A Corpus Investigation of the Right-hand Head Rule Applied to Japanese Affixes

Yayoi Miyaoka ²
Hiroshima University of Economics, Japan
Katsuo Tamaoka ³
Hiroshima University, Japan

Abstract: The present study investigates differences between Japanese prefixes and suffixes using editions of the Asashi Newspaper published between 1985 and 1998 (Amano & Kondo, 2000). The right-hand head rule (e.g., Kageyama, 1982; Kageyama, 1999; Namiki, 1982; Nishigauchi, 2004; Williams, 1981) predicts that prefixes would be attached to a wide variety of nouns while suffixes would be regularly attached to a smaller group of nouns. Twenty-four frequently-used affixes consisting of 12 prefixes and 12 suffixes were compared according to 7 corpus features, including printed-frequency, productivity, accumulative productivity, commonality, coalescence degree, Herdan’s logarithmic function of type-token ratio (log TTR), and entropy. Although a series of Mann-Whitney U-tests calculated for the six corpus features of printed-frequency, productivity, accumulative productivity, commonality, coalescence degree and log TTR did not reveal any differences between the 12 prefixes and the 12 suffixes, the t-test for entropy indicated a significant difference. This suggests that the prefixes were more randomly or chaotically attached to nouns than the suffixes. Although the present findings are limited only to the selected 24 affixes, the result supported the right-hand head rule.

Keywords: Japanese affix, prefix, suffix, right-hand head rule, coalescence degree, Herdan’s logarithmic function of type-token ratio (log TTR), entropy

1. Background of Japanese prefixes and suffixes

In general, there are three types of morphological word formation: compounding, derivation, and inflection (Kageyama, 1993). Morphological elements having independent units of meaning are called ‘bases’, while those that are connected to bases are called ‘affixes’. Furthermore, there are four types of affixes in the languages of the world: ‘prefixes’, which are placed at the head of the base (e.g. dis in ‘dislike’ in English); ‘infixes’, which appear within a word (e.g. um in ‘kumain’ for ‘ate’ in Tagalog), ‘suffixes’ which are added to the end of the base (e.g. er in ‘reporter’ in English), and ‘circumfixes’ (e.g. baik meaning ‘good’, and kebaikan ‘goodness’ in Indonesian). The only two types of affixes used in modern Japanese are prefixes and suffixes. For example, the Japanese prefix 不 meaning ‘un-’ is added to 自然

---

1 The present study was supported by the Research Institute of Science and Technology for Society (RISTEX) at the Japan Science and Technology Corporation (JST), Kawaguchi, Japan.
2 Address correspondence to: Yayoi Miyaoka, Hiroshima University of Economics, 37-1, 5-chome, Gion, Asaminami-ku, Hiroshima, Japan 731-0192. E-mail: y.miya8411@hue.ac.jp.
3 Katsuo Tamaoka, International Student Center, Hiroshima University 1-1, 1-chome, Higashihiroshima, Japan 739-8524. E-mail: ktamaoka@hiroshima-u.ac.jp.
meaning ‘natural’, creating the morphologically complex word 不自然 meaning ‘unnatural’. Similarly, another prefix 全 meaning ‘whole’ is added to the base 世界 meaning ‘world’, producing the word 全世界 meaning ‘whole world’. An example of a suffix is the which forms an adjective; it can be added to the end of the base 建設 meaning ‘construction’ to create a compound word 建設的 meaning ‘constructive’. Similarly, the suffix 性 forms a noun, so that when it is added to the end of the base 安全 meaning ‘safe’ becomes the compound word 安全性 meaning ‘safety’. Nomura (1977) explained that the base is the semantic core element of a word, while the affix element adds meaning to the base and determines its grammatical category. Thus, the base can stand alone as a word, but the affix cannot be a word by itself. In previous studies (e.g., Kageyama, 1999; Nomura, 1977), it has been shown that prefixes add meaning to the base without changing the grammatical category, whereas suffixes not only add meaning but may also change the grammatical category of the base. For instance, 全世界 meaning ‘whole world’ is composed of the prefix 全 ‘whole’ and the noun 世界 ‘world’: the attachment of a prefix to the noun does not result in a change in grammatical category. On the other hand, the suffix 性 in 建設的 changes the noun 建設 into an adjective. Thus, the hypothesis proposed by Williams (1981) that the right-hand side of a complex word determines the grammatical category of the word holds true not only for English but also for Japanese. This tendency is often referred to as the right-hand head rule (e.g., Kageyama, 1982; Kageyama, 1999; Namiki, 1982; Nishigauchi, 2004). This key difference between prefixes and suffixes may also affect the extent to which they co-occur with various nouns in printed-frequency. Prefixes do not change the grammatical category of the nouns to which they are attached, so they may be attached to a wide variety of nouns. By contrast, some suffixes change the grammatical category of a noun; this limits the range of nouns with which they may appear and results in the regular combination of a small group of certain suffixes and nouns.

