LIVIO GAETA – DAVIDE RICCA

Aspetti quantitativi della produttività morfologica

1. Frequenza e produttività nella formazione delle parole

Quanto spesso viene impiegato un certo procedimento di formazione delle parole (= PFP)? Da questa semplice domanda discende una serie di questioni connesse con la competenza lessicale dei parlanti. Innanzitutto cosa significa frequenza? Da un lato la frequenza può essere calcolata sulla base del numero di lessemi che un parlante ha a disposizione nel proprio lessico mentale (la cosiddetta *type frequency*, o numerosità secondo Thornton 1997: 386). Per orientarsi, si può ad esempio decidere di calcolare quante parole formate con un certo PFP sono attestate in un certo dizionario. Ma frequenza può anche essere intesa come numero di forme contenenti un certo PFP che ricorrono nell'uso dei parlanti (la cosiddetta *token frequency*, o semplicemente frequenza in Thornton 1997: 386). In quest'ultimo caso, i dizionari classici sono di poca utilità e bisogna ricorrere ad altri strumenti, ad esempi i dizionari di frequenza come il LIF (Bortolini *et al.* 1971) per l'italiano, che sono costruiti sulla base di corpus linguistici. Le due misure non necessariamente coincidono: un PFP può essere molto frequente ma scarsamente numeroso, o viceversa molto numeroso ma pochissimo frequente. Inoltre, la frequenza appare anche intuitivamente connessa con la capacità dinamica di arricchire il lessico mentale di un parlante, in altre parole con la produttività di un PFP. Tuttavia il rapporto tra frequenza (sia in termini di numerosità che di frequenza *stricto sensu*) e produttività non è immediato, come si vedrà nel prossimo paragrafo.

1.1 Produttività: concetto saliente ma sfuggente

Benché intuitivamente la produttività di un PFP sia una nozione estremamente saliente, tuttavia non è semplice definirla con precisione, dire cioè in che senso e in che termini un PFP sia produttivo, esplicitando il dominio lessicale al quale si applica. E parimenti non è facile stabilire con esattezza che cosa si debba misurare per dare una valutazione quantitativa della produttività di un PFP. Con il termine "produttività" in realtà si possono considerare almeno sei modi diversi di misurare l'impiego di un PFP (cfr. Rainer 1987, 1993: 29-34):

[1] a. l'insieme delle parole formate con un certo PFP;

b. l'insieme dei neologismi formati con un PFP in un certo lasso temporale;
c. la possibilità di formare nuove parole con un certo PFP;
d. la probabilità di formare nuove parole con un certo PFP;
e. l'insieme delle parole possibili (o generabili per regola) formate con un certo PFP;
f. il rapporto tra parole attestate e parole possibili.

Il primo uso di "produttività" è chiaramente inadeguato: un PFP può essere oggi ampiamente presente nel lessico con un numero significativo di formazioni attestate; la sua produttività, tuttavia, può essersi esaurita e il PFP può essere divenuto ormai improduttivo. Viceversa, PFP nati di recente non avranno una grande estensione in termini di formazioni attestate; ciò non esclude che possano essere considerati produttivi. D'altro canto, non è privo di interesse verificare la presenza nel lessico di un certo PFP, sia nei termini di *types*, informazione ricavabile dai dizionari, sia in termini di *tokens*, dato quest'ultimo non elicitabile dai lessici.

L'uso (1b) è stato in genere considerato in connessione con l'impiego di lessici, e per questo ne rimandiamo la discussione al paragrafo successivo.

Gli usi (1c) e (1d) mettono l'accento sulla capacità sincronica del parlante di formare nuove parole con un certo PFP. Da questo punto di vista, sono vicini alla famosa definizione di Schultink che già nel 1961 intendeva la produttività morfologica come "la possibilità sussistente per gli utenti della lingua di formare in maniera non intenzionale una quantità in linea di principio innumerevole di nuove formazioni per mezzo di un procedimento morfologico che sta alla base della corrispondenza tra forma e significato in alcune parole ad essi note" (Schultink 1961: 113; trad. L.G.).

Il vantaggio di questa definizione è quello di ancorare la nozione di produttività alla competenza sincronica attiva dei parlanti, che si esplicita nella capacità di creare neoformazioni. Si noti come la competenza sincronica attiva dei parlanti sia vincolata alla condizione di non intenzionalità, nel senso che non viene considerato un esempio affidabile di neoformazione un neologismo costruito intenzionalmente da un parlante ad esempio con fine ironico, come in casi del tipo: *la particolare copertura informativa che in America si chiama wifing (mogliegiamento?, mogliegiatura?)* ("La Stampa", 27-4-1996, 5). Tali esempi infatti possono violare le condizioni di accettabilità imposte dalla grammatica. D'altronde l'intenzionalità non è sempre facile da individuare, come messo in evidenza da Plag (1999: 13), che osserva come la coscienza metalinguistica vari da parlante a parlante, per cui quel che appare come intenzionale a qualcuno può passare inosservato ad un altro. Inoltre esistono settori lessicali come quelli connessi con le terminologie, in cui le creazioni neologiche sono intenzionali, e tuttavia non è detto che non possano corrispondere a modelli di formazione lessicale produttivi.

