Computational Linguistics and Chinese Language Processing
Vol. 8, No. 1, February 2003, pp. 49-76

Measuring and Comparing the Productivity of Mandarin Chinese Suffixes

Eiji Nishimoto*

Abstract

The present study attempts to measure and compare the *morphological productivity* of five Mandarin Chinese suffixes: the verbal suffix *-hua,* the plural suffix *-men*, and the nominal suffixes *-r*, *-zi*, and *-tou*. These suffixes are predicted to differ in their *degree of productivity*: *-hua* and *-men* appear to be productive, being able to systematically form a word with a variety of *base* words, whereas *-zi* and *-tou* (and perhaps also *-r*) may be limited in productivity. Baayen [1989, 1992] proposes the use of corpus data in measuring productivity in word formation. Based on *word-token* frequencies in a large corpus of texts, his *token-based* measure of productivity expresses productivity as the probability that a new word form of an affix will be encountered in a corpus. We first use the token-based measure to examine the productivity of the Mandarin suffixes. The present study, then, proposes a *type-based* measure of productivity that employs the *deleted estimation* method [Jelinek & Mercer, 1985] in defining *unseen* words of a corpus and expresses productivity by the ratio of *unseen word types* to *all word types*. The proposed type-based measure yields the productivity ranking "*-men, -hua, -r, -zi, -tou,*" where *-men* is the most productive and *-tou* is the least productive. The effects of corpus-data variability on a productivity measure are also examined. The proposed measure is found to obtain a consistent productivity ranking despite variability in corpus data.

Keywords: Mandarin Chinese word formation, Mandarin Chinese suffixes, morphological productivity, corpus-based productivity measure.

1. Introduction

1.1 Morphological Productivity

The focus of a study of *morphological productivity* is on *derivational affixation* that involves a *base* word and an affix [Aronoff, 1976], as seen in *sharp* + *-ness* → *sharpness*, *electric* + *-ity*

* Ph.D. Program in Linguistics, The Graduate Center, The City University of New York,
365 Fifth Avenue, New York, NY 10016, U.S.A.
e-mail: enishimoto@gc.cuny.edu

→ *electricity*, *child* + *-ish* → *childish*.[1] Native speakers of a language have intuitions about what are and are not acceptable words of their language, and if presented with non-existent, *potential* words [Aronoff, 1983], they accept certain word formations more readily than others [Anshen & Aronoff, 1981; Aronoff & Schvaneveldt, 1978; Cutler, 1980]. Most intriguing in the issue of productivity is that *the degree of productivity* varies among affixes, and many studies in the literature have been devoted to accounting for this particular aspect of productivity [see Bauer, 2001, and Plag, 1999, for an overview].

How the degree of productivity varies among affixes is best illustrated by the English nominal suffixes *-ness* and *-ity*, which are often considered "rivals" as they sometimes share a base word (e.g., *clear* → *clearness* or *clarity*). In general, -ness is felt to be more productive than *-ity*.[2] The word formation of -ity is limited, for example, by the *Latinate Restriction* [Aronoff, 1976: 51] that requires the base word to be of Latinate origin; hence, *purity* is acceptable but **cleanity* is not. In contrast, *-ness* freely attaches to a variety of base words of both Latinate and Germanic (native) origin; thus, both *pureness* and *cleanness* are acceptable. There are also some affixes that could be regarded as *unproductive*; for example, Aronoff and Anshen [1998: 243] note that the English nominal suffix *-th* (as in long → *length*) has long been unsuccessful in forming a new word that survives, despite attempts at terms like *coolth*. Varying degrees of productivity are also observed in Mandarin Chinese word formation. As will be discussed shortly, some Mandarin suffixes appear to be more productive than others.

1.2 Measuring the Degree of Productivity

Early studies on productivity mainly focused on restrictions on word formation and viewed the degree of productivity to be determined by such restrictions [Booij, 1977; Schultink, 1961; van Marle, 1985]. Booij [1977: 120], for example, considers the degree of productivity of a *word formation rule* to be inversely proportional to the amount of restrictions that the word formation rule is subject to. Although the view that productivity is affected by restrictions on word formation is certainly to the point, from a quantitative point of view, measuring productivity by the amount of restrictions on word formation is limited in that the restrictive weight of such restrictions is unknown [Baayen & Renouf, 1996: 87].

Baayen [1989, 1992] proposes a corpus-based approach to the quantitative study of productivity. His productivity measure uses word frequencies in a large corpus of texts to

---

[1] Excluded from the study of productivity are seemingly irregular word formations, or "oddities" [Aronoff, 1976: 20], such as *blendings* (e.g., *smoke* + *fog* → *smog*) and *acronyms* (e.g., *NATO*).

[2] *-ity* can be more productive than *-ness* depending on the type of base word; for instance, *-ity* is more productive than *-ness* when the base word ends with *-ile* as in *servile* [Aronoff, 1976: 36] or with *-ible* as in *reversible* [Anshen & Aronoff, 1981]. Still, overall, *-ness* is intuitively felt to be more productive than *-ity*.

express productivity as the probability that a new word form of an affix will be encountered in a corpus (see Section 3). Although Bauer [2001: 204] observes that a generally agreed measure of productivity is yet to be achieved in the literature, Baayen's corpus-based approach seems to be appealing and promising. Most importantly, since corpus data include productively formed words that are typically not found in a dictionary [Baayen & Renouf, 1996], corpus-based descriptions of productivity reflect how words are actually used.[3] The corpus-based approach is also timely, as linguists have growing interests in corpus data. The present study pursues the corpus-based approach to measuring productivity using a corpus of Chinese texts.

The outline of this paper is as follows. In Section 2, five Mandarin suffixes are introduced and are analyzed qualitatively based on observations in the literature. In Section 3, Baayen's *token-based* productivity measure is discussed, and the measure is applied to a corpus of Chinese texts to quantitatively analyze the productivity of the Mandarin suffixes. In Section 4, a *type-based* productivity measure is proposed, and its performance is evaluated. Also, some experiments are conducted to examine the effects of corpus-data variability on a productivity measure. Section 5 summarizes the findings.

2. Mandarin Chinese Suffixes

2.1 A Qualitative Analysis of Five Mandarin Suffixes

The present study examines the productivity of five Mandarin suffixes: the verbal suffix ***-hua***, the plural suffix ***-men***, and the nominal suffixes ***-r***, ***-zi***, and ***-tou***.

The verbal suffix ***-hua*** 化 functions similarly to English ***-ize*** (and ***-ify***):

(1) ***xiàndài*** 现代 'modern' → ***xiàndàihuà*** 现代化 'modernize'

Verbs formed with ***-hua*** can be used as nouns [Baxter & Sagart, 1998: 40], so ***xiàndàihuà*** 现代化 in (1) can also be interpreted as 'modernization'. Analogous to English ***-ize*** (and ***-ify***), ***-hua*** systematically attaches to a variety of base words to form verbs, such as ***gōngyèhuà*** 工业化 'industrialize', ***guójìhuà*** 国际化 'internationalize', and ***jìsuànjīhuà*** 计算机化 'computerize'.

The suffix ***-men*** 们 pluralizes a noun, as in the following example:

(2) ***xuésheng*** 学生 'student' → ***xuéshengmen*** 学生们 'students'

According to Packard's [2000] classification, ***-men*** is a ***grammatical affix***, whereas the other four suffixes that we examine are ***word-forming affixes***. If we use the standard terminology of

[3] But see also Plag [1999] for a discussion of how dictionary data can be useful in a study of productivity.

the field, *-men* could be viewed as an *inflectional affix*, and the other four suffixes could be considered *derivational affixes*. There are three major characteristics of *-men* that differentiate *-men* from the English plural suffix *-s* [Lin, 2001: 59; Norman, 1988: 159; Ramsey, 1987: 64]. First, *-men* attaches only to human nouns[4]; hence, ***zhuōzimen*** 桌子们 'desks' and ***diànnǎomen*** 电脑们 'computers' are not acceptable, unless they are considered animate as in a cartoon. Second, *-men* is obligatory with pronouns (e.g., *wǒ* 我 'I' → ***wǒmen*** 我们 'we') but not with nouns; for example, ***háizi*** 孩子 without *-men* can be interpreted as 'child' or 'children' depending on the context. Third, *-men* is not compatible with numeral classifiers; hence, ***sāngè xuéshengmen*** 三个学生们 'three students' is ungrammatical. Due to these characteristics, *-men* may not be as frequently used or "productive" [Lin, 2001: 58] as the English plural suffix *-s*. However, *-men* has many base words to which it can attach, for there are a variety of nouns in Mandarin (as in any language) designating human beings (e.g., ***jìzhěmen*** 记者们 'reporters', ***kèrénmen*** 客人们 'guests', ***shìzhǎngmen*** 市长们 'mayors').

The suffix *-r* 儿 forms a noun from a verb or an adjective, or *-r* can create a diminutive form [Ramsey, 1987: 63; Lin, 2001: 57–58]:

(3) *huà* 画 'to paint' → ***huàr*** 画儿 'painting'

(4) *niǎo* 鸟 'bird' → ***niǎor*** 鸟儿 'small bird'

The use of *-r* is abundant in the colloquial speech of local Beijing residents, and three distinct usages of *-r* by local Beijing residents are identified [Chen, 1999: 39]. First, *-r* can create a semantic difference:

(5) *xìn* 信 'letter' → ***xìnr*** 信儿 'message'

Second, a form with *-r* may be habitually preferred to a form without it:

(6) *huā* 花 'flower' → ***huār*** 花儿 'flower'

Third, *-r* may be attached to a word solely for a stylistic reason. The use of *-r* in the last category is the most frequent among local Beijing residents [Chen, 1999: 39]. In both Mainland China and Taiwan, the use of *-r* is not favored especially in broadcasting, and *-r* words are rarely incorporated into the standard [Chen, 1999: 39; Ramsey, 1987: 64].

The suffixes ***-zi*** 子 and ***-tou*** 头 typically appear in the following constructions:

(7) ***mào*** 帽 → ***màozi*** 帽子 'hat'

(8) ***mù*** 木 → ***mùtou*** 木头 'wood'

In these examples, *-zi* and *-tou* combine with a ***bound morpheme*** that does not constitute a

[4] In colloquial speech, *-men* can occasionally attach to some animal nouns (e.g., ***gǒurmen*** 狗儿们 'doggies').

word by itself (i.e., neither ***mào*** 帽 nor ***mù*** 木 is a word).

Historically, the word formation of ***-zi*** and ***-tou*** appeared in the course of two changes in Chinese: a shift from monosyllabic to disyllabic words and a simplification of the phonological system [Packard, 2000: 265–266]. According to Packard [2000: 265], the shift toward disyllabic words occurred as early as in the Zhou dynasty (1000–700 BC) and underwent a large scale development during and after the Han dynasty (206 BC–AD 220). The phonological simplification, which occurred around the same time [Packard, 2000: 266], caused syllable-final consonants to be lost, and many single-syllable words that were once distinct became homophones [Li & Thompson, 1981: 44]. One possible account of how the two changes occurred is that the phonological simplification preceded as a natural linguistic process of phonetic attrition, and the shift toward disyllabic words occurred as a solution to the increase of homophonous syllables [Li & Thompson, 1981: 44; Packard, 2000: 266]. The increase of homophonous syllables was particularly significant in Mandarin [Li & Thompson, 1981: 44], and ***-zi*** and ***-tou*** played a role in the disyllabification of Mandarin words.

