figura 6, sotto).

Figura 5. Spettro di frequenza dei composti *porta + NOME*

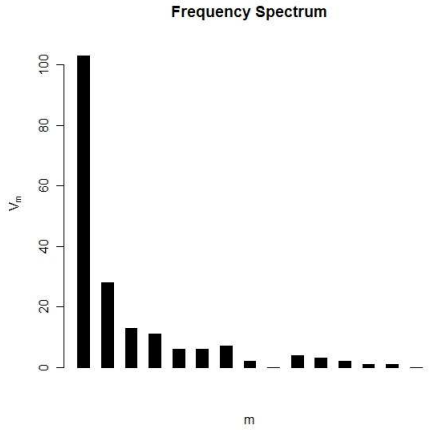
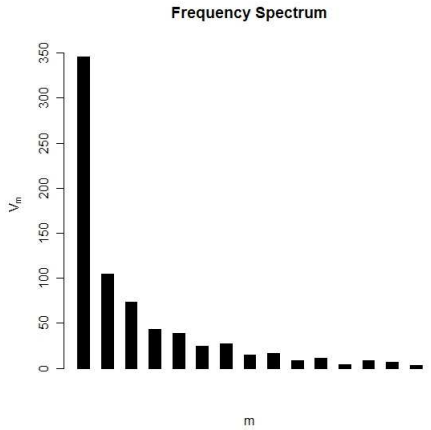


Figura 6. Spettro di frequenza dei verbi prefissati con *ri-*
(secondo BARONI, EVERT, 2006 : sezione 3.1)



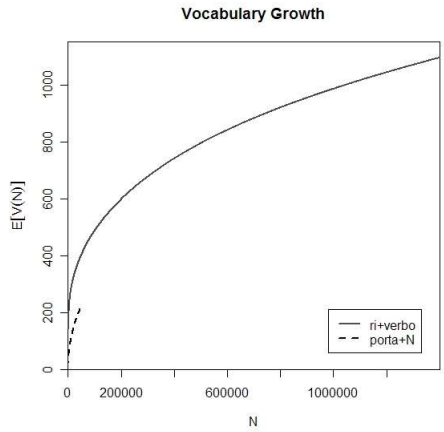
Tale spettro di frequenza, caratterizzato da Marco Baroni come « few giants, many dwarves » (BARONI, 2009 : 813), è molto importante per due motivi. Il primo è che esibisce un andamento unitario e molto diverso dallo spettro di frequenza delle categorie improduttive (ad esempio, le categorie lessicali chiuse come i pronomi, o gli affissi completamente indisponibili). Il secondo è di ordine matematico : essendo del tutto generale, le sue caratteristiche possono essere formalizzate matematicamente. Ed è infatti ciò che coglie la famosa legge di Zipf (per la formulazione cfr. BARONI, 2009 : 813 - 818), che sta alla base dei tre modelli statistici, menzionati sopra, elaborati per l'approccio quantitativo della produttività. La regolarità di tale distribuzione ci permette di affrontare anche il secondo problema che riguarda la variabilità nella dimensione del corpus.

4.2. Paragonare solo a parità di *N*

È ovvio che il computo di *V*, cioè il numero di lemmi, e il valore di *P* dipendono, in maniera diretta non solo dalla dimensione del corpus, ma anche dalla dimensione del campione (item sample), come osserva BAAYEN (1992 : 117) : « The productivity measure P has one disadvantage, however. Since P = n_1 [i.e. V_1 nella nostra notazione] / N is itself a function of N, its value depends on the size of the item sample. » Perciò, è importante paragonare le curve di accrescimento e i corrispondenti valori di *P* solo *a parità di N*, non a parità di *F* (cioè partendo dalla stessa dimensione dell'intero corpus (cfr. GAETA, RICCA, 2002 : 233, 246[9]). Così, i valori di *P* per i verbi prefissati con *ri-* e i composti *porta + NOME* (sempre prescindendo dal minimo interesse di paragonare questi due procedimenti) riportati sopra nelle tabelle, non sono paragonabili, come si vede ancora meglio se si confrontano (si veda la figura 7) le rispettive curve di accrescimento (le curve sono quelle interpolate).