Come dare un senso quantitativo all'idea "sincronicistica" di Schultink? Una prima risposta viene da Booij (1977: 120), che lega la produttività di un PFP al numero di restrizioni di tipo qualitativo cui esso soggiace. Quante più restrizioni saranno presenti nel dominio di un PFP, tanto più bassa sarà la produttività. Benché da un punto di vista qualitativo si possa in prima approssimazione concordare con quest'approccio, resta tuttavia nell'ombra se e come debba poi essere verificata quantitativamente la portata di un PFP. Poco praticabile appare cercare di avvalorare quantitativamente un concetto sfuggente quale quello di "parola potenziale", che sta alla base degli usi (1e) e ancor più (1f) che suggerisce di misurare la produttività di un PFP considerando il rapporto tra parole attestate con quel PFP e il numero di formazioni potenzialmente possibili con quel PFP (cfr. Aronoff 1976). Al di là della difficoltà di definire cos'è una parola "potenziale", questa misura conduce ad una serie di incongruenze, come ad esempio il fatto che per PFP con un numero di parole potenziali tendente a infinito, quindi intuitivamente molto produttivi, si ottiene una produttività tendente a zero. O viceversa, per PFP non più produttivi, che abbiano tuttavia sfruttato il loro bacino lessicale potenziale, la produttività risulta altissima.

1.2. L'uso dei dizionari per valutare la produttività

Come misurare le neoformazioni? Per rispondere a questa domanda, la pratica più comune seguita dagli studiosi è consistita nel far ricorso all'impiego dei dizionari. Ad esempio Neuhaus (1973) assume che il grado di produttività dei PFP sia direttamente connesso con la quantità di formazioni attestate in un certo lasso temporale, verificate sulle base di dizionari storici, il che corrisponde all'uso di "produttività" visto sopra in (1b). Ma anche l'approccio basato sulla competenza attiva del parlante "alla Schultink" può essere verificato con strumenti lessicografici, contando quanti neologismi sono riportati in un dizionario recente. Tuttavia l'impiego dei dizionari per verificare la produttività dei PFP non è esente da problemi, in parte legati al dizionario in quanto tale, in parte connessi con lo statuto teorico della produttività. Ad esempio si può mostrare (vedi 1.3) che quest'ultima è legata a parametri come la frequenza, per la quale i dizionari sono semplicemente inadeguati a fornire una risposta.

Un primo problema presentato dai dizionari è che per ragioni pratiche e commerciali spesso non mirano a fornire una documentazione completa delle forme trasparenti e produttivamente formate, ma piuttosto contengono termini frequenti e idiosincratici, specialmente nel caso di dizionari di piccola taglia. Chiaramente, questa strategia è perfettamente giustificata da un punto di vista lessicografico, in quanto l'utente medio per molti versi non ha bisogno di verificare parole complesse il cui significato sia interamente predicibile sulla base dei suoi elementi, come è per definizione il caso delle parole produttivamente formate. Inoltre, anche quando mira alla piena

esaustività, il lessicografo spesso trascura formazioni regolari, semplicemente perché … sono regolari! Ciò si verifica soprattutto per quei PFP il cui contenuto semantico non è particolarmente prominente come nomi d'azione, nomi di qualità, aggettivi di relazione, ecc.

Da un altro punto di vista, spesso i dizionari tendono invece ad essere delle enciclopedie, cioè a raccogliere parole appartenenti ai più disparati settori lessicali, come i repertori terminologici, che non necessariamente sono coperti dalla competenza lessicale dell'utente medio (e questa è una delle ragioni per cui i dizionari vengono consultati). D'altra parte, come si accennava sopra, non si può escludere che termini creati in specifici settori lessicali corrispondano a modelli produttivi, che in quanto tali contribuiscono a determinare la produttività in un certo PFP.

Un ulteriore problema presentato dai dizionari è che portano con sé un ampio serbatoio di forme complesse di formazione antica che possono distorcere un'analisi sincronica, in quanto sono residui di processi morfologici che hanno cessato di essere sincronicamente produttivi, oppure sono di provenienza alloglotta.

1.3. Produttività misurata su corpus lessicali

Si è accennato in precedenza come i dizionari siano assolutamente inadeguati per individuare il ruolo svolto dalla frequenza nella produttività dei PFP. Quest'ultimo è stato messo in evidenza in diverse indagini empiriche, come ad esempio il lavoro sull'inglese di Aronoff (1983), che mostra come la frequenza media dei derivati sia significativamente più alta per PFP meno produttivi rispetto a quella di PFP più produttivi. L'interpretazione di questo fatto fornita da Aronoff non potrebbe essere più esplicita: "the less productive W[ord] F[ormation] P[attern] is more remarkable and ... its members are therefore more likely to be lexicalized and assigned special meanings ... this lexicalization is reflected in frequency, for semantic complexity and frequency go hand in hand" (Aronoff 1983: 168). A ciò si aggiunge che i recenti studi psicolinguistici sull'accesso lessicale (cfr. per una rassegna Laudanna e Burani 1999) hanno puntato l'indice sul fatto che l'alta frequenza di un lessema ne favorisce l'accesso lessicale diretto, mentre la bassa frequenza ne favorisce l'analisi, e nel caso di parole complesse la scomposizione. È chiaro che solo la scomposizione è connessa con PFP produttivi.