In his seminal work, A Mathematical Theory of Communication (1948), American mathematician Claude Elwood Shannon (1916-2001) first developed the concepts of entropy and redundancy for information processing. Entropy is an index of the degree of disorder or chaos; redundancy refers to the degree of superfluousness. Since these two concepts can be applied to a wide range of corpus sizes, characteristics of prefix and suffix attachments to a variety of nouns can be directly compared (for details, see Hori, 1979 and Kaiho, 1989; for an example of an actual corpus study which applied these concepts, see Tamaoka, Miyaoka & Lim, 2003; Tamaoka, Lim & Sakai, 2004). Therefore, the present study utilizes the index of entropy to analyze co-occurring frequencies of affixes and nouns, hypothesizing that prefixes show higher entropy since they would be expected to be attached to a wider variety of nouns.

2. Selection of prefixes and suffixes

In the present study, 12 prefixes and 12 suffixes, all commonly used, were compared. The prefixes were 大 ‘big’, 不 ‘un-’, 無 ‘un-’, 新 ‘new’, 初 ‘first’, 非 ‘un-’, 全 ‘whole’, 再 ‘re-’, 超 ‘super’, 反 ‘anti’, 未 ‘not yet’, 毎 ‘every’. The suffixes were い ‘tive/-like’, 者 ‘person’, 性 ‘nature’, 学 ‘studies’, 化 ‘characteristic’, 論 ‘theory/-logy’, 家 ‘-ist’, 式 ‘manner/style’, 界 ‘world’, 風 ‘style’, 状 ‘state’, 用 ‘use’. Since these target affixes are the most commonly-seen items, they were sufficient to investigate the actual usage of Japanese affixes in a corpus. The simple printed-frequencies of the 24 kanji symbols used for the affixes are shown in Table 1.
3. Identifying prefix, suffix and base

There are some unresolved issues regarding the definitions of the prefix, suffix, and base in Japanese. The present study defines a base as an element that can be a single word by itself. Therefore, an element without an affix is presented in a single kanji. For instance the word 新世代 meaning ‘new generation’ is composed of the noun 世代 ‘generation’ attached to the prefix 新 ‘new’. In this case, 世代 is the base, as it can stand as a single word without the prefix. The base is easy to identify in this example, since 世代 is a two-kanji compound word. However, the base in an example such as 株式 ‘stock’ is more difficult to identify. Because the single kanji 株 can be found in the dictionary as a single word meaning ‘stump’ or ‘stock’, the present study interpreted the word 株式 in such a way that 式 is the suffix while 株 serves as the base. This is not to suggest that 式 can always be identified as a suffix, as by itself 式 also means ‘ceremony’. Accordingly, in the word 卒業式 meaning ‘graduation ceremony’, 式 is not considered to be a suffix. In the same way, the suffix 状 has two meanings of ‘state’ and ‘letter’ when standing alone; only 状 meaning ‘state’ is considered a suffix in the present study.

Some single-kanji affixes have more than one pronunciation. For instance, 大 meaning ‘big’ is pronounced in two ways: /dai/ in On-reading (a Chinese-originated sound), and /oR/ (/R/ refers to a long vowel) in Kun-reading (a Japanese-originated sound). The pronunciation of 大 varies depending on how the prefix is attached to bases. The word 大混乱 meaning ‘big confusion’ is pronounced /dai+koNraN/ (/N/ refers to a nasal) while the word 大津波 meaning ‘big tsunami’ or ‘big seismic sea wave’ is pronounced /oR+tunami/. The meaning of the prefix 大 remains the same in both instances. So the present study regards these two different pronunciations as belonging to the same affix 大.

