NOUN DISTRIBUTION IN NATURAL LANGUAGES

JUNYING LIANG AND HAITAO LIU
Zhejiang University
lhtzu@gmail.com

ABSTRACT

Previous research on word class distribution claimed that 37% of word tokens are nouns, suggesting that there might exist a certain regularity of noun proportion among human languages. To explore this possibility, we examined the proportion of noun and four other word classes within British and American English, and across seven languages in terms of different word frequency band. Results indicated that the noun proportion is evidently about or larger than 37%, and meanwhile increases with word rarity. Among frequent words, nouns increase as minor word classes decrease, whereas among rare words, the noun proportion remains a stable level.

KEYWORDS: Noun; word classes; distribution; word frequency; corpus; English.

1. Introduction

The distribution of word classes in languages has intrigued scholars in various fields. An early effort in this line proposed the lexical law (Ohno 1956), suggesting that the proportion of other word classes is all related with the noun proportion, and meanwhile, the noun proportion is also correlated with the text type. By making a step further, Hudson (1994a) compared the noun proportion in English news reports of Brown and LOB corpus, claiming that about 37% of word-tokens are nouns. Of interest, all nouns in their Chinese

1 We would like to thank the two anonymous Psicl reviewers for their insightful suggestions and comments. This research was partly supported by the National Social Sciences Foundation of China (No. 11&ZD188 and No. 10CYY002), Project of Humanities and Social Sciences, Ministry of Education (No. 09YJC740065), the Social Sciences Foundation of Zhejiang Province (No. 08CGWW006YBQ) and the Qianjiang Talent Project.
counterpart (Liu 2006: 132; Liu 2009a: 251) add up to a total of 7,872, accounting for 39.29% of all the word tokens. Despite a striking difference among languages and styles in these three corpora, the noun proportion seems quite similar. Is this distribution incidental or abiding by a certain law? As Hudson said (1994a: 338), there seems to be regularities in language of which most of us have been completely unaware—regularities which involve the statistical probability of any randomly selected word belonging to a particular word-class. At the end of another article in the same year, Hudson (1994b: 64) added, “Our descendants will know, however, that when they can explain why 47% of written English words are nouns, they will be near to The End”.

Whether it be 37% or 47%, notice that the noun proportion is steadily within this range. Again, is this noun proportion incidental or law-abiding? To uncover this mystery, it is vital to explore three issues: First, whether there does exist this regularity of noun proportion; second, whether this regularity is held true across different languages; third, if yes, whether there exists certain mechanism underlying this regularity.

Adapting the data from Hudson (1994a), we have found a distinct feature of noun distribution: the noun proportion in the New Testament Greek only reaches 32%, which is significantly smaller than 37% of the Brown-LOB. This differentiation seems to suggest that the 37% norm for writing may vary from languages.

Indeed, it is noteworthy that different languages witness variations in terms of word class. Schweers and Zhu (1991) examined the distribution of word classes of Latin, German (original and translated) and Chinese. Results show that the proportion of Latin nouns and pronouns is 36.48% and 14.93%; of the German translated texts – 25.6% and 13.3%; of the German originals – 19.2% and 15.3%; and of the Chinese – 20.16% and 12.38%. Except Chinese (with verbs taking 21.84%), the noun proportion in other three languages takes up the largest share of all word classes. Ziegler (1998, 2001) also investigated the Brazilian-Portuguese press texts with respect to their word class frequency distributions. It is shown that the noun proportion and the pronoun proportion clearly follow a regularity, with the former taking up around 20%, whereas the latter taking 10%. However, the noun+pronoun proportion is steadily within the range from 30% to 45%.