Per rimediare all'insufficienza di tale confronto, è necessario paragonare le due curve a parità di *N*, il che significa che è necessario ricondurre, per mezzo dell'interpolazione binomiale, il numero di occorrenze dei verbi in *ri-* a quello dei composti *porta + NOME* (*N* = 46289) ; le due curve che risultano (figura 8, sotto) sono senz'altro più informative dal momento che consentono il confronto di tutti i parametri coinvolti, cioè sia il numero di lemmi *V* e degli *hapax legomena* V_1, che il valore *P* (tabella 8, sotto).

Figura 7. Curve di accrescimento interpolate dei verbi in *ri-* e dei composti *porta + NOME*.



[9] Cfr. GAETA, RICCA, 2002 : 246 : « Risultati radicalmente diversi, e decisamente controintuitivi, si otterrebbero se si confrontassero le produttività non a parità di *N* bensì a parità di *corpus* [...]. Procedendo in questo modo, si favorirebbero infatti in modo schiacciante i suffissi poco frequenti (data la presenza di *N* al denominatore nella formula di *P*). »

Figura 8. Curve di accrescimento interpolate dei verbi in *ri-* e dei composti *porta + NOME* con *N* = 46289

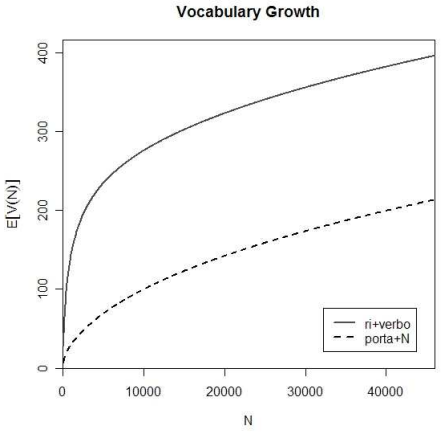


Tabella 8. Confronto del numero di lemmi *V*, di *hapax legomena* V_1 e dell'indice di produttività *P* dei verbi in *ri-* e dei composti *porta + NOME* a parità di *N* = 46289

RFP	*E(V)*	*N*	$E(V_1)$	$P\ (V_1 / N)$
ri + VERBO	396,4158	46289	102,2169	0,00220
RFP	*V*	*N*	V_1	$P\ (V_1 / N)$
porta + NOME	214	46289	103	0,00222

C'è però anche un secondo modo di unificare il valore *N* per due processi morfologici, che in alcuni casi può essere più importante. Può essere infatti più informativo *estrapolare* il valore *N* di un processo morfologico al valore più alto di un altro processo morfologico. In tal modo potremmo pensare di estrapolare *N* dei composti *porta + NOME* al valore *N* dei verbi in *ri-*.

Per farlo, si deve ricorrere ai tre modelli parametrici di distribuzione delle frequenze (*Large-Numbers-of-Rare-Events*, modelli LNRE), che sono già stati introdotti sopra. Ciascuno dei tre modelli offre un calcolo diverso dei parametri basato sullo spettro di frequenza di partenza ; i valori stimati sono di conseguenza diversi per ciascun modello ed è – pertanto – importante valutare la qualità dell'estrapolazione.

Per vedere le differenze, si può procedere direttamente a confrontare le curve di accrescimento basate sui valori stimati dei tre modelli presentati. Nella figura 9 vi è la prima curva di accrescimento dei verbi in *ri-* ; le altre tre curve corrispondono, rispettivamente, ai tre modelli LNRE (fZM, ZM e GIGP). Come si vede, ciascuna delle tre curve esibisce un andamento che si diversifica appena si supera un certo valore di *N*, cioè non appena l'estrapolazione oltrepassa un certo limite. Tale estrapolazione estrema porta dunque a risultati empiricamente inaffidabili, oltreché molto diversi tra di loro a seconda del modello. Per la grande differenza nei valori di *V*, V_1 e *P*, stimati in base ai tre modelli, si veda anche la tabella 9 qui sotto, in cui sono riportati i valori stimati per N = 1.000.000.

