Frequency and productivity in Italian derivation: A comparison between corpus-based and lexicographical data

Livio Gaeta & Davide Ricca

Although frequency is claimed to be a relevant parameter in order to evaluate the productivity and the availability of word formation processes, regrettably few studies primarily deal with it. Italian is no exception in this respect. In this paper, frequency data (both in tokens and types) are provided concerning 58 Italian derivational affixes totalling around 5,000,000 tokens and 30,000 types in a 75,000,000-tokens newspaper corpus. Moreover, two different kinds of hapax-based quantitative evaluations of productivity are applied to the corpus data. Type frequencies and productivities are then compared to those obtainable from lexicographical sources. On the whole, dictionary-based evaluations turn out to be much less reliable descriptors of the affix behaviour with respect to the corpus-based measurements.[*]

1. Introduction

In the current theoretical debate, the notion of frequency has acquired a central role to evaluate the status of word formation processes. Therefore, it should be indispensable to have at one's disposal systematic investigations on the frequency of single affixes. With the increasing availability of electronic corpora, this task is enormously facilitated. Nonetheless, it must be complained that only few studies primarily deal with this aspect. Especially for Italian, despite the increasing evidence supporting the role played by frequency in morphological analysis and production (see Laudanna & Burani 1999 for an overview), empirical research is scarce. In what follows we would like to assess the relation between the two notions of frequency and productivity on a dictionary-oriented and on a corpus-oriented basis.

1.1. The use of dictionaries in word formation studies: Questions and problems

A number of word formation studies rely on dictionaries to evaluate the 'vitality' of affixes. Thus, for instance, Neuhaus (1971:173) assumes that the degree of productivity of affixes is directly related to the number of derivatives in a given period, which is most conve-

(ricevuto nel giugno 2003)

niently counted by using historical dictionaries. However, there are several problems with such a measure, which has been repeatedly proposed in the literature (cf. Rainer 1987, Bauer 2001). On the one hand, the problems concern the dictionary in itself (cf. Plag 1999:96-99); on the other, it can be shown that for the measure of productivity, where other parameters such as frequency play a role, dictionaries are simply inadequate to provide answers.

A first problem with dictionaries is that for commercial and practical reasons, they do not aim at the comprehensive documentation of productively-formed, transparent forms, but rather cover the more frequent and idiosyncratic items, especially for small-sized dictionaries such as DISC (Sabatini & Coletti 1997), *Zingarelli 1998* (Dogliotti *et al.* 1998), and so on for Italian. This is of course meaningful from a lexicographical point of view, since "dictionary-users need not check those words whose meaning is entirely predictable from its elements, which by definition is the case with productive formations" (Plag 1999:96). Moreover, even when aiming at complete coverage, lexicographers often overlook new, regular formations, just because they are regular. This is especially true for those word formation processes whose semantic content is not particularly profiled, such as action nouns, quality nouns, relational adjectives, etc.

Secondly, dictionaries often tend to be enyclopedias, i.e. to list many words belonging to the most disparate lexical domains, such as the special language for architecture, economics, physics, etc., which do not necessarily belong to the average competence of the speaker (and this is a second reason why dictionaries are consulted). It must be added, however, that not necessarily all words belonging to special lexical domains of different sorts are unknown to the average competence of the speaker. In fact, there are differences within the special lexical domains as for the degree of specialization of the terminology – there is in other words a vertical variation within a special lexical domain – and special terms might have entered the common lexicon. Moreover, it cannot be *a priori* excluded that word formation processes, even though employed for special lexical domains, do not match more general productive patterns, and in this sense they do contribute to assess the productivity and the availability of a pattern. These problems can be at least partially overcome using a large-sized dictionary which also offers the possibility to make guided research on different sub-sectors of the lexicon. This is the reason why we employed for our investigation GRADIT (De Mauro 2000), which amounts to about 247,000 items, and allows one to perform sorted queries. Overall, in GRADIT about 58,000 items are marked for being of widespread usage. The total rises to about 163,000 when items belonging to specialistic domains are included, and to about 186,000 with the further addition of low-usage items.

A third problem presented by dictionaries is that they drag along a whole range of old complex forms that may distort the analysis because they are residues of morphological processes that have long ceased to be productive or because they are unanalyzed borrowings or reinterpreted as such. Also in this case, the use of a dictionary in which it is possible to achieve sorted queries allows one to overcome the problem by excluding archaic words from the sample. For instance, in GRADIT about 20,000 items are marked as obsolete.

Finally, dictionaries by definition can only provide data for the type frequency of a word formation process. Nothing can be gathered for the token frequency. As we will see, however, token frequency has been recently shown to play a major role for the measure of the productivity and the availability of word formation processes.

1.2. Corpus-based approaches to word formation

Already Aronoff (1983) draws the attention on the role played by frequency for the study of word formation, since the mean frequency of derivatives is shown to be significantly higher for less productive patterns with respect to more productive patterns. Thus, for the two English patterns *Xiveness* and *Xivity*, Aronoff found out that the mean token frequency of the word types for *Xiveness* is lower with respect to *Xivity* both for the base and for the derived words. Aronoff's (1983:168) interpretation is straightforward: "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". Moreover, it has already been mentioned that frequency plays a crucial role in lexical access.

In spite of the relevance of frequency to assess the status of word formation patterns, there are no systematic studies dealing with this particular aspect for Italian morphology with the remarkable exception of Thornton (1997). This gap is presumably due to the underestimation of quantitative work in theoretical linguistics, which is partially related to the dismissal of empirical investigations because more performance- than competence-oriented (in this respect cf. Dressler & Ladányi 2000). Only recently the attention has been drawn on the usefulness of large text corpora for quantitative investigations. The increase of linguistic material on electronic support and its easy access surely represent an important source to be exploited more and more in the future (see also Rainer 2003). However, Plag (1999:34) still complains that "[g]iven the advantages of the proposed measures, it is somewhat surprising that the work of Baayen and his collaborators has not yet lead [sic] to a proliferation of quantitative studies of productivity, inspite of the easy availability of large text corpora".

To fill this gap for Italian morphology, we elaborated a text corpus (see Gaeta & Ricca 2002 for details), which is constituted by three years (1996-1998) of the Italian newspaper *La Stampa* amounting to about 75 million tokens, available on compact disc and easily exportable on ASCII files to be treated with a text analysis software (DBT™ by E. Picchi - CNR Pisa).[1] The choice of a newspaper as text source for our corpus is not fully justified from a methodological point of view, since it is not carefully balanced for text types, speech registers, and so on, with respect to other text corpora as for instance the British National Corpus, or, to stay within Italian linguistics, to the data-base for the LIF (cf. Bortolini *et al.* 1971) and for the LIP (cf. De Mauro *et al.* 1993). However, the problem with the latter Italian corpora is their small size (respectively 1,500,000 tokens and 500,000 tokens), which makes them, as we will see below, useless for the research on word formation productivity developed by Baayen & Renouf (1996), where they adopt as a corpus several years of the *Times* amounting to about 80 million tokens. Moreover, even though not carefully balanced, a newspaper corpus at least presents a variety of text types and of speech registers distributed across several subjects (politics, culture, sport, and so on) normally occurring in a daily issue of a newspaper, which partially compensates for this shortcoming. Therefore, we do not claim our corpus to be thoroughly reliable; for the investigation of some phenomena it might turn out to be inadequate (obviously, for instance, to investigate the usage of the different verbal persons, given its written form). Nevertheless, we are confident that it provides quite a faithful picture of the 'ideal' competence of a (Northern) Italian (well-)educated speaker.

1.3. Quantitative dimensions of productivity in word formation

To grasp the quantitative dimension of word formation processes, several proposals have been taken into consideration. Besides the direct elicitation of the speakers' competence done on the basis of questionaries, which is however more suitable for the qualitative aspects of productivity (see for instance Rainer 1988), linguists usually rely on dictionaries to assess the vitality of word formation patterns. For instance, Bauer (2001:156) has suggested viewing productivity in terms of rate of additions. The basic idea is that the productivity of a word formation process can be equated with the average number of new words formed with that process that are used in the language within a specified time period, for which the appropriate data can be checked by means of a dictionary. However, to overcome the difficulties raised by the employment of dictionaries and discussed above, Bauer proposes a compromise, in which the dictionary is taken as a starting point for the investigation and the use of a (suitably large) corpus as a point of comparison. Therefore, the measurement *(a-b)* in (1), namely the number of words formed with a certain word formation process which are found in the corpus but not in the dictionary, should provide a measure of the increase of that process over the time period between the completion of the dictionary and the final date of material in the corpus. To be relevant, this measure should be compared to two (or more) similar measures for different processes:

(1) $$\frac{(a-b)_y}{(c-d)_z} > 1$$

When the inequality in (1) holds, we are allowed to say the process Y is more productive than the process Z.

A radically corpus-based approach to productivity has been recently developed by Baayen and his collaborators (cf. Baayen 1989, 1992, 1993, 2001; see also Baayen & Lieber 1991, Baayen & Renouf 1996, Plag *et al.* 1999). Baayen's idea is to consider productivity as the relation between the so-called hapax legomena *h*, i.e. words formed with a certain affix occurring with frequency 1 in the corpus, and the number *N* of tokens formed with that affix in the corpus:

(2) $$P = \frac{h}{N}$$

This index provides the probability that after counting *N* tokens of a certain affix a new formation *h* with that affix comes out. There is no space here to discuss the details concerning this approach, but undoubtedly, the index in (2) has proved useful for measuring productivity in word formation. In Gaeta & Ricca (2002, 2003, ms.), we proposed to modify Baayen's procedure by adopting what we called the variable-corpus approach. With respect to Baayen's procedure, where the number *N* in the denominator is given by the total token number sampled in the corpus for a given affix, in the variable-corpus approach affixes are compared at equal values of *N*, so that different subcorpora must be taken to compare *P*-values for affixes displaying different frequencies. In our opinion, the variable-corpus procedure provides an answer to the criticisms raised against Baayen's original procedure (cf. van Marle 1992), which in fact overestimates the values for low-frequency affixes with respect to the more frequent ones because of the decreasing character of the function *P(N)*, even tending to zero when *N* approaches infinity (cf. Baayen & Lieber 1991:837).[2]

Besides this productivity index, which is a probability measure, Baayen (1993) has suggested another way of expressing productivity for word formation processes, namely the mere number of words of the appropriate morphological category occurring just once in the whole corpus. With respect to the former measure, this hapax-conditioned degree of productivity is "particularly suited to ranking productive processes according to their degree of productivity" (Baayen 1993:205). Independently of the criticism that can be raised against the hapax-conditioned degree of productivity,[3] we will take it into consideration for our investigations, since it can at least be viewed as a quick and convenient way of quantifying how active a word formation process is.