Previous findings about the distribution of word classes within a language or across different languages converge on one point, that is, in most cases, the noun proportion takes a largest share in word classes, however, inconsistency still remains. Although the data from different languages consti-
Noun distribution in natural languages

tute a much larger sample, compared with Hudson’s work, the difference of noun proportion seems fairly noticeable. To tackle these issues, we need to bear two points in mind. To begin with, the difference of noun proportion in previous research might be attributed to word frequency of tested corpora, since the distribution of word classes differ with word frequency (Bybee and Hopper 2001; Bybee 2007). Secondly, simply examining the proportion of certain word class in not significant enough, since people are more concerned with whether there exists a certain distribution regularity of this word class in a text, and whether this regularity or constancy possesses a cross-language validity. If yes, then we could effectively explain and understand the distribution characteristics and universalities of this word class.

Thus, the goal of this present study is to verify that the noun proportion within a language or across languages follows a certain regularity. To carry out this goal, three manipulations were adopted here. Firstly, to examine the possibility that noun proportion might vary with different word frequency, we examined a huge corpus by looking into several bands based on word frequency, so as to check the probable difference of noun proportion at each frequency band. Secondly, to exclude the limitation of testing only one word class, we examined five major word classes for every corpus, such that a more refined and revealing distribution pattern of word classes could be tracked. Lastly, to compare with the short-text assumption (Hammer 1990; Best 1994, 1997), we adopted large corpora with a magnitude as large as or even larger than 100 million word tokens. All things taken together, to anticipate, the noun proportion within one specific language or across different languages would be limited within the range of 30–45%; and the noun proportion would vary with different word frequency band.

In the following sections, we begin with a brief introduction to the corpora we used and the method we adopted in the current research, and then present the distribution of word classes in American English, British English and six other languages, finally and also importantly elaborate on the distribution of word classes with word frequency.

2. Materials and method

The materials in the present study are mainly based on the Corpus of Contemporary American English (COCA), the British National Corpus (BNC), and the Wacky (Web as Corpus) project. These three corpora are chosen mainly because they are all released to the public and enjoy high reliability.
Besides, the good accessibility of these three corpora makes our results easier to be examined or replicated. Within the Wacky project, we used the corpora collected from the Internet, and here seven large representative corpora were included, namely English, Russian, Chinese, French, German, Italian and Japanese, with the search interface available from Leeds University, UK. As for the measurement, we adopted the data from Davies (2012) and Kilgarriff (2012).

3. Results and analysis

In this section, we will answer the following two questions: Is there a regularity of noun proportion in English or across languages? If this regularity exists, does it vary with the word frequency? As suggested by Hudson (1994a), nouns are (slightly) more common in more formal output by older people. Why? One possible explanation is that the noun proportion in general vocabulary increases with rarity – in other words, if a word is less common or of lower frequency, then this word is more likely to be a noun. As an a priori, this seems plausible, but it would be reassuring to have statistical evidence to support our hypothesis.

3.1. Distribution of word classes in American and British English

Shown in Figure 1 is the distribution of word classes based on Mark Davies’s Corpus of Contemporary American English (COCA) from Hudson (2009). Here, the vertical axis indicates the percentage of word classes, and the horizontal axis refers to the frequency band (hereafter, FB). For example, 1 in FB indicates the most frequent 256 lemmas, and 2 refers to the most frequent 512 lemmas, and so on so forth. As seen from the graph, firstly, the noun proportion is significantly larger than the proportion of other word classes; secondly, the noun proportion keeps rising steadily before 6, then is followed with a slight decrease till 9 and afterwards remains increasing. Likewise, the proportions of verbs and adjectives have witnessed a slight increase, however, the proportion of prepositions and adverbs keeps declining consistently.

For the sake of balance, a British English corpus is also considered. Adopting the word frequency table by Kilgarriff (2012), we have obtained the distribution of word classes in Figure 2. (Altogether, 939,028 word types were obtained from the total of 100.1 million tokens in the BNC).
Noun distribution in natural languages

Figure 1. Distribution of five word classes in the Corpus of Contemporary American English.