Figura 9. Curve di accrescimento dei composti *porta + NOME* estrapolate al valore di *N* = 1399898 mediante i tre modelli LNRE (fZM, ZM e GIGP), con la segnalazione del valore di *N* = 46289 dei composti *porta + NOME*

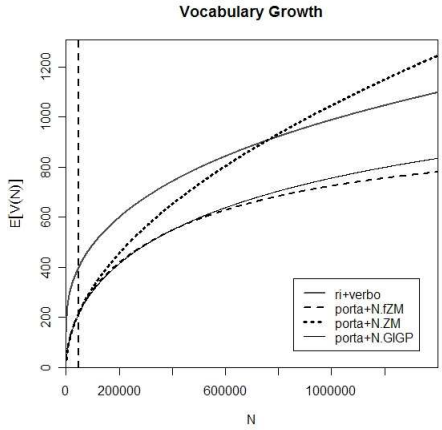


Tabella 9. Valori stimati di *V*, V_1 e *P* dei composti *porta + NOME* a seconda del modello LNRE con *N* = 1.000.000

RFP	*E(V)*	*N*	$E(V_1)$	$P\ (V_1 / N)$
ri + VERBO	988,4883	1000000	307,0424	0,000307
porta + NOME (fZM)	724,5855	1000000	176,6539	0,000176
porta + NOME (ZM)	1046,658	1000000	540,6853	0,000540
porta + NOME (GIGP)	756,457	1000000	234,9258	0,000234

L'estrapolazione che oltrepassa un certo limite diventa dunque inaffidabile. Per stabilire i limiti entro cui si può estrapolare, Stefan Evert e Marco Baroni (2006a) hanno condotto una serie di esperimenti in cui hanno sottoposto vari corpora (e varie categorie linguistiche) di varie dimensioni al test della qualità di estrapolazione. I loro risultati si possono riassumere come segue.

Nessuno dei tre modelli, se usato per l'estrapolazione, è in grado di dare risultati empiricamente affidabili. Tuttavia, il grado di affidabilità decresce in maniera netta man mano che cresce la distanza tra il valore di *N* di partenza e quello da raggiungere. La soglia critica pare essere quella del 25%, il che significa che se abbiamo, ad esempio, un campione di 25000 occorrenze e desideriamo estrapolare al valore N = 1.000.000, si può essere quasi certi che i risultati non potranno essere in alcun modo attendibili[10]: « ... often the step from 25% to 50% constitutes the boundary between extrapolation results that are clearly off-the-mark and results that are at least qualitatively plausible. (...) At 10 times the estimation size, the extrapolated vocabulary growth curves have little in common with the true growth curve any more (...). »

D'altra parte, però, i test di Evert e Baroni hanno anche mostrato che usando l'estrapolazione per i processi morfologici (nel loro caso i suffissi tedeschi *-bar* e

[10] Il nostro confronto tra *ri + VERBO* e *porta + NOME* è, quindi, da questo punto di vista, addirittura scorretto essendo la distanza tra *N(porta + NOME)=46193* e *N(ri + VERBO)= 1399898* troppo grande (il campione da raggiungere è 30 volte più grande, ossia il campione su cui estrapolare è appena il 3% di quello da raggiungere).

-*lich*), si possono ottenere risultati un po' più sicuri se ci si basa su un campione che rappresenta il 25% del valore di *N* estrapolato. Tra i modelli usati, inoltre, spicca ZM che offre risultati affidabili anche per l'estrapolazione dal 10% (cfr. EVERT, BARONI, 2006a : sezione 4.2.).

Infine, l'estrapolazione dei valori di *hapax legomena* V_1 rappresenta un'ulteriore complicazione nel senso che per la stima di V_1 i risultati ottenuti tendono ad essere ancora più inaffidabili di quanto lo siano per i valori di *V* (cfr. EVERT, BARONI, 2006a : sezione 6). Ciò accade soprattutto per i campioni di dimensioni piccole.

5. Conclusioni

Nonostante le difficoltà cui ho brevemente accennato, il presente approccio quantitativo alla produttività rappresenta oggigiorno un quadro teorico di massima importanza (cfr. BAAYEN, 2009) che ha già dato origine ad alcuni lavori di rilievo. Oltre a quelli già menzionati, che riguardano la comparazione – sincronica – di vari processi morfologici all'interno di una sola lingua[11], c'è da segnalare anche qualche tentativo di applicare l'approccio in prospettiva diacronica (cfr. LÜDELING, EVERT, 2005 ; ŠTICHAUER, 2009 ; *in corso di stampa* ; cfr. anche BAAYEN, 2009 : 909-910) : in effetti, il cambiamento della produttività attraverso periodi diversi è un fatto noto ed è quindi auspicabile che questo tipo di variazione diacronica venga ulteriormente studiato all'interno dell'approccio qui presentato.