Nonostante il rilevante ruolo della frequenza nella considerazione dei PFP, sono molto pochi gli studi che si occupano in maniera sistematica di quest'aspetto in relazione alla formazione delle parole (per l'italiano si veda Thornton 1997 e più di recente Gaeta e Ricca 2003a). Probabilmente, questa lacuna è da imputare più in generale alla diffidenza e alla sottovalutazione da parte di molti studiosi delle indagini quantitative, in quanto più legate a questioni di *performance* che di *competence* (cfr. Dressler e Ladányi 2000). Tut-

tavia, la crescente disponibilità di estese basi di dati su supporto elettronico di facile consultabilità è destinata a far crescere l'interesse generale verso gli studi quantitativi (si veda ad esempio l'indagine empirica di Rainer 2003 basata sul *Web*).

A tal proposito, abbiamo elaborato un corpus testuale costituito da tre annate (1996-1998) del quotidiano *La Stampa* di Torino, disponibili su CD-Rom e facilmente esportabili su *files* ASCII trattati successivamente per mezzo del *software* di analisi testuale DBT® di E. Picchi del CNR di Pisa. (per maggiori dettagli cfr. Gaeta e Ricca 2002)[1]. Nel seguito la nostra esemplificazione verrà fatta a partire da questo corpus. La scelta di un quotidiano non è inappuntabile da un punto di vista metodologico, in quanto il materiale testuale così ottenuto non è pienamente bilanciato per tipi testuali, registri stilistici e così via, rispetto ad altri corpus come il LIF o il LIP (cfr. De Mauro *et al.* 1993). Tuttavia, questi ultimi sono di dimensioni relativamente ridotte (circa 1.500.000 e 500.000 *tokens* rispettivamente), rispetto al nostro che comprende circa 75 milioni di *tokens*. Come si vedrà tra breve, la loro taglia ridotta li rende inutilizzabili per il tipo di indagine empirica sviluppata da Baayen e Renouf (1996), che adottano come corpus diverse annate del *Times* per un totale di circa 80 milioni di *tokens*. Inoltre, benché non adeguatamente bilanciato, un corpus giornalistico presenta in ogni caso una varietà testuale e di registri stilistici molto ampia, distribuita tra i diversi argomenti (politica, cultura, sport, ecc.) abitualmente trattati in un quotidiano, che almeno parzialmente compensa la mancanza di bilanciamento. Pertanto, benché non pienamente affidabile, il corpus fornisce un quadro abbastanza fedele della competenza "ideale" di un un parlante (colto) di italiano scritto (settentrionale).

In una serie di contributi recenti, H. Baayen (cfr. almeno Baayen 1992, 1993, 2001; Baayen e Lieber 1991, Baayen e Renouf 1996, Plag *et al.* 1999) ha proposto di ancorare la nozione di produttività alla misura del numero di *hapax legomena*, cioè parole con frequenza 1, presenti in un dato corpus. Come fanno osservare Baayen & Renouf (1996), benché gli *hapax legomena* non siano di per sé dei neologismi, è tuttavia molto plausibile che sia proprio tra le parole a frequenza più bassa che vadano ricercate le neoformazioni. Sul rapporto tra neologismi e *hapax legomena* in un corpus testuale si vedano anche Gaeta e Ricca (2002: 227-229), Dal (2003: 17-18).

Per misurare la produttività di un PFP, Baayen propone l'indice *P*, uguale al rapporto tra *h*, il numero di *hapax legomena* formati con un certo affisso

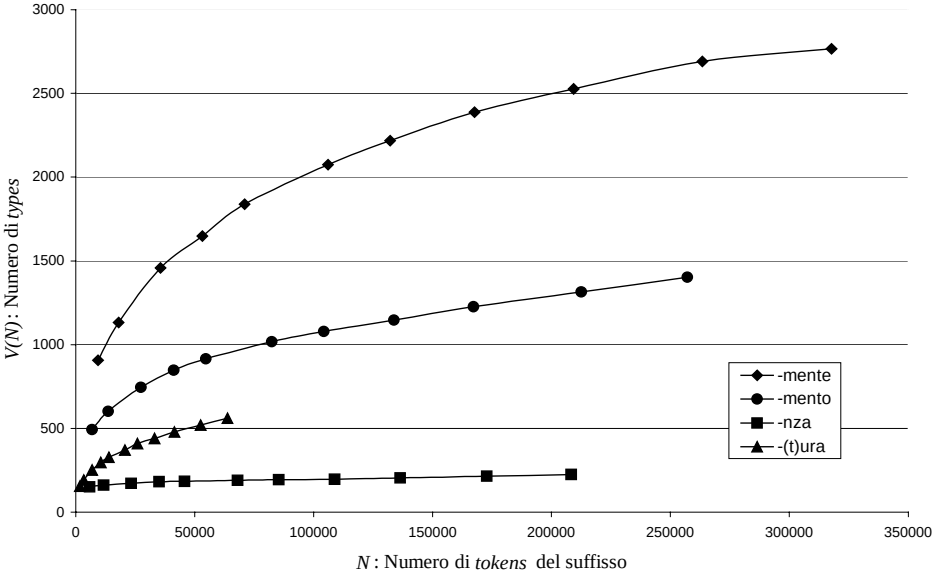
[1] Dal *data-base* ottenuto grazie al DBT, abbiamo estratto la lista completa delle forme di parola in ordine alfabetico diretto e inverso, con l'indicazione per ogni forma della sua frequenza. Da queste liste sono state estratte e lemmatizzate tutte le occorrenze di un dato affisso in modo da renderle disponibili per il calcolo dei *types* e dei *tokens*, dopo un inevitabile e dispendioso controllo manuale. Questo passaggio è ovviamente indispensabile per eliminare dal computo tutte le parole che non presentano affissi ma solo terminazioni omografe, e ancor di più per ripulire le liste dagli errori di stampa.

