Acta Linguistica Hungarica, Vol. 50 (1–2), pp. 89–108 (2003)

ITALIAN PREFIXES AND PRODUCTIVITY: A QUANTITATIVE APPROACH*

LIVIO GAETA – DAVIDE RICCA

Abstract

The quantitative approach to morphological productivity developed by Baayen and collaborators is crucially based on the count of *hapax legomena* in a given, very large textual corpus. In this paper, Baayen's main idea is applied to the little explored domain of Italian prefixation, on the basis of a 75,000,000-token newspaper corpus, and a significant improvement of his procedure is proposed by calculating productivity values at equal token numbers for different affixes. Consequently, variably-sized subcorpora must be sampled to compare affixes displaying different token frequencies. Following this approach, the Italian productive prefixes *ri-* and *in-* can be ranked by productivity within their respective derivational domains, and the impact of different derivational cycles on the measure of productivity can be dealt with satisfactorily.

1. Introduction

In a number of recent contributions, Baayen (1989; 1992; 1993; 2001; see also Baayen–Lieber 1991; Baayen–Renouf 1996; Plag et al. 1999) has suggested relating the notion of productivity to the number of hapax legomena, i.e., words with frequency 1, occurring in a sufficiently large corpus. The proposed measure of productivity P for a given affix is the ratio between the number h of hapax legomena derived by that affix and the number N of all tokens of that affix occurring in the corpus:

(1) $P = h/N$

In mathematical terms, it can be shown (Baayen 1989, 104) that the index (1) is the derivative at point N of the curve $V(N)$, which plots the type number

* This work, developed within the FIRB-project "Italian text and corpus linguistics", has also been partially funded by the Italian Ministry of Education, University and Research (MIUR). The whole paper, as well as the computational work, is the result of the close collaboration of both authors; however, for academic purposes, L.G. is responsible for sections **1**, **2**, **4.1–2** and D.R. for sections **3**, **4.3**, **5**, and **6**.

V for a given affix (i.e., the number of different words derived by that affix) as a function of the token number N of the same affix. To get a concrete illustration, four instances of the curve $V(N)$ are reported in Figure 1, taken from Gaeta–Ricca (2003): they refer to the Italian suffixes *-mente*, forming adverbs, and *-mento*, *-(t)ura* and *-nza*, forming action nouns, sampled from three years of the Italian newspaper *La Stampa*. Examples of the four derivations are given in (2) below:

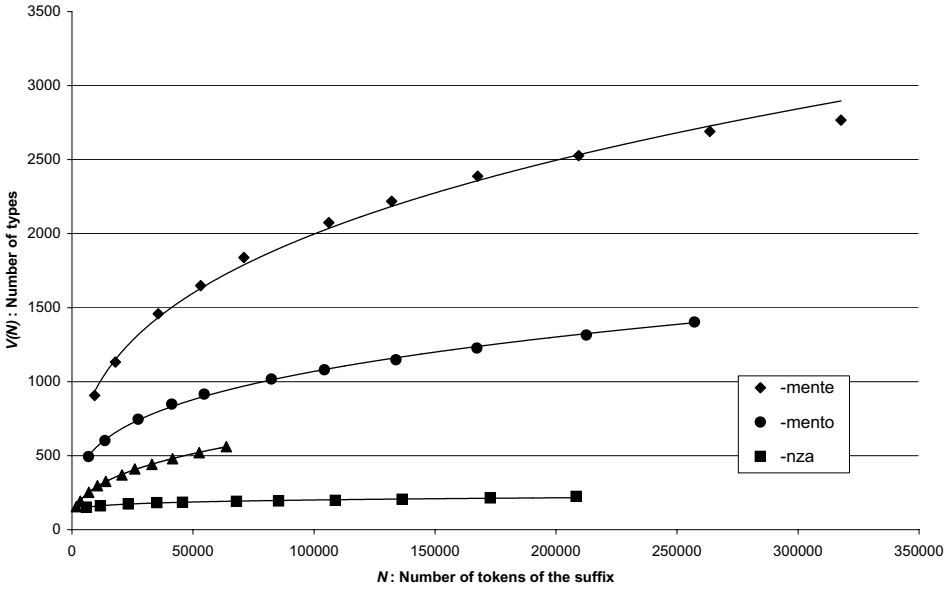

Fig. 1

La Stampa 1996–1998: types increasing curve as a function of N

(2)

	lento	→	lenta-mente	'slowly'
	cambiare	→	cambia-mento	'change'
	decadere	→	decade-nza	'decay'
	mappare	→	mappa-tura	'mapping'

In simpler terms, the ratio in (1) measures the probability of encountering a new type not attested before, i.e., a hapax legomenon, after N tokens of a given affix have been sampled (Baayen 1989, 99ff). The curve $V(N)$ in Figure 1 can be conceived as portraying the growth of the lexical inventory of an affix. The measure of the slope of the curve, i.e., the derivative at a certain point, gives the speed at which new types of a certain affix come

out from the sample. If an affix is even minimally productive, new types will be encountered: the value of V may only increase as N increases—mathematically it is a non-decreasing monotonic function. However, for every affix the increasing rate of $V(N)$ will decrease as we proceed in the sample, since it will become more and more probable that new tokens of the affix will be occurrences of already attested types. Hence, as also pointed out by Baayen–Lieber (1991, 837), productivity $P(N)$ is a monotonic decreasing function and even tends to zero for N tending to infinity.