2. Italian derivational affixes ranked by token frequency

Of the three quantitative parameters which will be considered in this study, token frequency is clearly the only one for which no comparison between corpus and lexicographical evidence is possible, for dictionaries have nothing to say in this respect. Ranking affixes by their token frequency is perhaps the most intuitive way to quantify how a derivational process 'weighs' in the language system under consideration: psycholinguistically, every occurrence of an affix in speech production involves its activation in the mental lexicon (at least under the assumption that the affix is indeed parsed as such). On the other hand, token frequency *per se* has little or nothing to say about productivity. Affixes which are unproductive even from a qualitative point of view can nevertheless exhibit relatively high token frequencies, if they have been productive in the past and several of their derivatives still are common words. Most evident instances of this phenomenon in Italian are suffixes like *–nza* and *–ore* (see Table 2 below).

A preliminary issue on any kind of quantitative data concerns corpus size. For data concerning token frequency, it can be safely stated that the dimension of our corpus is largely sufficient to yield stable results. Indeed, even a much smaller corpus would suffice. By way of illustration, Table 1 shows, for some of the affixes investigated, that their token frequency remains substantially stable while the chunk size increases from a minimal subcorpus of just two months up to the full 36-month corpus.

Table 1. Checking subcorpus uniformity.

Relative token frequency (‰) for different subcorpus sizes						
Suffix	2 months 4 162 397 tok.	4 months 8 302 320	6 months 12 535 480	12 months 24 915 369	24 months 49 485 568	36 months 74 917 798
-(z)ione	13.0	13.1	13.3	13.4	13.7	13.9
-mente	4.32	4.29	4.24	4.26	4.23	4.24
-nza	2.83	2.81	2.80	2.73	2.76	2.78
-(t)ura	0.82	0.84	0.84	0.83	0.84	0.85

To be sure, it is not to be expected that each lexical item is so evenly distributed throughout the corpus. Table 1, however, effectively shows that the averaging contribution of the different types belonging to the same affix leads very early to stable data concerning its overall token frequency.[4]

The results in Table 1 do not ensure us that the frequency data obtained could be automatically generalized to any kind of textual typology: the results necessarily reflect the language registers found in newspapers. For markedly different kinds of texts, i.e. for a corpus including only literary or scientific or legal works, results might differ to some extent. However, as said above, a newspaper corpus also provides a rich – even if not balanced – mixture of textual genres, from very informal styles up to articles of nearly literary character and others which include a fair amount of specialistic terminology (ranging from sports to science, medicine and so on).

A thorny point in evaluating rankings based on token frequency is related to the fact that for most affixes, the frequency distribution of the different types is very skewed: namely, a handful of derivatives (say, the ten or twenty most frequent ones) cover a great percentage of the overall token frequency of a given affix. Unfortunately, it often happens that several among those very frequent words are by no means ideal derivatives, as they display – in Dressler's (1985) terms – low morphosemantic and/or morphotactic transparency. In some cases it can be doubtful if the word in question can really be considered as containing the given affix, i.e. is still analyzed as a complex word by the native speaker. Similar problems occur whenever affix types have to be identified (cf. for English the discussion in Plag 1999:28); but, for the reason said above, their quantitative impact is particularly relevant when dealing with token frequency data. To enable data comparability, one should adopt criteria as coherent and explicit as possible to establish the cutoff point in both the lexicalization and the allomorphy continua. We will treat these questions very briefly here, as we already dealt more extensively with them elsewhere (see Gaeta & Ricca 2002, ms.).

The morphosemantic problem concerns first of all items like *sedimento* 'sediment' vs. *sedere* 'sit', *stazione* 'station' vs. *stare* 'stay', *temperatura* 'temperature' vs. *temperare* 'temper', *sentenza* 'verdict' vs. *sentire* 'hear, feel', *scuderia* 'stable' vs. *scudo* 'shield', *generoso* 'generous' vs. *genere* 'gender'. In these cases a morphotactically transparent lexeme is completely unrelated to the base from a semantic point of view, at least synchronically. Given their idiosyncratic meaning, it is questionable whether the occurring suffix is really being activated when such words are used.

Besides these fully lexicalized items, there are of course several cases where we can speak of regular polysemy in the sense of Apresjan's (1974) (cf. also Rainer 1993:136, Gaeta 1999, 2002:201ff.): take for instance *abitazione* 'house', *accampamento* '(military) camp', *ingranaggio* 'gear', *creatura* 'creature'. Although these latter items cannot be used at all as action nouns (*abitazione* cannot mean 'the fact of inhabiting' and *creatura* cannot mean 'creation'), the meaning shift from action to place or result has a systematic character both within and across languages. Furthermore, in many derivatives both meanings co-occur (e.g., *redazione* 'act of compiling' and 'editorial office', *trasmissione* 'act of broadcasting' and 'TV program', etc.). In the face of these different grades of morphosemantic opacity, it is far from obvious which items should still be included as types of the given affix, and which ones should not. As a general strategy, we assumed that wherever a polysemic chain could be identified, the affix in question could still be parsed as such in the mental lexicon. Accordingly, we only excluded fully lexicalized items like *sentenza* from the count.

The somehow specular issue of morphotactic transparency displays different grades as well. The end of the continuum is provided by items that we could call base-less derivatives. This group comprises words like *detrimento* 'detriment', *ovazione* 'ovation', *massaggio* 'massage', *cesura* 'interruption', *oratore* 'orator', *potabile* 'drinkable'. Synchronically, they are simplexes, since they cannot be related to any extant base: **detri-, *ova-, *massa-, *ces-, *ora-, *pota-* are not lexical morphemes in Italian (not even bound ones such as *idr-* 'water' in *idr-ante*, *idr-ico*, etc.), at least in the relevant meaning. However, the endings formally coincide with productive suffixes (action noun, agent noun, possibility adjective, etc.) whose meaning is clearly identifiable in the full word as well. There might be good theoretical (and/or psycholinguistic) reasons to include these lexemes in our counts, since they might induce the activation of the respective suffixes, thus influencing their availability in the mental lexicon.[5] On the other hand, in some cases it is rather difficult to discriminate between the examples mentioned above and other instances where at most the ending, but not the related semantics, can be identified. This is the case of items like *elemento* 'element', *dimensione* 'dimension', *equipaggio* 'crew', *figura* 'figure', *settore* 'sector', *labile* 'fleeting', etc. In view of these difficulties we preferred to exclude all base-less derivatives from the count.

Moving further along the continuum, we meet different levels of morphotactic opacity involving various kinds of allomorphies of the affix, the base, or both. In Italian, the most complex case in this respect is given by a group of very frequent derivational processes sharing the same behaviour (chiefly *-zione/-ione, -tura/-ura, -tore/-ore, -tivo/-ivo, -torio/-orio*) which, according to most analyses, display two main allomorphs in the affixes and three in the base, together with several further minor irregularities. For the details we refer to Scalise (1994:275-279, 1996), Thornton (1990-91), Rainer (2001), Gaeta (2002:66-79), Gaeta & Ricca (2002, ms.). It seems indisputable, however, that even such complex instances of allomorphy still allow the speakers to identify both the base and the derivational process. Accordingly, all derived words exhibiting allomorphies have been included in our calculations.

In (3-6) a glossed example for each affix investigated is provided, grouping the (nearly) synonymous affixes together. The affixes are given in their citation forms, but it is understood that throughout the paper the figures refer to lemmatized affixes, and therefore include all the relevant inflectional forms for each derived word.

(3) Deverbal affixes:

a. the suffixes *-mento, -(z)ione, -(t)ura, -aggio, -nza* forming action nouns:

cambiare → *cambiamento*	'change'
trasformare → *trasformazione*	'transformation'
mappare → *mappatura*	'mapping'
lavare → *lavaggio*	'washing'
decadere → *decadenza*	'decay'

b. the adjectival suffixes *-bile* '-able', *-evole*, *-(t)orio:*

lavare → *lavabile*	'washable'
mancare → *manchevole*	'faulty'
adulare → *adulatorio*	'flattering'

c. the prefix *ri-* 're-' (with its allomorph *re-*):
giocare → *rigiocare* 'play again'
dare → *ridare* 'give back'/ 'give again'

d. the suffixes *-(t)ore* and *-trice* forming masculine / feminine agent and instrument nouns and also deverbal adjectives with agentive semantics:[6]

giocare → *giocatore* / *giocatrice*	'player' / 'player (f.)'
calcolare → *calcolatore* 'computer' / *calcolatrice* 'pocket calculator'	
uno sguardo rivelatore	'a revealing (m.) glance'
un'osservazione rivelatrice	'a revealing (f.) observation'

e. the locative suffix *-toio* (with the fem. variant *-toia*):
mangiare → *mangiatoia* 'manger', *lavare* → *lavatoio* 'lavatory'