presenti nel corpus, e *N*, il numero di *tokens* formati con lo stesso affisso presenti nell'intero corpus:

[2] $P = h/N$

Si può dimostrare matematicamente (Baayen 1992: 115) che il valore *P* in [2] è un'approssimazione della derivata nel punto *N* della curva *V(N)*, che ha in ascissa il numero di *tokens*, *N*, e in ordinata il numero di *types*, *V*. Tale curva è raffigurata in Fig. 1 per quattro suffissi italiani, e per ogni suffisso rappresenta l'accrescimento del suo inventario lessicale man mano che si procede nello spoglio del corpus. L'indice *P*(*N*), che corrisponde alla pendenza della curva *V(N)* nel punto *N*, dà quindi una misura del ritmo di accrescimento della curva, e pertanto della propensione del suffisso ad incrementare il proprio inventario lessicale con nuove formazioni.

Fig. 1. *La Stampa* 1996-1998: curve di accrescimento dei *types* in funzione di *N*



Non c'è lo spazio in questa sede per discutere nei dettagli gli aspetti problematici di quest'approccio, per cui si rimanda a Gaeta e Ricca (2002, ms.). Si noti soltanto che le curve in Fig. 1 hanno lunghezza diversa, il che dipende dalla differente frequenza degli affissi nel corpus. A differenza di Baayen, per il quale il numero *N* nel denominatore della formula in [2] è dato dal numero totale di *tokens* di un affisso presenti nel corpus globale, la nostra proposta calcola l'indice *P* a parità di *N*. Per far ciò, bisogna estrarre i valori desiderati di *N* (diciamo ad esempio 50.000) da sottocorpus di dimensione variabile, data la diversa frequenza degli affissi nel corpus globale. Per rimanere agli esempi riportati in Fig. 1, per raggiungere 50.000 *tokens* del suffisso *-(t)ura* è necessario un sottocorpus di circa 58 milioni e mezzo di *tokens*, mentre per raggiungere lo stesso valore per il suffisso *-mento*, ne basta uno di 14 milioni e mezzo di *tokens*[2]. L'approccio a corpus variabile fornisce una risposta a varie critiche sollevate contro la procedura originaria di Baayen (cfr. van Marle 1992), che in effetti sovrastima i valori di *P* per affissi a bassa frequenza rispetto a quelli ad alta frequenza a causa dell'effetto dovuto al carattere decrescente della funzione *P(N)*, che tende addirittura a zero quando *N* tende a infinito (cfr. Baayen e Lieber 1991: 837).

Oltre all'indice di produttività *P*, che è una misura probabilistica e corrisponde pertanto all'accezione di produttività vista sopra in (1d), Baayen (1993: 193) suggerisce un'altra via di esprimere la produttività che risulta dal mero numero degli *hapax legomena* presenti nel corpus globale. Questo valore misura, come è stato obiettato (Bauer 2001: 155), il numero delle neoformazioni con un certo affisso rispetto alla totalità delle neoformazioni presenti nel corpus piuttosto che la proporzione delle neoformazioni tra le parole formate con un certo affisso. Tuttavia, questo "hapax-conditioned degree of productivity" può esser visto come una maniera veloce e conveniente per quantificare quanto sia attivo un PFP; inoltre, come si vedrà in § 4., nella maggioranza dei casi (anche se non in tutti), il semplice numero degli *hapax legomena* correla con la misura da noi proposta, mentre non correla affatto con l'indice *P* nell'interpretazione originaria di Baayen.

2. Quali parole vanno incluse nel conteggio di un affisso?

Qualunque approccio quantitativo alla morfologia si confronta necessariamente con una serie di decisioni operative non sempre facili su

[2] Per rendere fattibile l'approccio a corpus variabile, quest'ultimo dev'essere strutturato in singole porzioni di testo (per la precisione 36, una per ogni mese della *Stampa*; per i dettagli si veda Gaeta e Ricca 2002) computabili separatamente, in modo da fornire di volta in volta il sottocorpus che risponde al valore di *N* desiderato per i diversi affissi. Il confronto è reso possibile dal fatto che la frequenza degli affissi all'interno del corpus resta costante, garantendo l'uniformità dei dati estratti da sottocorpus di taglia diversa.

cosa contare e cosa no, perché non sempre è ovvio decidere se una data parola va analizzata in sincronia come morfologicamente complessa, e quindi le sue occorrenze contribuiscono come istanze dell'affisso derivazionale coinvolto. Infatti, dal momento in cui una parola formata derivazionalmente entra nel lessico, risulta soggetta, al pari di qualunque lessema non derivato, all'evoluzione semantica che ne può rendere meno composizionale il significato e a fenomeni fonetici che possono ridurre o anche cancellare del tutto la segmentabilità dei morfemi componenti. Sia la trasparenza semantica che quella morfottatica costituiscono dei *continua* la cui segmentazione comporta necessariamente margini di arbitrarietà. È però indispensabile mantenere per lo meno dei criteri coerenti se si vuole essere in grado di confrontare i dati per affissi differenti. I criteri adottati sono stati discussi estesamente in altre pubblicazioni (cfr. spec. Gaeta e Ricca 2002, ms.), a cui rimandiamo. In breve, abbiamo mantenuto un approccio abbastanza "largo" rispetto a entrambe le questioni. Per quanto riguarda la trasparenza morfosemantica, sono state escluse solo quelle parole in cui il rapporto tra base e derivato appare completamente opaco in sincronia (casi come *sedimento, temperatura, sentenza, generoso, scuderia*) e ha un forte carattere idiosincratico. Sono state invece conservate tra i *tokens* dei rispettivi suffissi le più numerose occorrenze in parole come *abitazione* o *creatura*, che non significano certo 'atto di abitare/creare', ma in cui il passaggio semantico può essere trattato in termini di polisemia regolare.