It is evident from Figure 1 that the curves $V(N)$ for the four suffixes increase at different rates, thus qualifying for different values of productivity. Whereas the curve of the suffix *-nza* immediately reaches almost the whole number of possible types and then remains stable, approximating a horizontal line, for the other suffixes the curve is clearly still increasing, although with different slopes, at the end of the sampling procedure. This is in essence the quality of the index P proposed by Baayen: investigating the increasing rate of new types formed with a certain affix in a corpus provides a clue for measuring the availability of a certain word formation rule.

The approach outlined above has been often discussed and diversely evaluated (cf. van Marle 1992; Plag 1999, 23ff; Bauer 2001, 150ff). In this paper, we will propose a revised procedure to calculate the productivity rates; then we will devote our attention to Italian prefixes and their ranking among productive Italian affixes, and especially we will discuss the impact of different derivational cycles on the measure of productivity.[1]

2. A variable-corpus approach

Most of the criticism raised against Baayen's approach is ultimately related to the presence of N in the denominator of (1), which results in underestimating the value of P for affixes with very high token frequency.[2] However, we argue

[1] We will not consider a further productivity measure proposed by Baayen (1993, 192), the so-called 'hapax-conditioned degree of productivity' P^*, which is basically given by the absolute number of hapaxes formed with a certain affix which occur in the whole corpus. As pointed out by Bauer (2001, 155), the main problem with P^* is that it "asks 'What proportion of new coinages use affix A?' rather than asking 'What proportion of words using affix A are new coinages?'. It is this latter which seems a more relevant question to ask".

[2] A very clear instance of such underestimation effect in Baayen's data is provided by the English suffix *-ly*, as discussed in Plag (1999, 113; for more details, see Gaeta–Ricca 2003).

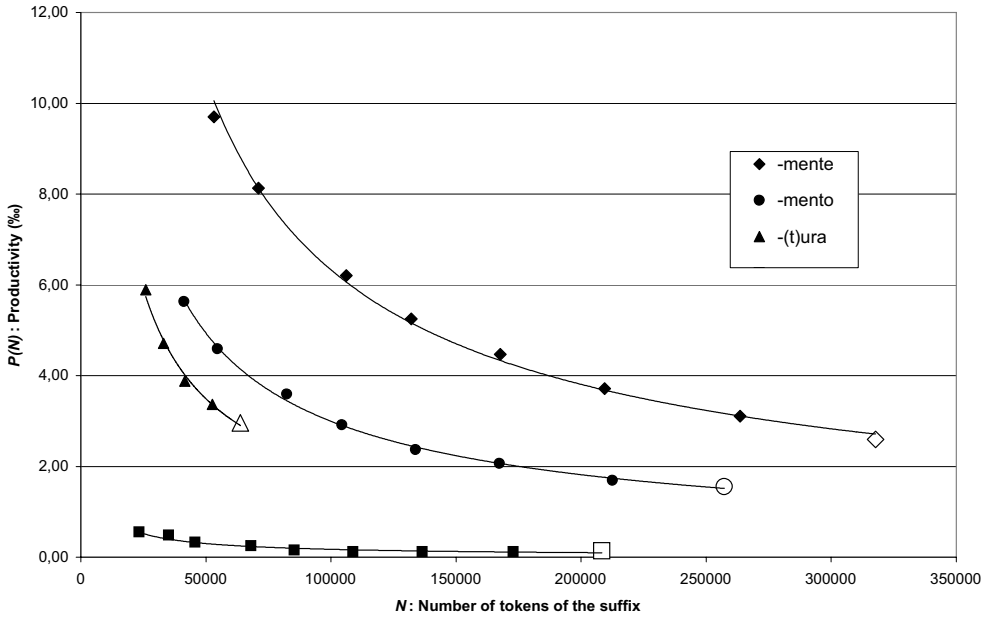

Fig. 2
Productivity as a function of N

that this underestimation effect is not related to the ratio (1) in itself, but rather to the way it is applied.

In fact, Baayen's data are always obtained by taking N as the number of affix tokens in the whole corpus, irrespective of the token frequency of the different affixes. Baayen's procedure can be graphically understood by referring to Figure 2, which displays $P(N)$ as a function of N for the four suffixes listed in (2). In Baayen's approach, the final values of the curves are compared: in Figure 2, they have been emphasized by the bigger size of the endpoints. However, these values lie on different points of the horizontal axis, due to the different token frequencies of the suffixes. Thus, for a rather infrequent suffix such as *-(t)ura*, the final value of the curve, corresponding to the sampling of the whole corpus, lies at a N value reached by a much more frequent suffix such as *-mento* after less than one year of its occurrences. For *-mento*, the final point of the curve lies much further in the horizontal axis, when the function $P(N)$ has further decreased. Therefore, very frequent