(4) Deadjectival affixes:

a. the suffixes *-ità/-età*, *-ezza*, *-*$^{(1)}$*erìa*, *-aggine*, *-izia*, *-igia*, *-ore* forming quality nouns:

vero → *verità*	'truth'
bello → *bellezza*	'beauty'
vigliacco → *vigliaccheria*	'cowardice'
insensato → *insensataggine*	'foolishness'
pigro → *pigrizia*	'laziness'
altero → *alterigia*	'haughtiness'
bianco → *biancore*	'whiteness'

b. the negative prefix *in-* 'un-' / 'in-' (with its allomorphs *il-*, *im-*, *ir-*):
utile → *inutile* 'useless'

c. the adverbializing suffix *-mente* '-ly': *fermo* → *fermamente* 'firmly'

d. the elative suffix *-issimo*:[7] *lungo* → *lunghissimo* 'very long'

(5) Denominal affixes:

a. the adjectival suffixes *-oso*, *-esco*, *-aneo* (mostly qualifying), and *-ale/-are*, *-aceo*, *-estre* ,*-iero*, *-izio* (mostly relational):

paura → *pauroso*	'fearful'
scimmia → *scimmiesco*	'monkeyish'
momento → *momentaneo*	'momentary'
regione → *regionale*	'regional'
perla → *perlaceo*	'pearly'
terra → *terrestre*	'earthly'
costa → *costiero*	'coastal'
impiegato → *impiegatizio*	'clerical'

b. the ethnic suffix *-ese*: *Torino* → *torinese* 'Turinese'

c. the deanthroponymic suffix *-iano*: *Kant* → *kantiano* 'Kantian'

f. the collective suffixes *-ame*, *-aglia*, *-ume*:

foglia → *fogliame*	'foliage'
soldato → *soldataglia*	'soldiery'
canaglia → *canagliume*	'rabble'

g. the female sex suffix *-essa*: *principe* → *principessa* 'princess'

h. the agentive suffixes -$^{(1)}$*aio*, *-iere*: *fiore* → *fioraio* 'florist'
banca → *banchiere* 'banker'

i. the locative suffixes *-eto/a*, -$^{(2)}$*aio/a*, -$^{(2)}$*erìa*, *-iera*, *-ificio*

olivo → *oliveto*	'olive grove'
pino → *pineta*	'pinewood'
formica → *formicaio*	'ant's nest'
riso → *risaia*	'rice-field'
libro → *libreria*	'bookshop'
polvere → *polveriera*	'powder magazine'
pasta → *pastificio*	'pasta factory'

l. the pejorative suffix *-accio*: *affare* → *affaraccio* 'bad affair, trouble'

(6) Denominal/deadjectival affixes:

a. the suffixes *-ista* '-ist' and *-ismo* '-ism', which take as input:

- common nouns:	*terrore* 'terror' → *terrorista, terrorismo*
- proper nouns:	*Marx* → *marxista, marxismo*
- adjectives:	*astratto* 'abstract' → *astrattista, astrattismo*

b. the pejorative suffix *-astro*:
poeta → *poetastro* 'bad poet', *rosso* → *rossastro* 'reddish'

c. the three verbalizing suffixes *-eggiare*, *-ificare*, *-izzare*:
rivale → *rivaleggiare* ‘to rival’, *bianco* → *biancheggiare* ‘to be white’
pane →*panificare* ‘to make bread’, *dolce* →*dolcificare* ‘to sweeten’
atomo → *atomizzare* ‘to atomize’, *civile* → *civilizzare* ‘to civilize’

d. the similative suffix *-oide*:
ameba → *ameboide* ‘amoeboid’, *geniale* _ *genialoide* ‘eccentric’

e. the evaluative prefixes *iper-*, *maxi-*, *mega-*, *micro-*, *mini-*, *super-*, *ultra-*:
mercato →*ipermercato* ‘hypermarket’, *realistico* →*iperrealistico* ‘hyperrealistic’
schermo → *maxischermo* ‘giant screen’
concerto → *megaconcerto* ‘mega-concert’
film → *microfilm* ‘microfilm’
gonna → *minigonna* ‘miniskirt’
uomo → *superuomo* ‘superman’, *leggero* → *superleggero* ‘super-light’
conservatore → *ultraconservatore* ‘ultraconservative’

There is no space to discuss all the details concerning these affixes, which actually constitute the major part of Italian derivation.[8] At any rate, notice that we distinguished two cases of homonymous affixes, namely -$^{(1)}$*erìa* and -$^{(2)}$*erìa*, cf. (4a) and (5i), and -$^{(1)}$*aio* and -$^{(2)}$*aio/a*, cf. (5h) and (5i), since the respective derivatives are clearly distinct either as for the selected inputs (-$^{(1)}$*erìa* only selects adjectives, while -$^{(2)}$*erìa* only nouns) or in semantic terms (-$^{(1)}$*aio* only produces agent nouns, while -$^{(2)}$*aio/a* only locative nouns). Moreover, they constitute quite distinct sets, not overlapping as a result of polysemic extension.

In (6) we reported affixes that systematically select as input both nouns and adjectives. In fact, nearly all affixes occur in some formations based on categories different from those exemplified in (3-5): cf. for instance *rocciatore* ‘rock climber’ from *roccia* ‘rock’, *pensoso* ‘thoughtful’ from *pensare* ‘to think’, *insuccesso* ‘failure’ from *successo* ‘success’ and so on. However, since these latter cases are sporadic, we did not list such affixes in (6) and we did not include the categorially deviant formations in our counts. A different pattern holds for unfrequent and scarcely productive affixes. In these cases, formations from different categories are often comparable in number (also because they are few in total): take for instance the collective suffix *-ume* occuring with nouns (*canaglia* → *canagliume* ‘rabble’), adjectives (*sudicio* _ *sudiciume* ‘dirt’), and verbs (*sfasciare* → *sfasciume* ‘wreck’). For these minor suffixes we included all analyzable material in the count.

The complete list of absolute and relative token frequencies is given in Table 2. Before discussing these data, a final remark has to be made: all figures in Table 2 refer to occurrences in the outmost derivational cycle only. Namely, a word like *nazionalizzazione* ‘nationalization’ has been counted among the tokens of *-(z)ione*, but not of *-izzare* or *-ale/-are*, even though also these latter suffixes, belonging to inner derivational cycles, are easily identifiable and contribute independently to the word meaning. There are good grounds for this choice, which has always been the standard option (for a discussion see Plag 1999:29) and is also operationally easier, at least for suffixes. Still, it is hard to justify such a choice from a psycholinguistic point of view; moreover, its implementation is not always straightforward. For an illustration of these issues, we refer again to Gaeta & Ricca (2003, ms.). Here we limit ourselves to notice that, similarly to the case of allomorphy, the conceivable inclusion of inner derivational cycles would have a very different impact depending on affixes. As shown in Gaeta & Ricca (2003, ms.), among affixes often occurring in inner position are *-bile*, *-izzare* and the two major prefixes *in-* and *ri-*. For other affixes, like *-mento*, *-ezza* or *-mente*, the contribution of inner cycles to the overall token frequency would be negligible or even absent.

Notice that the whole token number of the portion of Italian derivation considered amounts to about 5,000,000 tokens, which is more than 6% of the whole corpus token number. At first sight, these figures look impressive,[9] and it would be highly interesting to compare them with analogous evidence coming from other corpora such as LIP (De Mauro *et al.* 1993).

Coming to a detailed analysis, it is interesting to compare our data with the previous study by Thornton (1997), based on a much smaller corpus, about 1,500,000 tokens.

Thornton’s data are divided into logarithmic classes, taking the natural logarithm of the affix absolute frequencies, i.e. ln N. To make the comparison possible, our data have been similarly ordered, as shown in Table 3. Of course, since the size of Thornton’s corpus is fifty times smaller than ours, and token frequency is linear with respect to corpus size, our absolute frequencies N in Table 2 have to be divided by 50 before taking the logarithm to be comparable with Thornton’s. This is the same as calculating (ln N - ln 50) = (ln N - 3,91), which are the values according to which Table 3 is built.

Table 2. Token frequency for the Italian derivational affixes considered.

Affixes	token frequency (*N*)	relative token frequency (‰)	Affixes	token frequency (*N*)	relative token frequency (‰)
-(z)ione	1 043 979	14	*-(t)orio*	13 998	0.19
-ale / -are	734 725	9.8	-$^{(1)}$*aio*	12 025	0.16
-ità / -età	356 857	4.8	*-iera*	10 613	0.14
-mente	317 725	4.2	*-esco*	9 060	0.12
-(t)ore	273 706	3.7	*super-*	8 966	0.12
ri-	270 066	3.6	*-essa*	7 245	0.097
-mento	257 216	3.4	*-ificio*	6 965	0.093
-nza	208 365	2.8	*-toio / a*	5 740	0.077
-ista	160 318	2.1	*-izio*	5 721	0.076
in-	146 982	2.0	*-accio*	4 344	0.058
-oso	135 850	1.8	*-aneo*	4181	0.056
-ese	118 912	1.6	*-ame*	4 079	0.054
-bile	102 904	1.4	-$^{(1)}$*erìa*	3 207	0.043
-izzare	96 491	1.3	-$^{(2)}$*aio / a*	3 097	0.041
-ore	76 113	1.0	*micro-*	2 869	0.038
-ezza	69 090	0.92	*mini-*	1 830	0.024
-(t)ura	63 800	0.85	*iper-*	1 674	0.022
-ismo	63 295	0.84	*maxi-*	1 617	0.022
-iere	60 678	0.81	*-estre*	1 593	0.021
-issimo	51 636	0.69	*ultra-*	1 557	0.021
-izia	38 263	0.51	*mega-*	1 399	0.019
-iano	36 820	0.49	*-aglia*	1 190	0.016
-ificare	31 001	0.41	*-oide*	1 121	0.015
-eggiare	23 805	0.32	*-aggine*	914	0.012
-trice	23 780	0.32	*-ume*	912	0.012
-aggio	22 019	0.29	*-astro*	791	0.011
-evole	19 076	0.25	*-eto / a*	790	0.011
-$^{(2)}$*erìa*	18 021	0.24	*-aceo*	554	0.0074
-iero	15 871	0.21	*-igia*	492	0.0066
Tot.				4 955 908	66

The affixes considered overlap only partially: Thornton includes 36 affixes against our 58; she does not include any prefix nor any of the three verbal affixes *-izzare / -ificare / -eggiare*. Three among the quantitatively most relevant suffixes are also absent in Thornton's count, presumably for technical reasons, namely *-mente*, *-ità*, and *-ese*. On the other hand, Thornton includes some further low-frequency suffixes (as *-asta, -ense*, *-igno*, *-ingo*), which we did not consider main-

Table 3. Ordering affix token frequencies according to logarithmic classes.