The word formation of ***-zi*** and ***-tou*** is not limited to bound morphemes [Lin, 2001: 58–59; Packard, 2000: 84]:

(9) ***shū*** 梳 'to comb' → ***shūzi*** 梳子 'comb'

(10) ***xiǎng*** 想 'to think' → ***xiǎngtou*** 想头 'thought'

In these examples, ***-zi*** and ***-tou*** form a noun by attaching to a free morpheme (i.e., both ***shū*** 梳 and ***xiǎng*** 想 are independent words).

The term "productive" is sometimes used in the literature to describe the above-discussed suffixes. Ramsey [1987: 63] describes ***-tou*** to be much less productive than ***-zi***, while Li and Thompson [1981: 42–43] observe that ***-zi*** and ***-tou*** are both no longer productive. Lin [2001: 57] views ***-r*** to be the most productive Mandarin suffix. Unfortunately, the basis for these observations is left unclear. Some observations may be based on the number of word forms of a suffix found in a dictionary; for example, present-day Mandarin has by far more ***-zi*** word forms than ***-tou*** word forms, and this may lead to the view that ***-zi*** is more productive than ***-tou***. However, as Aronoff [1980] argues, of interest to linguists is the ***synchronic*** aspect of productivity (i.e., how words of an affix can be formed at a given point in time), rather than the ***diachronic*** aspect of productivity (i.e., how many words of an affix have been formed between two points in time). Concentrating on the synchronic aspect, if we associate productivity with regularity in word formation [Spencer, 1991: 49] or availability of base words with which a new word can be readily formed, we may predict ***-hua*** and ***-men*** to be productive, and ***-zi*** and ***-tou*** to be limited in productivity. The productivity of ***-r*** would likely depend on the context—if we focus on broadcasting, the productivity of ***-r*** may also be limited. Admittedly, these predictions are speculative, and the difficulty in describing the productivity

of an affix is where a quantitative productivity measure becomes important. In the following sections, the productivity of the Mandarin suffixes will be examined quantitatively.

3. Quantitative Productivity Measurement

3.1 Baayen's Corpus-Based Approach

Baayen [1989, 1992] proposes a corpus-based measure of productivity, formulated as:

(11) $$p = \frac{n_1}{N}$$

where given all word forms of an affix found in a large corpus of texts, n_1 is the number of word types of the affix that occur only once in the corpus, the so-called ***hapax legomena*** (henceforth, ***hapaxes***), N is the sum of word tokens of the affix, and p is the productivity index of the affix in question.[5] The measure (11) employs Good's [1953] probability estimation method (commonly known as the *Good-Turing* estimation method) that provides a mathematically proven estimate [Church & Gale, 1991] of the probability of seeing a new word in a corpus, based on the probability of seeing hapaxes in that corpus. The productivity index p expresses the probability that a new word type of an affix will appear in a corpus after N tokens of the affix have been sampled. One important characteristic of the measure (11) is that it is ***token-based***; that is, the measure relies on word-token frequencies in a corpus. The sum of word types of an affix in a corpus, represented by V, is not directly tied to the degree of productivity (see Section 4.1). In the remaining sections, the measure (11) will be referred to as the ***hapax-based*** productivity measure.[6]

While the hapax-based measure has been primarily used in the studies of Western languages, such as Dutch [e.g., Baayen, 1989, 1992] and English [e.g., Baayen & Lieber, 1991;

---

[5] A clear distinction has to be made between *word tokens* and *word types* in the context of a corpus study. To give the simplest example, if we have three occurrences of ***the*** in a small corpus, the token frequency of ***the*** is three, and the type frequency of ***the*** is one. In the case of affixation, we ignore the differences between singular and plural forms; for example, if we have a corpus that has {***activity***, ***activity***, ***activities***, ***possibility***, ***possibilities***}, the token frequency of ***-ity*** is five (the sum of all these occurrences of ***-ity***) while the type frequency of ***-ity*** is two (after normalizing the plural forms, we have two distinct ***-ity*** words, ***activity*** and ***possibility***). An exception to ignoring the plural suffix is when we are interested in the productivity of the plural suffix itself. In that case, if we have a corpus consisting of {***book***, ***books***, ***books***, ***student***, ***students***}, the token frequency of ***-s*** is three (i.e., ***books***, ***books***, and ***students***), and the type frequency of ***-s*** is two (we have two distinct ***-s*** forms, ***books*** and ***students***).

[6] For the purposes of this paper, the term ***hapax-based measure*** is used to express, in a shorthand manner, the fact that the measure defines new words based on hapaxes and that the measure is token-frequency-based. It should not be confused with the ***hapax-conditioned measure***, ***p****, discussed in Baayen [1993].

Baayen & Renouf, 1996], the measure was also used by Sproat and Shih [1996] in a study of Mandarin word formation. The focus of Sproat and Shih's study was on productivity in Mandarin *root compounding*, as seen in the nominal root *yǐ* 蚁 of *mǎyǐ* 蚂蚁 'ant' that forms many words of 'ant-kind', such as *yǐwáng* 蚁王 'queen ant' and *gōngyǐ* 工蚁 'worker ant'. By analyzing the degree of productivity of a number of Mandarin nominal roots, Sproat and Shih showed that, contrary to a claim in the literature, root compounding is a productive word-formation process in Mandarin. For example, while *shí* 石 'rock-kind' and *yǐ* 蚁 'ant-kind' had the productivity indices of 0.129 and 0.065, respectively, apparently unproductive *bīn* 槟 and *láng* 榔 of *bīnláng* 槟榔 'betel nut' were found to have zero productivity. Sproat and Shih's study shows that a corpus-based study of productivity in Chinese is fruitful.

3.2 A Corpus of Segmented Chinese Texts

A major difficulty in conducting a corpus-based study of productivity in Chinese is that Chinese texts lack word delimiters. Segmentation of Chinese text is, by itself, a contested subject [see Sproat, Shih, Gale, & Chang, 1996], and consequently, a large-size corpus of segmented Chinese texts is not as readily available as a large-size corpus of English texts. Sproat and Shih [1996] used a large-size Chinese corpus (40-million Chinese characters) in their study by running an automatic segmenter to segment strings that contained the Chinese characters of interest and manually processing some problematic cases where the segmentation was not complete.

The corpus of choice in the present study is a "cleaned-up" version of *the Mandarin Chinese PH Corpus* [Guo, 1993; hereafter, *the PH Corpus*] of segmented Chinese texts, made available in a study by Hockenmaier and Brew [1998].[7] The corpus contains about 2.4-million (2,447,719) words—or 3.7-million (3,753,291) Chinese characters—from *XinHua* newspaper articles between January 1990 and March 1991. The texts of the PH Corpus are originally encoded in *GB* (simplified Chinese characters), and to facilitate the processing of the texts in computer programs, we convert the texts into *UTF8* (*Unicode*) using an encoding conversion program developed by Basis Technology [Uniconv, 1999]. The size of the PH Corpus is relatively small by today's standards (cf. a corpus of 80-million English words used in Baayen & Renouf, 1996), but the PH Corpus is one of few widely available corpora of segmented Chinese texts. Another widely available corpus of segmented Chinese texts is *the Academia Sinica Balanced Corpus* [1998; hereafter, *the Sinica Corpus*] that contains 5-million words from a variety of text sources. The sentences of the Sinica Corpus are syntactically parsed, so the *part-of-speech* of each segmented word is identified. Although the Sinica Corpus is not

[7] The PH Corpus can be downloaded from the ftp server of the Centre for Cognitive Science at University of Edinburgh.

used in the present study, the use of the Sinica Corpus is certainly of interest.[8]

Certain words were filtered out as potentially relevant words of the Mandarin suffixes in question were collected from the PH Corpus. With *-r* and *-zi*, a criterion for distinguishing a suffix from a non-suffix is that *-r* and *-zi* as a suffix lose their tone [Liu, 2001, 57–58; Norman, 1988, 113–114]. This criterion helps identify and block many non-suffixal cases where *-r* and *-zi* denote 'son' or 'child', such as *yīng'ér* 婴儿 'baby', *fùzǐ* 父子 'father and son', and *xiàozǐ* 孝子 'filial son'.[9] We exclude *wénhuà* 文化 'culture' because it is never a verb, and according to Norman [1988: 21], the specific use of *wénhuà* 文化 to mean 'culture' was adopted from Japanese. Also excluded are some *-tou* words, such as *máotóu* 矛头 'spearhead', in which *-tou* is a bound morpheme denoting 'head'. In addition, all pronouns in *-men* are excluded, as suggested in Sproat [2002]. As discussed earlier, *-men* behaves differently between pronouns and nouns (i.e., it is obligatory only with pronouns), and it is *-men* attaching to open-class nouns, rather than closed-class pronouns, that we are currently interested in.

3.3 A Quantitative Analysis of the Mandarin Suffixes

The result of the hapax-based measure applied to the PH Corpus is shown in Table 1. Figure 1 presents a bar graph illustrating the productivity ranking of the suffixes based on the p values.

Table 1. The result of the hapax-based productivity measure applied to the PH Corpus

suffix	*V*	*N*	n_1	*p*
-r	35	184	14	0.076
-men	219	2324	101	0.043
-zi	177	2130	62	0.029
-hua	209	3366	93	0.028
-tou	36	600	6	0.010

Note. With all the occurrences of a suffix found in the corpus, V is the sum of types, N is the sum of tokens, n_1 is the number of hapaxes, and p is the productivity index of the suffix. The suffixes are sorted in descending order by p.

[8] The use of the PH Corpus in the present study is solely due to the fact that the computer programs currently used were written for the PH Corpus. It must be noted, however, that 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.

[9] Note in these examples that the tone of *-r* and *-zi* is retained (i.e., *-ér* and *-zǐ*, respectively). *-r* is originally *-ér*, and it becomes *-r* as a suffix, as a result of losing its syllabicity [Norman, 1988: 114].

Among the five suffixes, *-r* is found to be the most productive. The high productivity of *-r* is somewhat unexpected given the fact that the PH Corpus consists of newspaper texts. If the use of *-r* is not favored in broadcasting, we may also expect a limited use of *-r* in a newspaper context. In addition, the use of *-r* is often a mere phonological phenomenon as seen in the speech of local Beijing residents, and it is unlikely for such a phonological phenomenon to be represented in newspaper texts. In Table 1, the number of types (*V*) of *-r* does not reach the number of types of the least productive suffix *-tou*. However, the token frequency (*N*) of *-r* is lower than that of *-tou*, and *-r* has a larger number of hapaxes than *-tou*. Under the hapax-based measure, a high token frequency is associated with a high *degree of lexicalization of words* (i.e., the extent to which words are stored in the lexicon in their full form), and a high degree of lexicalization of words, in turn, is associated with a low degree of productivity [Baayen, 1989, 1992]. The rationale behind this mechanism is that if many words of an affix are lexicalized, the word formation rule of the affix needs to be invoked less often to form a word. What the present data of *-r* indicate, then, is that *-r* words are characterized by a low degree of lexicalization. The low degree of lexicalization of *-r* words and the relatively large number of hapaxes (as compared with *-tou*) suggest that the word formation rule of *-r* is active.