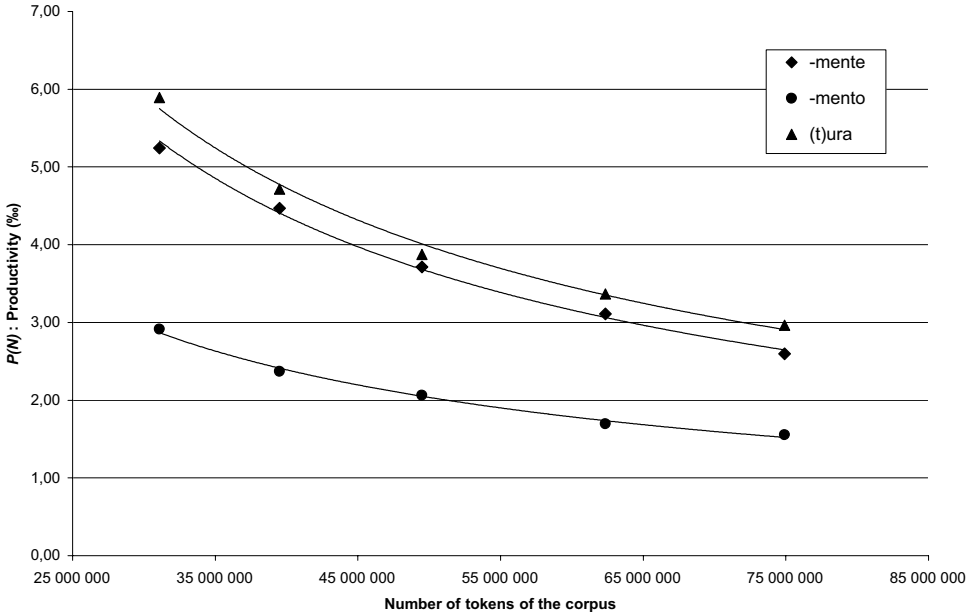

Fig. 3

Productivity as a function of fixed corpus chunks

affixes turn out to be disfavored because of the decreasing character of the function $P(N)$.

The distorting effect of this unbalanced comparison of $P(N)$ values can be seen very clearly in Figure 3, where P is plotted as a function of the total number of tokens of the corpus, and not of the single suffixes, thus representing graphically the data actually compared by following Baayen's procedure. The P-curve for a rather unfrequent suffix such as *-(t)ura* jumps over the curves for the much more frequent suffixes *-mento* and *-mente*.

For the reasons outlined above, when comparing values of productivity for affixes with different token frequencies, we did not adopt Baayen's procedure. Rather, we calculated $P(N)$ for equal values of N. Graphically, this means comparing the values of the curves in Figure 2 for the same values on the horizontal axis.

Of course, to implement our procedure the values of *P(N)* for different affixes have to be extracted from differently-sized corpora, whose size is inversely proportional to the token frequency values of the affixes. Thus, a necessary presupposition for the reliability of our method is that affix fre-

quencies remain stable throughout the sampling. To meet this requirement, we chose as a corpus three years (1996–1998) of the newspaper *La Stampa*, around 75,000,000 tokens in all. Our corpus is structured in 36 subcorpora of progressively increasing size (1 to 36 months), so that for each subcorpus the value of *P(N)* can be computed independently, selecting for each given affix the subcorpus best approaching the desired value for N. Values of $P(N)$ corresponding to the exact values of N can then be evaluated by linear interpolation.[3]

This structuring of our corpus easily allows us to check uniformity of affix frequencies. In Table 1, the data concerning the frequency of the suffixes mentioned above are reported measured on one single year, on two years and on the three years of the whole corpus.

Table 1

Token frequency in the corpus

SUFFIX	TOKEN FREQUENCY (‰)		
	***La Stampa* '96** 24 915 369	***La Stampa* '96–'97** 49 485 568	***La Stampa* '96–'98** 74 917 798
-(t)ura	0.8	0.8	0.9
-nza	2.7	2.8	2.8
-mento	3.3	3.4	3.4
-mente	4.3	4.2	4.2

Apart from minor fluctuations, the token frequency is fairly stable, as the sampling goes on. This makes our approach feasible, since comparing data extracted from different subcorpora is not—or only minimally—distorted by a non-uniform affix token distribution throughout the corpus.

As a further advantage, a newspaper corpus appears to be adequate for quantitative studies because it comprises different speech registers and different text types, as also argued by Baayen–Renouf (1996), whose *Times* corpus is fairly comparable to ours both in size and quality.

[3] To be sure, our procedure does not allow unrestricted comparison between any affix. From a linguistic point of view, comparing affixes with extremely divergent token frequencies might be questionable. In any case, data are not reliable if referred to subcorpus sizes below 6 million tokens (i.e., about 3 months), since under this threshold P-values become floating when calculated on different subcorpora of equal size (cf. Gaeta–Ricca in press). Practically, this means that given the current corpus size we are not able to compare affixes whose token frequency ratio is lower than 3 : 36 (i.e., 1 : 12).

3. Prefixation in Italian

The status of Italian prefixation is rather different from suffixation (cf. at least Bisetto et al. 1990 and Iacobini 1999 for an overall picture). This comes out quite clearly by a quantitative look at the token frequencies. While Italian suffixes which can be termed as qualitatively productive distribute rather smoothly on a wide frequency range, from rates of occurrence as low as one part on one hundred thousand to rates hundred times higher or even more (cf. Thornton 1998), qualitatively productive prefixes seem to fall in three rather sharply distinct categories, listed in (3) below. Simplifying somewhat, there are only two very frequent productive prefixes, namely verbal *ri-* 're-' and negative adjectival *in-* 'un-/in-'. A second group consists of the verbal parasynthetic prefixes (chiefly *in-*, *ad-* and *s-*), object of much theoretical debate (cf. Montermini 2002, 265ff for a survey): they will not be dealt with here. Finally, there is an interesting group of 'recent' prefixes, often of learned origin although nowadays pretty compatible with non-learned bases, of which the most common are listed in (3c):

(3) (a) HIGH-FREQUENCY PREFIXES:

ri- ('re-') V → V
in- ('in-'/'un-') A → A

(b) PARASYNTHETIC PREFIXES (A → V and N → V):

bello	'beautiful'	→	**im**bellire	'become beautiful'
			abbellire	'embellish'
vecchio	'old'	→	**s**vecchiare	'make less old'

(c) LOW-FREQUENCY PREFIXES:

EVALUATIVES:	**iper**protettivo, **maxi**schermo, **mega**concerto, **micro**criminalità, **mini**gonna, **super**leggero, **ultra**piatto, etc.
QUANTITATIVES:	**bi**motore, **mono**uso, **multi**culturale, **poli**sportivo, etc.
SPATIO-TEMPORAL:	**contro**corrente, **ex**-presidente, **neo**senatore, **post**coloniale, **vice**presidente, etc.
OTHERS:[4]	**anti**droga, **auto**distruzione

In the following sections, we will mainly deal with the two high-frequency prefixes *ri-* and *in-*. In section **6** we will briefly consider the low-frequency evaluative prefixes, which raise interesting questions on productivity.