Affixes	Class
-(z)ione, -ale / -are	10
-ità / -età, mente, -(t)ore, ri-, -mento	9
-nza, -ista, in-, -oso, -ese, -bile, -izzare	8
-ore, -ezza, -(t)ura, -ismo, -iere, -issimo, -izia, -iano	7
-ificare, -eggiare, -trice, -aggio, -evole, -$^{(2)}$*erìa, -iero, -(t)orio*	6
-$^{(1)}$*aio, -iera, -esco, super-, -essa, -ificio, -toio / a, -izio*	5
-accio, -aneo, -ame, -$^{(1)}$*erìa, -*$^{(2)}$*aio / a, micro-, mini-, iper-*	4
maxi-, -estre, ultra-, mega-, -aglia, -oide, -aggine, -ume, -astro	3
-eto / a, -aceo, -igia	2

ly due to their extremely low type frequency together with heavy transparency problems in most of their formations. We included some more transparent low-frequency items, however, in order to display the whole span of logarithmic classes occurring in Thornton's work.

Despite the lack of full uniformity, it is clear that the ranking pattern between Thornton's data and ours is more or less the same. There are little discrepancies worth signalling. First, there is a general shift upwards of most affixes, of about one logarithmic class. This is presumably due to the fact that Thornton's data refer to the citation form only (i.e the singular, and more restrictively the masculine singular for 1st class adjectives such as *-oso*), while ours have been lemmatized with respect to the relevant inflectional forms. For the suffixes *-(z)ione, -(t)ura, -(t)ore* mentioned above, a further reason why Thornton's figures are lower depends on her excluding all items formed with the allomorphs *-ione, -ura*, and *-ore* respectively (A. M. Thornton, p.c.). On the other hand, *-bile* is ranked slightly higher by Thornton (i.e. above the suffixes *-ista* and *-oso*), presumably because the particularly frequent occurrences of this suffix together with the negative prefix *in-* (as in *immangiabile* 'uneatable') have been included in her count, but not in ours being inner-cycle derivations. Finally, the relatively high ranking of *-aglia* by Thornton, which finds absolutely no match in our data, is really puzzling, but marginal in the whole context; perhaps it can be due to the inclusion of some frequent word with that ending that we judged as fully opaque.

The ranking in Table 3 shows fairly obvious results for the highest three classes, comprising only affixes which belong to the very core of Italian derivational morphology and are judged as highly productive by all qualitative standards and by all linguists' intuition.

The only exception is *-nza*, which exemplifies at best the case of a suffix heavily entrenched in the lexicon despite its synchronic marginality as a productive device. More blatant deviations come out in class 7, which hosts two absolutely unproductive suffixes, namely abstract *-ore* and *-izia*, simply because of the very high frequency of a couple of words which can still be analyzed as complex (e.g., *amore* 'love', *valore* 'value'; *giustizia* 'justice', *amicizia* 'friendship').

3. Type frequency in corpus and dictionaries: a comparison

Another kind of data to evaluate the relevance of different derivational processes in a language is given by the type frequency *V*. Again, a direct correlation between these data and (qualitative) productivity is not to be expected. Although qualitatively productive affixes usually display a high number of types, a non-negligible type frequency can also be found – much less than for tokens, however – for affixes which have been very productive in the past; on the other hand, affixes qualitatively productive in a very restricted domain (such as *-iera*, *-$^{(2)}$aio/a*, *-ificio*, *-eto/a* in our corpus) may display a low type frequency, as will be seen.

In our corpus, the type frequency for the affixes examined displays a very wide range, going from the lowest value of 7 for *-estre* and *-igia*[10] to the highest value of 2767 for *-mente*. The whole list of type frequencies is given in Table 6 together with the corresponding dictionary data and will be commented there.

Referring for the moment to corpus data only, it may be interesting to have a comparative look at type and token distributions. To do so, we grouped type frequencies in logarithmic classes, as we did before for the token frequencies, and again matching Thornton's (1997) procedure. The results are given in Table 4.

Globally, the logarithmic type frequency distribution is a bit less dispersed than the token frequency one (the standard deviations are 1,5 for the former and 2,1 for the latter[11]). More interestingly, it is immediately apparent that the two rankings diverge radically in some cases, although some very important affixes (*-mente*, *-(z)ione*, *-ità/-età*) rank at the top in both. Perhaps the most obvious contrast is given by the place of the very coherent group of evaluative prefixes, all ranking very low in tokens and very high in types. Two important quasi-inflectional suffixes, namely *-issimo* and *-iano*, rank also remarkably higher in the type- than in the token count, although they do not reach the top of the list either. Among the items which

Table 4. Ordering affix type frequencies according to logarithmic classes.

Affixes	Class
-mente, *-(z)ione*, *-ità / -età*	8
-issimo, *-ista*, *-(t)ore*, *-iano*, *-mento*, *-ismo*, *super-*, *-bile*, *-ale / -are*, *ri-*, *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

are substantially downgraded by the type ranking there are, quite expectedly, the fully unproductive *-izia* and *-ore*; in lesser measure and perhaps less expectedly as well, also *-nza* and *-ale / are*.

To get a general comparison of the two rankings, we may compare the two logarithmic distributions, assuming that they both approximate a normal distribution (cf. Thornton 1997:388). Since the mean value of ln N for the 58 affixes considered is 9.593, while the mean value for ln V is 5.375 (of course the total token number is much greater than the type number), in order to make the two distributions comparable we reduced them to the same mean, by subtracting (9.593-5.375) = 4.218 from the values of ln N for each affix. Clearly, such an operation does not modify the shape of the token distribution, simply shifting it backwards to superimpose it with the type distribution.

Then, for each affix we calculated the difference between the values for the two distributions, namely $\Delta_{N,V} = (\ln N - 4.218) - \ln V$. In Table 5 the affixes are listed in decreasing order of this value. A negative value of $\Delta_{N,V}$ should mean that the given affix is comparatively richer in types than in tokens, and conversely for a positive value.

Unfortunately, the results in Table 5 are not very revealing, because of the huge differences in token frequencies among the different affixes. Comparing token and type distributions as done here amounts to calculate – apart from the constant – the logarithm of N / V, i.e. the token/type ratio for each affix. However, the same problem which is well-known for text samples (cf. Baayen 2001:4) occurs for samples of single morphological processes as well: when token number increases, the ratio N/V increases steadily, even for productive affixes displaying a good number of types and new formations. For that reason all 'big' suffixes (*grosso modo* those with token number above 100,000 in the corpus) have positive values of $\Delta_{N,V}$ in Table 5, ranging from 2.3 for *-ale/-are* to 0.3 for *-bile*.

Nevertheless, Table 5 stresses at least the very special position of the evaluative prefixes (all coherently at the bottom of the list, as they exhibit a very high type frequency together with a very low token frequency), and on the other edge of the list, the position of some affixes absolutely marginal from the point of view of number of formations (not to speak of productivity), but still important at the token level (particularly *-ore* and *-izia*). But on the whole, a comparative look to Table 3 and Table 4 is more informative than the kind of evaluation tried in Table 5.

Table 5. Comparing type and token distributions.

Affixes	$\Delta_{N.V}$	Affixes	$\Delta_{N.V}$	Affixes	$\Delta_{N.V}$
-izia	**3.2**	*-izio*	0.9	*-issimo*	-0.8
-ore	**2.9**	*-izzare*	0.7	*-iano*	-1.0
-nza	**2.6**	*-essa*	0.6	*-esco*	-1.1
-ale/-are	**2.3**	*-mente*	0.5	*-eto/a*	-1.1
-(z)ione	1.9	*-ista*	0.5	-$^{(1)}$*erìa*	-1.2
-iere	1.6	*-(t)ura*	0.5	*-astro*	-1.2
-ificare	1.6	*-iera*	0.5	*-ume*	-1.3
-evole	1.5	*-eggiare*	0.4	*-aggine*	-1.6
-aneo	1.5	-$^{(2)}$*erìa*	0.4	*-accio*	-1.7
-ri-	1.4	*-ificio*	0.4	*-oide*	-1.8
-estre	1.2	-$^{(1)}$*aio*	0.4	*-aceo*	-1.9
-oso	1.2	*-bile*	0.3	*-toio/a*	-2.0
-ezza	1.1	*-ame*	0.3	***super-***	**-2.2**
-iero	1.1	*-igia*	0.0	***micro-***	**-2.3**
-ità	1.0	*-(t)orio*	-0.3	***ultra-***	**-2.6**
-(t)ore	1.0	*-ismo*	-0.4	***maxi-***	**-2.7**
-mento	1.0	-$^{(2)}$*aio/a*	-0.4	***iper-***	**-2.8**
in-	1.0	*-trice*	-0.6	***mega-***	**-3.0**
-ese	1.0	*-aglia*	-0.7	***mini-***	**-3.1**
-aggio	1.0				

For instance, Table 3 and Table 4 shed some light on a typical feature of Italian derivational morphology, namely the presence of several competing affixes nearly equivalent semantically. In many cases, they are judged as qualitatively productive and this kind of widespread redundancy looks somehow intriguing. Quantitative data both in types and in tokens, however, reduce significantly the relevance of this anomaly, since for several semantic domains one affix turns out to have a very clear primacy. This is the case of (i) verbal suffixes, where *-izzare* outnumbers *-ificare* and *-eggiare*; (ii) quality nouns, where *-ità / -età* takes very clearly the lead against its competitors like *-ezza*, *-ità*, *-*$^{(2)}$*erìa*, or *-aggine*; (iii) denominal agentive nouns, quantitatively dominated by *-ista* with respect to *-iere*, *-*$^{(1)}$*aio*.