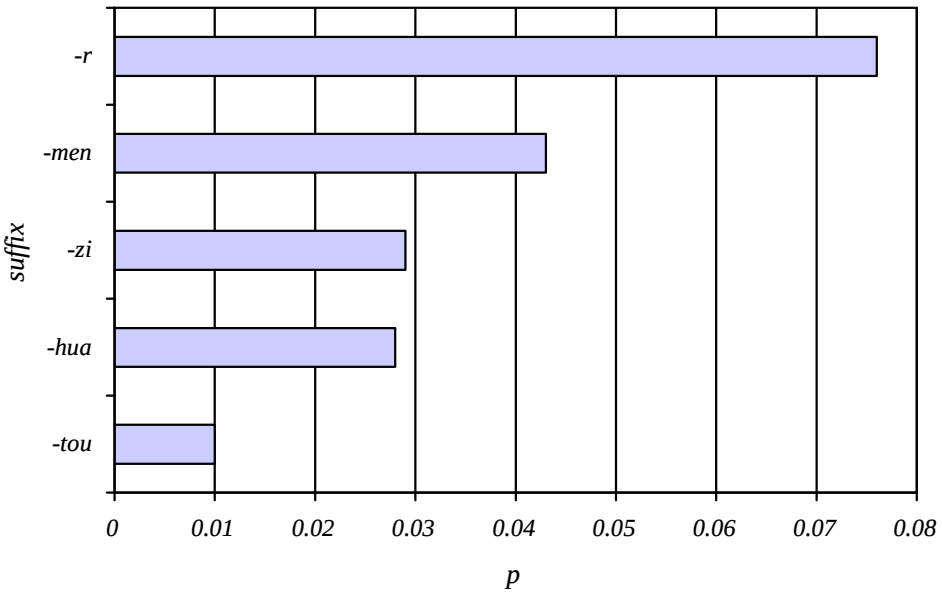

Figure 1 The productivity ranking of the Mandarin suffixes by the p values (the vertical axis lists the suffixes, and the horizontal axis shows the p values of the suffixes).

The productivity of *-hua* seems somewhat lower than what we may expect from the regularity in *-hua* word formation. Comparing *-men* and *-hua* in Table 1, we see that *-men* and *-hua* are similar with respect to both *V* and n_1, but the *p* value of *-hua* is lowered by the high token frequency (*N*) of *-hua*. The high token frequency of *-hua* could be attributed to the fact that the present analysis includes *-hua* words used as nouns. According to Baxter and Sagart [1998: 40], *-hua* words are formed as verbs first, and these verbs can be used as nouns. If this is the case, the word formation of *-hua* is also relevant in *-hua* nouns. However, the uniform treatment of *-hua* verbs and *-hua* nouns may not be appropriate for the hapax-based measure. It could be the case, for example, that some *-hua* words are typically used as nouns with high token frequencies while other *-hua* words are typically used as verbs with low token frequencies. It is, therefore, necessary to make a more detailed analysis of the word frequency distribution of *-hua* by separating *-hua* nouns from *-hua* verbs. Distinguishing nouns from verbs is unfortunately not available in the PH Corpus due to lack of syntactic information. A clearer description of the productivity of *-hua* could be achieved with a syntactically parsed corpus such as the Sinica Corpus.

4. Type-Based Deleted Estimation

4.1 Type-Based Measures

The present study explores a *type-based* measure of productivity. It has been argued that the sum of types of an affix in a corpus, *V*, alone often leads to some unintuitive results in measuring productivity [Baayen, 1989, 1992; Baayen & Lieber, 1991].[10] For example, Baayen and Lieber [1991: 804] point out that the type frequencies of *-ness* and *-ity* in their corpus (497 and 405, respectively) do not adequately represent the fact that *-ness* is intuitively felt to be much more productive than *-ity*. If the number of types in a corpus can be misleading with respect the degree of productivity, how can we make use of type frequencies in a productivity measure?

An early attempt at a type-based measure of productivity was made by Aronoff [1976: 36], in which he proposed that the degree of productivity of an affix could be measured by the ratio of the number of actual words of the affix to the number of *possible words* of the affix. The measure is described by Baayen [1989: 28] as:

(12) $$I = \frac{V}{S}$$

where *V* is the number of actual words with the relevant affix, *S* is the number of possible words with the affix, and *I* is the productivity index of the affix. Baayen [1989: 28] argues that

[10] See Baayen [1992] and Baayen and Lieber [1991] for a discussion of the *global productivity* of an affix (expressed as *P**) based on a two-dimensional analysis of *p* and *V*.

the measure lacks specification on how to obtain *V* and *S*. Moreover, he argues that the measure can be interpreted to express, ironically, the degree of "unproductivity" of an affix because the number of possible words (*S*) would be, in theory, increasingly large (hence, the productivity index *I* would be increasingly small) for a very productive affix [Baayen, 1989: 30].

Baayen [1989, 1992] defines *V* and *S* based on corpus data. *V* is (as before) the sum of types of the relevant affix found in a corpus, and *S* (expressed as $\hat{S}$) is statistically estimated for an infinitely large corpus; that is, $\hat{S}$ is the number of possible word types of the relevant affix to be expected when the corpus size is increased infinitely.[11] The measure that Baayen [1989: 60] proposes:

(13) $$I = \frac{\hat{S}}{V}$$

is the inverse of (12) and expresses the *potentiality of word formation rules*, the extent to which the number of actual word types of an affix exhaust the number of possible word types of the affix [Baayen, 1992: 122]. The measure (13), however, is not considered an alternative measure of the degree of productivity [Baayen, 1992: 122].

What does not appear to have been explored so far is the question of what *new words* would mean under a type-based measure. One major appeal of the hapax-based measure is that it centers on the formation of new words, and we may wish to try focusing on the formation of new words under a type-based measure. However, a problem with taking a type-based approach is that we can no longer rely on the Good-Turing estimation method. In the next section, we will discuss another method of defining new words of a corpus.

4.2 The Deleted Estimation Method

To define new words of a corpus in a type-based manner, we can employ the *deleted estimation* method [Jelinek & Mercer, 1985] used in language engineering. In a probabilistic language model, given a training corpus and a test corpus, we process words in the test corpus based on the probabilities of word occurrence in the training corpus. Since not all words of the test corpus appear in the training corpus, we need a method of assigning an appropriate probability mass to the *unseen words* in the test corpus. The main task involved here is to adjust the probabilities of word occurrence in the training corpus so that non-zero probability can be assigned to unseen words of the test corpus. A method used in this probability adjustment, if incorporated into a productivity measure, can tell us the probability of encountering unseen words in a corpus. The Good-Turing estimation method underlying the

[11] The statistical techniques for obtaining $\hat{S}$, which involve an extended version of Zipf's law, are beyond the scope of this paper. For more details, the reader is referred to Baayen [1989, 1992].

hapax-based measure is widely used in probabilistic language modeling, and its successful performances are reported in the literature [Chen & Goodman, 1998; Church & Gale, 1991]. While the Good-Turing estimation method is a ***mathematical*** solution to the task of probability adjustment, where the needed probability adjustment is mathematically determined, the deleted estimation method is an ***empirical*** solution, where the needed adjustment is determined by comparing discrepancies in word frequency between corpora [Church & Gale, 1991; Manning & Schütze, 1999].

The deleted estimation method, when incorporated into a type-based productivity measure, proceeds as follows. We begin by preparing two corpora of the same size and text type. The easiest way to have two such corpora is to split a large corpus in the middle into two sub-corpora, which we will call ***Corpus A*** and ***Corpus B***.[12] Comparing word types that appear in Corpus A against word types in Corpus B, ***unseen word types*** (or ***unseen types***) in Corpus A are defined as those word types that do not appear in Corpus B. Likewise, unseen types in Corpus B are those that are absent in Corpus A. We obtain the average number of unseen types between Corpus A and Corpus B. Defining ***all word types*** (or ***all types***) in a corpus as all the word types found in that corpus,[13] we also obtain the average number of all types between the two sub-corpora. The ratio of the average number of unseen types to the average number of all types expresses the extent to which word types of an affix are of an unseen type. With an assumption that unseen types are good candidates for new word types, the degree of productivity expressed in this manner comes close to Anshen and Aronoff's [1988: 643] definition of productivity as "the likelihood that new forms will enter the language."

The type-based deleted estimation productivity measure is formulated as follows:

Given Corpus A and Corpus B of the same size and text type, and all word types of an affix found in these corpora,

$$(14)\quad P_{tde}(A,B) = \frac{"\textit{unseen types in A given B}" + "\textit{unseen types in B given A}"}{"\textit{all types in A}" + "\textit{all types in B}"}$$

where ***all types*** of a corpus are all the word types found in that corpus, ***unseen types*** in one corpus are those that are absent in the other corpus, and P_{tde} is the degree of productivity of the affix in question (***tde = type-based deleted estimation***). In calculating P_{tde} by the measure (14), we can first average the unseen types in the nominator and the all types in the denominator. This will conveniently give us the average number of unseen types and the average number of all types, which are both of interest by themselves, before examining the ratio of the two (as

---

[12] These sub-corpora would be labeled ***retained*** and ***deleted*** (hence the term ***deleted estimation***) under the original deleted estimation method. However, in the present context, we can simplify the argument by using the labels *Corpus A* and *Corpus B*.

[13] The number of ***all types*** is essentially the same as V.

will be seen later in Table 2). In the remaining sections, the measure (14) will be referred to as the P_{tde} measure. Using a Venn Diagram, Figure 2 illustrates elements involved in the P_{tde} measure.

Given $\mathbf{A} = \{a_1, ..., a_m\}$ *from Corpus A, and* $\mathbf{B} = \{b_1, ..., b_n\}$ *from Corpus B, where* a_i *and* b_i *are word types of an affix found in the two corpora,*

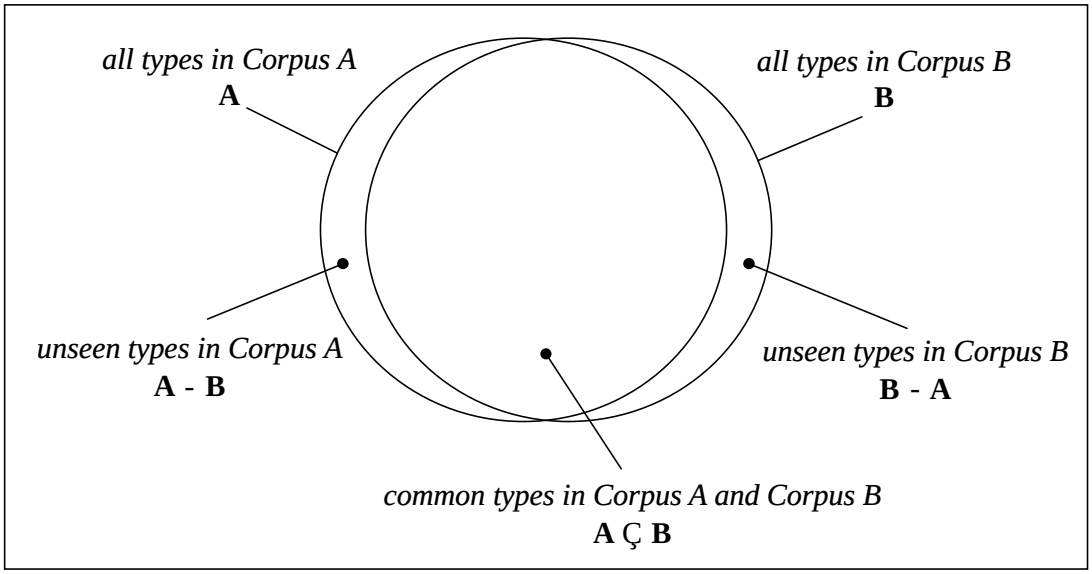

Figure 2 An illustration of elements involved in the P_{tde} measure (all types in a corpus are all the word types found in that corpus, unseen types in one corpus are those that are absent in the other corpus, and common types are the word types shared by the two corpora).