Analogamente, per quanto riguarda la trasparenza morfotattica, tutti i casi di allomorfia della base e/o del suffisso in cui i due elementi siano comunque individuabili (si pensi in particolare alla complesse allomorfie che coinvolgono base e suffisso nei nomi in *-zione*/*-ione* come *delusione, assunzione, emissione, gestione, guarigione*, ecc., cfr. Rainer 2001: 386-389, Gaeta 2002: 66-71) sono stati inclusi nei conteggi. Si sono invece escluse parole come *ovazione, massaggio, potabile* in cui, benché le terminazioni possano essere identificate e si possa anche loro attribuire il significato usuale, le "basi" non sono possibili morfemi lessicali dell'italiano (neppure morfemi lessicali non autonomi, a differenza dei primi elementi neoclassici del tipo di *idr-* in serie come *idrante, idrico* ecc.). Si tratta delle parole che Corbin (1987: 463) identifica come "mots complexes non construits" e considera come unità lessicali non derivate.

Un ulteriore problema rilevante per le scelte di inclusione nei conteggi è posto dai morfemi che compaiono in cicli derivazionali interni. Ad esempio, una parola come *nazionalizzazione* conta sicuramente come *token* di *-zione*; ma occorre anche considerarla un *token* di *-ale* e di *-izzare*? È evidente che anche questi due suffissi sono chiaramente identificabili nella parola in questione e contribuiscono indipendentemente al suo significato; tuttavia la prassi consueta nelle ricerche quantitative sulla produttività è stata quella di tener conto soltanto dei cicli derivazionali esterni (cfr. Plag 1999 : 29). Tra le ragioni di questa scelta, oltre alla maggiore praticabilità, c'è il fatto che in

questo modo le popolazioni dei diversi suffissi corrispondono a sottoinsiemi disgiunti, e perciò statisticamente indipendenti, di parole del corpus (cfr. Baayen 1992: 200). I dati riportati nei paragrafi successivi sono tutti riferiti ai soli cicli esterni. In Gaeta e Ricca (2003b, ms), tuttavia, si è indagato anche quale sia l'impatto quantitativo dei cicli interni: si è dimostrato come esso possa essere molto rilevante per alcuni affissi in termini di *tokens* e *types* (ad esempio per *-bile*, *-izzare*, *ri-/re-* e *in-*, mentre è del tutto trascurabile per altri come *-mento* o *-ezza*). Nel contempo, si è verificato come in generale il calcolo della produttività non appaia seriamente modificato dall'inclusione dei cicli interni, il che giustifica la loro esclusione in questa sede.

3. *Type* e *token frequency* dei morfemi derivazionali italiani

In Gaeta e Ricca (2003a) si è fornito un quadro complessivo della derivazione italiana rispetto ai tre parametri fondamentali – indipendenti tra loro – della frequenza, della numerosità e della produttività. Per quanto riguarda frequenza e numerosità, i dati sono confrontabili con quelli di Thornton (1997), anche se il nostro lavoro si basa su un corpus 50 volte più ampio e prende in considerazione un numero maggiore di affissi (58 contro 36), che includono anche tipi di formazioni completamente assenti in Thornton (1997), in particolare la suffissazione verbale (*-izzare*, *-ificare*, *-eggiare*) e varie istanze di prefissazione. Lo spazio consente qui solo alcuni commenti d'insieme sui risultati ottenuti; per una disamina più dettagliata si rimanda appunto a Gaeta e Ricca (2003a).

Un primo dato concerne la dimensione complessiva della derivazione in italiano. I 58 affissi considerati, infatti (per la lista si vedano più avanti le tabelle 1 e 2), pur non esaurendo certo il quadro,[3] comprendono la maggior parte degli affissi di alta e media frequenza dell'italiano.[4] Il totale di

[3] Una stima complessiva del numero degli affissi derivazionali in italiano può aversi dal numero di affissi posti a lemma in De Mauro (2000), e cioè 261 suffissi e 94 prefissi. Si noti peraltro che in questo numero rientrano tutti gli affissi propri esclusivamente delle tassonomie tecnico-scientifiche, numerosi affissi molto rari e completamente improduttivi e non poche varianti allomorfiche. Un altro dato di riferimento è in Thornton (1997: 389), che indica in 180 il numero di suffissi (prefissi dunque esclusi) presenti in almeno una parola del Vocabolario di Base dell'italiano.