RIFERIMENTI BIBLIOGRAFICI

BAAYEN, Harald R. (1992), Quantitative aspects of morphological productivity, in BOOIJ, G., MARLE, J. VAN (eds.), *Yearbook of Morphology 1991*, Dordrecht, Kluwer, pp. 109-149.

BAAYEN, Harald R. (2001), *Word frequency distributions*, Dordrecht, Kluwer.

BAAYEN, Harald R. (2008), *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*, Cambridge, Cambridge University Press.

BAAYEN, Harald R. (2009), Corpus linguistics in morphology : Morphological productivity, in LÜDELING, Anke, MERJA, Kytö (eds.), *Corpus Linguistics. An International Handbook*, Berlin, Mouton de Gruyter, vol. 2, article 41, pp. 899-919.

BAAYEN, Harald, RENOUF, Antoinette (1996), Chronicling *the Times*. Productive Lexical Innovations in an English Newspaper, *Language*, 72, pp. 69-96.

BARONI, Marco (2007), I sensi di *ri-*. Un'indagine preliminare, in MASCHI, R., PENELLO, N., RIZZOLATTI, P. (eds.), *Miscellanea di studi linguistici offerti a Laura Vanelli*, Udine, Forum, pp. 163-171.

BARONI, Marco (2009), Distributions in text, in LÜDELING, Anke, MERJA, Kytö (eds.), *Corpus Linguistics. An International Handbook*, Berlin, Mouton de Gruyter, vol. 2, article 37, pp. 803-822.

[11] Per l'italiano : GAETA, RICCA, 2002 ; 2003 ; 2006. Per l'inglese: BAAYEN, RENOUF, 1996 ; PLAG, 1999 ; 2006. Per il francese : DAL et al., 2007.

BARONI, Marco, EVERT, Stefan (2006), The *zipfR* package for lexical statistics : A tutorial introduction. Disponibile su : http://zipfr.r-forge.r-project.org/.

BAUER, Laurie (2001), *Morphological Productivity*, Cambridge, Cambridge University Press.

CARSTAIRS-MCCARTHY, Andrew (1992), *Current Morphology*, London and New York, Routledge.

CORBIN, Danielle (1987), *Morphologie dérivationnelle et structuration du lexique*, 2 voll., Tübingen, Niemeyer.

DAL, Georgette (2003), Productivité morphologique : définitions et notions connexes, *Langue française*, 140, pp. 3-23.

DAL, Georgette, FRADIN, B., GRABAR, N., LIGNON, S., NAMER, F., PLANCQ, C., YVON, F., ZWEIGENBAUM, P. (2007), Linguistic prerequisites to the calculation of morphological productivity and first results. Relazione presentata alle *Journées ATALA*, Paris, November 10, 2007.

DISC - *Dizionario Italiano Sabatini-Coletti* Compact versione 1.1. Milano, Giunti, 1997.

EVERT, Stefan (2004), A simple LNRE model for random character sequences, *Proceedings of JADT 2004*, pp. 411-422.

EVERT, Stefan, BARONI, Marco (2006a), Testing the extrapolation quality of word frequency models. Proceedings of Corpus Linguistics 2005, disponibile su http://www.corpus.bham.ac.uk/PCLC/.

EVERT, Stefan, BARONI, Marco (2006b), The *zipfR* library : Words and other rare events in R. Relazione presentata all'*useR! 2006 : The Second R User Conference*, Vienna, Austria.

EVERT, Stefan, BARONI, Marco (2007), *zipfR* : Word frequency distributions in R, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session*, Prague, Czech Republic, pp. 29-32.

GAETA, Livio, RICCA, Davide (2002), Corpora testuali e produttività morfologica : i nomi d'azione in due annate della *Stampa*, in BAUER, R. - GOEBL, H. (a cura di). *Parallela IX. Testo – variazione – informatica. Text – Variation – Informatik*, Wilhelmsfeld, Gottfried Egert Verlag, pp. 223-249.

GAETA, Livio, RICCA, Davide (2003), Frequency and productivity in Italian derivation : A comparison between corpus-based and lexicographical data, *Italian Journal of Linguistics / Rivista di Linguistica* 15, 1, pp. 63-98.