As a byproduct, the P_{tde} measure also identifies *common types*, word types that are shared by two sub-corpora, as shown in Figure 2. One possible interpretation of these common types is that they represent attested words, where attested words are defined as those words that are familiar to the majority of speakers. Although an approximation,[14] common types may be good candidates for attested words because unseen types, which are less likely to be familiar to the majority of speakers, are maximally excluded. As the corpus size increases, the number of common types may begin to provide a good estimate of the range of word types that are

[14] Strictly speaking, any word type with the token frequency of two or more in the original whole corpus has a chance to be shared by the two sub-corpora after the corpus is split. Thus, a word that appears only twice in a large corpus could be identified as a common type.

shared by the majority of speakers. Such a range of word types differs from the range of word types in a dictionary. Common types will not be pursued in the present study, but they may be worth further investigation in future research.

4.3 Performance of the P_{tde} Measure

The result of the P_{tde} measure applied to the PH Corpus is shown in Table 2. Figure 3 presents a bar graph that illustrates the productivity ranking of the suffixes based on the P_{tde} values.

Table 2. The result of the P_{tde} measure applied to the PH Corpus

suffix	*(average) all types*	*(average) unseen types*	P_{tde}
-men	149	70	0.470
-hua	144	65	0.451
-r	24.5	10.5	0.429
-zi	130.5	46.5	0.356
-tou	29.5	6.5	0.220

Note. The PH Corpus is split in the middle into two sub-corpora. *All types* in a sub-corpus are all the word types that appear in that sub-corpus. The second column shows the average number of all types between the two sub-corpora. *Unseen types* are those that appear in one sub-corpus but are absent in the other sub-corpus. The third column shows the average number of unseen types between the two sub-corpora. The tenths place in the second and third columns is due to the averaging. P_{tde} is the ratio of (*average*) *unseen types* to (*average*) *all types*. The suffixes are sorted in descending order by P_{tde}.

In Table 2, we find that *-r* is not as highly productive as under the hapax-based measure, though it still appears to be grouped with the more productive suffixes. Here, we may wonder why we examine the ratio of unseen types to all types, instead of examining the number of unseen types only. If productivity is determined by the number of unseen types only, *-r* would be among the less productive suffixes. However, comparing the number of unseen types alone is not satisfactory because an affix with a low frequency of use would generally be found to be less productive. The P_{tde} measure must be able to capture the possibility that an affix with a low frequency of use can nevertheless be productive when it is used to form a word. With respect to the present data, the ratio of unseen types to all types is relatively high for *-r*, indicating that a large proportion of *-r* word types are of an unseen type, or a potentially new type.

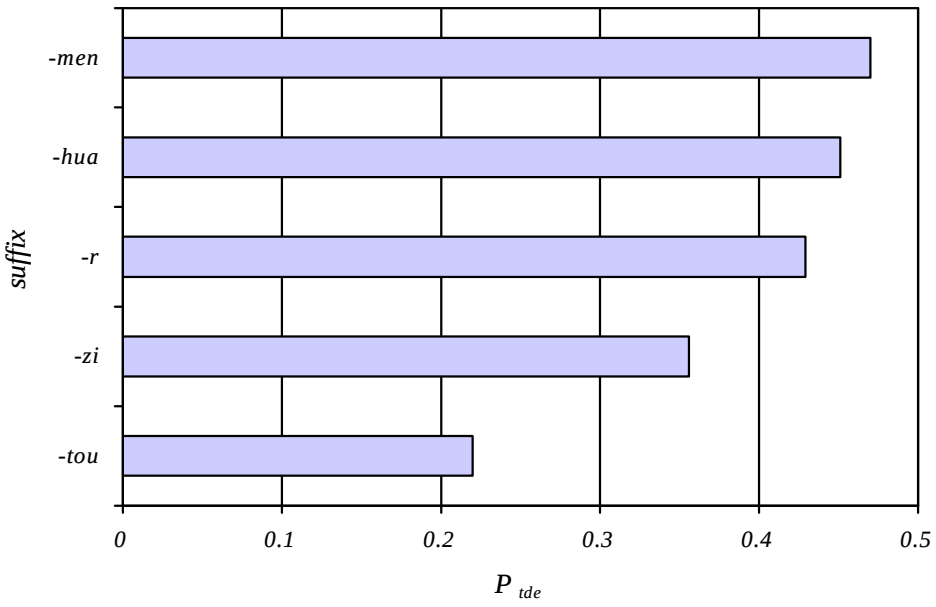

Figure 3 The productivity ranking of the Mandarin suffixes by the P_{tde} values (the vertical axis lists the suffixes, and the horizontal axis shows the P_{tde} values of the suffixes).

As was the case under the hapax-based measure, *-men* is found to be highly productive and *-tou* is found to be the least productive. The uniform treatment of *-hua* verbs and *-hua* nouns does not seem to pose a problem, though it is also of interest to investigate the effect of separating *-hua* nouns from *-hua* verbs under the P_{tde} measure.

The P_{tde} measure defines unseen types irrespective of word-token frequencies; that is, an unseen type in a corpus is "unseen" as long as it is absent in the other corpus, regardless of how many times the word is repeated in the same corpus. Figure 4 shows the word-token frequency distribution of unseen types in Corpus A and Corpus B. The labels used for the word-token frequency categories are: n_1 = words occurring once, n_2 = words occurring twice, ..., n_{5+} = words occurring five times or more.

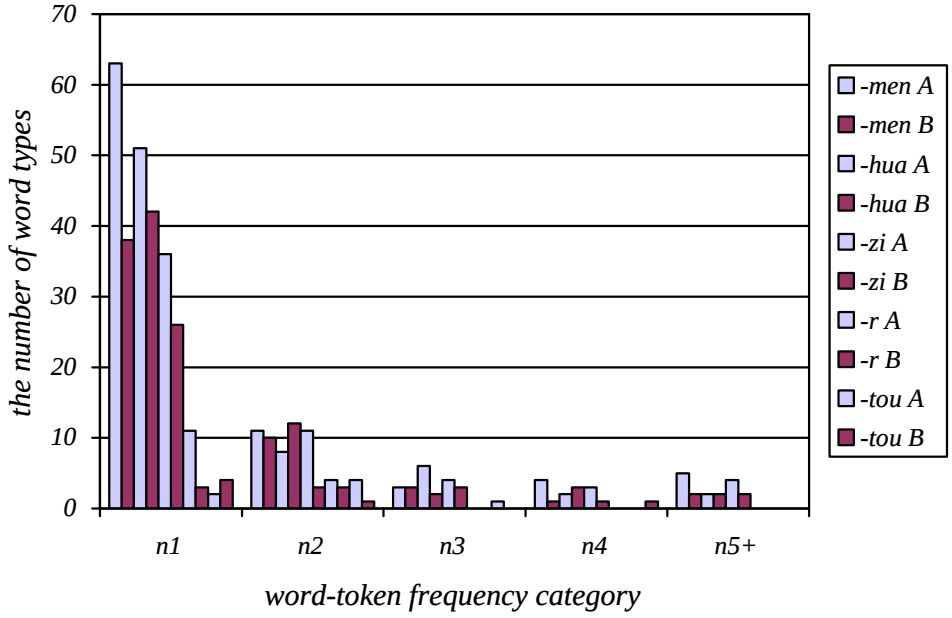

Figure 4 The word-token frequency distribution of unseen types in the two sub-corpora of the PH Corpus, Corpus A and Corpus B (the horizontal axis shows the word-token frequency category, and the vertical axis shows the number of word types in each frequency category; the letter following each suffix in the legend indicates from which sub-corpus the data are drawn; the order of the suffixes in the legend (from top down) corresponds to the order of bars in each frequency category (from left to right)).

We find in Figure 4 that the majority of unseen types are hapaxes. There are, nonetheless, unseen types that appear more than once in a corpus—some unseen types appear even five times or more (n_{5+}). We also notice gaps between the two sub-corpora in the word frequency of the unseen types (e.g., compare the number of *-men* hapaxes). Variability between two corpora will be the topic of discussion in the next section.

4.4 Variability in Corpus Data

Under the P_{tde} measure, a corpus is split in the middle to create two sub-corpora. So far, we have made the assumption that splitting a corpus in the middle would create two sub-corpora that are similar with respect to the text type. However, we must be cautious about this assumption. Baayen [2001] discusses how the texts and word frequency distribution of a

corpus can be non-uniform.[15] One way to reduce variability between split halves of a corpus is to randomize words of the corpus before splitting the corpus into two. Randomization of words can be accomplished by shuffling words; that is, given a corpus of n words, we exchange each i-th word (i = 1, 2, ..., n) with a randomly chosen j-th word ($1 \leq j \leq n$). If we repeat the "random split" of a corpus (i.e., randomizing words of a corpus and splitting the corpus in the middle) for a large number of times, say 1,000 times, and compute the mean of the relevant data, we should be able to obtain a stable, representative result of a productivity measure.[16] Table 3 shows the result of the hapax-based measure applied to the two sub-corpora of the PH Corpus, with and without randomization of words.

Table 3. The result of the hapax-based productivity measure applied to the two sub-corpora of the PH Corpus, Corpus A and Corpus B, with and without randomization of words

(a) Without randomization, a single split									
Corpus A					*Corpus B*				
suffix	*V*	*N*	n_1	*p*	*suffix*	*V*	*N*	n_1	*p*
-r	29	113	13	0.115	*-r*	20	71	6	0.085
-men	165	1183	84	0.071	*-zi*	119	841	53	0.063
-hua	148	1599	72	0.045	*-men*	133	1141	60	0.053
-zi	142	1289	57	0.044	*-tou*	29	256	8	0.031
-tou	30	344	5	0.015	*-hua*	140	1767	55	0.031
(b) With randomization, the mean of 1000 splits									
Corpus A					Corpus B				
suffix	*V*	*N*	n_1	*p*	*suffix*	*V*	*N*	n_1	*p*
-r	26	93	12	0.133	*-r*	26	91	12	0.130
-men	158	1164	77	0.067	*-men*	157	1160	77	0.066
-zi	138	1075	54	0.050	*-zi*	137	1055	54	0.051
-hua	154	1680	71	0.042	*-hua*	152	1686	69	0.041
-tou	31	303	8	0.025	*-tou*	31	297	8	0.027

Note. Each value in Part (b) is the mean of 1,000 random splits. The suffixes in each section are sorted in descending order by *p*. In Corpus B of Part (a), the *p* values of *-tou* and *-hua* expressed to the fourth decimal place are 0.0313 and 0.0311, respectively.

[15] See Baayen [2001] for an in-depth discussion of techniques for measuring variances among segments of a corpus.

[16] The procedure described here is thanks to suggestions by Baayen [personal communication].

In Part (a) of Table 3, the difference in *V* between Corpus A and Corpus B is almost significant,[17] which suggests variability in texts between the two sub-corpora, and a different productivity ranking is obtained in each sub-corpus. However, if we turn to Part (b) of Table 3, the productivity ranking becomes consistent between the two sub-corpora.[18] Interestingly, the productivity ranking in Part (b) of Table 3 is the same as one obtained earlier in Table 1 in Section 3.3. The *p* values in Part (b) of Table 3 are overall higher than those in Table 1, but this is an expected outcome, for *p* is dependent on the size of a corpus [Baayen, 1989, 1992; Baayen & Lieber, 1991]. We find that the hapax-based measure can achieve stability by means of a large number of random splits of a corpus.