At least two instances of 'fair' competion remain, however: (i) action nouns, where *-(z)ione* and *-mento* are both among the top scorers, and *-(t)ura* is not much far away, and (ii) the set of evaluative prefixes *maxi-, ultra-, super-, iper-, mega-* and so on, all displaying a fairly parallel and very peculiar behaviour, as discussed above.

One major difference between the token and the type level is that the latter allows for a comparison between corpus-based and lexicographical data. As said in § 1.1, there are many factors of distortion to be expected when data from dictionaries are considered. Since dictionaries include a fair amount of very rare and obsolete words, they should enhance the values of type frequency for those affixes which have been most productive in earlier stages of the language. Moreover, dictionaries (especially large ones) tend presumably to overstate the impact of special language terminology with respect to the mental lexicon. Therefore, it can be interesting to see to what extent lexicographical data on affix types diverge from textual ones, and more particularly, for which affixes they are fundamentally reliable or have instead to be discarded throughout.

This comparison can be made in detail due to the possibility of resorting to a recent and very large dictionary of Italian such as GRADIT, which provides every entry with indication of its level(s) of use, by means of several labels. Three of them (FO, AD, AU) identify items belonging to the core vocabulary, and a fourth one (CO) the items of common use: the entries carrying at least one of the four labels mentioned above can be estimated to constitute a first lexical stratum ("common words") to be compared with our corpus. We also considered three further extended sets of words. The first one adds to the basic stratum the items restricted to technical and specialistic domains (labelled TS); the second one includes the words labelled "of low use" (BU), but not the TS ones; finally the third one is the maximal set comprising both extensions. Obsolete, strictly literary and regional words (each receiving a separate label in GRADIT) have been left out throughout.[12]

It must immediately be stated that the two sorts of data are not fully homogeneous, for several reasons. First, dictionary counts include only those items which are given as derived in the etymological section. This means that, at least in principle, all words which have an attested Latin – or sometimes French – antecedent that could function as a model are excluded. In our corpus counts we have instead included such words whenever their morphological structure is sufficiently transparent in Italian to assume that they are analyzed as complex words by the native speaker (as is often the case).

Moreover, in the dictionary a suffix is identified and therefore counted (provided that the word is not a borrowing) irrespective of the syntactic category of the base. In our counts, on the contrary, all categorially deviant formations for major affixes have been discarded, as said in § 2.

Finally, as seen above, we judged indispensable to consider as homonymous some pairs of suffixes treated as polysemic in GRADIT, presumably on purely etymological grounds. This latter problem, however, can be easily overcome by summing the data for the two homonymous suffixes (this is the case for *-erìa* and *-aio* in Table 3) or separating manually the dictionary data (which has been done for the rather extreme case of *-ore* vs. *-(t)ore*).

The other two discrepancies affect the different suffixes in a more or less relevant way. The first one, tending to lower dictionary figures, is probably most relevant for *-(z)ione*, while the second one, which on the contrary enhances them, mainly affects *-(t)ura*, *-aggio* and *-nza*. At any rate, it is unlikely that these shortcomings could seriously trouble the overall picture.

In Table 6 the affixes considered have been ordered by decreasing type frequency as it results from our corpus, while the other columns give the number of types included in the four lexicographical counts. To make the major deviations more evident, we have printed in boldface all instances where the dictionary type frequency amounts to more than twice the one from the corpus, and in italics all instances in which the former is less than half than the latter (this is slightly less than a distance of one logarithmic class in Table 4).

The first observation to be made concerns the evaluation of the full totals of Table 6. They make it clear that a large textual corpus like ours captures a very significant portion of the derivational formations of the language, extending well above the level of words of common use.[13] Only when both specialistic and low use words are

Table 6. Comparing type frequencies in the corpus and in GRADIT.

Affixes	*V* (corpus)		GRADIT common words	GRADIT common + specialistic	GRADIT common + low use (excl TS)	GRADIT maximal count
-mente	2767		2996	3470	4260	4734
-(z)ione	2363		*716*	1643	*998*	1925
-ità / -età	1962		1314	1806	1719	2221
-issimo	1697		*9*	*15*	*9*	*15*
-ista	1482		*553*	1787	764	1998
-(t)ore	1480		*634*	1545	1403	2314
-iano	1415		*125*	*695*	*159*	729
-mento	1403		1069	1510	2548	2989
-ismo	1375		*474*	2049	810	2385
super-	1147		*87*	*212*	119	*244*
-bile	1117		1023	1168	1269	1414
-ale / -are	1063		*311*	1201	*438*	1328
ri -	934		1156	1226	1675	1745
in-	767		402	476	645	719
-izzare	717		*293*	649	470	826
-ese	657		**6958**	**6978**	**6999**	**7019**
-trice	645		*34*	512	*43*	521
-oso	626		*293*	495	539	741
mini-	612		*20*	*37*	*20*	*37*
-(t)ura	561		597	**1480**	1146	**2029**
micro-	437		*24*	521	*31*	528
mega-	426		*9*	*114*	*11*	*116*
-esco	405		*167*	237	401	471
iper-	390		*27*	415	*57*	445
maxi-	365		*9*	*11*	*9*	*11*
-$^{(2)}$*erìa*	182	332	210	316	436	542
-$^{(1)}$*erìa*	150					
-accio	334		*34*	*66*	*42*	*74*
-ezza	324		312	320	549	557
ultra-	302		*20*	*82*	*22*	*84*
-(t)orio	292		*86*	165	176	255
-eggiare	227		166	226	450	**510**
-nza	225		164	281	228	345
-$^{(1)}$*aio*	128	193	196	380	380	**564**
-$^{(2)}$*aio / a*	65					
-iere	183		109	236	152	279
-aggio	114		87	269	125	**307**

Affixes	*V* (corpus)	GRADIT common words	GRADIT common + specialistic	GRADIT common + low use (excl TS)	GRADIT maximal count
-iera	99	102	193	133	**224**
-oide	96	*17*	**356**	*40*	**379**
-ificare	94	*28*	73	*43*	88
-iero	76	46	76	57	87
-ificio	67	45	82	49	86
-aggine	66	75	79	**195**	**199**
-essa	61	45	59	88	102
-evole	61	59	63	149	**153**
-ore	61	*16*	*16*	33	33
-toio/a	61	58	**223**	105	**270**
-aceo	56	*12*	41	*25*	54
-ume	48	43	53	**167**	**177**
-ame	44	38	77	**106**	**145**
-astro	39	23	54	30	61
-eto/a	36	50	**75**	69	**94**
-aglia	34	25	28	64	67
-izio	33	*15*	28	23	36
-izia	22	*2*	*4*	*4*	*6*
-aneo	14	*3*	*4*	*4*	*5*
-estre	7	*2*	*2*	*2*	*2*
-igia	7	*3*	5	*3*	5
Tot.	30424	21391	34184	30491	43294

included, the total lexeme inventory from the dictionary substantially outnumbers the corpus inventory. Notice that at this level we are much beyond what can be estimated to be the lexical competence of a cultivated speaker: the global total of words (derived and underived) registered in GRADIT under the labels of "basic", "high use", "common", "specialistic" and "low use" amounts to 186,000 entries.

As a further overall result, the lexicographical data which best approach those obtained from our corpus, from the point of view of the affix ranking, are those in the fourth column of Table 6. In fact, when both common and specialistic words are considered, but low use words are ruled out, instances of major deviations concern only a minority of suffixes. Therefore, we will refer below to this extended set unless otherwise specified.

A wholly idiosyncratic behaviour is found with the suffix *-ese*. All ethnic adjectives (also those referring to very little Italian towns and villages unknown to anyone apart inhabitants and close neighbours) have been introduced in GRADIT and – strangely enough – labelled as common words (De Mauro 2000: xx). Since *-ese* is by far the dominant suffix in this domain, and the only one included in our list, its abnormally high type frequency in lexicographical data (indeed, by far the highest of all!) is readily understood. At the same time, it constitutes a strong argument in favour of preferring type frequency data taken from corpora: in fact, no speaker community in Italy shares the use of such a high number of ethnic adjectives, and their impact in real linguistic interaction is clearly much better estimated by the figures obtained in a large textual corpus.

Similar considerations, on a different scale, may hold for suffixes like *-(t)ura* and *-oide*, which are widely employed in the specialistic vocabulary: a substantial part of this vocabulary is known or even interpretable only by a little minority of speakers. The same enhancement effect is found, to a lesser extent, for some other suffixes preferred in the specialistic domain, like *-ismo*, and quite surprisingly, -*toio* (a locative/instrumental suffix of limited use nowadays, but largely present in the often old-fashioned terminology of agriculture and handicraft). Clearly, a newspaper corpus also provides texts (and words) of specialistic character, but presumably it does not overstate their weight.

Equally interesting are the instances in which lexicographical data underestimate type frequency. A paradoxical case is given by the elative *-issimo* and by the only highly frequent evaluative suffix in our corpus, namely *-accio*. Items formed with these suffixes are not felt by lexicographers as separate lexemes, except in cases of strong lexicalization.[14] Not very far from these limiting cases is the group of evaluative prefixes already discussed above. Although all of them have a learned origin and many are very common in specialistic terminologies, only in two cases (*micro-* and *iper-*) lexicographical type frequency matches the one found in the corpus, when specialistic words are included: in all other cases, it remains much lower. Notice also that if only words labelled as common had been included, the figures would have been fairly irrelevant. This confirms that this subsystem of contemporary Italian derivation really has fully peculiar properties: the very high type frequency in the corpus consists of very few firmly established words (the only ones reported in the dictionary), like *minigonna* 'miniskirt', *maxischermo* 'giant screen' or *microcriminalità* 'micro-criminality', and almost exclusively of a huge amount of nonce formations like *megacena* 'mega-dinner', *megaorologio* 'mega-watch', *mini-emirato* 'mini-emirate', *mini-epurazione* 'miniepuration' and so on. In their attitude to combine very freely – but also rather loosely – with any sort of bases, these prefixes nearly border on syntax (for some more discussion, cf. Gaeta & Ricca 2003).

