Taiwan Journal of Linguistics
Vol. 10.2, 89-118, 2012

ON THE PRODUCTIVITY OF THE CHINESE SUFFIXES −兒 −R, −化 −HUÀ AND −頭 −TOU [*]

Giorgio Francesco Arcodia and Bianca Basciano
University of Milano-Bicocca
University of Verona

ABSTRACT

The notion of 'productivity' is an essential one in the study of linguistic morphology, but its definition is indeed challenging, and there are different ways to measure different aspects of the productivity of a morphological process. In this paper we shall adopt Baayen's *P* measure of productivity for a corpus-based study of the productivity of three Mandarin derivational suffixes, namely the nominalizer/diminutive −兒 −*r*, −化 −*huà* '−ise, −ify' and −頭 −*tou*, a 'dummy' nominal suffix (Lin 2001:82), in order to assess how this index relates to our received knowledge about the productivity of such forms, and, also, to compare our results with a previous study by Nishimoto (2003) on a small corpus of Modern Chinese. Moreover, in a diachronic perspective, we shall compare data from the *Academia Sinica Tagged Corpus of Early Mandarin Chinese* and from the *Academia Sinica Balanced Corpus of Modern Chinese.* We shall show that our *P* values mostly reflect what descriptive works tell us about the productivity of the affixes considered here in two different periods of the history of the language; when corpus data for previous stages of a language are available, they appear as a better basis for assessments on the profitability of a morphological process than dictionary data.

Key words: Chinese, morphology, derivation, productivity

---

[*] An earlier version of this paper ("Measuring the morphological productivity of two Chinese affixes") was presented at the 20th International Conference on Historical Linguistics (Osaka, Japan, July 2011); the authors would like to thank the participants to the discussion, as well as Sergio Scalise and Vito Pirrelli, for their insightful comments. Traditional Chinese characters and the *Pinyin* romanization system are used throughout the paper. For academic purposes, Giorgio F. Arcodia is responsible for sections 3.1, 3.2 and 4.2, Bianca Basciano is responsible for sections 1, 2, 4.1 and 5.

1. INTRODUCTION

Derivation is a morphological process which results in the creation of a new word from an existing word/lexical morpheme, most often with the addition of an affix (Beard 1998:55; Naumann and Vogel 2000:929). Thus, the English words *enlist* and *enshrine* are derived from *list* and *shrine*, and *peripheralize* and *Clintonize* are derived from *peripheral* and *Clinton*; however, it seems that no more words may be formed with the addition of the prefix *en−*, whereas *−ize* is "happily" employed to form new verbs in English (Plag 2006b:537). We usually say that the prefix *en−* (or, depending on the theoretical framework, the 'word formation rule', the 'word formation schema', etc.; see Booij 2010) is unproductive, whereas the suffix *−ise* is productive, i.e. it can be used to form new words.

The notion of productivity, however, "is among the least clear concepts in linguistics", as suggested by Mayerthaler (1981, quoted in Bauer 2001:1). In its most general sense, 'productivity' is rule-based creativity, be it building a new sentence or a new word; we refer to the latter as morphological productivity. As to morphological productivity, there are both qualitative and quantitative approaches to this notion (Plag 2006a-b), and different ways to measure different aspects of productivity, but there is no real consensus on the methodology, and all procedures have their weak points (Bauer 2001:207; see below, section 3.1).

In this paper, we chose Baayen's hapax-based *P* measure of productivity (Baayen 1989, 1992, Baayen and Lieber 1991) to assess the productivity of three Chinese derivational suffixes, namely the nominalizer/diminutive −兒 *−r*, −化 *−huà* '−ise, −ify' and −頭 *−tou*, a 'dummy' nominal suffix (Lin 2001:82), in order to check the validity of this measure by comparing it to previous work on the topic (Nishimoto 2003) and by relating it to what descriptive works on the Chinese lexicon and morphology tell us on the ability of those affixes to form new words. Our data will be drawn for the *Academia Sinica Tagged Corpus of Early Mandarin Chinese* and from the *Academia Sinica Balanced Corpus of Modern Chinese*, two medium-sized tagged corpora of Early Mandarin

and Modern Mandarin, thus providing also an evaluation of the shifts in the productivity of the formants considered through time.

This paper is organized as follows. Firstly, we shall introduce some key aspects of word formation in Chinese, focussing on the items to be analysed here. Secondly, we shall discuss the issue of the definition(s) of morphological productivity, and we shall illustrate the productivity index we adopted in our study, namely Baayen's hapax-based *P* measure. We shall then present our methodology and data, highlighting the significance of our results for further studies of Chinese morphology. In the last section of this paper we shall summarize our main conclusions.

2. WORD FORMATION IN MANDARIN CHINESE

While many scholars believe that Old Chinese had subsyllabic morphology (see e.g. Baxter and Sagart 1998), in Modern Chinese each morpheme consists (at least) of a syllable, which corresponds to a character in the written language:

(1)	火	貓	書
	huǒ	*māo*	*shū*
	fire	cat	book

The 'typical' Mandarin word, however, is not monomorphemic; according to one estimate (Xing 2006), around 80% of Modern Mandarin words are multimorphemic:

(2)	火山	熊貓	書店
	huǒ-shān	*xiǒng-māo*	*shū-diàn*
	fire-mountain	bear-cat	book-shop
	volcano	panda	bookshop

The basic units of Chinese word formation are lexical morphemes, many of which are bound morphs; to date, there is no consensus among linguists as to whether derivation in Mandarin Chinese is an independent process of word formation, distinct from compounding (see e.g. Pan, Ye

and Han 2004, Arcodia 2012). For instance, Mandarin speakers can build a noun for a member of any ethnic group, a citizen of any country, territory, region or city by adding the (free) morpheme 人 *rén* 'person' to the name of the ethnic group, country or city (Dong 2004; our examples):

(3)	回族人	英國人	上海人
	Húizú-rén	*Yīngguó-rén*	*Shànghǎi-rén*
	ethnic Hui	Briton	Shanghainese

The meaning of the words in (3) are the sum of the meanings of their constituents, they are compositionally transparent; moreover, the pattern appears to be 'fully productive', i.e. it can be used any time a new word with the above described meaning is needed (as e.g. 羅安達人 *Luóāndárén* 'person from Luanda'). In such cases, which are quite common in Mandarin word formation, it is a matter of debate whether to regard the pattern as derivation or compounding (see e.g. Ma 1995, Packard 2000, Dong 2004). Pan, Ye and Han (2004:77ff.) analysed a sample of 14 works dealing with Chinese morphology published between 1932 and 1982, and they report that no less than 340 different morphemes have been analysed as affixes (or affixoids) at least once, but only 16 among those may be found in the majority of the works considered; on the other hand, 223 morphemes (around two thirds of the total) were labelled as affixes / affixoids only by one author. It thus appears that there is no consensus on what processes of word formation (if any) are to be regarded as derivation in Modern Chinese.