What will be the effects of corpus-data variability on the P_{tde} measure? To examine this, we need to temporarily simplify the P_{tde} measure so that the value of P_{tde} will be obtained for each individual sub-corpus (without averaging unseen types and all types between two sub-corpora). That is, under the simplified measure, P_{tde} for Corpus A, P_{tde}(A), will be the ratio of "unseen types in A given B" to "all types in A"; and similarly, P_{tde}(B) will be the ratio of "unseen types in B given A" to "all types in B." Table 4 shows the result of the simplified P_{tde} measure applied to the two sub-corpora of the PH Corpus, with and without randomization of words.

The simplified P_{tde} measure is found to be quite vulnerable to corpus-data variability. In Part (a) of Table 4, the difference between Corpus A and Corpus B is almost significant in *all types* and *unseen types*, and the P_{tde} values differ significantly between the two sub-corpora.[19] However, if we turn to Part (b) of Table 4, the productivity ranking becomes consistent between the two sub-corpora.[20] Similarly to the hapax-based measure, the P_{tde} measure can achieve stability through a large number of random splits of a corpus.

---

[17] A paired *t*-test reveals that the difference in *V* approaches significance [$t(4) = 2.595$, $p = .06$], though the difference is not significant in other elements: *N* [$t(4) = .905$, $p > .10$], n_1 [$t(4) = 2.046$, $p > .10$], and *p* [$t(4) = .555$, $p > .10$].

[18] The correlation coefficient between Corpus A and Corpus B improves in *p* after the random splits: *p* [$r(5) = (.850 \rightarrow) 1.0$, $p < .01$].

[19] A paired *t*-test shows that the difference approaches significance in *all types* [$t(4) = 2.595$, $p = .06$] and in *unseen types* [$t(4) = 2.595$, $p = .06$] and the difference is significant in P_{tde} [$t(4) = 2.869$, $p < .05$].

[20] The correlation coefficient between Corpus A and Corpus B improves in P_{tde} after the random splits: P_{tde} [$r(5) = (.753 \rightarrow) 9.99$, $p < .01$].

Table 4. The result of the simplified P_{tde} measure applied to the two sub-corpora of the PH Corpus, Corpus A and Corpus B, with and without randomization of words

(a) Without randomization, a single split							
Corpus A				*Corpus B*			
suffix	*all*	*unseen*	P_{tde}	*suffix*	*all*	*unseen*	P_{tde}
-men	165	86	0.521	*-hua*	140	61	0.436
-r	29	15	0.517	*-men*	133	54	0.406
-hua	148	69	0.466	*-r*	20	6	0.300
-zi	142	58	0.408	*-zi*	119	35	0.294
-tou	30	7	0.233	*-tou*	29	6	0.207
(b) With randomization, the mean of 1000 splits							
Corpus A				*Corpus B*			
suffix	*all*	*unseen*	P_{tde}	*suffix*	*all*	*unseen*	P_{tde}
-men	158	62	0.394	*-men*	157	61	0.389
-hua	154	57	0.372	*-hua*	152	55	0.364
-r	26	9	0.356	*-r*	26	9	0.342
-zi	138	40	0.291	*-zi*	137	39	0.287
-tou	31	5	0.160	*-tou*	31	5	0.163

Note. Each value in Part (b) is the mean of 1,000 random splits. The suffixes in each section are sorted in descending order by P_{tde}.

Figure 5 shows the word-token frequency distribution of unseen types averaged over the 1,000 random splits. We see in Figure 5 that unseen types with higher token frequencies (e.g., n_4 and n_{5+}) are almost absent. What this indicates is that as a result of randomizing words of a corpus, it became unlikely for unseen types to include word types that are repeated many times in a corpus. As compared with what we saw earlier in Figure 4, the greater majority of unseen types are now hapaxes, and variances between Corpus A and Corpus B are also reduced.

We now consider the P_{tde} measure in its original state (as in Section 4.2, with the averaging of unseen types and all types between two sub-corpora). Comparing Table 2 and Part (b) of Table 4, we find that the original P_{tde} measure achieves a result that is highly correlated with the result obtained with the 1,000 random splits.[21] Note in particular that the

[21] Comparing the elements of Table 2 and the elements of Corpus A in Part (b) of Table 4, the correlation coefficient is significant in all elements: *all types* [$r(5) = 1.0$, $p < .01$], *unseen types* [$r(5) = 1.0$, $p < .01$], and P_{tde} [$r(5) = 1.0$, $p < .01$]. Likewise, the correlation coefficient is significant in all elements when we compare the elements of Table 2 and the elements of Corpus B in Part (b) of Table 4: *all types* [$r(5) = 1.0$, $p < .01$], *unseen types* [$r(5) = 1.0$, $p < .01$], and P_{tde} [$r(5) = .999$, $p < .01$].

productivity ranking is consistent between Table 2 and Part (b) of Table 4. The P_{tde} measure seems to reduce the effects of corpus-data variability by averaging unseen types and all types between two sub-corpora. This is an advantage and makes the P_{tde} measure handy, for a large number of random splits of a corpus can be computationally expensive, especially when the corpus size is large.

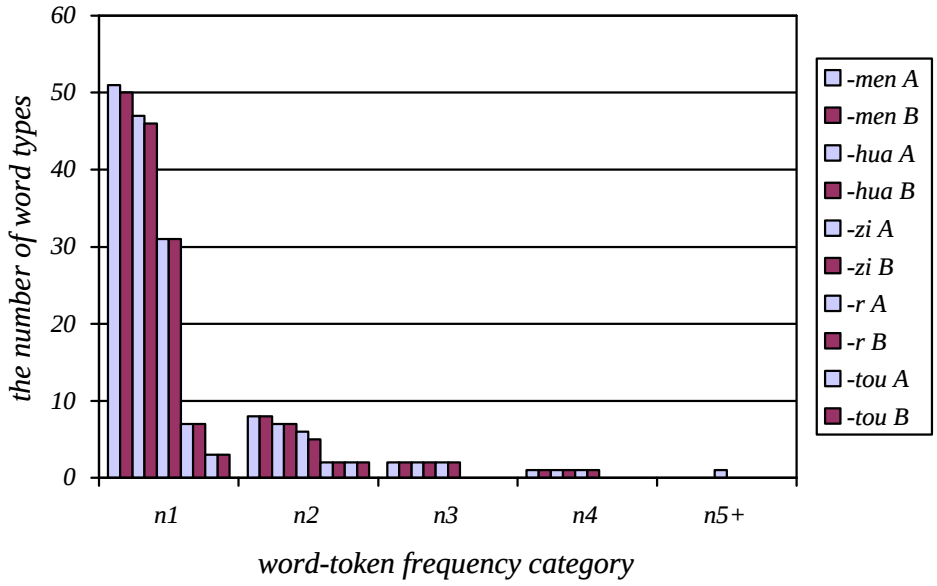

Figure 5. The word-token frequency distribution of unseen types in the two sub-corpora of the PH Corpus, Corpus A and Corpus B, averaged over 1000 random splits (the horizontal axis shows the word-token frequency category, and the vertical axis shows the number of word types in each frequency category; the letter following each suffix in the legend indicates from which sub-corpus the data are drawn; the order of the suffixes in the legend (from top down) corresponds to the order of bars in each frequency category (from left to right)).

5. Conclusion

The present study has proposed a type-based measure of productivity, the P_{tde} measure, that uses the deleted estimation method [Jelinek & Mercer, 1985] in defining unseen word types of a corpus. The measure expresses the degree of productivity of an affix by the ratio of unseen word types of the affix to all word types of the affix. If the ratio is high for an affix, a large proportion of the word types of the affix are of an unseen type, indicating that the affix has a great potential to form a new word.

We have tested the performance of the P_{tde} measure as well as the hapax-based measure of Baayen [1989, 1992] in a quantitative analysis of the productivity of five Mandarin suffixes: *-hua*, *-men*, *-r*, *-zi*, and *-tou*. The P_{tde} measure describes *-hua*, *-men*, and *-r* to be highly productive, *-zi* to be less productive than these three suffixes, and *-tou* to be the least productive, yielding the productivity ranking "*-men*, *-hua*, *-r*, *-zi*, *-tou*." The P_{tde} measure and the hapax-based measure rank the suffixes differently with respect to *-hua* and *-r*. The relatively low productivity of *-hua* under the hapax-based measure could be attributed to the inclusion of *-hua* nouns in the present analysis. *-r* is assigned a larger productivity score under the hapax-based measure. The two measures agree on the high productivity of *-men* and the low productivity of *-tou*. The different results of the two measures are likely due to the type-based/token-based difference of the measures. The result of each measure requires an individual evaluation, for the knowledge that we can obtain from the result of each measure is different; for example, while the hapax-based measure takes into consideration the degree of lexicalization of words of an affix, the P_{tde} measure does not consider such an issue.

We have also examined how corpus-data variability affects the results of a productivity measure. It was found that a large number of random splits of a corpus adds stability to both the P_{tde} measure and the hapax-based measure. Moreover, it was found that even without randomization of words, the averaging of unseen types and all types under the P_{tde} measure reduces the effects of corpus-data variability. This is an advantage of the P_{tde} measure, considering the computational cost involved in randomizing words repeatedly, especially when the corpus is large.

With an assumption that unseen words of a corpus are good candidates for new words, a corpus-based productivity measurement can be regarded as a search for unseen words in a corpus. The apparent paradox is that the words that we seek are "unseen." Baayen's hapax-based measure achieves a mathematical estimate of the probability of seeing unseen words in a corpus by the Good-Turing estimation method. The deleted estimation method provides another way of defining unseen words of a corpus by comparing discrepancies in word frequency between two corpora, and the method also enables defining unseen words in a type-based context. It is hoped that words identified as unseen by the P_{tde} measure are also good candidates for new words, and this requires further investigation in future research. The implication of the successful result of the P_{tde} measure presented in this paper is that, in addition to what has been proposed by Baayen [1989, 1992, and subsequent works], there appear to be possibilities for capturing and exploiting elements in corpus data that are relevant to the quantitative description of productivity. The study of morphological productivity will be enriched by exploring such possibilities in the corpus-based approach to measuring productivity.

Acknowledgments

The author wishes to thank Harald Baayen, Richard Sproat, Martin Chodorow, and the anonymous reviewers for their insightful comments on the first draft of this paper. Any errors are the responsibility of the author.

References

Academia Sinica Balanced Corpus (Version 3.0) [CD-ROM]. Taipei, Taiwan: Academia Sinica, 1998.

Anshen, F., & Aronoff, M. "Morphological Productivity and Phonological Transparency." *Canadian Journal of Linguistics, 26,* 1981, 63–72.

Anshen, F., & Aronoff, M. "Producing Morphologically Complex Words." *Linguistics, 26,* 1988, 641–655.

Aronoff, M. *Word Formation in Generative Grammar.* Cambridge, MA: MIT Press, 1976.

Aronoff, M. "The Relevance of Productivity in a Synchronic Description of Word Formation." In J. Fisiak (Ed.), *Historical Morphology.* The Hague: Mouton, 1980, 71–82.

