Frequency Effects in Chinese Morphology:
Diachronic Evidence from a Synchronic Corpus*

Chao-Jan Chen (陳超然)
National Chi Nan University

This paper explores the type frequency effect in Chinese compounding. We observe that the distribution of character productivity in compound words follows a power law. Based on relevant research on network connectivity, I suggest that the power-law distribution be regarded as the consequence of a positive feedback effect. Such a “rich-get-richer” effect is brought about by the type frequency effect of characters in compounding morphology during the diachronic development of Chinese compound lexicon. The study shows how diachronic evidence can be provided by a large synchronic corpus with statistical analysis. This work also provides empirical evidence for the usage-based hypothesis in the framework of Cognitive Grammar.

Key words: frequency effects, morphological productivity, compounding, power-law, network model

1. Introduction

The ideas of frequency and entrenchment play a fundamental role in the usage-based model of the framework of Cognitive Grammar (see Langacker 1987, 2000, 2008, and Barlow & Kemmer 2000). Different kinds of frequency effects have been explored in various fields of linguistic research, especially in morphology and phonology (e.g. Bybee & Hopper 2001, and Bybee 1985, 2007). Regarded as “a major factor determining the degree of productivity of a construction” (Bybee 1985, 2007), type frequency is supposed to play an important role in compounding. However, as Brinton & Traugott (2005:34) remarked, the original productivity in compounding might be obscured over time due to some phonological and other changes. Such obscurity of participant morphemes in words thus makes it difficult, if not impossible, to trace the morphological productivity of certain morphemes in the course of history.

* This research was partially supported by Grants NSC (NSC 97-2410-H-260-036 and NSC 98-2410-H-260-044).
As a language extremely rich in compound words, Chinese is supposed to be an ideal target for the study of the type frequency effect in compounding. What is even more desirable for research on frequency effects is that the component morphemes in Chinese compound words generally remain transparent and thus perfectly discernable as units of \textit{zi} (or Chinese characters\footnote{Throughout this paper, the term “character” will be used to refer to the monosyllabic semantic unit, almost always a morpheme, rather than a mere grapheme, in the word writing system of Chinese.}) in the long course of history. However, quantitative studies about the overall morphological productivity of characters in compounding have rarely been the focus of research in Chinese morphology. For linguists who work under the generative paradigm, quantitative study involving frequency is regarded as a performance rather than competence issue and is thus neglected. On the other side, for computational linguists, who surely do quantitative research, the focus of morphological productivity is often placed on certain highly productive affixes (or fixed-placed roots). Therefore, researchers of Chinese morphology seldom pay attention to the productivity of “ordinary” morphemes in word formation, let alone study their productivity distribution pattern.

This paper thus aims to explore the type frequency effect in Chinese compounding. Instead of doing case studies on the morphological productivity for individual characters in forming compounds, we will focus directly on the overall distribution pattern of character productivity and explore its significance in diachronic linguistics under the theoretic framework of Cognitive Grammar, specifically in the usage-based model. Related research in dynamic network structures will be used to help build a usage-based model of compounding. The data we will use in this study is collected from a large corpus (containing five million word tokens) of contemporary Chinese, the Academia Sinica Balanced Corpus version 3.0 (abbreviated as ASBC). As two-character compounds are the only type of compound concerned in this paper, from now on, for the sake of simplicity, they will simply be referred to as “compounds”.

2. Theoretic framework

2.1 Usage-based model and compounding template

According to Langacker’s (1987, 2000, 2008) usage-based model, linguistic symbolic units can be schematized from usage events. In this theoretic framework, language structure is supposed not to be autonomous from the usage. Rather, the structure emerges from language use. Patterns can be abstracted from similar instances and entrenched through frequency to form schemas (or constructions in Construction Grammar’s term), which can later serve to form new instances.
Adopting this dynamic viewpoint of constructions, we propose, in the case of Chinese compounding, a hypothesis that the production of compounds can be template-based: existing compounds can serve as templates (or “schemas”, to use Langacker’s term) to motivate the creation of new compounds, when related semantic constraints are satisfied. For example, an existing compound X-Y (with X and Y as component characters) can serve respectively as two templates [( )-Y] and [X-( )] to motivate the creation of new compounds such as W-Y or X-Z.