Nevertheless, according to the data from Pan, Ye and Han referred to above, there actually are a few items which are regarded as derivational affixes in most descriptive works on Mandarin. They include, among others, the three formants considered in our study: namely, the nominal (and weakly diminutive) suffix −兒 −*r*, verb-forming −化 −*huà*, roughly corresponding to Eng −*ise*, −*ify*, and the 'dummy' nominal suffix −頭 −*tou*:

(4)	花兒	現代化	想頭	
	huā-r	*xiàndài-huà*	*xiǎng-tou*	flower-R
	modern-HUA	think-TOU		
	flower	modernise	idea	

The suffix −兒 *−r* is the only exception to the principle outlined above, according to which there are no subsyllabic morphemes in Mandarin (Li and Thompson 1981:39); the suffix merges with the morph to its left and, if a consonantal coda is present, then −兒 *−r* substitutes it, as in 根兒 *gēnr* 'root' ([n] is dropped). Originally a diminutive suffix, it is now attached to a root to express "a sense of smallness, intimacy, familiarity, colloquialism and/or casualness" (Lin 2001:57), as in 老頭兒 *lǎotour* 'old man' (compare 老頭子 *lǎotóuzi* 'old fogy/codger'); it is a typical feature of colloquial Beijing Mandarin. According to Chen (1999:39), −兒 *−r* has three different uses; it can be used to actually build a new word from an existing one:

(5)	白面	白面兒
	báimiàn	*báimiànr*
	(white) flour	heroin

In other instances, the addition of −兒 *−r* does not produce a new word, but if this suffix is not present the word sounds "unnatural and stilted":

(6)	盆	盆兒
	pén	*pénr*
	basin	

Also, −兒 *−r* is sometimes used in Beijing casual speech as a 'substitute' for other syllables:

(7) 多少錢 多兒錢
duōshao qián *duōr qián*
how.much money how.much money
'how much money'

We shall consider −兒 *−r* as a suffix only in instances like those in (5) and (6); the 'rhotacization' of syllables as exemplified in (7) is a syntactic phenomenon typical of informal speech, rather than a morphological process. Also, since our study concerns −兒 *−r* as a nominal affix, we shall not take into consideration those (few) instances of verbs as 玩兒 *wǎnr* 'to play, have fun'. Note that some −兒 *−r* words are regarded as non-standard; according to Chen (1999:39), the use of −兒 *−r* word forms has decreased in radio and television broadcasting since the nineties.

The bound morph −化 *−huà* '−ise, −ify, −en' is also often regarded as a derivational suffix; it can attach to words belonging to any major word class (mostly, adjectives and nouns):

(8) 神化 軟化 崇敬化
shén-huà *ruǎn-huà* *chóngjìng-huà*
god-HUA soft-HUA respect-HUA
'deify' 'soften' 'respectify' (?)

Although −化 *−huà* is commonly described as a verbalizing suffix, many −化 *−huà* derived words are actually ambiguous between verbal and nominal usage (Baxter and Sagart 1998), as 世界化 *shìjiè-huà* (world-HUA) 'universalise' / 'universalisation'. It is generally agreed that this suffix is functionally equivalent to suffixes as Eng. *−ise* / *−ify*, French *−iser* / *−ifier*, Italian *−izzare* / *−ificare*; −化 *−huà* is claimed to have been 'imported' from Europe through the mediation of Japanese (following the May Fourth Movement of 1919; Wang 1980:311). Later on, −化 *−huà* developed independently from the original model in Chinese, and it began to be used to create new words by analogy (Steffen Chung 2006:202). *huà* can also be used as a free morpheme, meaning 'to melt', 'to make disappear', etc. (as in 化了 *huà le* 'melt-PFV = melted')

The nominal suffix −頭 −*tou* is called a "dummy affix" by Lin (2001:82) because it gives no semantic contribution to the base it is combined with; its only function is that of bearing nominal word class, and it can attach to nominal, verbal and adjectival roots[1]:

(9)	石頭	想頭	苦頭
	shí-tou	*xiǎng-tou*	*kǔ-tou*
	stone-TOU	think-TOU	bitter-TOU
	'stone'	'idea'	'suffering'

What does the literature tell us about the productivity (in its broadest sense, as the possibility to form new words) of such forms? In Wang (1980), it is argued that −化 −*huà* became productive only in the XXth century; Nishimoto (2003) suggests that the "regularity" of this suffix in word formation leads to the expectation that it be very productive. Nishimoto also remarks that Li and Thompson (1981) regard −頭 −*tou* as no longer productive, whereas Lin (2001) claims that −兒 −*r* is the most productive among Mandarin suffixes, and −頭 −*tou* is less productive; however, "the basis for these observations is left unclear" (Nishimoto 2003:53). What do we exactly mean when we say that an affix (or a rule) is productive? Is productivity (only) something which an affix either has or has not, or is it a gradual property? This will be the topic of the next section.

[1] As pointed out by an anonymous reviewer, it is not entirely correct to say that −頭 −*tou*, which derives from the lexeme 頭 *tóu* 'head', is a completely empty morpheme; it means something like 'a concentration, crystallization, gathering in one place of', and is still used in nominal V−頭 −*tou* constructions, as e.g. 玩頭 *wántou* 'have.fun-TOU = fun'.

3. DEFINING AND MEASURING PRODUCTIVITY

3.1 Qualitative *vs.* Quantitative Approaches to Productivity

As mentioned before, productivity is a composite notion, which contains different aspects. In general terms, productivity "deals with the number of new words that can be coined using a particular morphological process", and may be understood in at least two distinct senses, namely "availability" and "profitability" (Bauer 2001:205-211). We say that a morphological process is available if it can still be used in a synchronic stage of a given language to build new words; either a process is available, as −*ise* in English, or it is not, as *en*− (see section 1). When we deem a certain pattern as no longer productive, it amounts to saying that it is unavailable, in Bauer's sense. What is more interesting, in our perspective, is the other sense in which the word 'productivity' may be understood, namely profitability (Bauer 2001:201 and 207):

> "The profitability of a morphological process reflects the extent to which its availability is exploited in language use (...).
>
> (…) there are various ways of measuring productivity in this profitability sense, both direct and indirect, but with no general agreement on how it should be done and no genuinely problem-free procedure available"

In what follows, we shall understand productivity in this latter sense; 'fully productive' and 'unproductive', hence, represent the two extremes of the profitability scale (Plag 2006a)[2].