Aronoff, M. "Potential Words, Actual Words, Productivity and Frequency." *Proceedings of the International Congress of Linguists, 13,* 1983, 163–171.

Aronoff, M., & Anshen, F. "Morphology and the Lexicon: Lexicalization and Productivity." In A. Spencer & A. M. Zwicky (Eds.), *The Handbook of Morphology.* Oxford, UK: Blackwell Publishers, 1998, 237–247.

Aronoff, M., & Schvaneveldt, R. "Testing Morphological Productivity." *Annals of the New York Academy of Sciences, 318,* 1978, 106–114.

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

Baayen, R. H. "Quantitative Aspects of Morphological Productivity." In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1991.* Dordrecht: Kluwer, 1992, 109–149.

Baayen, R. H. "On Frequency, Transparency and Productivity." In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1992.* Dordrecht: Kluwer, 1993, 181–208.

Baayen, R. H. *Word Frequency Distributions.* Dordrecht: Kluwer, 2001.

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

Baayen, R. H., & Renouf, A. "Chronicling the Times: Productive Lexical Innovations in an English Newspaper." *Language, 72,* 1996, 69–96.

Bauer, L. *Morphological Productivity.* Cambridge, UK: Cambridge University Press, 2001.

Baxter, W. H., & Sagart, L. "Word Formation in Old Chinese." In J. L. Packard (Ed.), *New Approaches to Chinese Word Formation: Morphology, Phonology and Lexicon in Modern and Ancient Chinese.* Berlin: Mouton de Gruyter, 1998, 35–76.

Booij, G. E. *Dutch Morphology: A Study of Word Formation in Generative Grammar.* Dordrecht: Foris, 1977.

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

Chen, S. F., & Goodman, J. *An Empirical Study of Smoothing Techniques for Language Modeling* (Tech. Rep. No. 10-98). Cambridge, MA: Harvard University, Center for Research in Computing Technology, 1998.

Church, K. W., & Gale, W. A. "A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams." *Computer Speech and Language, 5,* 1991, 19–54.

Cutler, A. "Productivity in Word Formation." *Papers from the Sixteenth Regional Meeting of the Chicago Linguistic Society.* Chicago, IL: Chicago Linguistic Society, 1980, 45–51.

Good, I. J. "The Population Frequencies of Species and the Estimation of Population Parameters." *Biometrika, 40,* 1953, 237–264.

Guo, J. "PH: A Chinese Corpus." *Communications of COLIPS, 3* (1), 1993, 45–48.

Hockenmaier, J., & Brew, C. "Error-Driven Learning of Chinese Word Segmentation." In J. Guo, K. T. Lua, & J. Xu (Eds.), *12th Pacific Conference on Language and Information.* Singapore: Chinese and Oriental Languages Processing Society, 1998, 218–229.

Jelinek, F., & Mercer, R. "Probability Distribution Estimation for Sparse Data." *IBM Technical Disclosure Bulletin, 28,* 1985, 2591–2594.

Li, C., & Thompson, S. A. *Mandarin Chinese: A Functional Reference Grammar.* Berkeley, CA: University of California Press, 1981.

Lin, H. *A Grammar of Modern Chinese.* LINCOM EUROPA, 2001.

Manning, C. D., & Schütze, H. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press, 1999.

Norman, J. *Chinese.* Cambridge University Press, 1988.

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

Plag, I. *Morphological Productivity: Structural Constraints in English Derivation.* Berlin: Mouton de Gruyter, 1999.

Ramsey, R. S. *The Languages of China.* Princeton, NJ: Princeton University Press, 1987.

Schultink, H. "Produktiviteit als Morfologisch Fenomeen." *Forum der Letteren, 2,* 1961, 110–125.

Spencer, A. *Morphological Theory: An Introduction to Word Structure in Generative Grammar.* Cambridge, UK: Cambridge University Press, 1991.

Sproat, R. "Corpus-Based Methods in Chinese Morphology." Tutorial given at COLING, Taipei, Taiwan, 2002.

Sproat, R., & Shih, C. "A Corpus-Based Analysis of Mandarin Nominal Root Compound." *Journal of East Asian Linguistics, 5,* 1996, 49–71.

Sproat, R., Shih, C., Gale, W., & Chang, N. "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese." *Computational Linguistics, 22* (3), 1996, 66–73.

Uniconv [Computer Software]. Cambridge, MA: Basis Technology, 1999.

Van Marle, J. On the Paradigmatic Dimension of Morphological Productivity. Dordrecht: Foris, 1985.

Appendix: Words of the Mandarin Suffixes in the PH Corpus

Below are the words of the Mandarin suffixes and their token frequencies in the PH Corpus.

-hua

变化 *biànhuà* 495 – 现代化 *xiàndàihuà* 473 – 深化 *shēnhuà* 323 – 自由化 *zìyóuhuà* 167 – 一体化 *yītǐhuà* 138 – 强化 *qiánghuà* 131 – 恶化 *èhuà* 122 – 优化 *yōuhuà* 99 – 消化 *xiāohuà* 71 – 石化 *shíhuà* 68 – 国产化 *guóchǎnhuà* 59 – 转化 *zhuǎnhuà* 54 – 社会化 *shèhuìhuà* 53 – 正常化 *zhèngchánghuà* 52 – 美化 *měihuà* 51 – 净化 *jìnghuà* 50 – 自动化 *zìdònghuà* 50 – 电气化 *dàinqìhuà* 45 – 机械化 *jīxièhuà* 42 – 制度化 *zhìdùhuà* 41 – 标准化 *biāozhǔnhuà* 33 – 工业化 *gōngyèhuà* 29 – 氧化 *yǎnghuà* 25 – 电化 *dàinhuà* 25 – 系列化 *xìlièhuà* 22 – 民主化 *mínzhǔhuà* 22 – 科学化 *kēxuéhuà* 21 – 液化 *yèhuà* 21 – 商品化 *shāngpǐnhuà* 19 – 火化 *huǒhuà* 18 – 演化 *yǎnhuà* 18 – 革命化 *gémìnghuà* 17 – 生物化 *shēngwùhuà* 15 – 简化 *jiǎnhuà* 14 – 融化 *rónghuà* 14 – 国际化 *guójìhuà* 14 – 老化 *lǎohuà* 13 – 农机化 *nóngjīhuà* 13 – 激化 *jīhuà* 13 – 专业化 *zhuānyèhuà* 12 – 产业化 *chǎnyèhuà* 11 – 沙漠化 *shāmòhuà* 11 – 多元化 *duōyuánhuà* 10 – 裂化 *lièhuà* 10 – 军事化 *jūnshìhuà* 10 – 煤气化 *méiqìhuà* 9 – 良种化 *liángzhǒnghuà* 8 – 硬化 *yìnghuà* 8 – 生化 *shēnghuà* 8 – 法制化 *fǎzhìhuà* 8 – 分化 *fēnhuà* 8 – 林网化 *línwǎnghuà* 7 – 工厂化 *gōngchǎnghuà* 7 – 系统化 *xìtǒnghuà* 6 – 模式化 *móshìhuà* 6 – 集团化 *jítuánhuà* 6 – 大众化 *dàzhònghuà* 6 – 恐龙化 *kǒnglónghuà* 6 – 企业化 *qǐyèhuà* 6 – 殖民化 *zhímínhuà* 5 – 规模化 *guīmóhuà* 5 – 全球化 *quánqiúhuà* 5 – 活血化 *huóxuèhuà* 5 – 硫化 *liúhuà* 4 – 立体化 *lìtǐhuà* 4 – 家庭化 *jiātínghuà* 4 – 形象化 *xíngxiànghuà* 4 – 中华化 *zhōnghuáhuà* 4 – 智能化 *zhìnénghuà* 4 – 软化 *ruǎnhuà* 4 – 表面化 *biǎomiànhuà* 4 – 物化 *wùhuà* 4 – 白热化 *báirèhuà* 3 – 程序化 *chéngxùhuà* 3 – 焦化 *jiāohuà* 3 – 牙齿化 *yáchǐhuà* 3 – 纯化 *chúnhuà* 3 – 气化 *qìhuà* 3 – 园林化 *yuánlínhuà* 3 – 合作化 *hézuòhuà* 3 – 异化 *yìhuà* 3 – 风化 *fēnghuà* 3 – 焚化 *fénhuà* 3 – 资源化 *zīyuánhuà* 3 – 僵化 *jiānghuà* 3 – 作物化 *zuòwùhuà* 3 – 固化 *gùhuà* 3 – 数字化 *shùzìhuà* 3 – 歧化 *qíhuà* 2 – 遗愿化 *yíyuànhuà* 2 – 西化 *xīhuà* 2 – 集约化 *jíyuēhuà* 2 – 板化 *bǎnhuà* 2 – 化学化 *huàxuéhuà* 2 – 商业化 *shāngyèhuà* 2 – 丑化 *chǒuhuà* 2 – 反自由化 *fǎnzìyóuhuà* 2 – 区域化 *qūyùhuà* 2 – 群众化 *qúnzhònghuà* 2 – 法律化 *fǎlǜhuà* 2 – 国有化 *guóyǒuhuà* 2 – 乳化 *rǔhuà* 2 – 水利化 *shuǐlìhuà* 2 – 产品化 *chǎnpǐnhuà* 2 – 法规化 *fǎguīhuà* 2 – 基地化 *jīdìhuà* 2 – 驯化 *xúnhuà* 2 – 信息化 *xìnxīhuà* 2 – 水化 *shuǐhuà* 2 – 煤化 *méihuà* 2 – 孵化 *fūhuà* 2 – 极化 *jíhuà* 2 – 植物化 *zhíwùhuà* 2 – 中文化 *zhōngwénhuà* 2 – 资本主义化 *zīběnzhǔyìhuà* 2 – 计算机化 *jìsuànjīhuà* 2 – 电脑化 *diànnǎohuà* 1 – 短期化 *duǎnqīhuà* 1 – 赔偿仪化 *péichángyíhuà* 1 – 组织化 *zǔzhīhuà* 1 – 类型化 *lèixínghuà* 1 – 实体化 *shítǐhuà* 1 – 集体化 *jítǐhuà* 1 – 林带化 *líndàihuà* 1 – 华东化 *huádōnghuà* 1 – 湿化 *shīhuà* 1 – 鱼粉化 *yúfěnhuà* 1 – 联合化 *liánhéhuà* 1 – 批量化 *pīliànghuà* 1 – 概念化 *gàiniànhuà* 1 – 集成化 *jíchénghuà* 1 – 碱化 *jiǎnhuà* 1 – 民族化 *mínzúhuà* 1 – 管道化 *guǎndàohuà* 1 – 网络