A similar phenomenon, to a lesser extent, can be identified for the deanthroponymic suffix *-iano*. This suffix applies to a very large domain (person proper nouns) and its meaning borders on inflection, as in many (not all) instances it reduces to the categorial shift to a relational adjective without further semantic addition, more or less like a genitive. Clearly, most of such formations are disregarded – and reasonably so – in the dictionary, which tends to report only those deanthroponymic adjectives which refer to worldwide known celebrities (*bachiano*, *kantiano* and so on), or those which have acquired some idiosyncratic meaning referring to clichés and the like (*freudiano*, *lapalissiano*).

Thus, curiously enough, the two main suffixes taking proper nouns as input, *-ese* for places and *-iano* for persons, are treated exactly the opposite way in GRADIT; and in both cases their impact on real language fully escapes the dictionary data, while it can presumably be estimated rather safely through corpus type frequency.

A further mainly deanthroponymic suffix is *-esco*. This suffix, which displays a somehow unexpected high type frequency in our corpus data, usually adds also semantic content (mostly of derogatory or jocular character) to the derived adjective. Dictionary entries underrate it to a lesser extent than *-iano*, also because it also takes common nouns as input; figures are sensibly higher when low use words are taken into account.

Finally, the negligible figures for the suffixes at the bottom of the table (*-izia*, *-aneo*, *-estre* and *-igia*), still lower than those obtained from the corpus, are easily explained: the dictionary does not treat complex words like *pigrizia* 'laziness' as derivatives, which is correct, because they have not been formed in Italian, but are borrowings from Latin. In this extreme case, it is the corpus count which becomes misleading: at least for *-izia*, its figures seem to allow for a (very marginal) word formation rule which actually never existed in Italian, apart from a couple of sporadic and clearly analogic formations.

4. Dictionary-based vs. corpus-based insights into productivity: a comparison

In this section, we will deal with the question of productivity, as discussed in § 1.3. Basically, dictionary-based approaches rely upon the evaluation of the enrichment of an affix domain in terms of types within a given time span. On the other hand, the quality of the productivity index relating to a text corpus seen above in (2) is to give some hint at the probability of forming a new derivative with a certain affix. Therefore, the two measures cannot be directly comparable. As shown above, the probability index involves token frequency, but not type frequency. On the contrary, in a dictionary-based approach only the latter plays a role, although in a usage-based dictionary such as GRADIT it is possible to get some information about token frequency via the usage labels illustrated in § 3.

A more reasonable comparison can probably be made with the hapax-conditioned degree of productivity, which only measures the rough number of hapax legomena occurring in the corpus with a certain affix. Given the property of hapaxes of being rare words, and to a good extent new formations if the corpus is large enough (cf. Baayen & Renouf 1996, and Gaeta & Ricca 2002 for Italian), a direct comparison between the amount of hapaxes and the amount of new formations reported in a dictionary for a given time span looks promising.

Let us consider the data obtained following the different methods. First, we will look at the probability index, in order to provide a global ranking for the affixes considered. Here, a number of caveats are in order. In fact, the whole framework requires a more thorough illustration than the sketchy picture given in § 1.3. However, for lack of space we will limit ourselves to a few necessary remarks, and refer the reader to Gaeta & Ricca (2002, 2003, ms.).

A first point concerns the method followed here. As explained in § 1.3, we are basically adopting Baayen's productivity index, but computed according to what we call the variable-corpus procedure. This means that we compared the productivity values of the affixes investigated at the same token number, which implies – as affixes display different token frequencies – extracting the data from subcorpora of different size, sampled in a progressive way.

A second remark is connected with the minimal corpus size. To avoid instability occurring with too little subcorpora, we placed a lower threshold for their size at about 6 million tokens. Given this limit, not all the affixes can be compared directly together because of the sharp differences in token frequency which can be gathered from Table 2 above. First, all low-frequency affixes (under a token frequency of 19,000) could not be compared with the most important ones and do not appear at all in Table 7.[15] Second, to get information for a sufficiently wide range of affixes (i.e. nearly all those with token frequencies from 19,000 upwards[16]), we had to compute productivity values for three values of *N*, as reported in the first three columns of

Table 7. The value N = 50,000 is the most suitable to embrace the greatest number of affixes. The value N = 100,000 allows us to include the top frequency affixes *-(z)ione* and *-ale / -are*, while the value N = 19,000 makes it possible to take into account several further affixes which do not reach a total token frequency of 50,000. At the same time, the count for three values of N is useful to verify the stability of the ranking order when the sample size changes.[17]

Table 7. Italian derivational affixes ranked by productivity and by hapax number in the corpus.

	$P(N) \cdot 10^3$			*h* in the whole corpus
Affixes	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-		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

The ranking based on the P-index suggests three groups of affixes. The first group contains affixes characterized by what can be reasonably called a borderline status: the elative suffix *-issimo*, the adverbializer *-mente*, and the 'genitival' suffix *-iano*, very productive especially with proper nouns. As can be expected given their quasi-inflectional character, their behaviour approximates the high productivity of inflectional rules (cf. Dressler 1989, Gaeta in press). Immediately after these borderline affixes, the big bulk of productive derivational affixes is found, starting with the very productive *-ismo* up to *-aggio*. This group contains core derivational affixes, like the nominalizers (-*mento*, *-(z)ione*, *-ità/-età*, *-(t)ura*, *-(t)ore*, *-trice*, *-ista*, *-ismo*, *-aggio*, *-ezza*), the adjectivalizers (*-ale* / *-are*, *-bile*, *-oso*), and the two prefixes *ri-* and *in-*. These affixes may display rather different productivity values, which also reflect the number of qualitative restrictions on their applicability. For instance, among the less productive ones *-aggio* is chiefly restricted to specialistic domains, *-ezza* attaches productively only to adjectival formations in *-to* and *-evole,* and to underived bisyllabic bases (cf. Rainer 1989:299). At the bottom of Table 7 we find the scarcely or non-productive affixes *-nza*, *-ificare*, *-ore*, *-evole*, *-izia*.

A similar affix order obtains by taking into consideration the hapax-conditioned degree of productivity, namely the rough number *h* of hapaxes formed with a certain affix occurring in the whole corpus, reported in the fourth column of Table 7. For productive affixes, this measure essentially expresses the number of new formations.[18] Since *h* is not immediately related to a probability evaluation, we expect to find some variation in the affix order with respect to the P-index. However, the main tendencies should be confirmed. This is in fact what we found out.

The nearly inflectional affixes (*-mente*, *-iano*, *-issimo*) and the non-productive ones (*-nza*, *-ificare*, *-ore*, *-evole*, *-izia*) are clearly distinguished from the core word formation processes. Within the core derivational processes some variation betweeen the two rankings occurs. The most striking difference concerns the feminine agent suffix *-trice*, whose low frequency is also partly reflected in a lower hapax number. In this way, however, the nice correlation with the masculine agent suffix *-(t)ore* gets lost.

The picture changes radically when we address our attention to dictionaries. Exploiting the possibility offered by GRADIT of making sorted queries by selecting items belonging to different sub-dictionaries, we will report data for the four lexical strata as illustrated in § 3. To make a reasonable comparison with the data extracted from corpora, we decided to compare the number of hapaxes to the number of new formations dated by GRADIT from 1950 onwards. Also in this case, the goal of our exploration was to understand if at least the same general tendencies are mirrored in both sorts of data. Adding to the caveats on the limits of comparability between corpus and lexicographical data already expressed in § 3, it must be stressed that the figures for GRADIT in Table 8 crucially rely on the accuracy of the datings reported in the dictionary.[19]

The picture in Table 8 shows the effects and the limits of a typic-

Table 8. Comparing hapaxes in the corpus with recent formations in GRADIT.

Affixes	*h* in the whole corpus	GRADIT 1950- common use	GRADIT 1950- common + specialistic	GRADIT 1950- common + low use (excl. TS)	GRADIT 1950- maximal count
-mente	825	155	217	262	324
super-	667	56	107	73	124
-issimo	643	6	7	6	7
-iano	615	40	241	49	250
-ità / -età	544	321	529	422	630
-(z)ione	486	207	593	277	663
-ista	470	195	682	272	759
-(t)ore	461	97	452	190	545
-ismo	448	109	588	215	694
-bile	409	204	241	249	286
-mento	402	119	233	252	366
mini-	383	12	24	12	24
ri-	312	144	178	194	228
-izzare	280	101	248	153	300
micro-	276	12	191	14	193
iper-	276	10	165	26	181
mega-	252	4	37	4	37
-ese	244	3012	3015	3021	3024
maxi-	230	9	10	9	10
-trice	224	12	191	13	192
ultra-	197	7	22	7	22
-esco	195	20	32	61	73
-(t)ura	189	88	365	151	428
-ale / -are	155	65	382	99	416
in-	148	54	73	102	121
-accio	140	2	6	2	6
-oso	127	23	76	73	126

Affixes	*h* in the whole corpus	GRADIT 1950- common use	GRADIT 1950- common + specialistic	GRADIT 1950- common + low use (excl. TS)	GRADIT 1950- maximal count
-(t)orio	107	20	41	41	62
-eggiare	93	15	22	58	65
-$^{(1)}$*erìa*+-$^{(2)}$*erìa*	90	29	54	71	96
-ezza	70	12	12	38	38
-$^{(1)}$*aio*+-$^{(2)}$*aio/a*	48	16	61	33	78
-iere	37	11	32	20	41
-oide	37	6	125	16	135
-nza	29	13	46	25	58
-aggio	29	28	95	39	106
-aggine	28	7	7	19	19
-aceo	28	0	7	2	9
-ificio	25	22	34	26	38
-ume	24	3	3	20	20
-ificare	20	5	24	9	28
-essa	19	3	4	8	9
-toio/a	18	1	22	6	27
-iera	17	17	36	28	47
-ame	17	5	10	16	21
-iero	15	10	21	14	25
-astro	12	1	4	1	4
-ore	9	0	0	2	2
-eto/a	9	8	16	14	22
-aglia	8	0	1	5	6
-evole	6	2	2	7	7
-izio	4	2	10	6	14
-aneo	1	0	1	0	1
-igia	1	0	0	0	0
-izia	0	0	0	0	0
-estre	0	0	0	0	0

al dictionary, as discussed in § 1.1. First, the borderline affixes (*-issimo*, *-mente*, and *-iano*) are strongly lowered in the ranking. This is to be expected, given that usually dictionaries tend to neglect productively-formed, transparent forms, and rather cover the more frequent and idiosyncratic items. This is most heavily reflected in the irrelevant figures for the elative suffix *-issimo*: this suffix is fully disregarded in the dictionary, since it is considered as inflectional according to the Italian grammatical tradition (cf. fn. 7). Similar observations also apply to *-mente* and *-iano*: many regular new formations go unnoticed, although their derivational character is not doubted for Italian.