[2] Actually, labelling a process as 'fully productive' or 'unproductive' is not as straightforward as it seems. As remarked by Plag (2006a), a seemingly unproductive affix/process as Eng. −*th* may occasionally be used to coin new words (*greenth*); it is unclear whether these forms are the product of some rule or of simple analogy, and the

How do we assess the profitability of a word formation process? Several methods have been proposed in the literature (see the summary in Plag 2006a and 2006b). One simple way of measuring productivity is that of counting the number of words containing a given affix in an unabridged dictionary; this is a type-based measure of the profitability of a process. By such procedure, however, what one actually measures is the productivity of a process in the past; there is a high number of words containing the suffix *−ment* in the lexicon of Present Day English, but a significant share of those terms were introduced between the XVIth and the XIXth century, and nowadays the suffix is virtually unproductive (Plag 2006a:122).

Another approach is that of counting the neologisms built according to a certain derivational pattern in a given period; if different periods are analysed, the changes in the profitability of a process become visible. However, the problem with this method is that dictionary data are not fully reliable in this respect. Firstly, the fact that a word is not listed in a dictionary (no matter how big) does not necessarily mean that it does/did not exist; it may just have been left out or gone unnoticed by the compilers (Plag 2006b:541). Moreover, unabridged historical dictionaries such as the *Oxford English Dictionary* are available only for a very small number of languages and, thus, this dictionary-based approach is not viable for the vast majority of the languages of the World (Plag 2006a:123). Note also that the abundance of types of a derivational affix in a dictionary may also be interpreted as an index of *low* productivity: Packard (2000:71-73) argues that a very productive affix yields a number of derived words which is so high that they cannot be exhaustively listed; hence, according to him, it is more likely for words built according to unproductive or not very productive derivational processes to be accepted in a dictionary. This is especially true for very transparent derivational affixes with a general meaning, as, say, Eng. *−ly*; since the meaning of these derivatives is normally predictable, it is not

---

very nature of rules, processes, analogy etc. are understood differently in the various theoretical framework.

necessary to list them all, whereas the opposite may hold for unproductive (or almost unproductive) affixes. We shall get back to this point later.

The method we chose to adopt to measure the profitability of morphological processes, following Nishimoto (2003), is the hapax-based *P* index developed by Baayen (1989, 1992 and Baayen and Lieber 1991). The assumption behind this method is that if an affix is very productive, we expect to find many *hapax legomena* containing that affix in a large text corpus: it is just among hapaxes that we typically find "the higher proportion of neologisms", and thus "the number of hapaxes of a given morphological category correlates with the number of neologisms in that category"; as pointed out before, many neologisms are indicative of high productivity (Plag 2006a:123; see also Renouf and Baayen 1996). The use of corpora rather than dictionaries as a source of data is motivated by the fact that in a corpus we may find productively formed derivates which are not listed in dictionaries, and thus "corpus-based descriptions of productivity reflect how words are actually used" (Nishimoto 2003:51). However, in a small corpus many hapaxes may actually be just 'ordinary' words of the language; the larger the corpus, the higher the number of neologisms one finds among *hapax legomena* (Plag 2006b:542-543).

Baayen's *P* index is obtained by the formula below:

$$P = \frac{n_1}{N}$$

Where n_1 stands for the number of *hapax legomena* with a given affix and *N* stands for the number of tokens of the same affix in the corpus considered. Differently from the measures discussed above, types play no role in the calculation of the *P* index; by such index, we measure the synchronic aspect of productivity, rather than historical profitability (Nishimoto 2003:53). However, as pointed out in the introduction, by comparing hapax-based measures of productivity at different points in time we may assess the change in the profitability of a word formation process.

If all of the words found in a text sample are hapaxes, the *P* index will be 1, i.e. maximal productivity, whereas many high frequency word increase the value of n_1 and, hence, lead to a low *P* productivity index. Productivity is thus understood as the likelihood that a new derived word with a certain affix will be found in a text corpus, after *N* tokens have been sampled. In this model, high token frequency is connected with a high degree of lexicalization (as *storage in the lexicon*) and low productivity, and vice versa; *hapax legomena* are often unfamiliar words, but they are understandable for the hearer or reader if the process / rule which created them is still 'active'. As Plag (2006a:123) puts it,

> "Productive processes are therefore characterized by large numbers of low-frequency words and small numbers of high-frequency words. The many low-frequency words keep the rule alive, because they force speakers to segment the derivatives and thus strengthen the existence of the affix. Unproductive morphological categories will, in contrast, be characterized by a preponderance of words with rather high frequencies and by a small number of words with low frequencies."[3]

This hapax-based measure of productivity has some known shortcomings. It is not a fixed measure of productivity, since figures are comparable only in corpora of roughly the same size; also, it can produce nonsensical results (Bauer 2001:150-153). For instance, in the 1 million word Wellington Corpus of Written New Zealand English, the suffix *–iana* occurs only once (in the word *Victoriana*), leading to *P* = 1 (1 token / 1 hapax), i.e. full productivity; according to Bauer, such paradoxical results may be avoided by using a larger sample, although "there is not enough information available to be able to give a precise estimate of the size of the sample that would be required to give a reliable statistic in

[3] See also Bauer (2001:151): "(...) with a widely generalised but unproductive process, each type should have a high token frequency, which should keep the productivity index low".

this case" (2001:151). Generally speaking, an increase in the size of the sample leads to increased accuracy in calculating the *P* index.