GAETA, Livio, RICCA, Davide (2006), Productivity in Italian word formation : A variable-corpus approach, *Linguistics* 44, 1, pp. 57-89.

LÜDELING, Anke, EVERT, Stefan (2005) The Emergence of Non-Medical -itis. Corpus Evidence and Qualitative Analysis, in KEPSER, S., REIS, M. (eds.), *Linguistic evidence. Empirical, Theoretical, and Computational Perspectives*. Berlin, Mouton de Gruyter, pp. 315-333.

PLAG, Ingo (1999), *Morphological Productivity. Structural Constraints in English Derivation*, Berlin, Mouton de Gruyter.

PLAG, Ingo (2006), Productivity, in AARTS, B., MCMAHON, A., *The Handbook of English Linguistics*, Oxford, Blackwell, pp. 537-557.

RADIMSKÝ, Jan (2006), *Les composés italiens actuels*, Paris, Cellule de Recherche en Linguistique.

RICCA, Davide (2005), Al limite tra sintassi e morfologia : i composti aggettivali V-N nell'italiano contemporaneo, in GROSSMANN, M. - THORNTON, A. M. (a cura di) *La formazione delle parole*, Atti del XXXVII congresso della Società di Linguistica Italiana, Roma, Bulzoni, pp. 465-486.

RICCA, Davide (2008), VN compounds in Italian : Data from corpora and theoretical issues. Comunicazione presentata al convegno CompoNet Congress on Compounding, Bologna, 6-7 giugno 2008.

SCALISE, Sergio (1994), *Morfologia*, Bologna, il Mulino.

ŠTICHAUER, Pavel (2007), *Tvoření slov v současné italštině*, Praha, Karolinum.

ŠTICHAUER, Pavel (2009), Morphological productivity in diachrony: the case of the deverbal nouns in *-mento*, *-zione* and *-gione* in Old Italian from the 13th to the 16th century, in MONTERMINI, F., BOYÉ, G., TSENG, J. (eds.), *Selected Proceedings of the 6th Décembrettes*. Somerville, MA: Cascadilla Proceedings Project, 2009, pp. 138-147. Disponibile su : http://www.lingref.com/cpp/decemb/6/abstract2241.html

ŠTICHAUER, Pavel (*in corso di stampa*), *La produttività morfologica in diacronia : i suffissi* -mento, -zione *e* -gione *in italiano antico dal Duecento al Cinquecento*, Praha, Karolinum.

SUMMARY

This article aims at presenting the quantitative approach to morphological productivity based mainly on Baayen's work. The discussion starts out from the widely accepted distinction between a qualitative and a quantitative approach. It is argued that there are two main quantitative approaches: one based on dictionaries, the other on large text corpora. While the dictionary-based investigation is limited to measures based on type frequency (V), the corpus-based research requires another variable: the token frequency (N). The main idea behind the relation of type frequency and token frequency is that the former (V) can be viewed as a function of the latter (N). The increasing value of N (given by the corpus size) will lead to the increasing value of V. This relation gives rise to the definition of *vocabulary growth curve* (BAAYEN, 1992 ; 2008). Two additional measures are also presented. The rate at which the vocabulary grows can in fact be captured by the proportion of *hapax legomena* (V_1), the types that occur precisely once, to the overall number of tokens of the formations with a given affix. The notion of *vocabulary growth rate* ($P = V_1 / N$) (BAAYEN, 1992) is thus introduced. Finally, a third statistical tool of modelling the relation of V, N and V_1, put forward by BARONI, EVERT, 2006, is presented. It is the *frequency spectrum*, which is a specific object that views the number of types (V) as a function of a *frequency rank* (m) assigned to every type according to its token frequency. Some problems typical of this quantitative approach are also discussed, namely the difficult relation between *hapax legomena* and neologisms, and the role of the number of tokens for the assessment of the productivity. As far as the role of the number of tokens is concerned, it is shown that – in the light of the evident fact that the measure depends directly on the corpus size – it is not possible to compare corpora of different sizes using this measure (cfr. BAAYEN, 1992 : 117). In order to overcome this problem, two main techniques are presented: binomial interpolation and extrapolation. Especially, three modes of extrapolation are introduced. In conclusion, something is said about the research and particular studies being conducted within this framework.