On the other hand, as discussed in § 3, GRADIT systematically reports all *-ese* ethnic derivatives. Since the main sources for such ethnic adjectives date back to the sixties, this suffix turns out to display a huge productivity. This distortion is clearly due to the dictionary, and makes the data for *-ese* useless.

As for the different usage levls, the number of hapaxes in the corpus is far higher than the number of new formations also labelled as common in the dictionary. Much closer is the matching with the other sorted lists containing specialistic and low-usage words. This is quite reasonable, since the hapaxes may be words not well established in the lexicon.

The resulting ranking once again identifies the subset of productive derivational affixes (*-ismo*, *-ista*, *-ità*, *-(z)ione*, *-(t)ore*), as well as the group of scarcely or non-productive affixes (*-nza*, *-ificare*, *-ore*, *-evole*, *-izia*). However, the figures for the adjectival suffix *-bile* are considerably lower with respect to the corpus-based measurements, presumably because it approximates the behaviour of a verbal participle being productive with every transitive verb (cf. Ricca in press). Also for the suffix *-trice* the dictionary reports a lower number of new formations: further evidence that the dictionary displays a clear bias against formations approaching inflectional behaviour (one should recall in this respect the agreement function of *-trice* shown in (3d)).

A lower number of new formations is also found for the abstract suffix *-mento* with respect to its counterpart *-(z)ione*. Besides the observation that simply regular formations may not be reported in dictionaries, a further consideration plays a role here. Namely, *-mento* is not particularly suited for specialistic terminologies, while it is highly productive for the mere nominalization function, as can be gathered from its higher *P*-index in Table 7. On the other hand, *-(z)ione* is widely used in terminologies and in specialistic domains, as well as the suffix *-(t)ura* which in dictionary data appears even more productive than *-mento*. Similar considerations hold also true for -*ale*/*-are* and *-aggio*, which are much overrepresented in GRADIT with respect to their respective hapax number in the corpus. Finally, both prefixes *in-* and *ri-* are clearly underrepresented compared to their respective hapax number: presumably, this again goes back to the high regularity shown by these formations expressing a fairly grammatical meaning. Moreover, they are mostly not amenable to specialistic domains.

Turning our attention to less frequent affixes not reported in

Table 7, the general outline looks similar. For instance, *-oide* joins the group of affixes belonging to specialistic domains which are strongly enhanced in GRADIT. On the other hand, the 'genitival' suffix *-esco* behaves similarly to *-iano*, and the evaluative suffixes *-accio* and *-astro* match the elative suffix *-issimo*, displaying a much smaller number of new formations in GRADIT, due to the attitude typical of lexicographers of neglecting evaluative derivatives, felt as not forming 'new lexemes'. This attitude is also reflected in the huge differences between hapaxes in the corpus and new formations in GRADIT for all evaluative prefixes, and in particular *super-*, *ultra-*, *maxi-*, *mini-* and *mega-*. For the other two prefixes *micro-* and *iper-*, the number of new formations reported in GRADIT is to a large extent due to their usage in specialistic domains, as seen in § 3.

As for *-(t)orio*, the result may be surprising, since it is usually considered to belong to specialistic domains. Thus, we would have expected a higher number of new formations in GRADIT than in the corpus. The opposite however obtains. This may be due to a peculiar feature developed by this suffix in recent years and correspondingly reflected in the newspaper, namely a derogatory connotation, which makes the suffix largely available also for verbal bases of common use: e.g., *sussulto castratorio* 'castrating impulse', *contenuto dileggiatorio* 'mocking content', *ipotesi smembratoria* 'dismembering hypothesis', etc. (cf. Ricca in press for more details). Presumably, this relevant extension of the suffix domain has not yet been captured by lexicographers, and explains why *-(t)orio* ranks lower in dictionary data.

5. Conclusion

The aim of this paper was essentially empirical: namely, to give a comprehensive picture of Italian derivation from a quantitative point of view. From a large newspaper corpus comprising 75,000,000 tokens, a substantial amount of frequency data (both in tokens and types) for derivational affixes has been extracted: the total of the derived words accounted for amounts to around 5,000,000 tokens and 30,000 types. Moreover, a quantitative evaluation of productivity has been given for about a half of the affixes considered (i.e. all those exhibiting the highest frequencies) following Baayen's productivity index, as modified adopting the variable-corpus approach discussed elsewhere (cf. Gaeta & Ricca 2003). The corpus figures for types and hapaxes have been systematically compared with data available from GRADIT, the most comprehensive and up-to-date lexicographical source for Italian which allows for electronic queries. We are confident to have shown that in this domain corpus-based data are on the whole much more reliable, since they escape some well known distortions in lexicographers' practice, which understandably tends to disregard many new formations among the more general and regular processes, and conversely to overstate the impact of specialistic terminologies, with respect to what is found in real language interaction.

Address of the Authors:

Dipartimento di Scienze del Linguaggio, Università di Torino, Via S. Ottavio 20 - 10124 Torino <livio.gaeta@unito.it>, <davide.ricca@unito.it>

Notes

[*] This work, developed within the FIRB-project "L'italiano nella varietà dei testi", co-ordinated by Carla Marello, 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 §§ 1 and 4 and D.R. for §§ 2 and 3.

[1] From the data-base obtained via DBT, we extracted the complete list of word forms in direct and inverse alphabetical order, with each word form carrying its token frequency. From these lists, all the occurrences of a given affix were extracted and lemmatized, and finally made ready for type/token calculations, after an unavoidable and much time-consuming manual check. This last stage is necessary to eliminate all endings which are not suffixes and to group all misprints together with their correct type.

[2] To make the variable-corpus approach feasible, the corpus must be structured in single text chunks that can be computed separately, providing subcorpora matching the required *N* value for different affixes. So, for instance for a very frequent suffix like *-er* a much smaller subcorpus will be needed to sample 50,000 tokens than for a much less frequent suffix like *-ee*. The comparison is made possible by the overall constant frequency of the affixes throughout the whole corpus (see Table 1). Such a design underlies our corpus, which has been structured in 36 subcorpora, each corresponding to one-month issues of *La Stampa*.

[3] As observed by Bauer (2001:155), the problem with the hapax-conditioned degree of productivity 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".

[4] Low frequency affixes may be slightly more sensitive to corpus size, because possible idiosyncrasies in the distribution of a single lexical item become more relevant. A check was made for the medium-frequency suffix *-trice* and the low-frequency suffix *-essa*, giving fairly stable results. In fact, for a 12-months subcorpus the relative token frequency of *-essa* is 0.091 to compare with the full-corpus value of 0.097; for *-trice* the 12-months count gives 0.33 to compare with 0.32 in the full corpus.

[5] For instance, Corbin (1987:188) labels such items as “mots complexes non construits”, and comments: “Linguistiquement, ce sont des mots non construits, qui ont néanmoins une structure interne”. But she also concludes listing them as lexical non-derived units (1987:463).

[6] In this respect, Dressler (1989:7) speaks of gender agreement by derivation.

[7] Although *-issimo* is usually considered inflectional according to the Italian grammatical tradition, we also included it into our counts, given the interesting number of restrictions shown by this suffix (cf. Rainer 1983, 2003). Independently of any theoretical assumption about its status, its behaviour can be straightforwardly compared to other suffixes such as *-mente* or *-iano* displaying similar properties, as will be shown.

[8] Since our corpus is not tagged, it is impossible to treat conversions, affixes with too little phonetic substance (e.g, suffixes like *-ìo*, *-ìa* or prefixes like *a-* and *s-*, especially in parasynthetic verbs) and affixes displaying widespread homonymy with other word forms, and particularly the verbal participle. This is the case of the deverbal suffix *-ata*, and the relational adjectival suffix *-ato*. Similarly, discriminating between the different homonymous suffixes ending in *-ino*, even though theoretically possible, would have been too much time-consuming.

[9] The percentage of derived words taken into account becomes still more remarkable if one considers the weight of functional words in the whole corpus: taking only the functional words occurring among the 100 most frequent word forms, they amount to more than 31,000,000 tokens, about 42% of all tokens!

[10] Needless to say, there are suffixes which exhibit still a lower type frequency in our corpus and in dictionaries as well. If it is uncertain whether a single-instance ending can be considered a suffix, even in case of full transparency of the putative base (cf. *caduco* ‘ephemeral’ from *cadere* ‘to fall’), it is normally assumed that two instances may suffice. Clearly, there is no question of productivity here.

[11] It must be emphasized, however, that great caution must be taken in mentioning statistic parameters like standard deviation with reference to the distributions in Table 3 and 4, since the 58 affixes considered obviously do not exhaust Italian derivation, and it remains unclear to what extent they can be taken as a reliable sample for the entire population.