化 *wǎngluòhuà* 1 – 氨化 *ānhuà* 1 – 整体化 *zhěngtǐhuà* 1 – 渠网化 *qúwǎnghuà* 1 – 健康化 *jiànkānghuà* 1 – 神化 *shénhuà* 1 – 本地化 *běndìhuà* 1 – 欧洲化 *ōuzhōuhuà* 1 – 合理化 *hélǐhuà* 1 – 馆化 *guǎnhuà* 1 – 规格化 *guīgéhuà* 1 – 贵族化 *guìzúhuà* 1 – 模块化 *mókuàihuà* 1 – 个性化 *gèxìnghuà* 1 – 原生动物化 *yuánshēngdòngwùhuà* 1 – 普及化 *pǔjíhuà* 1 – 成人化 *chéngrénhuà* 1 – 硬朗化 *yìnglanghuà* 1 – 欧共体化 *ōugòngtǐhuà* 1 – 氰化 *qínghuà* 1 – 定量化 *dìngliànghuà* 1 – 氯苯化 *lǜběnhuà* 1 – 电器化 *diànqìhuà* 1 – 龄化 *línghuà* 1 – 氯化 *lǜhuà* 1 – 官僚化 *guānliáohuà* 1 – 氯磺化 *lǜhuánghuà* 1 – 政治化 *zhèngzhìhuà* 1 – 关怀化 *guānhuáihuà* 1 – 档案化 *dàngànhuà* 1 – 磷化 *línhuà* 1 – 凝固化 *nínggùhuà* 1 – 质化 *zhìhuà* 1 – 溶化 *rónghuà* 1 – 皂化 *zàohuà* 1 – 尘化 *chénhuà* 1 – 藻类化 *zǎolèihuà* 1 – 元首化 *yuánshǒuhuà* 1 – 园田化 *yuántiánhuà* 1 – 腐化 *fǔhuà* 1 – 关系化 *guānxìhuà* 1 – 塑化 *sùhuà* 1 – 艺术化 *yìshùhuà* 1 – 国家化 *guójiāhuà* 1 – 足迹化 *zújìhuà* 1 – 炼化 *liànhuà* 1 – 棉花化 *miánhuahuà* 1 – 通用化 *tōngyònghuà* 1 – 渍化 *zìhuà* 1 – 行政化 *xíngzhènghuà* 1 – 越南化 *yuènánhuà* 1 – 蠕虫化 *rúchónghuà* 1 – 模硫化 *móliúhuà* 1 – 量化 *liànghuà* 1 – 时装化 *shízhuānghuà* 1 – 部门化 *bùménhuà* 1 – 理想化 *lǐxiǎnghuà* 1 – 省城化 *shěngchénghuà* 1 – 党化 *dǎnghuà* 1 – 战略化 *zhànlüèhuà* 1 – 全能化 *quánnénghuà* 1 – 催化 *cuīhuà* 1 – 数量化 *shùliànghuà* 1 – 空心化 *kòngxīnhuà* 1 – 纤化 *xiānhuà* 1 – 羽化 *yǔhuà* 1 – 套路化 *tàolùhuà* 1 – 平面化 *píngmiànhuà* 1 – 雪化 *xuěhuà* 1 – 生活化 *shēnghuóhuà* 1 – 动物化 *dòngwùhuà* 1 – 程控化 *chéngkònghuà* 1 – 氮化 *dànhuà* 1 – 谱化 *pǔhuà* 1 – 庸俗化 *yōngsúhuà* 1

-men

人们 *rénmen* 734 – 代表们 *dàibiǎomen* 175 – 专家们 *zhuānjiāmen* 117 – 委员们 *wěiyuánmen* 109 – 工人们 *gōngrénmen* 75 – 同志们 *tóngzhìmen* 72 – 孩子们 *háizimen* 64 – 战士们 *zhànshìmen* 59 – 职工们 *zhígōngmen* 39 – 同学们 *tóngxuémen* 32 – 队员们 *duìyuánmen* 31 – 姑娘们 *gūniangmen* 26 – 客人们 *kèrenmen* 24 – 记者们 *jìzhěmen* 23 – 科学家们 *kēxuéjiāmen* 23 – 老人们 *lǎorénmen* 23 – 农民们 *nóngmínmen* 22 – 学生们 *xuéshengmen* 21 – 分析家们 *fēnxījiāmen* 21 – 姐妹们 *jiěmèimen* 19 – 朋友们 *péngyoumen* 18 – 艺术家们 *yìshùjiāmen* 16 – 干部们 *gànbùmen* 16 – 市民们 *shìmínmen* 15 – 市长们 *shìzhǎngmen* 14 – 居民们 *jūmínmen* 14 – 首脑们 *shǒunǎomen* 14 – 村民们 *cūnmínmen* 13 – 演员们 *yǎnyuánmen* 13 – 旅客们 *lǚkèmen* 12 – 同事们 *tóngshìmen* 12 – 小伙子们 *xiǎohuǒzimen* 11 – 医生们 *yīshēngmen* 10 – 行家们 *xíngjiāmen* 10 – 议员们 *yìyuánmen* 10 – 大学生们 *dàxuéshēngmen* 10 – 官兵们 *guānbīngmen* 9 – 运动员们 *yùndòngyuánmen* 9 – 观察家们 *guānchájiāmen* 9 – 同行们 *tóngxíngmen* 8 – 经理们 *jīnglǐmen* 8 – 师生们 *shīshēngmen* 7 – 常委们 *chángwěimen* 7 – 企业家们 *qǐyèjiāmen* 7 – 外长们 *wàizhǎngmen* 7 – 指战员们 *zhǐzhànyuánmen* 7 – 船员们 *chuányuánmen* 6 – 列车员们 *lièchēyuánmen* 6 – 部长们 *bùzhǎngmen* 6 – 作家们 *zuòjiāmen* 6 – 建设者们 *jiànshèzhěmen* 6 – 工友们 *gōngyǒumen* 6 – 青年们 *qīngniánmen* 6 – 党员们 *dǎngyuánmen* 5 – 顾客们 *gùkèmen* 5 – 干警们 *gànjǐngmen* 5 – 学者们 *xuézhěmen* 5 – 娘们 *niángmen* 5 – 劳模们 *láomómen* 5 – 教师们 *jiàoshīmen* 5 – 营业员们 *yíngyèyuánmen*

4 – 团员们 *tuányuánmen* 4 – 成员们 *chéngyuánmen* 4 – 子女们 *zǐnǚmen* 4 – 队友们 *duìyǒumen* 4 – 妇女们 *fùnǚmen* 4 – 乘客们 *chéngkèmen* 4 – 侨胞们 *qiáobāomen* 4 – 伙伴们 *huǒbànmen* 4 – 来宾们 *láibīnmen* 4 – 儿女们 *érnǚmen* 3 – 军人们 *jūnrénmen* 3 – 将军们 *jiāngjūnmen* 3 – 父母官们 *fùmǔguānmen* 3 – 乘务员们 *chéngwùyuánmen* 3 – 护士们 *hùshimen* 3 – 大师们 *dàshīmen* 3 – 儿孙们 *érsūnmen* 3 – 戏迷们 *xìmímen* 3 – 小学生们 *xiǎoxuéshēngmen* 3 – 文艺家们 *wényìjiāmen* 3 – 观众们 *guānzhòngmen* 3 – 球迷们 *qiúmímen* 3 – 司长们 *sīchángmen* 3 – 领导们 *lǐngdǎomen* 3 – 教练员们 *jiàoliànyuánmen* 2 – 爷们 *yémen* 2 – 人员们 *rényuánmen* 2 – 女工们 *nǚgōngmen* 2 – 摄影家们 *shèyǐngjiāmen* 2 – 板报员们 *bǎnbàoyuánmen* 2 – 老板们 *lǎobǎnmen* 2 – 老汉们 *lǎohànmen* 2 – 状元们 *zhuàngyuanmen* 2 – 会员们 *huìyuánmen* 2 – 州长们 *zhōuzhǎngmen* 2 – 女士们 *nǚshìmen* 2 – 友人们 *yǒurénmen* 2 – 大家们 *dàjiāmen* 2 – 师傅们 *shīfumen* 2 – 创作者们 *chuàngzuōzhěmen* 2 – 喇嘛们 *lǎmamen* 2 – 经济学家们 *jīngjìxuéjiāmen* 2 – 支持者们 *zhīchízhěmen* 2 – 老师们 *lǎoshīmen* 2 – 儿子们 *érzimen* 2 – 祖辈们 *zǔbèimen* 2 – 少女们 *shàonǚmen* 2 – 学员们 *xuéyuánmen* 2 – 书画家们 *shūhuàjiāmen* 2 – 选手们 *xuǎnshǒumen* 2 – 妈妈们 *māmamen* 2 – 同胞们 *tóngbāomen* 2 – 员工们 *yuángōngmen* 2 – 亲戚们 *qīnqimen* 2 – 选民们 *xuǎnmínmen* 2 – 天文学家们 *tiānwénxuéjiāmen* 2 – 儿童们 *értóngmen* 2 – 法官们 *fǎguānmen* 1 – 行人们 *xíngrénmen* 1 – 歹徒们 *dǎitúmen* 1 – 高徒们 *gāotúmen* 1 – 瘾君子们 *yǐnjūnzǐmen* 1 – 贵宾们 *guìbīnmen* 1 – 厨师们 *chúshīmen* 1 – 台胞们 *táibāomen* 1 – 老伙伴们 *lǎohuǒbànmen* 1 – 勇士们 *yǒngshìmen* 1 – 车迷们 *chēmímen* 1 – 支委们 *zhīwěimen* 1 – 孙子们 *sūnzimen* 1 – 夫妇们 *fūfùmen* 1 – 配水员们 *pèishuǐyuánmen* 1 – 伤员们 *shāngyuánmen* 1 – 囚犯们 *qiúfànmen* 1 – 客户们 *kèhùmen* 1 – 军官们 *jūnguānmen* 1 – 士兵们 *shìbīngmen* 1 – 巾帼们 *jīnguómen* 1 – 助手们 *zhùshǒumen* 1 – 留学生们 *liúxuéshēngmen* 1 – 设计师们 *shèjìshīmen* 1 – 局长们 *júzhǎngmen* 1 – 老工人们 *lǎogōngrénmen* 1 – 渔工们 *yúgōngmen* 1 – 副市长们 *fùshìzhǎngmen* 1 – 侦察员们 *zhēnchányuánmen* 1 – 观察员们 *guāncháyuánmen* 1 – 设计者们 *shèjìzhěmen* 1 – 家属们 *jiāshǔmen* 1 – 检察官们 *jiǎncháguānmen* 1 – 体育迷们 *tǐyùmímen* 1 – 女生们 *nǚshēngmen* 1 – 革命先烈们 *gémìngxiānlièmen* 1 – 飞行员们 *fēixíngyuánmen* 1 – 老头子们 *lǎotóuzimen* 1 – 海外侨胞们 *hǎiwàiqiáobāomen* 1 – 炮制者们 *pàozhìzhěmen* 1 – 服务员们 *fúwùyuánmen* 1 – 推销员们 *tuīxiāoyuánmen* 1 – 太太们 *tàitaimen* 1 – 伐木者们 *fámùzhěmen* 1 – 劳动模范们 *láodòngmófànmen* 1 – 水兵们 *shuǐbīngmen* 1 – 使节们 *shǐjiémen* 1 – 歌唱家们 *gēchàngjiāmen* 1 – 主任们 *zhǔrènmen* 1 – 个体户们 *gètǐhùmen* 1 – 演说家们 *yǎnshuōjiāmen* 1 – 音乐家们 *yīnyuèjiāmen* 1 – 亲友们 *qīnyǒumen* 1 – 功臣们 *gōngchénmen* 1 – 职员们 *zhíyuánmen* 1 – 姐姐们 *jiějiemen* 1 – 司机们 *sījīmen* 1 – 制造商们 *zhìzàoshāngmen* 1 – 英雄们 *yīngxióngmen* 1 – 画家们 *huàjiāmen* 1 – 外商们 *wàishāngmen* 1 – 患者们 *huànzhěmen* 1 – 村邻们 *cūnlínmen* 1 – 卫士们 *wèishìmen* 1 – 大臣们 *dàchénmen* 1 – 技术员们 *jìshùyuánmen* 1 – 图者们 *túzhěmen* 1 – 教员们 *jiàoyuánmen* 1 – 老大娘们 *lǎodàniángmen* 1 – 法学家们 *fǎxuéjiāmen* 1 – 研究者们