The token frequencies for *ri-* and *in-* are reported in Tables 2 and 3, jointly with those of the main Italian derivational suffixes in the deverbal

[4] In (3c) we basically follow the classification proposed in Montermini (2002, 105), keeping apart, however, the two prefixes *anti-* and *auto-*, which are the least easy to classify semantically and also are the comparatively most frequent among those listed under (3c).

and deadjectival domains (the latter data are taken from Gaeta–Ricca 2003). They have been calculated on the whole corpus of three years; however, as remarked above, the values are very stable much below that size.

Table 2

Frequency data for *ri-* compared to some major deverbal suffixes
(full corpus of 36 months—75,000,000 tokens)

DEVERBAL AFFIX	N OF TOKENS	FREQ. (‰)
-(z)ione	1 043 979	13.9
ri- (all cycles)	500 912	6.7
ri- (outmost cycle)	270 066	3.6
-mento	257 216	3.4
-nza	208 365	2.8
-bile	102 904	1.4
-(t)ura	63 800	0.9

Table 3

Frequency data for *in-* compared to some major deadjectival suffixes
(full corpus of 36 months—75,000,000 tokens)

DEADJECTIVAL AFFIX	N OF TOKENS	FREQ. (‰)
-ità/-età	356 857	4.8
-mente	317 725	4.2
in- (all cycles)	202 744	2.7
in- (outmost cycle)	146 982	2.0
-ezza	69 090	0.9
-issimo	51 636	0.7

Notice that the for the prefixes *ri-* and *in-* the tables report two markedly different values depending on whether inner-cycle derivations are incuded or not: this issue will be dealt with in section **4.3**. For all the suffixes, only outmost-cycle values are given. From Tables 2 and 3, one can observe that *ri-* and *in-* belong to the core of Italian derivational strategies from the point of view of their token frequency. In the following sections we will try to assess if the same holds true for productivity as well, applying the methodology for calculating the productivity illustrated in section **2** above.

4. What counts as a token/type of a given affix?

Before presenting the results of our investigations, we have to make it clear what we counted as a token/type of the affixes in question. The issue is not

as trivial as it could seem, and has been the object of much debate in the literature (cf. Plag 1999, 108; Bauer 2001, 151). We cannot enter into much detail here (cf. Gaeta–Ricca 2003); we will limit our discussion to the main difficulties concerning the two prefixes under investigation.

4.1. Questions about allomorphy

Issues concerning allomorphy and segmentation are not really problematic. Both prefixes display a little amount of non-automatic, i.e., not strictly phonology-driven, allomorphy, especially when compared with the relevance of allomorphy in cases like *-(z)ione* or *-(t)ura* (cf. Thornton 1990–1991; Rainer 2001; Gaeta–Ricca in press). The main instances of allomorphy are shown in (4) below:

(4) (a) **in-**
- progressive assimilation of /n/: **il**legale 'illegal', **ir**rilevante 'irrelevant'
- other minor allomorphies: scusabile 'forgivable' → in**e**scusabile 'unforgivable'

(b) **ri-**
- lowering before /i/: **re**idratare, **re**imporre, **re**inventare
- other minor allomorphies: incontrare 'meet' → **r**incontrare 'meet again'

The corresponding items clearly count as types of the prefixes and have been accordingly included into our counts.

4.2. Polysemy and lexicalization

The semantic problems are more thorny, especially for *ri-*. The latter prefix displays in fact an extended polysemy, which can be described in terms of three basic meanings: the repetitive *ri-* meaning 'again' which is the most common and the most typical meaning for the new formations; the reversal/repair *ri-* 'back' which implies the restoring of a preceding situation; finally, the intensive *ri-*, which is in fact a very vague label and embraces rather divergent cases, ranging from instances where the semantic contribution of the prefix is nearly zero (*tornare/ritornare*), to cases of intensification proper (*chiedere/richiedere*), up to cases of marked lexicalization and semantic drift (*guardare/riguardare*). Examples are given in (5):

(5) (a) repetition *ri-*: *giocare* 'play' → *rigiocare* 'play again',
leggere 'read' → *rileggere* 're-read'

(b) reversal/repair *ri-*: *spedire* 'send' → *rispedire* 'send back'
conquistare 'conquer' → *riconquistare* 'reconquer'

(c) 'intensive' *ri-*: *tornare* ≈ *ritornare* 'come back'
chiedere 'ask' → *richiedere* 'request'
guardare 'watch' → *riguardare* 'regard, concern'

Reversal and repair readings have been included under the same label since their selection basically depends on the semantics of the base predicate (verbs involving movement vs. change-of-state). Moreover, the reversal/repair reading can be further subsumed under the more general meaning 'repetition' (as argued by Rainer 1993, 361 referring to Spanish), provided that repetition is meant to refer to the process only, without implying identity of participants. The 'intensive' meaning, on the contrary, stands clearly apart and turns out to occur with a limited amount of bases (cf. Iacobini in press).