[12] It must be noticed that the usage labels mentioned so far largely rely on lexicographers’ subjective evaluation with scarce empirical support except for the core vocabulary. However, this is an inherent feature of dictionary-based evidence, and we assume that, on the average, lexicographers’ intuition in these matters can be considered informative, despite unavoidably idiosyncratic judgements for single items.

[13] The similarity in total type frequency does not mean, of course, that there is full overlap with respect to single types. The corpus reports many new formations not registered in the dictionary, while many everyday words escape it.

[14] Notice that the few entries of *-issimo/a* in GRADIT mainly refer to its marginal denominal use, as in *governissimo* ‘very strong government’, *rigorissimo* ‘indisputable penalty’, etc.

[15] Apart from technical problems, one wonders whether the P-index is reasonable for affixes displaying very low token frequencies. There is no space here to tackle this question, and we have to refer the reader to Gaeta & Ricca (2003).

[16] The only affix with token frequency above 19,000 which does not occur in Table 7 is *-iere*, and this for technical reasons: the separation of its word-forms from those of *-iera* and *-iero* for each subcorpus (and nearly every word!) would have been far too much time-consuming.

[17] The dark grey cells correspond to values of *N* which are too high for the least

frequent affixes, with no available data; the light grey ones, on the contrary, correspond to values of *N* which would be too low to be reliable for the most frequent affixes, since they would refer to a subcorpus under the threshold of 6 millions tokens.

[18] For scarcely or non-productive affixes such as *-evole*, *-nza* or *-ore* some hapaxes are new formations as well (although often wih some jocular connotation), but other instances turn out to be very rare or archaic words and not new coinages at all: e.g. *perdonanza* 'forgiveness', *bisognevole* 'needy', *buiore* 'darkness'.

[19] As an illustration of the uncertainty remaining in this domain, we had regrettably to leave out of consideration all entries carrying only the rough indication "20th century".

References

ALBANO LEONI Federico, Daniele GAMBARARA, Stefano GENSINI, Franco LO PIPARO & Raffaele SIMONE, eds. (1997), *Ai limiti del linguaggio*, Bari, Laterza.

APRESJAN Julius D. (1974), "Regular polysemy", *Linguistics* 12:5-32.

ARONOFF Mark (1983), "Potential Words, Actual Words, Productivity and Frequency", in HATTORI & INOUE (1983:163-171).

BAAYEN Harald (1989), *A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation*, PhD. Diss. Vrije Universiteit, Amsterdam.

BAAYEN Harald (1992), "Quantitative aspects of morphological productivity", in BOOIJ & VAN MARLE (1992:109-149).

BAAYEN Harald (1993), "On frequency, transparency and productivity", in BOOIJ & VAN MARLE (1993:181-208).

BAAYEN Harald (2001), *Word-Frequency Distributions*, Dordrecht, Kluwer.

BAAYEN Harald & Rochelle LIEBER (1991), "Productivity and English word-formations: a corpus-based study", *Linguistics* 29:801-843.

BAAYEN Harald & Annette RENOUF (1996), "Chronicling the Times: Productive lexical innovations in an English newspaper", *Language* 72:69-96.

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

BAUER Roland & Hans GOEBL, eds. (2002), *Parallela IX. Testo variazione informatica / Text Variation Informatik. Atti del IX Incontro italo-austriaco dei linguisti*, Wilhelmsfeld, Egert.

BENINCÀ Paola, Alberto MIONI & Laura VANELLI, eds. (1999), *Fonologia e Morfologia dell'italiano e dei dialetti d'Italia. Atti del XXXI Congresso Internazionale di Studi della Società di Linguistica Italiana*, Roma, Bulzoni.

BENINCÀ Paola, Guglielmo CINQUE, Tullio DE MAURO & Nigel VINCENT (eds.) (1996), *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio C. Lepschy*, Roma, Bulzoni.

BITTNER Andreas, Frans PLANK & Patrick STEINKRÜGER, eds. (in press), *On Inflection. In Memory of Wolfgang U. Wurzel*, Berlin / New York, Mouton de Gruyter.

Booij Geert & Jaap van Marle (eds.) (1988), *Yearbook of Morphology*, Dordrecht, Kluwer.

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

Booij Geert & Jaap van Marle (eds.) (1993), *Yearbook of Morphology 1992*, Dordrecht, Kluwer.

Bortolini Umberta, Carlo Tagliavini & Antonio Zampolli (1971), *Lessico di frequenza della lingua italiana contemporanea*, Milano, IBM Italia.

Corbin Danielle (1987), *Morphologie dérivationelle et structuration du lexique*, Tübingen, Niemeyer.

Dardano Maurizio, Wolfgang U. Dressler & Gudrun Held, eds. (1983), *Parallela. Atti del 2° convegno italo-austriaco*, Tübingen, Gunter Narr.

De Mauro Tullio (2000), *GRADIT – Grande Dizionario Italiano dell'Uso*, Torino, UTET.

De Mauro Tullio, Federico Mancini, Massimo Vedovelli & Miriam Voghera (1993), *Lessico di frequenza dell'italiano parlato*, Milano, Etas Libri.

Dietrich Wolf (ed.) (1987), *Grammatik und Wortbildung romanischer Sprachen*, Tübingen, Gunter Narr.

Dogliotti Miro & Luigi Rosiello (1998), *Zingarelli 1998 in CD-Rom. Vocabolario della lingua italiana di Nicola Zingarelli*, 12th ed., Bologna, Zanichelli.

Dressler Wolfgang U. (1985), "On the predictiveness of Natural Morphology", *Journal of Linguistics* 21:321-337.

Dressler Wolfgang U. (1989), "Prototypical Differences between Inflection and Derivation", *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 42:3-10.

Dressler Wolfgang U. & Maria Ladányi (2000), "Productivity in word formation (WF): a morphological approach", *Acta Linguistica Hungarica* 47:103-144.

Gaeta Livio (1999), "Polisemia e lessicalizzazione: un approccio naturalista", *Italienische Studien* 20:7-27.

Gaeta Livio (2002), *Quando i verbi compaiono come nomi. Un saggio di morfologia naturale*, Milano, Franco Angeli.

Gaeta Livio (in press), "Inflectional morphology and productivity: Considering qualitative and quantitative approaches", in Bittner *et al.* (in press).

Gaeta Livio & Davide Ricca (2002), "*Corpora* testuali e produttività morfologica: i nomi d'azione italiani in due annate della *Stampa*", in Bauer & Goebl (2002:223-249).

Gaeta Livio & Davide Ricca (2003), "Italian prefixes and productivity: a quantitative approach", *Acta Linguistica Hungarica* 50:93-112.

Gaeta Livio & Davide Ricca (ms.), "Productivity in Italian word formation: A variable-corpus approach", ms., University of Turin.

Grossmann Maria & Franz Rainer, eds. (in press), *La formazione delle parole in italiano*, Tübingen, Niemeyer.

Hattori Shirô & Kazuko Inoue (eds.) (1983), *Proceedings of the XIII International Congress of Linguists*, Tokyo, Permanent International Committee on Linguistics.

LAUDANNA Alessandro & Cristina BURANI (1999), "I processi lessicali: come è rappresentata la struttura morfologica delle parole?", in BENINCÀ *et al.* (1999:613-626).

NEUHAUS H. Joachim (1973), "Zur Theorie der Produktivität von Wortbildungssystemen", in TEN CATE & JORDENS (1973:305-317).

PLAG Ingo (1999), *Morphological productivity. Structural constraints in English Derivation*, Berlin / New York, Mouton de Gruyter.

PLAG Ingo, Christiane DALTON-PUFFER & Harald BAAYEN (1999), "Morphological productivity across speech and writing" *English Language and Linguistics* 3:209-228.

RAINER Franz (1983), "L'intensificazione di aggettivi mediante '-issimo'", in DARDANO *et al.* (1983:94-102).

RAINER Franz (1987), "Produktivitätsbegriffe in der Wortbildungslehre", in DIETRICH (1987:187-202).

RAINER Franz (1988), "Towards a theory of blocking: the case of Italian and German quality nouns", in BOOIJ & VAN MARLE (1988:155-185).

RAINER Franz (1989), *I nomi di qualità nell'italiano contemporaneo*, Vienna, Braunmüller.

RAINER Franz (1993), *Spanische Wortbildungslehre*, Tübingen, Niemeyer.

RAINER Franz (2001), "Compositionality and paradigmatically determined allomorphy in Italian word-formation", in SCHANER-WOLLES *et al.* (2001:383-392).

RAINER Franz (2003), "Studying restrictions on patterns of word-formation by means of the Internet", this volume.

RICCA Davide (in press), "Aggettivi deverbali", in GROSSMANN *et al.* (in press).

SABATINI Francesco & Vincenzo COLETTI (1997), *DISC Compact – Dizionario Italiano Sabatini Coletti. Edizione in CD-Rom*, Firenze, Giunti.

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

SCALISE Sergio (1996), "Preliminari per lo studio di un affisso: *-tore* o *-ore*?", in BENINCÀ *et al.* (1996:291-307).

SCHANER-WOLLES Chris, John RENNISON & Friedrich NEUBARTH (eds.) (2001), *Naturally! Linguistic studies in honour of Wolfgang Ulrich Dressler presented on the occasion of his 60th birthday*, Torino, Rosenberg & Sellier.

TEN CATE Abraham P. & Peter JORDENS (eds.) (1973), *Linguistische Perspektiven: Referate des VII Linguistischen Kolloquiums*, Tübingen, Niemeyer.

THORNTON Anna M. (1990-91), "Sui deverbali italiani in *-mento* e *-zione* (I-II)", *Archivio Glottologico Italiano* 75:169-207 and 76:79-102.

THORNTON Anna M. (1997), "Quali suffissi nel "Vocabolario di base"?", in ALBANO LEONI *et al.* (1997:385-396).

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 (1992:151-163).