[4] Le assenze più rilevanti, dovute in parte a varie complicazioni tecniche (soprattutto la scarsa sostanza grafica e/o problemi di omonimia troppo laboriosi da discriminare manualmente) sono: (a) tutti i derivati per conversione, per motivi evidenti; (b) i suffissi valutativi, tranne *-accio*, e in particolare il suffisso *-ino* con i suoi diversi valori (valutativo, formante di aggettivi di relazione, agentivo); (c) i verbi parasintetici; (d) il suffisso di nomi d'azione *-ata*, a causa dell'omonimia con il participio passato femminile; (e) vari formanti minori di aggettivi denominali e di aggettivi etnici (di questi ultimi il solo incluso è *-ese*); (f) molti tra i prefissi come *anti-*, *auto-*, *neo-*, *multi-*, *pseudo-*, oggi frequenti anche in parole non tecnico-scientifiche, dei quali si è incluso solo il sottoinsieme dei prefissi valutativi.

occorrenze di questi 58 affissi è di circa 5 milioni. Poiché la distribuzione in *tokens* degli affissi, come quella delle parole in generale, è molto fortemente asimmetrica (v. oltre), si può stimare che la popolazione considerata comprenda ben più della metà dei *tokens* di parole derivate presenti nel corpus: il totale di 5 milioni è dunque di per sé molto indicativo. Se si tiene conto che dei 75 milioni di *tokens* dell'intero corpus, circa la metà sono da attribuire a parole funzionali,[5] si può ipotizzare, sia pure come stima grossolana, una percentuale intorno al 20% di occorrenze di parole derivate rispetto al totale di *tokens* appartenenti alle categorie lessicali maggiori. Il dato appare intuitivamente alto, anche se non potrà estendersi automaticamente a qualunque tipologia testuale e tantomeno al parlato.

Per quanto riguarda il totale complessivo dei *types*, esso ammonta a 30.424. Questo dato è interessante se comparato all'insieme dei lemmi derivati con gli stessi affissi registrati in un dizionario di grandi dimensioni come De Mauro (2000) (per il confronto in dettaglio si veda Gaeta e Ricca (2003a: 83-84). Il totale limitato alle parole etichettate come "di uso comune" (cioè provviste di una delle quattro marche d'uso CO, AD, AU e FO) è di molto inferiore, 21.391; e lo sarebbe ancora di più espungendo gli etnici in *-ese*, tutti contrassegnati come CO in De Mauro (2000), che da soli contribuiscono per un terzo dei lemmi, mentre costituiscono appena il 2% dei *types* derivati presenti nel corpus. Per raggiungere nel lemmario di De Mauro (2000) il totale dei *types* presenti nel corpus occorre affiancare alle parole comuni anche i termini tecnico-specialistici. Ciò significa che un corpus delle dimensioni considerate è un buon indicatore – almeno per le parole derivate – della dimensione complessiva dell'inventario lessicale di una lingua. Naturalmente non c'è piena sovrapposizione tra il lemmario di un dizionario e quello estratto da un corpus, che da un lato non comprenderà certo *tutte* le parole di uso comune registrate nel dizionario e dall'altro include molte neoformazioni ed occasionalismi che non hanno (ancora) trovato consacrazione lessicografica.

Le tabelle 1 e 2, riprese da Gaeta e Ricca (2003a),[6] riassumono in modo compatto l'ordinamento dei 58 affissi considerati secondo i due parametri della frequenza (Tab. 1) e della numerosità (Tab. 2). Analogamente a Thornton (1997: 389), si sono raggruppati i suffissi in classi logaritmiche. Nella classe 10 della Tab. 1, ad esempio, sono inclusi gli affissi di frequenza *N* tale per cui ln *N* è compreso tra 9,5 e 10,5 (all'interno della stessa classe gli affissi sono disposti in ordine di frequenza decrescente da sinistra a destra). Due coppie di affissi omonimi sono state distinte: *-*$^{(1)}$*erìa*, formante

[5] Un conto esplicito è stato fatto per le forme di parole funzionali aventi rango da 1 a 100: la somma delle loro occorrenze ammonta già a più di 31 milioni di *tokens*, circa il 42%.

[6] La tabella 1 risulta modificata non nella sostanza, ma nel valore numerico delle classi, perché in Gaeta e Ricca (2003a: 77) si faceva riferimento non a ln *N*, bensì a ln (*N*/50) = ln *N*- 3,91, in modo da sovrapporre le classi ottenute a quelle di Thornton (1997). Si ha quindi uno slittamento sistematico di quasi 4 classi per tutti i suffissi.

di nomi di qualità come *vigliaccheria*, e *-(2)erìa*, formante di collettivi e nomi di luogo, specie negozi e simili, come *libreria*; *-(1)aio* nei nomi d'agente come *fioraio* e *-(2)aio/a* nei nomi di luogo come *pollaio, risaia.*

L'organizzazione in classi logaritmiche rende meno appariscente il fatto che le distribuzioni in frequenza e in numerosità degli affissi sono estremamente sbilanciate. In particolare, per quanto riguarda la frequenza in *tokens*, l'affisso più frequente, *-(z)ione*, presenta nel corpus oltre un milione di occorrenze, pari da solo a un quinto del totale complessivo dei 58 affissi, e ben al 13,9 ‰ di tutti i *tokens* del corpus. Il secondo affisso, *-ale/-are*, scende già a 734.725 occorrenze, e il gruppo degli altri affissi della classe 13 presenta valori di frequenze, tra loro ravvicinati, intorno a 300.000.