As shown in Plag *et al.* 1999 and Plag 2002 (see the summary in Plag 2006b:544-546), if one assesses the productivity of derivational affixes with the measures described above, it is possible that the same affix scores high for one measure and low for another, thus having different productivity rankings for different measures; this is because each of those measures "highlights a special aspect of productivity" (Plag 2006a:123). To provide an example of such discrepancies, we shall quote Plag *et al.*'s (1999) data on the English suffixes −*wise* and −*ness*, drawn from the "written language" section of the British National Corpus (version 1.0, 100 million tokens) and from the above mentioned *Oxford English Dictionary* (neologisms of the 20th century):

Table 1. Measures of productivity for −*wise* and −*ness* (adapted from Plag 2006a:124)

Suffix	*V* (types)	n_1 (hapaxes)	*N* (tokens)	*P*	*OED* neologisms
−*wise*	183	128	2091	0.061	12
−*ness*	2466	943	106957	0.0088	279

As can be seen in Table 1, −*ness* has a number of types, hapaxes, neologisms and tokens much higher than −*wise*; however, the *P* measure for −*ness* (0.0088) is significantly lower than that for −*wise* (0.061). Thus, whereas −*wise* appears more productive than −*ness* if one considers the hapax-based *P* index, all the other measures indicate that the former is actually *less* productive than the latter. This apparent inconsistency is explained by Plag (2006a:123-124) as such:

> "−*Wise* has a small number of types *V* and a small number of hapaxes n_1, which indicates that the suffix is not used very often, neither in terms of different derivatives nor in terms of new formations. Nevertheless, among all tokens with that suffix (i.e.,

> N^{aff} [our *N*]), the number of hapaxes is quite high, leading to a high value of productivity in the narrow sense *P*. This is a sign of the suffix's potential to be easily used for the coinage of new forms, if need be.
>
> The suffix *−ness*, on the contrary, scores very high in terms of type frequency *V* and also has many OED neologisms. Its *P* value is, however, significantly lower than that of *−wise*, because many *−ness* words are also quite frequently used (e.g., *happiness*), leading to a large number of tokens N^{aff} and thus an overall decrease of P"

Again, this shows that the hapax-based *P* index measures the synchronic aspect of productivity, i.e. the possibility of using a certain affix to build new words in the present stage of the language. This is connected to the ability of the language user to understand a new (or unfamiliar) word: if the process (/rule) by which this word has been built is still available, the speaker will be able to segment it into its constituent morphemes and to 'reconstruct' its meaning (see the quotation from Plag 2006a:123 above).

3.2 The *P* Index and Text Corpora

The *P* measure of productivity has previously been applied by Sproat and Shih (1996) in a study of Mandarin root compounding; as to derivation, Nishimoto (2003) measured the *P* index of productivity of five Mandarin suffixes, namely the 'plural' suffix −們 *−men* (see fn. 5), the nominal suffix −子 *−zi*, −兒 *−r*, −化 *−huà* and −頭 *−tou*. In his study, Nishimoto used a "cleaned-up" version of the Mandarin Chinese PH Corpus (see Nishimoto 2003:55 for the details), a 2.4 million words / 3.7 million characters[4] corpus of *XinHua* newspaper articles, collected between January 1990 and March 1991. This corpus is relatively small (compare the 100-million-word British National Corpus mentioned above), and it is very homogeneous, since all of the texts come from

[4] On the relationship between word and character in Chinese, see above, section 2.

newspaper articles. The latter aspect is particularly relevant since, as pointed out by Plag (2006a:124), "it is well known that certain affixes are more commonly found in certain types of texts than in others"; as highlighted before (section 2.), −兒 *−r* is a feature of the colloquial speech of Beijing, and is thus expected to be much less common in written texts, especially in those from an official media outlet such as *XinHua*. Moreover, the PH corpus has no part of speech tags, thus making it difficult e.g. to distinguish verbal from nominal −化 *−huà* derived words (Nishimoto 2003:58).

Because of these limitations of the PH corpus, Nishimoto suggests that the *Academia Sinica Balanced Corpus* (*of Modern Chinese*) be used, since it is much larger, it is made of texts of different kinds and its words are tagged for part of speech, making it possible to isolate only the output class one needs for words containing a certain constituent; as Nishimoto (2003:56, fn. 7) points out,

> "(...) findings from a larger, more balanced corpus do not necessarily minimize findings from a smaller, less balanced corpus.
>
> Findings from both the PH Corpus (a small corpus of newspaper texts) and the Sinica Corpus (a large corpus of a variety of texts) are of interest because corpora of different types enable a comparison of findings by the corpus type."

We picked Nishimoto's suggestion and performed his measurements on three affixes from his sample, namely −兒 *−r*, −化 *−huà* and −頭 *−tou*[3], in the *Academia Sinica Balanced Corpus of Modern Chinese* (henceforth: SCMC), version 3.0. The SCMC is slightly more than twice the size of the PH corpus (approx. 5 million words / 8 million characters),

[3] Since we had to deal with a greater quantity of data than Nishimoto, we considered three affixes only. We chose to exclude −們 *−men* because it is not, strictly speaking, a word forming affix (it acts as a marker of 'collective', rather than plural) and −子 *−zi* because it is particularly hard to separate its affixal uses from the related (and homograph) morpheme 子 *zǐ* 'child', 'egg', etc. (as e.g. in 魚子 *yúzǐ* 'fish roe'); the latter has a third tone, whereas the former has a neutral tone, but this is not indicated in written texts.

and it is composed of a variety of written and oral texts, balanced for topic. The categories according to which the texts of the SCMC are classified are presented in table 2:

Table 2. Classification of texts in the SCMC[4]

Parameter	Categories
Mode	written, written-to-be-read, written-to-be-spoken, spoken, spoken-to-be-written
Style	narration, argumentation, exposition, describe [sic!]
Medium	newspaper, general magazine, academic journal, textbook, reference book, thesis, general book, audio/visual medium, conversation/interview, elsewhere
Topic	philosophy, natural science, social sciences, arts, general/leisure, literature

We must however remark that although the SCMC contains a variety of text types, oral texts account for approximately 10% only of the whole corpus; hence, the difference between the PH corpus and the SCMC mostly lies in the range of variation of textual types and styles, but the share of actual spoken Chinese in the latter is in fact quite modest[5].

Besides comparing our findings with Nishimoto's, we also extracted the same data from the *Academia Sinica Tagged Corpus of Early Mandarin Chinese* (henceforth: SCEMC), a segmented corpus of approximately 3.7 million words (4.4 million characters), tagged for part-of-speech, from seven novels and three collections of theatrical texts dating from the 13th to the 19th century (the period termed 近代漢語 *Jìndài Hànyǔ* in the Chinese linguistic tradition). The SCMC and the SCEMC are obviously not readily comparable, both because of the smaller size of the latter and, especially, because of the huge difference in the variety of texts; however, provided that vernacular novels and plays should reflect both written and spoken varieties much more than

[4] From http://tinyurl.com/bngdzg7; see also http://tinyurl.com/btkzdsh.
[5] We would like to thank an anonymous reviewer for pointing this out to us.

other kinds of documents, we do believe that the comparison is indeed significant[6]. The analysis and comparison of data on the productivity of −兒 *−r*, −化 *−huà* and −頭 *−tou* in two different historical stages of Chinese will also be a testing ground for our received knowledge on the history of the Chinese lexicon.