Despite this admittedly wide semantic range, all words belonging to any of these three categories have been included into the counts, for two kinds of reasons. First, we believe that apart from extreme cases of lexicalization, even terms like *riguardare* are still able to activate the prefix *ri-* in the mental lexicon. If it is so, they should give their contribution to the token amount of the prefix. Second, the three meanings reported in (5) constitute a polysemic chain, since they are not always easily separable and often co-occur with the same base (especially the 'again' and 'back' meanings). In particular, the repetitive meaning appears to be available in practically any case and is attested in our corpus even when it has to confront with a frequent and entrenched non-compositional meaning, as is the case for *richiedere* or *riguardare*. For instance, in our corpus sentences like the following are easily found:

(6) Se riguardiamo i cinegiornali fine Anni Sessanta
'If we go back to the newsreels of the late sixties' (*La Stampa*, 20-5-'97, p. 24)

On the other hand, for both prefixes we excluded from our counts three small classes of items: (a) those without any identifiable base, at least synchronically; (b) a few cases in which the prefixes select a lexical category different from their main domain (verbs for *ri-* and adjectives for *in-*); and finally, (c) some really extreme cases of opaque lexicalizations. Examples are given in (7):

(7) (a) *incolume* 'unhurt', *insulso* 'dull'
ripetere 'repeat', *ricordare* 'remember'

(b)

denominal *in-*:	N → N:	*in-azione*	'inactivity',	*in-successo*	'failure'
	N → A:	*in-forme*	'shapeless',	*in-colore*	'colourless'
denominal *ri-*:	*ri-esame*	're-examination',		*re-ingresso*	're-entering'
parasynthetic *ri-*:	*ri-modernare*	'refurbish',		*ri-bassare*	'lower'

(c) *infermo* 'sick' vs. *fermo* 'steady, motionless'
rilevare 'notice' vs. *levare* 'take away, remove'

Concerning (7c), we excluded only those cases which could clearly not be dealt with in terms of polysemic chains like (5), although we are aware that a certain amount of arbitrariness cannot be avoided. At any rate, the items which have not been included into our count amount to relatively few types: for both prefixes, they constitute about 10% of a maximal choice including nearly all verbal items beginning with *ri-* or items beginning with *in-* and carrying some negative meaning. It is true that some of the excluded items do have a high token frequency and could therefore lower significantly the P values if included into the count. However, the problems of delimiting the field of items to be included are on the whole less serious than for many important Italian suffixes, and the unavoidable margin of arbitrariness which still remains is unlikely to affect the quantitative results heavily.

4.3. Inner-cycle derivations

The prefixes *ri-* and *in-* are challenging for a quantitative evaluation of productivity from another point of view. They both occur in many derived words where they do not constitute the outmost derivational cycle. In fact, from *ri-*verbs one can easily further derive, for instance, action and agent nouns and verbal adjectives; from *in-*adjectives there is plenty of derivation of quality nouns and manner adverbs:

(8)	(a)	***ri-*:**	action nouns	[*ri*[*fonda*]]*zione*	'refound-ation'
			agent nouns	[*ri*[*fonda*]]*tore*	'refound-er'
			possibility adjectives	[*ri*[*fonda*]]*bile*	'refound-able'
	(b)	***in-*:**	quality nouns	[*in*[*util*]]*ità*	'useless-ness'
			manner adverbs	[*in*[*util*]]*mente*	'useless-ly'

It is not clear whether the words in (8) should be considered as tokens of the *ri-* and *in-* prefixes respectively. This has never been done for counts on suffixes: *nationalization* has not been counted as a token of *-ize*, and so on, as observed by Plag (1999, 29), who points out the problem. One

could argue that a different approach might be adopted for prefixes, given their more salient position at the word beginning. Notice that for prefixes the choice of including all inner-cycle derivations is also easier from an operational point of view, since it amounts to include all words beginning with the prefix under investigation. However, if we want to compare the productivity rates of prefixes and suffixes, we should take the same attitude towards both of them: either limiting our counts to the outmost cycle, or including inner derivations.

Moreover, the possibility of these two options (counting and not counting inner derivations) raises an interesting empirical question: does the one or the other choice make a great difference in the quantitative results? If it does, this would cast some doubts on the reliability of the whole method. We will try to give an empirical answer to this question concerning the two major Italian prefixes. The results obtained by applying Baayen's procedure and ours are shown in Tables 4 and 5:

Table 4

Comparing P-values for *ri-* obtained by applying the fixed and the variable-corpus approach

	ri-				
	N tokens	V types	h hapaxes	Baayen's P (calculated on the whole 36-months corpus)	$P(N = 270066)$ (calculated on a 19-months subcorpus plus interpolation)
all cycles	500 912	989	325	0.7	**1.1**
outmost cycle only	270 066 (53.9%)	935	312	**1.2**	

Table 5

Comparing P-values for *in-* obtained by applying the fixed and the variable-corpus approach

	in-				
	N tokens	V types	h hapaxes	Baayen's P (calculated on the whole 36-months corpus)	$P(N = 146982)$ (calculated on a 26-months subcorpus plus interpolation)
all cycles	202 744	779	140	0.7	**0.9**
outmost cycle only	146 982 (72.5%)	767	148	**1.0**	

The inner-cycle contribution turns out to be quite relevant in both cases in terms of tokens. Much less so for the types: as can be expected, only a little amount of prefixed items with *ri-* and *in-* do occur in inner-cycle derivations only (to give a concrete example, if we find a word like *rifondabile* in the corpus, it is highly probable that we will find the word *rifondare* as well: as for the *ri-*prefix, the two words belong to the same type). The same happens *a fortiori* for the hapaxes (indeed, their number can even be reduced by the inclusion of inner cycle derivations, as is the case for *in-*). Consequently, the value of Baayen's P is sensibly lower if one includes also inner cycle derivations in the count, especially for *ri-* where they amount to about half of the tokens.