Tabella 1 - Affissi derivazionali italiani ordinati per classi logaritmiche di frequenza (in *tokens*)

Affissi	Classe (ln *N*)
-(z)ione	14
-ale/-are, -ità/-età, -mente, -(t)ore, ri-/re-	13
-mento, -nza, -ista, in-, -oso, -ese, -bile	12
-izzare, -ore, -ezza, -(t)ura, -ismo, -iere, -issimo, -izia, -iano	11
-ificare, -eggiare, -trice, -aggio, -evole, -(2)erìa, -iero, -(t)orio	10
-(1)aio, -iera, -esco, super-, -essa, -ificio, -toio/a, -izio	9
-accio, -aneo, -ame, -(1)erìa, -(2)aio/a, micro-, mini-	8
iper-, maxi-, -estre, ultra-, mega-, -aglia, -oide, -aggine, -ume, -astro, -eto/a	7
-aceo, -igia	6

All'opposto, gli affissi "piccoli" delle ultime quattro classi hanno tutti frequenze comprese tra alcune centinaia e qualche migliaio di occorrenze. La distribuzione complessiva può essere confrontata con quella riportata in Thornton (1997: 389), e non se ne discosta in modo sostanziale, anche se in Tab. 1 fanno ovviamente la loro comparsa nelle classi alte alcuni importanti affissi non inclusi in Thornton (1997), in particolare *-mente*, *-ità/-età*, *-ese* e i due prefissi *ri-/re-* e *in-*.

Anche la distribuzione degli affissi in termini di numerosità è fortemente asimmetrica, benché la dispersione della corrispondente distribuzione logaritmica sia un po' più bassa (la deviazione standard è di 1,5 contro 2,1): si va dai 2.767 e 2.363 *types* per *-mente* e *-(z)ione* rispettivamente, agli appena 7 *types* per *-estre* ed *-igia* (e naturalmente non sarebbe difficile aggiungere altri affissi nelle due classi inferiori della Tab. 2).

Tabella 2 - Affissi derivazionali italiani ordinati per classi logaritmiche di numerosità

Affissi	Classe

	(ln *V*)
-mente, -(z)ione, -ità/-età	8
-issimo, -ista, -(t)ore, -iano, -mento, -ismo, super-, -bile, -ale/-are, ri-/re-, in-, -izzare	7
-ese, -trice, -oso, mini-, -(t)ura, micro-, mega-, -esco, iper-, maxi-, -accio, -ezza, ultra-, -(t)orio	6
-eggiare, -nza, -iere-, -(2)erìa, -(1)erìa, -(1)aio, -aggio, -iera, -oide, -ificare	5
-iero, -ificio, -aggine, -(2)aio/a, -essa, -evole, -ore, -toio/a, -aceo, -ume, -ame, -astro, -eto/a, -aglia	4
-izio, -izia, -aneo	3
-estre, -igia	2

Uno sguardo comparativo alle due distribuzioni in frequenza e numerosità delle tabelle 1 e 2 mostra prevedibilmente come i suffissi più importanti, unanimememente considerati molto produttivi anche su un piano qualitativo, si collochino nelle classi più alte rispetto ad entrambi i parametri. È però più interessante rilevare come la correlazione tra *token e type frequency* venga meno completamente in altri casi. Da un lato alcuni suffissi manifestamente improduttivi come *-izia* e *-ore* (formante di nomi astratti come *spessore*), pur avendo come ci si può aspettare una numerosità molto bassa, si collocano nella fascia alta per quanto riguarda la frequenza in *tokens*, a causa della presenza di alcune parole di altissima frequenza che possono essere analizzate come derivate (*giustizia, amicizia*; *amore, valore*). Il caso opposto è rappresentato dal gruppo molto coerente dei prefissi valutativi (si veda Gaeta e Ricca 2003b), che hanno una frequenza in *tokens* molto bassa, e una alta numerosità (altissima per *super-*). Il dato dipende dal fatto che questi prefissi danno luogo a pochissime formazioni stabilizzate nel lessico (come *minigonna, supermercato, maxischermo, microcriminalità*), ma si combinano molto liberamente con ogni sorta di basi nominali e aggettivali in formazioni occasionali (come *megacena, mini-epurazione* ecc.) che incrementano la numerosità senza incidere grandemente sul totale dei *tokens*. È questo un caso in cui la numerosità ricavabile da un corpus dà risultati molto diversi (e molto più elevati) da quelli che si otterrebbero dal computo dei lemmi di un dizionario: infatti per formazioni di questo tipo moltissimi *types* (la metà circa) risultano anche *hapax legomena*, per cui la numerosità nel corpus tende a riflettere più la produttività della formazione che il suo consolidamento nel lessico. Per un confronto sistematico tra le numerosità ricavabili da dati testuali e lessicografici, si rimanda ancora una volta a Gaeta e Ricca (2003a: 81-86).

Infine, un aspetto della morfologia italiana su cui i dati delle tabelle 1 e 2 danno utili informazioni è quello della sua forte ridondanza funzionale, cioè della presenza, per molti processi derivazionali, di molti affissi in competizione, praticamente equivalenti dal punto di vista semantico, e non di rado ritenuti tutti qualitativamente produttivi. Si può dire che il punto di

vista quantitativo riduce significativamente la portata di tale ridondanza, dato che per diversi domini semantici un affisso acquista una chiara preminenza. È il caso dei nomi di qualità, in cui *-ità* prevale largamente, in *tokens* e in *types*, su *-ezza*, *-aggine*, *-*$^{(1)}$*erìa*; dei suffissi verbali, con il primato di *-izzare* su *-eggiare* e *-ificare*; e dei nomi denominali di agente/mestiere, quantitativamente dominati da *-ista* rispetto a *-iere* e *-*$^{(1)}$*aio*.