4. OUR DATA: PRESENTATION AND ANALYSIS

4.1 Modern Chinese

As illustrated in the preceding sections, we analysed all the occurrences of the suffixes −兒 *−r*, −化 *−huà* and −頭 *−tou* in the SCMC and the SCEMC, counting types (*V*), tokens (*N*), hapaxes (N_1) and the *P* index for each of them. First, let us present our data on Modern Chinese, summarised in table 3 (affixes ranked according to their *P* index[7]):

Table 3. Measures of productivity for −兒 *−r*, −化 *−huà* and −頭 *−tou* in the SCMC

Suffix	*V* (types)	n_1 (hapaxes)	*N* (tokens)	*P*
−兒 *−r*	139	62	783	0.079
−化 *−huà*	464	232	4516	0.051
−頭 *−tou*	98	24	1593	0.015

The suffix with the lowest *P* index is −頭 *−tou*, as expected; as remarked earlier (section 2), Lin (2001) suggest that this suffix is not

[6] Generally speaking, the vernacular language of those times (traditional 白話 *báihuà*, as opposed to 20th-century 白話 *báihuà*, i.e. the language of early modern Chinese literature; Chen 1999:69) was not "purely" vernacular, but rather "a mixture of the literary and spoken languages" (Norman 1988:111). Nevertheless, as far as grammar and vocabulary are concerned, it was much closer to the spoken 'standard' than classical literary Chinese, 文言 *wényán*, modelled after the written language of the age between the Spring and Autumn period and the Eastern Han Dynasty (i.e. between the 8th century BCE and the 3rd century CE), which gradually became divorced from any common parlance (Chen 1999:67).

[7] Figures are rounded to the third decimal place, both in our statistics and in Nishimoto's.

very common, whereas Li and Thompson (1981) claim that it is no longer productive (but see above, fn. 1). As to −兒 −*r*, Lin proposes that it is perhaps the most productive suffix in Mandarin; we also expected a high productivity index for −化 −*huà*, and this is what we found. It may be interesting to remark that for −化 −*huà* we have well above three times more types and *hapax legomena* than for −兒 −*r* but, also, almost six times the tokens; the situation is similar to that described above (3.1) for −*ness* and −*wise* in English: the data shows that −兒 −*r* is apparently used much less often than −化 −*huà*, but the potential to form new words in higher for the former than for the latter. As in the case of English −*ness*, many −化 −*huà* words are also quite frequently used, e.g. 自動化 *zìdònghuà* 'automatization' (329 occurrences), leading to a large number of tokens and, hence, an overall decrease of P.

However, it is still surprising to have such a high productivity index for −兒 −*r*, since it is said to be a feature of the colloquial speech of Beijing, thus mostly limited to a specific modality and to a diatopic variety (albeit arguably the most important one in the Chinese-speaking world) and, also, its use is said to be declining in the standard; Li and Thompson (1981:40) suggest that already at the beginning of the eighties −兒 −*r* was less common in current Standard Mandarin (普通話 *Pǔtōnghuà*) than in the Mandarin described in the textbooks of the time. We would also expect that −兒 −*r* be even less productive in the PH corpus, since it is made of newspaper articles, as pointed out by Nishimoto (2003:53; see above, 3.2). Let us then compare our results to Nishimoto's:

Table 4. Measures of productivity for −兒 *−r*, −化 *−huà* and −頭 *−tou* in the SCMC and in the PH corpus (Nishimoto 2003)[8]

Suffix	*V* (types)	n_1 (hapaxes)	*N* (tokens)	*P*
−兒 *−r*	139	62	783	0.079
	35	14	184	0.076
−化 *−huà*	464	232	4516	0.051
	209	93	3366	0.028
−頭 *−tou*	98	24	1593	0.015
	36	6	600	0.010

The ranking of suffixes in terms of *P* is the same in both corpora; also, for −兒 *−r* and −頭 *−tou* the indices are very close (0.079 and 0.015 *vs.* 0.076 and 0.010). As to −化 *−huà*, the difference between our figures and Nishimoto's appears as more significant. Let us discuss these similarities and discrepancies in more detail.

Nishimoto (2003:53) predicted that −化 *−huà* (and −們 *−men*) be productive, −頭 *−tou* (and −子 *−zi*) be "limited in productivity" and that the profitability of −兒 *−r* be dependent on the context; in the PH corpus, as said before, it was expected to be limited. However, −兒 *−r* turned out to be the suffix with the highest *P* index by far in Nishimoto's sample (the second one being −們 *−men* with 0.043). Nishimoto remarks that, although the *P* index for −兒 *−r* is the highest among the affixes he analysed, the number of types is even lower than that of −頭 *−tou*, the least productive one, but the token frequency of the former is significantly lower than that of the latter (184 *vs.* 600), while the number of *hapax legomena* for −兒 *−r* is much higher. The high number of tokens is interpreted by this author as a sign of the higher degree of lexicalization of −頭 *−tou*, whereas the comparatively high number of hapaxes for −兒 *−r* indicates that those derived words "are characterized by a low degree of lexicalization", suggesting that the "rule" for −兒 *−r* is still productive (Nishimoto 2003:57). This is what was argued also for

[8] For each suffix, SCMC figures are given on the first line, whereas Nishimoto's PH figures are on the second.

Eng. −*wise* above. It is also interesting to remark that, although the SCMC is roughly twice the size of the PH corpus, the number of types and tokens of −*兒* −*r* in our sample is about four times the figures provided by Nishimoto; hence the prediction that −*兒* −*r* be less frequent in a newspaper-only corpus is fundamentally correct. Nevertheless, the *P* index is roughly the same in both samples (only slightly higher in the SCMC), and we believe that this may be interpreted as a sign of the reliability of this measure of productivity; the same holds for −*頭* −*tou* derived words, although for this formant the difference in the *P* index is a little bigger.