yánjiūzhěmen 1 – 游人们 *yóurénmen* 1 – 元首们 *yuánshǒumen* 1 – 娃娃们 *wáwamen* 1 – 青少年们 *qīngshàoniánmen* 1 – 力士们 *lìshìmen* 1 – 售货员们 *shòuhuòyuánmen* 1 – 教练们 *jiàoliànmen* 1 – 采购员们 *cǎigòuyuánmen* 1 – 女们 *nǚmen* 1 – 游客们 *yóukèmen* 1 – 烈士们 *lièshìmen* 1 – 西藏史学家们 *xīzàngshǐxuéjiāmen* 1 – 老奶奶们 *lǎonǎinaimen* 1 – 大夫们 *dàifūmen* 1 – 气象学家们 *qìxiàngxuéjiāmen* 1 – 工作者们 *gōngzuòzhěmen* 1 – 县太爷们 *xiàntàiyémen* 1 – 商贩们 *shāngfànmen* 1 – 松们 *sōngmen* 1 – 亲人们 *qīnrénmen* 1 – 老朋友们 *lǎopéngyoumen* 1 – 家长们 *jiāzhǎngmen* 1 – 夫妻们 *fūqīmen* 1 – 学子们 *xuézǐmen* 1 – 东道主们 *dōngdàozhǔmen* 1 – 省长们 *shěngzhǎngmen* 1 – 同仁们 *tóngrénmen* 1 – 山水画家们 *shānshuǐhuàjiāmen* 1 – 战略家们 *zhànlüèjiāmen* 1 – 董事长们 *dǒngshìzhǎngmen* 1

-r

这儿 *zhèr* 32 – 会儿 *huìr* 30 – 哪儿 *nǎr* 18 – 劲儿 *jìnr* 13 – 事儿 *shìr* 12 – 点儿 *diǎnr* 9 – 那儿 *nàr* 8 – 伙儿 *huǒr* 7 – 个儿 *gèr* 7 – 活儿 *huór* 5 – 鸟儿 *niǎor* 5 – 块儿 *kuàir* 4 – 花儿 *huār* 3 – 法儿 *fǎr* 3 – 风儿 *fēngr* 2 – 字儿 *zìr* 2 – 条儿 *tiáor* 2 – 味儿 *wèir* 2 – 片儿 *piànr* 2 – 玩儿 *wánr* 2 – 弯儿 *wānr* 2 – 样儿 *yàngr* 1 – 轧伙儿 *yàhuǒr* 1 – 脸儿 *liǎnr* 1 – 干劲儿 *gānjìnr* 1 – 头儿 *tóur* 1 – 万儿 *wànr* 1 – 话儿 *huàr* 1 – 抠儿 *kōur* 1 – 犟劲儿 *jiàngjìnr* 1 – 信儿 *xìnr* 1 – 塞儿 *sèr* 1 – 主儿 *zhǔr* 1 – 芯儿 *xīnr* 1 – 当儿 *dāngr* 1

-tou

势头 *shìtou* 133 – 码头 *mǎtou* 99 – 街头 *jiētóu* 96 – 石头 *shítou* 33 – 罐头 *guàntou* 30 – 镜头 *jìngtóu* 26 – 年头 *niántóu* 20 – 拳头 *quántou* 18 – 馒头 *míntou* 16 – 炕头 *kàngtóu* 14 – 老头 *lǎotóu* 12 – 心头 *xīntóu* 11 – 木头 *mùtou* 9 – 骨头 *gǔtou* 9 – 源头 *yuántóu* 8 – 口头 *kǒutóu* 8 – 苗头 *miáotou* 7 – 地头 *dìtóu* 7 – 指头 *zhǐtou* 7 – 锄头 *chútou* 5 – 桥头 *qiáotóu* 5 – 部头 *bùtóu* 4 – 枕头 *zhěntou* 3 – 斧头 *fǔtou* 2 – 先头 *xiāntóu* 2 – 脚趾头 *jiǎozhǐtou* 2 – 里头 *lǐtou* 2 – 风头 *fēngtou* 2 – 手指头 *shǒuzhǐtóu* 2 – 犁头 *lítóu* 2 – 滩头 *tāntóu* 1 – 丫头 *yātou* 1 – 窝窝头 *wōwōtóu* 1 – 关头 *guāntóu* 1 – 眉头 *méitóu* 1 – 两头 *liǎngtóu* 1

-zi

孩子 *háizi* 457 – 种子 *zhǒngzi* 146 – 儿子 *érzi* 131 – 日子 *rìzi* 129 – 妻子 *qīzi* 112 – 班子 *bānzi* 105 – 路子 *lùzi* 63 – 篮子 *lánzi* 58 – 伙子 *huǒzi* 53 – 房子 *fángzi* 50 – 帽子 *màozi* 37 – 一下子 *yíxiàzi* 29 – 样子 *yàngzi* 27 – 辈子 *bèizi* 25 – 饺子 *jiǎozi* 23 – 贩子 *fànzi* 22 – 担子 *dànzi* 21 – 孙子 *sūnzi* 20 – 牌子 *páizi* 20 – 肚子 *dùzi* 19 – 步子 *bùzi* 18 – 村子 *cūnzi* 18 – 一揽子 *yīlǎnzǐ* 16 – 桔子 *júzi* 16 – 脖子 *bózi* 15 – 身子 *shēnzǐ* 14 – 竹子 *zhúzi* 12 – 汉子 *hànzi* 11 – 侄子 *zhízi* 10 – 车子 *chēzi* 10 – 钉子 *dīngzi* 10 – 屋子 *wūzi* 10 – 厂子 *chǎngzi* 10 – 册子 *cèzi* 9 – 鼻子 *bízi* 9 – 茄子 *qiézi* 9 – 粒子 *lìzi* 8 – 苗子 *miáozi* 8 – 裙子 *qúnzi* 8 – 脑子 *nǎozi* 8 – 林子 *línzi* 8 – 椅子 *yǐzi* 8 – 鸽子 *gēzi* 8 – 被子 *bèizi* 8 – 鞋子 *xiézi* 7 – 沙子 *shāzi* 7 – 西门子 *xīménzi* 7 – 幌子 *huǎngzi*

6 – 绳子 *shéngzi* 6 – 袋子 *dàizi* 6 – 金子 *jīnzi* 6 – 影子 *yǐngzi* 6 – 例子 *lìzi* 6 – 枪杆子 *qiānggānzi* 6 – 斧子 *fǔzi* 6 – 口子 *kǒuzi* 6 – 梆子 *bāngzi* 5 – 底子 *dǐzi* 5 – 袜子 *wàzi* 5 – 膀子 *bǎngzi* 5 – 嗓子 *sǎngzi* 5 – 桌子 *zhuōzi* 5 – 票子 *piàozi* 5 – 胡子 *húzi* 5 – 话匣子 *huàxiázi* 5 – 圈子 *quānzi* 4 – 摊子 *tānzi* 4 – 棍子 *gùnzi* 4 – 杆子 *gānzi* 4 – 园子 *yuánzi* 4 – 院子 *yuànzi* 4 – 炉子 *lúzi* 4 – 果子 guǒzi 4 – 筷子 *kuàizi* 4 – 豹子 *bàozi* 4 – 片子 *piànzi* 4 – 刀子 *dāozi* 4 – 箱子 *xiāngzi* 3 – 匣子 *xiázi* 3 – 裤子 *kùzi* 3 – 褥子 *rùzi* 3 – 瓶子 *píngzi* 3 – 胆子 *dǎnzi* 3 – 豆子 *dòuzi* 3 – 个子 *gèzi* 3 – 点子 diǎnzi 3 – 狮子 shīzi 3 – 阵子 zhènzi 3 – 小子 xiǎozi 3 – 老头子 *lǎotóuzi* 3 – 台子 *táizi* 3 – 叶子 *yèzi* 3 – 杯子 *bēizi* 3 – 帘子 *liánzi* 2 – 梯子 *tīzi* 2 – 烂摊子 *làntānzi* 2 – 毯子 *tǎnzi* 2 – 瞎子 *xiāzi* 2 – 毽子 *jiànzi* 2 – 燕子 *yànzi* 2 – 兔子 *tùzi* 2 – 袖子 *xiùzi* 2 – 椰子 *yēzi* 2 – 瘤子 *liúzi* 2 – 猴子 *hóuzi* 2 – 盒子 *hézi* 2 – 虫子 *chóngzi* 2 – 蝎子 *xiēzi* 2 – 案子 *ànzi* 2 – 句子 *jùzi* 2 – 模子 *mózi* 2 – 空子 *kòngzi* 2 – 鞭子 *biānzi* 2 – 命根子 *mìnggēnzi* 2 – 曲子 *qǔzi* 2 – 法子 *fǎzi* 1 – 窗子 *chuāngzi* 1 – 谷子 *gǔzi* 1 – 哨子 *shàozi* 1 – 靶子 *bǎzi* 1 – 麂子 *jǐzi* 1 – 兜子 *dōuzi* 1 – 尖子 *jiānzi* 1 – 岔子 *chàzi* 1 – 游子 *yóuzi* 1 – 老样子 *lǎoyàngzi* 1 – 褂子 *guàzi* 1 – 乱子 *luànzi* 1 – 苇子 *wěizi* 1 – 坝子 *bàzi* 1 – 空架子 *kōngjiàzi* 1 – 银子 *yínzi* 1 – 阀子 *fázi* 1 – 丸子 *wánzi* 1 – 笛子 *dízi* 1 – 棚子 *péngzi* 1 – 辫子 *biànzi* 1 – 栗子 *lìzi* 1 – 柿子 *shìzi* 1 – 链子 *liànzi* 1 – 头子 *tóuzi* 1 – 蹄子 *tízi* 1 – 梭子 *suōzi* 1 – 骡子 *luózi* 1 – 骗子 *piànzi* 1 – 柚子 *yòuzi* 1 – 锤子 *chuízi* 1 – 石磙子 *shígǔnzi* 1 – 箕子 *jīzi* 1 – 槽子 *cáozi* 1 – 锭子 *dìngzi* 1 – 两口子 *liǎngkǒuzi* 1 – 椽子 *chuánzi* 1 – 单子 *dānzi* 1 – 剪子 *jiǎnzi* 1 – 档子 *dàngzi* 1 – 沙苑子 *shāyuànzǐ* 1 – 面子 *miànzi* 1 – 缨子 *yīngzi* 1 – 号子 *hàozi* 1 – 皮夹子 *píjiāzi* 1 – 镯子 *zhuózi* 1 – 卒子 *zúzi* 1 – 橙子 *chéngzi* 1 – 集子 *jízi* 1 – 鼓子 *gǔzi* 1 – 扇子 *shānzi* 1 – 桶子 *tǒngzi* 1 – 桃子 *táozi* 1 – 脚脖子 *jiǎobózi* 1 – 叔子 *shūzi* 1 – 庄子 *zhuāngzi* 1 – 胖子 *pàngzi* 1 – 杏子 *xìngzi* 1 – 狍子 *páozi* 1 – 台柱子 *táizhùzi* 1 – 份子 *fènzi* 1