4. Selection of words incorporating the 24 selected affixes

The present study used a lexical corpus of the Asahi Newspaper printed from 1985 to 1998, produced by Amano and Kondo (2000). This corpus contains 341,771 types and 287,792,797 tokens. All words co-appearing with the selected 24 affixes were extracted from the Asahi Newspaper corpus using the software called EasySrch (Amano & Kondo, 2003). For example, 42 compound word types co-appearing with the prefix 再 meaning ‘re-’ were found, including 再検討 ‘reconsideration’ (3,859 tokens), 再確認 ‘reconfirm’ (3,109 tokens), 再開発 ‘redevelopment’ (2,704 tokens), 再構築 ‘reconstruction’ (2,012 tokens), 再評価 ‘reevaluation’ (1,069 tokens), and so on. This word selection process was applied to all 24 selected affixes. The base elements could be any type of Sino-Japanese word (wa-go), Chinese-originated word (kan-go), or loanword (commonly called gairai-go or katakana-hyooki-go). Nine corpus features were calculated for each affix using the lexical frequency index of the Asashi Newspaper corpus.

5. Calculating corpus features for the 24 affixes

Using the data collection process explained in Section 4, the present study calculated nine different corpus features, including a simple addition of printed-frequency, productivity, accumulative productivity and a more complex calculation of entropy. The calculation proces-
<table>
<thead>
<tr>
<th>Affix Type</th>
<th>Kanji</th>
<th>Printed-freq.</th>
<th>Productivity</th>
<th>Acc. Prod.</th>
<th>Affix-used ratio</th>
<th>Commonality</th>
<th>log TTR</th>
<th>Entropy</th>
<th>H-maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prefixes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>大</td>
<td>2,103,545</td>
<td>359</td>
<td>192,138</td>
<td>0.091</td>
<td>0.132</td>
<td>0.484</td>
<td>5.395</td>
<td>8.488</td>
<td></td>
</tr>
<tr>
<td>不</td>
<td>446,439</td>
<td>158</td>
<td>85,223</td>
<td>0.191</td>
<td>0.093</td>
<td>0.446</td>
<td>5.272</td>
<td>7.304</td>
<td></td>
</tr>
<tr>
<td>無</td>
<td>222,077</td>
<td>150</td>
<td>60,732</td>
<td>0.272</td>
<td>0.098</td>
<td>0.455</td>
<td>5.361</td>
<td>7.229</td>
<td></td>
</tr>
<tr>
<td>新</td>
<td>767,768</td>
<td>132</td>
<td>49,477</td>
<td>0.064</td>
<td>0.368</td>
<td>0.452</td>
<td>3.804</td>
<td>7.044</td>
<td></td>
</tr>
<tr>
<td>初</td>
<td>302,843</td>
<td>90</td>
<td>11,189</td>
<td>0.037</td>
<td>0.215</td>
<td>0.483</td>
<td>4.246</td>
<td>6.492</td>
<td></td>
</tr>
<tr>
<td>非</td>
<td>108,303</td>
<td>58</td>
<td>31,364</td>
<td>0.290</td>
<td>0.193</td>
<td>0.392</td>
<td>3.624</td>
<td>5.858</td>
<td></td>
</tr>
<tr>
<td>全</td>
<td>555,403</td>
<td>44</td>
<td>16,876</td>
<td>0.030</td>
<td>0.298</td>
<td>0.389</td>
<td>3.541</td>
<td>5.459</td>
<td></td>
</tr>
<tr>
<td>再</td>
<td>190,873</td>
<td>42</td>
<td>20,871</td>
<td>0.110</td>
<td>0.184</td>
<td>0.376</td>
<td>4.013</td>
<td>5.392</td>