4. Due misure di produttività per i morfemi derivazionali italiani

Come illustrato in § 1.3, nei nostri lavori abbiamo adottato come misura della produttività il rapporto $P = h/N$ calcolato per i diversi affissi ad uguali valori di N. Tale misura, per essere operativamente affidabile, richiede che i suffissi in questione abbiano frequenze in *tokens* non certo identiche ma confrontabili, in modo che le dimensioni dei sottocorpus variabili a cui fare riferimento non divergano eccessivamente. In Tab. 3 – ripresa da Gaeta e Ricca (2003a: 77) – si riporta una lista di valori di produttività per un sottoinsieme dei 58 affissi presentati nelle tabelle 1 e 2, e precisamente per gli affissi di frequenza medio-alta, compatibili con almeno uno dei valori N = 19.000, 50.000 e 100.000. Le caselle vuote nella tabella sono in parte dovute ad assenza di dati, quando il valore di N per un dato affisso è troppo alto per essere raggiunto anche alla fine dello spoglio di tutto il corpus triennale; oppure, al contrario, per i suffissi ad altissima frequenza, al fatto che il valore di N corrisponde a una dimensione troppo piccola dei sottocorpus, che rischia di distorcere i dati. Nella quarta colonna di Tab. 3 è riportata una seconda possibile valutazione della produttività anch'essa discussa in § 1.3, e cioè semplicemente il numero di *hapax legomena* dell'affisso nel corpus globale. Si può vedere come i due ordinamenti complessivi non mostrino grandissime divergenze, ma non coincidano nel dettaglio.

Entrambe le misure permettono di individuare con chiarezza due regioni estreme nell'ordinamento. Da un lato, i tre suffissi a capo della lista (*-issimo*, *-mente* e *-iano*) si trovano esattamente dove ci si aspetta. Si tratta infatti di suffissi che manifestano una generalità e un livello di trasparenza semantica che li pone al limite tra flessione e derivazione, ed è quindi confortante per la plausibilità linguistica della misura scelta che la loro produttività sia maggiore di quella degli affissi più prototipicamente derivazionali che seguono nella lista. Dall'altro lato, i suffissi in fondo alla Tab. 3 sono unanimemente giudicati del tutto improduttivi (è il caso di *-ore*, *-evole* ed *-izia*) o assai marginalmente produttivi (*-nza* e *-ificare*): anche in questo caso, quindi, la misura quantitativa di P coincide con le aspettative qualitative.

Nella zona centrale, quella degli affissi produttivi, le intuizioni dei linguisti risultano parimenti confermate in alcuni casi di suffissi rivali che

competono sugli stessi domini, in particolare per *-età/-età* rispetto ad *-ezza*, e per *-izzare* rispetto ad *-eggiare*. In questa regione l'allineamento tra le due misure, *P(N)* e il valore assoluto di *h*, non è più completo, e si osserva una tendenza della misura *h* a privilegiare gli affissi ad alta frequenza: così, in termini di *h*, *-(z)ione* risulta superiore a *-mento*, *-età/-età* passa largamente al di sopra di *-bile* ed *-ismo*, e il suffisso più produttivo in assoluto risulta *-mente* anziché *-issimo*. Quest'ultimo risultato non appare molto convincente, tenendo conto che *-issimo* è generalmente considerato flessivo e *-mente* derivazionale (benché entrambi siano certo esponenti non prototipici delle rispettive classi). La discrepanza maggiore tra i due ordinamenti in Tab. 3 è però quella che coinvolge la coppia *-(t)ore/-trice*, e anche in questo caso i valori di *P*, che collocano sullo stesso piano i due suffissi, funzionalmente molto simili e paralleli, si lasciano a nostro avviso preferire rispetto a quelli risultanti dalla considerazione di *h*, che assegnano al meno frequente *-trice* un valore di produttività più che dimezzato rispetto a *-(t)ore*.

Tabella 3 - Alcuni affissi derivazionali dell'italiano ordinati secondo due misure di produttività

Affissi	*P(N)* $\cdot 10^3$			*h in tutto il corpus*
	N = 19.000	*N = 50.000*	*N = 100.000*	
-issimo	25.8	12.9	-	643
-iano	24.3	-	-	615
-mente	-	10.1	6.4	825
-ismo	15.2	8.2	-	448
-bile	11.3	6.3	4.1	409
-ità/-età	-	6.3	3.7	544
-ista	11.3	6.2	3.8	470
-trice	10.8	-	-	224
-(t)ore	-	5.0	3.2	461
-mento	-	4.9	3.1	402
-(z)ione	-	-	2.7	486
ri-/re-	-	3.8	2.3	312
-izzare	7.6	3.8	-	280
-ese	-	3.6	2.2	244
-(t)ura	6.6	3.5	-	189
-ale/-are	-	-	1.9	155
in-	4.1	2.1	1.3	148
-eggiare	4.1	-	-	93
-oso	3.7	1.6	1.0	127
-ezza	2.7	1.3	-	70
-aggio	1.5	-	-	29
-nza	0.7	0.3	0.2	29
-ificare	0.6	-	-	20
-ore	0.4	0.2	-	9
-evole	0.3	-	-	6

-izia	0.0	-	-	0

Per concludere, i materiali presentati in questo capitolo mostrano come la prospettiva quantitativa in morfologia derivazionale offra spunti di ricerca interessanti, sia di rilievo teorico (come si è visto a proposito delle diverse opzioni, probabilistiche e no, per quantificare la produttività), sia di rilievo empirico, fornendo non da ultimo materia di riflessione per quanti si occupano di questioni relative all'accesso al lessico mentale.