By the same token, the production of the compound X-Y can be motivated by template [( )-Y], by template [X-( )], or by both of them. Such a way of template-motivated compounding can also be regarded as analogical creation (Chen 2005) or modification of prefabricated structures (Barlow 2000).2

2.2 Type frequency effect and potential positive feedback

According to Bybee (2007), type frequency is “a major factor determining the degree of productivity of a construction”, which means “constructions that apply to a high number of distinct items also tend to be highly applicable to new items” (cf. Bybee 1985). While type frequency means the frequency of word types that conform to a schema, the productivity of a schema is argued to be a function of type frequency with the following hypothesis: “The productivity of a schema is a function of the type frequency of the instances of the schema.” (Bybee 1985:132-134, see also Bybee & Hopper 2001, and Langacker 2008).

Following Bybee’s hypothesis of the type frequency effect, in our usage-based model of Chinese compounding, we would expect that existing compounds serve as compounding templates, of which the type frequency will lead to the entrenchment of the compounding template (the schema in our case) as a frequency effect. To be more specific, we propose that the type frequency of the compounding template can determine, in a certain way, the probability of the derived compounds to be produced. A theoretical consequence of this proposal should then be positive feedback: the characters used more frequently in existing compounds will be more likely to be used in new compounds and thus become more productive afterward. In the following sections, we will use data in the Chinese corpus ASBC to test this hypothesis.

2 Barlow (2000:318) argues that “the creativity or the expressive dimension to language comes in large part from the modification of prefabricated structures, rather than the novel combination of lexical categories.”
3. Character productivity: data and observation

3.1 Data collection

As our study of the type frequency effect is based on the framework of Cognitive Grammar (also abbreviated as CG) and focuses on two-character compounds in Chinese, we regard the compound word as a symbolic unit, combining two morpheme-functioning characters, which is well entrenched through language use. We are then allowed to avoid the problem, brought about by the traditional definition of compounds which involves confusing and controversial criteria about the distinction between bound and free morphemes, and between phrases and compound words in Chinese (see, for example, Dong 2002:25-27). As Li & Thompson (1981:45) remarked, “no matter what criteria one picks, there is no clear demarcation between compounds and non compounds”. In other words, there is no clear-cut between compounding and derivation in Chinese morphology.

The CG-based definition of compounds also allows us to use data in the ASBC corpus for our work, as the criteria of wordhood in ASBC generally are compatible with our criterion of compounds above (cf. the technical report no.95-02/98-04 of CKIP, the creating and maintaining group of the ASBC corpus). As multi-character morphemes are extremely rare in Chinese, virtually all the two-character words in ASBC can thus be considered as compounds. Since our study of character productivity in compounding is focused on two-character compound words only, we take as analysis data of compounds all the two-character words in the ASBC corpus. This simple criterion for compoundhood has its practical advantage. It allows us to do our corpus-based statistic analysis in a large-scale and automatic way and to avoid manual distinction of compounds among huge data, which is not quite feasible. Surely the two-character words in the corpus include few single-morpheme words. However, such relatively rare exceptions can be considered as noise technically tolerable in our statistical analysis because our task is mainly to observe an overall statistical pattern and therefore would not be sensitive to this trivial data noise.

3.2 An aristocratic world of productivity

Containing about five million word tokens, the ASBC corpus version 3.0 includes 66,722 (two-character) compound types, which are formed by 5,486 character types. Simple statistics show that the compounding productivity is extremely variable among the characters: while a great number of characters (1,070 or about 19.5%) only appear in one compound (see Figure 1, which lists only the partial distribution for the number of characters that appear in under 20 compound types), the most productive character
can appear in up to 607 compounds (occupying almost 1% of compound types) (see Table 1).