As to −*化* −*huà*, in both samples it is by far the suffix with the highest number of types, hapaxes and tokens among the three considered here; however, its *P* index is significantly lower in the PH corpus, and the suffix ranks fourth among the five affixes analysed by Nishimoto, even below the supposedly unproductive −*子* −*zi*. Nishimoto (2003:58) suggests that the unexpectedly low *P* index could be explained by the fact that, as pointed out before, −*化* −*huà* suffixed nouns and verbs were lumped together due to limitations in the corpus, and "[i]t could be the case, for example, that some −*huà* words are typically used as nouns with high token frequencies while other −*huà* words are typically used as verbs with low token frequencies"; thus, the high token frequency of some −*化* −*huà* nouns would result in a low *P* index. However, we separated nouns from verbs in our sample and we saw no significant difference, as shown in table 5:

Table 5. Measures of productivity for −*化* −*huà* verbs and nouns in the SCMC

Word class	*V* (types)	n_1 (hapaxes)	*N* (tokens)	*P*
V	499	223	4448	0.050
N	15	9	6	0.132
V+N	564	232	4516	0.051

What can be seen is that not only the *P* measure of productivity of all −*化* −*huà* words (nouns and verbs) is not lower than that of −*化* −*huà*

verbs alone, but it is even slightly higher (0.051 *vs.* 0.050), and the ratio of *hapax legomena* to tokens for −化 *−huà* nouns is comparatively high, leading to $P = 0.132$; this goes against Nishimoto's (2003) hypothesis quoted above. Thus, we can conclude that −化 *−huà* appears as much more productive in our sample, and considering verbs only does not produce any significant difference.

However, if we look at the figures other than *P*, we notice that the number of types and *hapax legomena* for −化 *−huà* in our sample is approximately 2.2 times and 2.5 times respectively that of −化 *−huà* words in Nishimoto's sample, but the number of tokens of the same affix is only about 1.3 times bigger. A possible explanation, then, is that in the PH corpus there were some −化 *−huà* words with a particularly high token frequency. Since in Nishimoto's paper a list of the words found in his sample with the number of tokens for each one is available (in the appendix), we looked at the −化 *−huà* words with the highest frequency, and we realised that 變化 *biànhuà* 'change' (both noun and verb) scored first with 495 tokens; this, however, is not a −化 *−huà* derived word, but rather a coordinate compound, composed of two constituents both meaning '(to) change' (although only 變 *biàn* is a free morpheme in modern usage). Eliminating 變化 *biànhuà* from the count lowers the number of tokens to 2871, yielding $P = 0.032$. Also, the word which has the second highest token frequency is 現代化 *xiàndàihuà* 'modernise' (ex. 4 above), with 473 occurrences; in our sample we had only 221 tokens for 現代化 *xiàndàihuà*, i.e. less than half, despite the fact that the SCMC is twice the size of the PH corpus. If we were to delete 現代化 *xiàndàihuà* from both samples, we would have a *P* index of 0.039 for the PH corpus and 0.053 for the SCMC, with a difference of 0.014 only. Thus, the distance between the *P* measure for −化 *−huà* words in our sample and in Nishimoto's appears smaller if the non-derived form 變化 *biànhuà* 'change' is eliminated, and even smaller if the word with the highest count 現代化 *xiàndàihuà* is also deleted. Incidentally, we may point out that all of the occurrences of 現代化 *xiàndàihuà* in the SCMC are verbs, but this word is also commonly used as a noun ('modernisation', as in the famous 四個現代化 *sì ge xiàndàihuà* 'four modernisations'); we may hypothesize that part of the occurrences of 現

代化 *xiàndàihuà* in Nishimoto's sample are nouns, although this claim may not be verified.

In short, the results which we obtained for the three suffixes considered are mostly comparable to Nishimoto's, notwithstanding the differences in the text samples used, and this may be regarded as evidence of the reliability of the hapax-based *P* measure of productivity; nevertheless, we found a significant difference in the productivity index of −化 −*huà* derived words. We also showed that this discrepancy becomes less significant if the two items with the highest frequency in Nishimoto's sample are eliminated. Whereas the deletion of 現代化 *xiàndàihuà* 'modernisation' may be disputable, the most frequent word, 變化 *biànhuà* 'change' must incontrovertibly be eliminated, as it is not a derived word.

4.2 Early Mandarin

Whereas Nishimoto's study was limited to Modern Chinese, we chose to extract the same data on the productivity of the suffixes −兒 −*r*, −化 −*huà* and −頭 −*tou* in the SCEMC, a corpus consisting of novels and plays in Early Mandarin Chinese, i.e. the vernacular language from the 13th to the 19th century. We believe that the *P* measure, as well as the other indices of productitivity considered here, may be fruitfully employed to analyse both synchronic and diachronic data. While the comparison of the *P* index for different affixes in a corpus is a valid tool to assess differences in the profitability for a synchronic stage of a language, the comparison of productivity measures for the same affix(es) in different stages of a language is particularly significant, in our opinion; as we shall see, the differences between productivity figures for −兒 −*r* and −化 −*huà* in Modern Chinese may be best explained in a historical perspective.

One major problem we encountered is that the SCEMC returns a maximum of 5000 hits, but the occurrences of −兒 −*r* were actually more; what we did, then, was searching for this suffix in a subcorpus of 4 literary works (out of 10), namely *A collection of Thirty Yuan Dynasty Dramas* (元刊雜劇三十種 *Yúan Zájù Sānshízhǒng*), *Three Dramas from the Yongle Encyclopedia* (永樂大典戲文三種 *Yónglè Dàdiǎn Xìwén*

Sānzhǒng), *Journey to the West* (西遊記 *Xīyóujì*) and *The Dream of the Red Chamber* (紅樓夢 *Hóng Lóu Mèng*), two theatrical texts and two novels covering a time span ranging from the 13th to the 18th century. For −化 −*huà* and −頭 −*tou* we had less than 5000 tokens, and thus we were able to perform the calculations on the entire corpus; therefore, the figures for these two affixes are not readily comparable with those for −兒 −*er*, but they may nonetheless be compared to those from the SCMC. Our results are summarised in table 6:

Table 6. Measures of productivity for −兒 −*r*, −化 −*huà* and −頭 −*tou* in the SCEMC

Suffix	*V* (types)	n_1 (hapaxes)	*N* (tokens)	*P*
−化 −*huà*	10	6	26	0.231
−兒 −*r*	860	482	4009	0.120
−頭 −*tou*	107	45	2083	0.022

Our data show that −化 −*huà* derived words (all verbs in this sample) were extremely uncommon throughout the period of Early Mandarin Chinese; this suffix has a low number both of types *V* and of hapaxes n_1, indicating that the suffix was not used very often, neither in terms of different derivatives, nor in terms of new formations. Nevertheless, among all tokens with that suffix, the number of hapaxes is quite high, leading to a high *P* index: this is a sign of the potential of −化 −*huà* to be used for the coinage of new forms, if needed. Such figures, however, should be taken with a pinch of salt: when the numbers are so low, one may receive the false impression that a rare affix is actually very productive, as with the *Victoriana* case seen above (3.1). We have said before that in works dealing with the history of the Chinese lexicon (see e.g. Wang 1980), it is claimed that −化 −*huà* became productive at the beginning of the 20th century; however, the pattern existed already before, although is was very uncommon (Arcodia 2012); our data seem to support such course of historical development in the use of −化 −*huà* as a verb-deriving suffix analogous to Eng. −*ise*, −*ify*. The very high *P* index for −化 −*huà* in the Early Mandarin corpus may be interpreted as a

measure of the propensity of this suffix to the formation of new words; the many words that were created between the Early Mandarin and the Modern period resulted in the high type frequency of −化 −*huà* in the SCMC.