<td></td>
</tr>
<tr>
<td>超</td>
<td>91,843</td>
<td>40</td>
<td>9,879</td>
<td>0.108</td>
<td>0.217</td>
<td>0.401</td>
<td>3.500</td>
<td>5.322</td>
<td></td>
</tr>
<tr>
<td>反</td>
<td>265,758</td>
<td>30</td>
<td>6,957</td>
<td>0.024</td>
<td>0.329</td>
<td>0.384</td>
<td>2.924</td>
<td>4.907</td>
<td></td>
</tr>
<tr>
<td>未</td>
<td>74,563</td>
<td>19</td>
<td>6,430</td>
<td>0.036</td>
<td>0.225</td>
<td>0.336</td>
<td>3.295</td>
<td>4.248</td>
<td></td>
</tr>
<tr>
<td>每</td>
<td>41,715</td>
<td>16</td>
<td>4,739</td>
<td>0.768</td>
<td>0.369</td>
<td>0.258</td>
<td>2.443</td>
<td>4.000</td>
<td></td>
</tr>
<tr>
<td><strong>Suffixes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>的</td>
<td>833,709</td>
<td>342</td>
<td>158,849</td>
<td>0.191</td>
<td>0.054</td>
<td>0.487</td>
<td>6.382</td>
<td>8.418</td>
<td></td>
</tr>
<tr>
<td>者</td>
<td>1,080,160</td>
<td>310</td>
<td>415,725</td>
<td>0.355</td>
<td>0.129</td>
<td>0.443</td>
<td>5.461</td>
<td>8.276</td>
<td></td>
</tr>
<tr>
<td>性</td>
<td>453,110</td>
<td>216</td>
<td>130,676</td>
<td>0.238</td>
<td>0.481</td>
<td>0.456</td>
<td>3.731</td>
<td>7.755</td>
<td></td>
</tr>
<tr>
<td>学</td>
<td>751,148</td>
<td>178</td>
<td>36,747</td>
<td>0.047</td>
<td>0.125</td>
<td>0.493</td>
<td>5.533</td>
<td>7.476</td>
<td></td>
</tr>
<tr>
<td>化</td>
<td>578,810</td>
<td>167</td>
<td>172,339</td>
<td>0.298</td>
<td>0.129</td>
<td>0.424</td>
<td>5.146</td>
<td>7.384</td>
<td></td>
</tr>
<tr>
<td>论</td>
<td>265,332</td>
<td>141</td>
<td>5,298</td>
<td>0.020</td>
<td>0.184</td>
<td>0.577</td>
<td>5.177</td>
<td>7.140</td>
<td></td>
</tr>
<tr>
<td>家</td>
<td>488,437</td>
<td>113</td>
<td>101,770</td>
<td>0.208</td>
<td>0.280</td>
<td>0.410</td>
<td>3.861</td>
<td>6.820</td>
<td></td>
</tr>
<tr>
<td>式</td>
<td>229,697</td>
<td>40</td>
<td>23,460</td>
<td>0.102</td>
<td>0.922</td>
<td>0.367</td>
<td>0.669</td>
<td>5.322</td>
<td></td>
</tr>
<tr>
<td>界</td>
<td>374,754</td>
<td>35</td>
<td>11,468</td>
<td>0.035</td>
<td>0.458</td>
<td>0.380</td>
<td>2.809</td>
<td>5.129</td>
<td></td>
</tr>
<tr>
<td>凰</td>
<td>102,317</td>
<td>22</td>
<td>956</td>
<td>0.009</td>
<td>0.270</td>
<td>0.450</td>
<td>3.443</td>
<td>4.459</td>
<td></td>
</tr>
<tr>
<td>须</td>
<td>186,124</td>
<td>13</td>
<td>809</td>
<td>0.034</td>
<td>0.575</td>
<td>0.383</td>
<td>1.943</td>
<td>3.700</td>
<td></td>
</tr>
<tr>
<td>用</td>
<td>465,611</td>
<td>6</td>
<td>2,721</td>
<td>0.005</td>
<td>0.680</td>
<td>0.227</td>
<td>1.337</td>
<td>2.585</td>
<td></td>
</tr>
</tbody>
</table>

*Note 1*: 'Acc. Prod.' refers to accumulative productivity.

*Note 2*: 'log TTR' refers to a type-token ratio calculated by legalism of type frequency divided by legalism of token frequency.
ses are explained in the following sub-sections. The figures for the 12 prefixes and 12 suffixes with regard to each of the nine features are presented in Table 1.