Since the average number of compound types formed by the characters is 24.42, we might expect that “normally” a character would appear in around 24 compound words. However, it is not the case. In fact, as we can see in Figure 1, the distribution of character productivity in compounding is not “normal” at all: it is not a bell-shaped distribution as that which is called normal distribution in statistics. Contrary to our expectations, the world of morphological productivity we face here is so “aristocratic” instead of “equalitarian” (terms borrowed from Buchanan 2002:119): while most of the characters are “impoverished” in compounding productivity, the top ranking ones are extremely “rich”.

**Figure 1:** Distribution of character productivity (i)

**Table 1:** Distribution of character productivity (ii): Top 10 and Bottom 10

<table>
<thead>
<tr>
<th></th>
<th>Top 10</th>
<th></th>
<th>Bottom 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>#comp</td>
<td>#char</td>
<td>#comp</td>
<td>#char</td>
</tr>
<tr>
<td>607</td>
<td>1</td>
<td>415</td>
<td>1</td>
</tr>
<tr>
<td>553</td>
<td>1</td>
<td>392</td>
<td>1</td>
</tr>
<tr>
<td>513</td>
<td>1</td>
<td>383</td>
<td>1</td>
</tr>
<tr>
<td>482</td>
<td>1</td>
<td>340</td>
<td>1</td>
</tr>
<tr>
<td>418</td>
<td>1</td>
<td>322</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>99</td>
<td>5</td>
<td>191</td>
</tr>
<tr>
<td>9</td>
<td>108</td>
<td>4</td>
<td>244</td>
</tr>
<tr>
<td>8</td>
<td>142</td>
<td>3</td>
<td>351</td>
</tr>
<tr>
<td>7</td>
<td>142</td>
<td>2</td>
<td>527</td>
</tr>
<tr>
<td>6</td>
<td>162</td>
<td>1</td>
<td>1,070</td>
</tr>
</tbody>
</table>
4. Power-law distribution and its significance

4.1 A power-law distribution

If we present the productivity distribution of the characters in ordinary coordinates, we can see its decline tendency through the curve in the scatter plot (see Figure 2).

![Figure 2: Distribution of character productivity (iii)](image1)

If we present this distribution in another way, namely in a plot in log-log coordinates, we will see that the productivity distribution of the characters turns out to be a straight line, instead of a curve, which is the decline tendency of the curve in Figure 2 (see Figure 3).

![Figure 3: Distribution of character productivity (iv) (log-log)](image2)
By the method of least-square regression, we obtain a straight line as its best fitting curve:

\[ \log N(k) = 3.028 - 1.003 \log k \quad (r = -0.9846, \quad r^2 = 0.9694), \]

where \( N(k) \) is the number of characters participating in \( k \) compounds

The formula of productivity distribution can also be represented as follows:

\[ N(k) = 1066 k^{-1.003} \]

This formula clearly shows that the character productivity in compounding follows a power-law distribution.\(^3\)

**4.2 Significance of power-law distribution**

In fact, the power-law distribution has been found in many kinds of real or virtual scale-free networks, including the World Wide Web, academic coauthor networks, the food chain web, airport networks, and the synonym web (see also Buchanan 2002, Barabási 2003, Barabási et al. 2003, Steyvers & Tenenbaum 2005, and Newman et al. 2006). In such networks, the number of nodes \( N \) is a function of the number of links according to the following formula:

\[ N(k) = C k^{-a} \]

According to Barabási & Albert (1999), a power-law distribution can be regarded as the signature of two properties in a network: (1) incremental growth of nodes and links, and (2) preferential attachment. In other words, the network following a power-law distribution is evolving with growing nodes and links; new nodes are linked to existing nodes by the principle of preferential attachment. That means the more links a node already has, the more likely it will get new links, which brings about a “rich-get-richer” effect. For example, a website linked by more websites will be more likely to be linked by new websites, a more frequently cited paper will be more likely to be cited again by other new papers, and so on.