To be sure, a lower value for the all-cycle count might make sense linguistically, since with these two prefixes inner derivations are overwhelmingly found with the most entrenched and lexicalized words, as for instance *indennizzare* 'to indemnify' from *indenne* 'unharmed', *immobiliare* 'building (society)' from *immobile* 'immovable', etc. Therefore, it is legitimate to predict a lowering effect on productivity. However, if the two counts diverge too sharply, it becomes hard to link the prefix under investigation with a single well-defined quantitative value which could rank it consistently among other derivational affixes.

The impact of internal cycles on productivity values is much lower if we follow the procedure outlined in section **2** above. In this case, we have to compare the P's for the same value of N. The maximum value available for N to compare the two counts—all-cycles and outmost-cycle only—is the one reached by the outmost-cycle count when the full corpus is sampled. We should then make the all-cycle count on a suitably sized subcorpus, such as to get a value of N near to the one reached on the full corpus when only the outmost cycle is taken into account. This is well approximated with 19 months for *ri-* and 26 months for *in-*, to which a tiny correction by linear interpolation is added to reach the value of P corresponding to the exact value of N. Comparing Tables 4 and 5, it can be seen that within the variable-corpus approach the results—printed in boldface in the tables—show a substantial alignment of the data for the two counts. Summing up, whereas the inner cycles strongly influence the productivity calculated with Baayen's procedure (i.e., on the full 3-year corpus), the gap between the two counts is markedly reduced by considering the values of P for equal values of N.

5. Main deverbal and deadjectival affixes ranked by productivity

We are now ready for a final assessment of the productivity rates of the two prefixes investigated. In Tables 6 and 7, they are compared with the other high-frequency deverbal and deadjectival affixes listed in Tables 2–3 above:

Table 6

P-values for deverbal affixes

DEVERBAL AFFIXES (N= 100 000)	
AFFIX (outmost cycle only)	P (N=100 000) ‰
-bile	4.0
-mento	3.1
-(z)ione	2.8
ri-	2.3
-nza	0.1

Table 7

P-values for deadjectival affixes

DEADJECTIVAL AFFIXES (N= 50 000)	
AFFIX (outmost cycle only)	P (N=50 000) ‰
-issimo	12.7
-mente	10
-ità	6.6
in-	2.1
-ezza	1.2

A useful value for N has been chosen in order to maximize the number of affixes which can be compared in both cases. The lower value of $N = 50{,}000$ chosen for the deadjectival ranking allows us to include two more interesting suffixes, namely *-ezza* (*bello* 'beautiful' → *bellezza* 'beauty') and *-issimo* (*bello* 'beautiful' → *bellissimo* 'very beautiful'), whose total frequency values do not reach $N = 100{,}000$. This means that the two rankings in Tables 6 and 7 cannot be directly compared, as $P(N)$ is a steadily decreasing function, and therefore its values for $N = 50{,}000$ are globally higher than those for $N = 100{,}000$. Most affixes in both tables, however, could be directly compared without difficulty by selecting a common value of N.

The comparison with the elative suffix *-issimo* is particularly interesting, since this suffix is notoriously at the border between inflection and deriva-

tion,[5] and should therefore display the highest productivity among the affixes considered, which is indeed the case. The second-ranking affix is another borderline suffix, namely *-mente*, which some analyses would even assign to inflection (cf. e.g., Haspelmath 1996, 49f on its close English equivalent *-ly*; for a discussion see Ricca 1998).

Tables 6 and 7 show that both *ri-* and *in-* are to be included within the productive segment of Italian derivation, although their relevance in productivity is less high than in token frequency, especially for *in-*. The values for *ri-* place the prefix relatively near to the highly productive suffixes for action nouns, while *in-* falls clearly below the main adjectival formations, though doubling the productivity of a still productive suffix like *-ezza*. As for the comparison between the two prefixes, the higher productivity of *ri-* with respect to *in-* can be inferred from Tables 6 and 7, taking into account the decreasing character of the function $P(N)$, since the value for *in-* is lower than the one for *ri-* even if the latter is calculated for a value of N which is twice higher. More explicitly, making a proper comparison at equal N (not reported in the tables), we get *ri-* values clearly above *in-* values. At $N = 50{,}000$, P is 3.8 for *ri-* against 2.1 for *in-*, and at $N = 100{,}000$, P is 2.3 for *ri-* against 1.4 for *in-*. The lower value for *in-* with respect to *ri-* matches linguists' expectation since the former has a learned flavour and undergoes relevant semantic restrictions (cf. Iacobini in press). Looking at the list of low-frequency items for *in-*, its productivity—which is nevertheless considerable—comes out as being mainly due to its combination with the deverbal *-bile* adjectives, in its turn a very productive derivational process in Italian, and partly with past participles in *-to* (among the hapaxes in our corpus, we found *incapibile* 'un-understandable', *inaccoglibile* 'un-receivable', *impraticato* 'un-practised', *inabituato* 'un-accustomed', etc.).