5.1 Printed-frequency, productivity and accumulative productivity

Since the three corpus features of printed-frequency, productivity and accumulative productivity are simple frequency counts, a nonparametric analysis of the Mann-Whitney $U$-test was used to compare the 12 prefixes and the 12 suffixes. The mean rank and the sum of the ranks are shown in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Kanji</th>
<th>Mean and Sum</th>
<th>Printed-freq.</th>
<th>Productivity</th>
<th>Acc. Prod.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prefixes</td>
<td>Rank Mean</td>
<td>10.417</td>
<td>11.792</td>
<td>12.417</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sum of the Ranks</td>
<td>125.000</td>
<td>141.500</td>
<td>149.000</td>
</tr>
<tr>
<td></td>
<td>Suffixes</td>
<td>Rank Mean</td>
<td>14.583</td>
<td>13.208</td>
<td>12.583</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sum of the Ranks</td>
<td>175.000</td>
<td>158.500</td>
<td>151.000</td>
</tr>
</tbody>
</table>

Table 2
Comparisons between prefixes and suffixes by Mann-Whitney $U$-test

Note: A sample size of prefixes and suffixes is 12 each.

The printed-frequencies for the 24 different kanji symbols used for the affixes considered in this study were derived from the lexical corpus of the Asashi Newspaper (Amano & Kondo, 2000). The printed-frequency of the 12 prefixes (rank mean = 10.417, sum of the ranks = 124.000) was not significantly larger than that of the 12 suffixes (rank mean = 14.583, sum of the ranks = 175.000) [$U = 47.000, p = .160, n.s.$]. This implies that simple frequencies of kanji symbols encoding affixes do not distinguish between prefixes and suffixes.

The productivity feature indicates in how many words each affix is combined with nouns (i.e., type frequency of words with the given affix). As shown in Table 1, the prefix 大 had the greatest productivity at 359 words, while the suffix 用 had the greatest productivity at 342 words. The prefix 每 had the lowest productivity at 16 words, while the suffix 用 had the lowest productivity at only 6 words. The rank mean of productivity for the 12 prefixes (rank mean = 11.79, sum of the ranks = 141.50) was not significantly larger than that of the 12 suffixes (rank mean = 13.21, sum of the ranks = 158.50) [$U = 63.500, p = .630, n.s.$]. In other words, these 12 prefixes are attached to words as often as the 12 suffixes.

The accumulative productivity was calculated by summing the printed-frequencies of all the words with the given affix. In other words, productivity is type frequency whereas accumulative productivity is token frequency. Similarly to productivity, a U-test for accumulative productivity indicates no difference between the 12 prefixes (rank mean = 12.42, sum of the ranks = 149.00) and the 12 suffixes (rank mean = 12.58, sum of the ranks = 151.00) [$U = 71.00, p = .977, n.s.$].

A nonparametric rank-order correlation coefficient of Spearman’s rho was computed for all the 24 affixes together based on the three features mentioned above, since there were no differences between the 12 prefixes and the 12 suffixes. Correlation coefficients of all combinations of the three variables were significant; the correlation between printed-frequency and productivity [$r_s(24) = .704, p < .001$], the correlation between printed-frequency and accumulative productivity [$r_s(24) = .609, p < .01$], and the correlation between productivity
and accumulative productivity \( r_s(24) = .813, p < .001 \). These frequency features have strong interrelations.

5.2 Coalescence degree, commonality and Herdan’s type-token ratio (log TTR)

Since simple type and token frequencies did not show any difference between prefixes and suffixes, somewhat more complex features of their behavior in the corpus were also calculated. The coalescence degree was calculated by dividing accumulative productivity by printed-frequency for a given affix. For example, the suffix 家 was printed 488,437 times (including proper nouns in the Asashi Newspaper\(^4\)) and accumulative productivity 101,770 times. Thus, coalescence degree was 0.028 (101,770 divided by 488,437). The rank mean of the coalescence degree for the 12 affixes (rank mean=13.50, sum of the ranks=162.00) was not significantly larger than that of the 12 suffixes (rank mean=11.50, sum of the ranks=138.00) \( U=60.00, p=.514, \text{n.s.} \). Thus, there was no difference between prefixes and suffixes in the number of times that the kanji were used for affixes.