\(^3\) According to the definition of power-law distribution, when \( Y(x) = C X^{-a} \), we have a power-law distribution. We can see a straight line with a slope \(-a\) in its corresponding log-log plot. As \( \log Y = \log C - a \log X \), such a straight line in a log-log plot is thus regarded as the most salient feature of a power-law distribution.
5. A network model of compounding and its linguistic significance

In fact, as proposed in our previous approach (Chen 2005), we can regard compounding as a link: for a compound X-Y, there exists a link between X and Y. Therefore such compounding connectivity forms a network, in which characters are nodes and compounding relations are links, as shown in Figure 4.

![Character network of compounding](image)

**Figure 4:** Compounding as links between characters

Viewing compounding as a network of characters, we can also find in this network the two critical factors accounting for the power-law distribution: the growth of the network and preferential attachment. Evidently, according to the diachronic development of a compound lexicon in Chinese, the modeling network should be growing. What is more important, the preferential attachment principle holds as a theoretical consequence under the Cognitive Grammar’s hypothesis: the more used characters become more likely to be used again, thus being more productive in compounding (a “rich get richer” effect), which is exactly what the type frequency effect tells. In other words, when a character is more ‘productive’ in existing words, it is also more ‘productive’ for potential words.

However, in our compounding network model for two-character compounds, only the links grow, while the nodes (the characters) remain the same over time, which is slightly different from the dynamic mathematical models for network growth proposed in previous research such as that of Barabási & Albert (1999) (see also Newman et al. 2006 and Durrett 2007 for related mathematical models). Our previous research (Chen 2012) therefore sets up a stochastic model of compounding based on this feature. In this dynamic model, the probability of a compound X-Y being produced is proportional to the type frequency of template [X-( )] and to that of template [( )-Y]. The compounds thus produced can go on to contribute to the type frequency of related templates. The results
of the simulated evolution of a compound lexicon show that a power-law distribution can also be observed in such a model, as expected of other dynamic models with preferential attachment.\(^4\)

Therefore, our evolving network model for Chinese compounding shows that the power-law distribution of character productivity can be accounted for by the type frequency effect proposed by the usage-based framework of Cognitive Grammar. Such a significant distribution pattern would not be otherwise satisfactorily explained linguistically. In other words, the power-law distribution of character productivity strongly suggests that the morphological mechanism of Chinese compounding is under the type frequency effect, which provides evidence for the usage-based framework of Cognitive Grammar.

6. Concluding remarks

In this paper, we have tried to show the effect of type frequency on Chinese compounding. To account for the fact that the distribution of character productivity in forming compounds follows a power law, we propose a network model of compounding, which is composed of characters and links between every pair of component characters in compounds. We have argued that such a network model can give rise to a power-law degree distribution, which is, according to related research in the structure and dynamics of various networks in virtual or real worlds, considered as the signature of a “rich-get-richer” growing network with preferential attachment. In the dynamic model we proposed, preferential attachment is actually the effect produced by a stochastic mechanism applying the hypothesis of type frequency effect.

Though our model is rather simplistic, it still allows us to capture the essence of the positive feedback effect in the diachronic development of the Chinese lexicon and to account for the power-law distribution we observed in character productivity. A broader theoretical significance of our work is that the effect of type frequency on compound production provides evidence for the usage-based model in the theoretical framework of Cognitive Grammar. The results of our study also suggest that compounding can be an analogical creation that is motivated by templates abstracted from existing compounds. Moreover, the study shows how diachronic evidence can be provided by a large synchronous corpus with statistical analysis. The application of such a global statistic method in diachronic linguistic research allows us to bypass the problem brought about by individual words whose historical development is unclear and difficult to trace.

\(^4\) As the main aim of this paper is to explore the linguistic explanation rather than to provide an elaborated mathematic model for the power-law distribution of character productivity, we are contented with such a simple and plausible model capturing the essence of the rich-get-richer effect without getting into mathematical details.
References

Taipei: Academia Sinica.