One should probably expect still a higher value for *ri-*, nearer to the other most productive derivational processes listed in Tables 6 and 7. A factor limiting its productivity may be the fact that *ri-* is the only verbal affix taken into account: verbs are on the whole less easy to form than nouns and adjectives, as can be seen from the size of the respective type inventories in any large dictionary.

[5] While the Italian grammatical tradition usually recognizes *-issimo* as the exponent of the inflectional category of gradation, other linguists treat it more or less along with evaluative suffixation and therefore place it on the derivational side (cf. Rainer 1983; in press).

6. The case of low-frequency prefixes

Among the low-frequency prefixes mentioned in (3c), we investigated the evaluative group in detail. With the exception of *super-*, these items are around one-hundred times less frequent than the two prefixes considered above, and therefore cannot be directly compared with them (see fn. 3). Their token frequencies are reported in Table 8, together with the number of their types and hapaxes:

Table 8

Frequency data for evaluative prefixes

AFFIX	N OF TOKENS	FREQ. (‰)	V (TYPES)	h (HAPAXES)
super-	8966	0.120	1147	667
micro-	2869	0.038	437	276
mini-	1830	0.024	612	383
iper-	1675	0.022	389	276
maxi-	1617	0.022	365	230
ultra-	1557	0.021	302	197
mega-	1399	0.019	426	252

However, one could consider the possibility of using a medium-frequency affix as a bridge to fill the gap. A good candidate is *-issimo*, which is also semantically akin to the evaluative set. The suffix *-issimo* is about seven times more frequent than *super-* and can thus be compared with it. On the other hand, *super-* can be compared with the other—still much less frequent—evaluative prefixes, which can thus also be ranked, at least indirectly, with respect to *-issimo* itself. The somehow astonishing result is given in Table 9 for the subset of augmentative/meliorative prefixes:

Table 9

The elative *-issimo* compared with some low-frequency evaluative prefixes

AFFIX	P (N= 8966) ‰	P (N=1400) ‰
mega-		180
iper-		174
super-	74.4	165
maxi-		151
ultra-		129
-issimo	41.2	

The prefix *super-* displays a productivity rate which is nearly two times the already very high value for *-issimo*. The other evaluative prefixes in Table 9, when compared at equal N, show a remarkable uniformity in their productivity values, all ranging within $\pm 10\%$ of the P value for *super-*, except for *ultra-* whose P is slightly (20%) lower. It should be remembered that *-issimo* was the affix ranking highest among all those discussed until now. Does it make sense, linguistically, that such low-frequency items exhibit a top value in productivity? Indeed, it could also be the case that such methods are simply unreliable if applied to affixes of too low frequency, even when data are calculated on huge corpora. As a matter of fact, these items have such a high value precisely because they are in a way 'newcomers' to the lexicon. While they occur in very few firmly established words, like *minigonna* 'miniskirt', *maxischermo* 'maxi-screen' or *microcriminalità* 'micro-criminality', they combine very freely—but also rather loosely—with any sort of bases, giving raise to a huge amount of nonce formations like *megacena* 'mega-dinner', *mega-friggitrici* 'mega-fryer', *megaorologio* 'mega-watch', *mini-emirato* 'mini-emirate', *mini-proibizionismo* 'mini-prohibitionism', *mini-epurazione* 'mini-epuration' and so on.[6] The peculiar character of this group of items with respect to most word-formation processes is also confirmed by two further well-known properties, extensively discussed by Montermini (2002, 170ff). First, they can be factorized in coordinate structures as in (9):

(9) collegamenti internet su maxi e mini schermi (*La Stampa* 5-12-'97, 23)
'Internet connections on maxi- and mini-screens'

in un super o ipermercato (24-9-'97, 24)
'in a super- or hyper-market'

Moreover, they can occur as free forms in adjectival position with the very same meaning they have as prefixes:

(10) *una serie davvero* mega (25-8-'98, 22; cf. *megaserie* 'mega-serial TV')
il bagagliaio è mini (22-2-'96, 34; cf. *minibagagliaio* 'mini-boot')
i concorsi continueranno ad essere maxi (10-12-'97, 5; cf. *maxiconcorso* 'maxi-competition')
la criminalità micro *e* macro (24-6-'97, 1; cf. *microcriminalità* 'micro-criminality')

Examples like (9) and (10), all taken from our corpus, further support the idea that such items do not fully behave as derivational items, but rather bor-

[6] To have a quantitative idea, notice that for the prefixes listed in Table 8 the number of types whose token frequency in our corpus exceeds 1:1,000,000 is extremely low: *super-* 13, *micro-* 8, *mini-* 2, *iper-* 3, *maxi-* 4, *ultra-* 3, *mega-* 2.

der on syntax, and therefore their productivity cannot be straightforwardly compared with the one displayed by core instances of bound derivational processes. At any rate, we would like to leave the question open for further research.

7. Conclusion

To sum up, in our contribution we hope to have proposed a significant improvement of the quantitative approaches on productivity which rely on the counting of hapaxes in a wide text corpus and are mainly linked to the name of Baayen and collaborators.

The key point is the suggestion of comparing productivity values across affixes for equal values of their token number. In this way, those inconsistencies are avoided which come up when affixes with different token frequency are compared with reference to a corpus of fixed size: the latter procedure unavoidably results in a heavy underestimation of the productivity for the most frequent affixes. The variable-corpus procedure, on the contrary, allows a consistent ranking by productivity of affixes within a given derivational domain.