Let us compare the figures for −兒 −*r*, −化 −*huà* and −頭 −*tou* derived words in Early and Modern Mandarin:

Table 7. Measures of productivity for −兒 −*r*, −化 −*huà* and −頭 −*tou* in the SCEMC and the SCMC

Suffix	Corpus	*V* (types)	n_1 (hapaxes)	*N* (tokens)	*P*
−兒 −*r*	SCEMC	860	482	4009	0.120
	SCMC	139	62	783	0.079
−化 −*huà*	SCEMC	10	6	26	0.231
	SCMC	464	232	4516	0.051
−頭 −*tou*	SCEMC	107	45	2083	0.022
	SCMC	98	24	1593	0.015

The suffix −兒 −*r* appears as more productive and more used in Early Mandarin by all measures, despite the fact that the SCEMC corpus is much smaller and less varied than the SCMC; moreover, as pointed out above, the subcorpus we used for −兒 −*r* consists of 4 literary works out of 10, and thus is even considerably smaller. The decrease in profitability from the Early Mandarin to the Modern period, hence, is even more remarkable; although figures from the SCEMC and the SCMC are not readily comparable (see above, 3.2), our data suggest that the decline in the usage of −兒 −*r* as a nominal suffix highlighted by Li and Thompson (1981) began much earlier than the last quarter of the 20th century.

Another *caveat* is necessary. The four works on which the statistics for −兒 −*r* in Early Mandarin are based are not representative of a unitary language in the same way as Modern Mandarin Chinese is: they cover a very long time span, and the local elements they contain varies (on the 'construction' of the standard language, see Chen 1999 and

Coblin 2000); however, we believe that these figures cannot but indicate the high profitability of −兒 *−r* in those times.

As to −頭 *−tou*, the suffix appears to be slightly more productive in the SCEMC (*P* = 0.022 *vs.* 0.015). The number of types is actually about the same in both corpora (107 in the SCEMC *vs.* 98), but the number of *hapax legomena* in the Early Mandarin sample is almost twice that of SCMC; however, the very high number of tokens in the SCEMC (2083 *vs.* 1593) keeps the *P* index at a comparatively low level. We noticed that there was one single word, namely 丫頭 *yātou* 'servant girl' (or just 'girl'), which occurred 1198 times in the Early Mandarin sample, roughly 57.5% of all tokens for −頭 *−tou*; if this word is deleted from the count, the *P* index for this affix rises to 0.050. To a (much) lesser extent, this is what happens for −化 *−huà* derived words in the PH corpus, as argued above (4.1): if the two words with the highest token frequency are deleted, a dramatic rise in the *P* index for this suffix occurs.

To sum up, both −兒 *−r* and −頭 *−tou* appear as much more productive in the Early Mandarin corpus, which was not unexpected, whereas the number of −化 *−huà* derived words in this sample is perhaps too small to calculate a reliable *P* index; nevertheless, the figures for −化 *−huà* as well provide support for the claims found in the literature on the development of this pattern of word formation. In the Early Mandarin corpus, −化 *−huà* words (specifically, verbs) are very few, but among those the number of hapaxes is extraordinarily high (6 out of 10 types); taken by themselves, these data are not particularly significant since, as remarked above, very low figures are not always reliable in this respect. However, if we compare these data to those from the SCMC, we may confidently infer that in the period between the end of the Early Mandarin literature and the first decades of the twentieth century a large number of −化 *−huà* words were created, as said before. Such comparison makes the *P* measure for −化 *−huà* in the SCEMC become meaningful: the suffix had a potential for building new words in those times, and actually a large number of derived verbs were built; eventually, the profitability of the suffix declined, although it seems to be still fairly productive nowadays.

5. CONCLUDING REMARKS

The data presented in this paper suggest that Baayen's *P* index may be a reliable measure of the profitability of a morphological process or, rather, of one fundamental aspect of productivity, namely the 'readiness' with which an affix is used to build a new word in a synchronic stage of a language. Most of our findings were analogous to Nishimoto's, notwithstanding the fact that his corpus was considerably smaller and much more limited in variety than ours, which has also tags for parts of speech; this is evidence of the validity of the *P* index, in our opinion. It is also interesting to remark that our data mostly seem to support the claims found in the literature on the productivity of the affixes analysed here, even though the works considered did not suggest any empirical basis for their observations, as already remarked by Nishimoto. The main discrepancy between our data and Nishimoto's is the significantly higher *P* index which we obtained for –化 *–huà* derived words. However, we also showed that by eliminating the two words with the highest token frequency, the difference of *P* becomes much smaller; this is the same kind of frequency effect which we found for –頭 *–tou* in the Early Mandarin corpus, as more than half of the tokens of this affix were of a single word, strongly influencing the results.

The historical data which we collected for –兒 *–r* and –頭 *–tou* gave us some interesting results; although Early Mandarin data for –化 *–huà* was too limited to provide a comparable hapax-based measure of its productivity, nevertheless the figures we have confirm what we know from the history of the Chinese lexicon. This highlighted again a known shortcoming of this method, namely that a large sample is needed to provide significant results, and even in a large sample it might be the case that there are too few types and tokens of an affix, as for –化 *–huà* in the SCEMC. Nevertheless, the fact that the figures for –化 *–huà* in Early Mandarin suggest a high productivity is not, in principle, incompatible with the notion of productivity as understood in the present work. Corpus data for previous historical stages of a language, when

available, appear to be a better basis for assessments on the profitability of a morphological process than dictionary data.