<table>
<thead>
<tr>
<th>Kanji</th>
<th>Mean and Sum</th>
<th>Coalescence Degree</th>
<th>Commonality</th>
<th>log TTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prefixes</td>
<td>Rank Mean</td>
<td>13.50</td>
<td>11.33</td>
<td>11.58</td>
</tr>
<tr>
<td></td>
<td>Sum of the Ranks</td>
<td>162.00</td>
<td>156.00</td>
<td>159.00</td>
</tr>
<tr>
<td>Suffixes</td>
<td>Rank Mean</td>
<td>11.50</td>
<td>13.67</td>
<td>13.42</td>
</tr>
<tr>
<td></td>
<td>Sum of the Ranks</td>
<td>160.00</td>
<td>164.00</td>
<td>161.00</td>
</tr>
<tr>
<td>Results of U-test</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td></td>
</tr>
</tbody>
</table>

Note: 'n.s.' refers to not significant.

The commonality refers to how often the most frequently-used word with a target affix occupy the total accumulative frequencies of all the words with a target affix. In the case of the prefix 不, the most frequently-used compound noun was 不十分 consisting of the prefix 不 and the two-kanji compound word 十分 meaning ‘sufficient’. This word appeared 7,965 times in the corpus. Since the total accumulative productivity (i.e., the token frequency of all words with the prefix 不) was 85,223, the commonality was calculated by dividing the frequency of the most frequently-used word by the accumulative productivity. Thus, commonality for 不 was 0.093 (7,965 divided by 85,223). The rank mean of commonality for the 12 affixes (rank mean = 11.33, sum of the ranks = 136.00) was not significantly larger than that of the 12 suffixes (rank mean = 13.67, sum of the ranks = 164.00) \( U=58.000, p = .443, \text{n.s.} \).

The log type-token ratio (log TTR) quantifies how many words with the target affix are in the corpus and how frequently they are used. However, the result of simple TTR is almost always equal to zero in a corpus. In the present study, as proposed by Wimmer and Altmann (1999) as one of the candidate calculations, a logarithmic function of TTR by Herdan (1960) was utilized to compare the 12 prefixes and the 12 suffixes. The calculation is simply:

---

\(^4\) Tamaoka and Makioka (2004) calculated frequencies for the 1,945 basic Japanese kanji without proper nouns in order to avoid biases from specific popular incidents or events reported in the Asashi Newspaper.
Right-hand head rule applied to Japanese affixes

\[ TTR = \frac{\ln V}{\ln N} \]

where V is the number of words (i.e. productivity or type frequency) and N is the number of all accumulative frequencies of words (i.e. accumulative productivity or token frequency). For example, the suffix 界 was attached to 35 words (productivity) which appeared 11,468 times (accumulative productivity) in the corpus. Thus, the log TTR for 界 becomes 0.380 (log35 divided by log11,468). The rank mean of log TTR for the 12 affixes (rank mean = 11.58, sum of the ranks = 139.00) was not significantly larger than that of the 12 suffixes (rank mean = 13.42, sum of the ranks = 161.00) \( U = 61.000, p = .551, \text{n.s.} \)\(^5\).

Spearman’s rho was computed for all the 24 affixes together. Rank-order correlation coefficients between coalescence degree and log TTR \( r_s(24) = -.839, p < .001 \) was significant. However, the other two correlations between coalescence degree and commonality \( r_s(24) = -.383, p = .064, \text{n.s.} \) and between commonality and log TTR \( r_s(24) = .064, p = .765, \text{n.s.} \) were not significant. Coalescence degree and log TTR seem to indicate similar features while commonality differs.

Simple calculations based on type and token frequency manipulations did not indicate any differences between the 12 prefixes and the 12 suffixes. Thus, a more complex mathematical concept of entropy was applied to compare them.

5.3 Entropy

The feature entropy refers to how randomly a single affix is combined to various base words. It is calculated using the following formula.

\[ H = -\sum_{j=1}^{J} p_j \log_2 p_j \]

---

\(^{5}\) A test of difference between the 12 prefixes and the 12 suffixes using Herdan’s TTR, the following asymptotic formula, should be used

\[ z = \frac{TTR_{prefix} - TTR_{suffix}}{\sqrt{\frac{1}{12^2} Var(S_{prefix}) + \frac{1}{12^2} Var(S_{suffix})}} \]

which has a standard normal distribution. For this calculation, each affix variance should be calculated by the following:

\[ Var(TTR) = \frac{N^2 \sigma^2}{V^3 \mu^2 \ln^2 N} \]