Moreover, the procedure suggested here seems to be particularly suitable for treating those prefixes, like *ri-* and *in-* in Italian, which display a great amount of inner-cycle derivations. Referring to a fixed number of tokens succeeds in minimizing the lowering impact that the inclusion of inner cycle derivations would otherwise have on the count.

References

Baayen, Harald R. 1989. A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation. Ph.D. dissertation, Vrije Universiteit, Amsterdam.

Baayen, Harald R. 1992. Quantitative aspects of morphological productivity. In: Booij – van Marle (1992, 109–49).

Baayen, Harald R. 1993. On frequency, transparency and productivity. In: Geert Booij – Jaap van Marle (eds) Yearbook of Morphology 1992, 227–54. Kluwer, Dordrecht.

Baayen, Harald R. 2001. Word-frequency distributions. Kluwer, Dordrecht.

Baayen, Harald R. – Rochelle Lieber 1991. Productivity and English word-formations: a corpus-based study. In: Linguistics 29 : 801–43.

Baayen, Harald R. – Antoinette Renouf 1996. Chronicling the Times: productive lexical innovations in an English newspaper. In: Language 72 : 69–96.

Bauer, Laurie 2001. Morphological productivity. Cambridge University Press, Cambridge.

Bisetto, Antonietta – Rossella Mutarello – Sergio Scalise 1990. Prefissi e teoria morfologica. In: Berretta Monica – Piera Molinelli – Ada Valentini (eds) Parallela 4. Morfologia, 29–41. Gunter Narr, Tübingen.

Booij, Geert – Jaap van Marle (eds) 1992. Yearbook of Morphology 1991. Kluwer, Dordrecht.

Gaeta, Livio – Davide Ricca 2003. Productivity in Italian word formation: a variable-corpus approach. Manuscript. University of Turin.

Gaeta, Livio – Davide Ricca in press. Corpora testuali e produttivita morfologica: i nomi d'azione italiani nelle annate della Stampa. In: Roland Bauer – Hans Goebl (eds) Parallela IX. Testo – variazione – informatica / Text – Variation – Informatik (Salzburg 1–4 November 2000). Egert, Wilhelmsfeld.

Haspelmath, M.?? 1996. Word-class-changing inflection and morphological theory. In: Geert Booij – Jaap van Marle (eds) Yearbook of Morphology 1995, 43–66. Kluwer, Dordrecht.

Iacobini, Claudio 1999. I prefissi dell'italiano. In: Paola Beninca – Alberto M. Mioni – Laura Vanelli (eds) Fonologia e morfologia dell'italiano e dei dialetti d'Italia. Atti del XXXI Congresso della Societa di Linguistica Italiana, 369–99. Bulzoni, Roma.

Iacobini, Claudio in press. Prefissazione. In: Maria Grossmann – Franz Rainer (eds) La formazione delle parole in italiano. Niemeyer, Tübingen.

van Marle, Jaap 1992. The relationship between morphological productivity and frequency: a comment on Baayen's performance-oriented conception of morphological productivity. In: Booij – van Marle (1992, 151–63).

Montermini, Fabio 2002. Le systeme préfixal en italien contemporain. Ph.D. dissertation, Université de Paris X, Paris.

Plag, Ingo 1999. Morphological productivity. Structural constraints in English derivation. Mouton de Gruyter, Berlin & New York.

Plag, Ingo – Chris Dalton-Puffer – Harald R. Baayen 1999. Morphological productivity across speech and writing. In: English Language and Linguistics 3 : 209–28.

Rainer, Franz 1983. L'intensificazione di aggettivi mediante '-issimo'. In: Maurizio Dardano – Wolfgang U. Dressler – Gudrun Held (eds) Parallela. Atti del 2° convegno italo-austriaco, 94–102. Gunter Narr, Tübingen.

Rainer, Franz 1993. Spanische Wortbildungslehre. Niemeyer, Tübingen.

Rainer, Franz 2001. Compositionality and paradigmatically determined allomorphy in Italian word-formation. In: Chris Schaner-Wolles – John Rennison – Friedrich Neubarth (eds) Naturally! Linguistic studies in honour of Wolfgang Ulrich Dressler presented on the occasion of his 60th birthday, 383–92. Rosenberg & Sellier, Torino.

Rainer, Franz in press. Internettissimo. Internet come strumento di lavoro per il morfologo: le restrizioni di *-issimo*. In: Franz Rainer – Achim Stein (eds) Die neuen Medien als Instrument linguistischer Forschung. Peter Lang, Bern.

Ricca, Davide 1998. La morfologia avverbiale tra flessione e derivazione. In: Giuliano Bernini – Pierluigi Cuzzolin – Piera Molinelli (eds) Ars Linguistica. Studi offerti da colleghi ed allievi a Paolo Ramat in occasione del suo 60° compleanno, 447–66. Bulzoni, Roma.

Thornton, Anna M. 1990–1991. Sui deverbali italiani in *-mento* e *-zione* (I–II). In: Archivio Glottologico Italiano 75 & 76 : 169–207 & 79–102.

Thornton, Anna M. 1998. Quali suffissi nel vocabolario di base? In: Federico Albano Leoni – Daniele Gambara – Stefano Gensini – Franco Lo Piparo – Raffaele Simone (eds) Ai limiti del linguaggio. Vaghezza, significato e storia, 385–97. Laterza, Bari.

Address of the authors: Livio Gaeta – Davide Ricca
Dipartimento di Scienze del Linguaggio
University of Turin
Via S. Ottavio 20
I-10124 Turin
Italy
{livio.gaeta|davide.ricca}@unito.it