We also tried to show that although the *P* index appears as a very useful tool for measuring the synchronic profitability of a process, provided that enough data are available, to obtain a complete picture one must take into consideration all the relevant statistics, including the number of types, tokens, neologisms, etc.; we hope that scholars will make use of this corpus-based methodology when dealing with productivity issues, as it appears that the judgements found in the descriptions of the morphology of Chinese, as well as of other languages, are often based on unclear methodologies, or even on impressionistic data.

REFERENCES

Arcodia, Giorgio F. 2012. *Lexical Derivation in Mandarin Chinese.* Taipei: Crane.

Baayen, R. Harald 1989. *A Corpus-Based Study of Morphological Productivity: Statistical Analysis and Psychological Interpretation.* PhD dissertation, Free University, Amsterdam.

Baayen, R. Harald. 1992. Quantitative Aspects of Morphological Productivity. *Yearbook of Morphology* 1991, ed. by Geert Booij and Jaap van Marle, 109-149. Dordrecht/London: Kluwer.

Baayen, R. Harald and Rochelle Lieber. 1991. Productivity and English Word-Formation: A Corpus-Based Study. *Linguistics* 29:801–843.

Bauer, Laurie. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.

Baxter, William H. and Laurent Sagart. 1998. Word formation in Old Chinese. *New Approaches to Chinese Word Formation*, ed. by Jerome Packard, 35-76. Berlin and New York: Mouton de Gruyter.

Beard, Robert 1998. Derivation. *Handbook of Morphology*, ed. by Andrew Spencer and Arnold M. Zwicky, 44-56. Oxford: Blackwell.

Booij, Geert. 2010. *Construction Morphology*. Oxford: Oxford University Press.

Chen, Ping 1999. *Modern Chinese: History and Sociolinguistics.* Cambridge: Cambridge University Press.

Coblin, W. South. 2000. A brief history of Mandarin. *Journal of the American Oriental Society* 120.4:537-552.

Dong, Xiufang. 2004. 汉语的词库与词法 *Hanyu de ciku yu cifa* (Chinese lexicon and morphology). Beijing: Beijing Daxue Chubanshe.

Li, Charles and Sandra A. Thompson. *Mandarin Chinese. A Functional Reference Grammar*. Berkeley: University of California Press.

Lin, Hua. 2001. *A Grammar of Mandarin Chinese.* München: Lincom Europa.

Ma, Qingzhu. 1995. 现代汉语词缀的性质、范围和分类 Xiandai Hanyu cizhui de xingzhi, fanwei he fenlei (Nature, scope and classification of Modern Chinese affixes). 中国语言学报 Zhongguo Yuyanxuebao 6:101-137.

Mayerthaler, Willi. 1981. *Morphologische Natürlichkeit*. Wiesbaden: Athenaion.

Naumann, Bernd and Petra M. Vogel. 2000. Derivation. *Morphologie-Morphology,* ed. by Geert Booij, Christian Lehmann and Joachin Mugdan, 929-943. Berlin-New York: Mouton de Gruyter.

Nishimoto, Eiji. 2003. Measuring and Comparing the Productivity of Mandarin Chinese Suffixes. *Computational Linguistics and Chinese Language Processing*, 8.1:49-76.

Norman, Jerry. 1988. *Chinese*. Cambridge: Cambridge University Press.

Packard, Jerome L. 2000. *The Morphology of Chinese. A Linguistic and Cognitive Approach*. Cambridge: Cambridge University Press.

Pan, Wenguo, Buqing Ye and Yang Han. 2004. 汉语的构词法研究 *Hanyu de goucifa yanjiu* (Research on word formation in Chinese). Shanghai: Huadong Shifan Daxue Chubanshe.

Plag, Ingo, Christiane Dalton-Puffer and R. Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3:209-228.

Plag, Ingo. 2002. The role of selectional restrictions, phonotactics and parsing in constraining suffix ordering in English. *Yearbook of morphology* 2001, ed. by Geert Booij and Jaap van Marle, 285–314. Dordrecht/London: Kluwer.

Plag, Ingo. 2006a. Productivity. *Encyclopedia of Language and Linguistics*, Second Edition, Vol. 10, ed. by Keith Brown, 121-128. Oxford: Elsevier.

Plag, Ingo. 2006b. Productivity. *The Handbook of English Linguistics*, ed. by Bas Aarts and April McMahon, 537-556. Oxford: Blackwell Publishing.

Renouf, Antoinette and R. Harald Baayen 1996. Aviating among the hapax legomena: Morphological grammaticalisation in current British newspaper English. *Explorations in Corpus Linguistics*, 181-189. Amsterdam and Atlanta, GA: Rodopi.

Sproat, Richard and Chilin Shih. 1996. A Corpus-Based Analysis of Mandarin Nominal Root Compound. *Journal of East Asian Linguistics* 5:49–71.

Steffen Chung, Karen. 2006. *Mandarin Compound Verbs.* Taiwan Journal of Linguistics, Book Series in Chinese Linguistics. Taipei: Crane.

Wang, Li 1980[1957]. 汉语史稿 *Hanyu Shigao* (Draft History of the Chinese Language). Beijing: Zhonghua Shuju.

Xing, Janet Z. 2006. *Teaching and Learning Chinese As a Foreign Language: A Pedagogical Grammar.* Hong Kong : Hong Kong University Press.

URL of Corpora:

Academia Sinica Balanced Corpus of Modern Chinese:
http://db1x.sinica.edu.tw/kiwi/mkiwi/
Academia Sinica Tagged Corpus of Early Mandarin Chinese:
http://db1x.sinica.edu.tw/cgi-bin/kiwi/pkiwi/pkiwi.sh

Giorgio Francesco Arcodia
Department of Humanities
University of Milano-Bicocca
Milan, Italy
giorgio.arcodia@unimib.it

Bianca Basciano
Department of Philology, Literature and Linguistics
University of Verona
Verona, Italy
bianca.basciano@gmail.com

論漢語「－兒」、「－化」與「－頭」詞綴的能產性

馬振國　　白夏儂
米蘭比可卡大學
維羅納大學

能產性作為形態學的基本概念之一，其定義引起了不少爭論，學者们测量能產性的方法和項目也不盡相同。本文基於中央研究院的「近代漢語標記語料庫」與「現代漢語平衡語料庫」，採用 Baayen 的 *P* 指數來測量漢語中「－兒」、「－化」與「－頭」三個詞綴的共時與歷時能產性，並將其結果與 Nishimoto (2003) 的結論進行比較。研究結果表明，運用不同時代的語料來測量某一詞綴的歷時能產性，可以視為非常有效的研究方法。

關鍵字：漢語，形態，派生構詞法，能產性