The variance of the mean for each of the 12 prefixes and the 12 suffixes is computed by the following.

\[ Var(S) = \frac{1}{12^2} \sum_{i=1}^{12} Var(TTR_i) \]

However, in the present study, we judged that Mann-Whitney U-test is good enough for computing Herdan’s TTR for the 12 prefixes and the 12 suffixes.
In the present study, the entropy of affixes was calculated according to the base to which they were attached. For example, the prefix 超 ‘super’ appeared to be attached to 40 different nouns in the newspaper corpus. The total number of times that the prefix 超 appeared with base nouns was 9,879 times, as seen with the highest frequency of 超伝導 ‘superconductivity’ counted 2,140 times, the second highest frequency of 超対立 ‘nonpartisan’ at 1,794 times, and the third highest frequency of 超大国 ‘super-power nation’ at 1,736 times. The ‘p’ in the formula stands for the relative frequency of occurrence of a specific compound word among all compounds attached to affixes. In the case of the highest frequency of 超伝導, ‘p’ is 0.217, as calculated by dividing 2,140 by 9,879. The formula \( \log_2 p_j \) for this word is evaluated as \( \log_2 0.217 = -2.207 \). Then, \( p_j \log_2 p_j \) for the 超伝導 is -0.479 (the result of 0.217 \times -2.207). The values for the remaining 39 compound nouns were also calculated in the same manner. The entropy of 超 was finally determined as 3.500 by adding all the scores of \( \log_2 p_j \) and multiplying by -1.

The variance of entropy is calculated by

\[
V(H) = \frac{1}{N} \left( \sum_{j} p_j \log^2 p_j - H^2 \right).
\]

In this formula, \( N \) is the sum of all frequencies, \( \log_2 \) is the logarithm to base 2 and \( \log^2 \) is \((\log p)^2\). The standard deviation of means is given by

\[
\sigma_{\bar{H}} = \sqrt{\frac{V(H)}{n}}.
\]

The \( t \)-test for entropy is then calculated as

\[
t = (\bar{H}_{\text{pref}} - \bar{H}_{\text{suf}}) / \sqrt{\sigma^2_{\bar{H}_{\text{pref}}} + \sigma^2_{\bar{H}_{\text{suf}}} / n}.
\]

A \( t \)-value of the above formula was 2.096. The difference in entropy between the 12 prefixes and the 12 suffixes is significant \( t(22) = 2.096, p < .05 \). Thus, entropy, referring to the degree of affix attachment disorder, reveals a significant difference between the 12 prefixes and the 12 suffixes. The mean entropy of the 12 prefixes was 3.952 while the mean entropy of the 12 suffixes was 3.791. The result of the \( t \)-test suggested that the prefixes were more randomly or chaotically attached to nouns than the suffixes. Although this finding is limited to the selected 24 affixes, the result supported the right-hand head rule (Kageyama, 1982; Kageyama, 1999; Namiki, 1982; Nishigauchi, 2004; Williams, 1981).

6. Conclusion

The present corpus study assumed that Japanese affixes would generally follow the right-hand head rule, which predicts that prefixes would be attached to a wide variety of nouns while suffixes would be regularly attached to a smaller group of nouns. Twenty-four frequently-used affixes (12 prefixes and 12 suffixes) were compared with regard to seven corpus features. A series of Mann-Whitney \( U \)-tests calculated for the first six corpus features (printed-
Right-hand head rule applied to Japanese affixes

frequency, productivity, accumulative productivity, commonality, coalescence degree and log TTR) did not reveal any differences between the 12 prefixes and the 12 suffixes. Simple frequency counts and their ratios seem not to be able to distinguish between characteristics of the prefixes and suffixes attached to nouns. However, the t-test for entropy indicates a significant difference. This result suggested that the prefixes were more randomly or chaotically attached to nouns than the suffixes. Although the present findings are limited to the selected 24 affixes, this result supported the right-hand head rule proposed by various linguists (e.g., Kageyama, 1982; Kageyama, 1999; Namiki, 1982; Nishigauchi, 2004; Williams, 1981).

References


---

6 Japanese titles and journals are transliterated in the Hepburn Style of Romanization with the same vowel repeated for long vowels.