Figure 2. Distribution of five word classes in the BNC.
The horizontal axis is the number of words contained in the frequency list, as 3k indicates the most frequent 3000 types in the frequency list; the vertical axis shows various proportions of word classes with a specified frequency. (This figure is calculated in the same way as Figure 1.)

Figure 2 shows that with the decrease of word frequency, the noun proportion steadily increases, but keeps stabilized while reaching a high point around 61%, and remains this level ever since. Meanwhile, the adjective proportion is much smaller than the noun proportion, but with a similar trend. The verb proportion reaches the highest at the most frequent 9,000 word types, and then shows a downward trend. Overall, the proportions of adverbs, prepositions, and pronouns clearly display a declining trend. Probably, it can be reasoned that nouns, adjectives, and verbs belong to the open-class words, whereas the rest belongs to the closed-class words.

3.2. Distribution of word classes across different languages

Does the high noun proportion we obtained in both American and British English have a cross-language generalization? Put differently, is the noun proportion comparable across different languages? To anticipate, if we use different language corpora with similar genres, then the noun proportion shall be virtually the same. Here, as stated earlier, seven large corpora collected from the Internet are included, namely English, Russian, Chinese, French, German, Italian and Japanese, using the search interface available from Leeds University, UK.

As displayed in Figure 3, the noun proportion of seven languages approximates between 31–34%. Here, the noun proportion means the proportion of noun-like words, which include nouns and pronouns. Further, the chi-square test revealed that there existed no significant difference of noun proportion among these languages $\chi^2(6)=3.16, p=0.788$. Since the corpora of seven languages are collected from the Internet using the same method, and importantly the size of each corpus is large enough (except that the corpus of Italian is above 10 million word tokens, the other corpora are all above 140 million), we may sum up that the high proportion of nouns has a cross-language generalization.

Given the above findings, we can safely conclude that the regularity of noun proportion not only exists within one language, but also has a cross-language generalization.
Table 1. The number of words of each word class across seven languages
(with the unit as the number of each million tokens).

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Russian</th>
<th>Chinese</th>
<th>French</th>
<th>German</th>
<th>Italian</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun</td>
<td>287244.79</td>
<td>242019.26</td>
<td>251240.88</td>
<td>226497.88</td>
<td>239781.51</td>
<td>252267.78</td>
<td>343804.25</td>
</tr>
<tr>
<td>adjective</td>
<td>71563.94</td>
<td>107876.06</td>
<td>45194.68</td>
<td>66061.88</td>
<td>91358.44</td>
<td>89657.99</td>
<td>10040.91</td>
</tr>
<tr>
<td>verb</td>
<td>140893.58</td>
<td>139866.04</td>
<td>193190.45</td>
<td>138931.62</td>
<td>125074.64</td>
<td>125136.07</td>
<td>99121.8</td>
</tr>
<tr>
<td>adverb</td>
<td>40246.76</td>
<td>22174.32</td>
<td>58509.92</td>
<td>52623.85</td>
<td>49929.88</td>
<td>50977.01</td>
<td>15003.37</td>
</tr>
<tr>
<td>preposition</td>
<td>128670.4</td>
<td>78664.11</td>
<td>30502.29</td>
<td>141741.96</td>
<td>82631.37</td>
<td>91968.64</td>
<td>208357.19</td>
</tr>
<tr>
<td>pronoun</td>
<td>42616.71</td>
<td>97897.44</td>
<td>59894.05</td>
<td>86928.26</td>
<td>84795.25</td>
<td>63714.45</td>
<td>0</td>
</tr>
<tr>
<td>noun+pronoun</td>
<td>329861.5</td>
<td>339916.7</td>
<td>311134.93</td>
<td>313426.14</td>
<td>324576.76</td>
<td>315982.23</td>
<td>343804.25</td>
</tr>
<tr>
<td>Size</td>
<td>181376006</td>
<td>156534391</td>
<td>281660631</td>
<td>185102375</td>
<td>187789449</td>
<td>12113422</td>
<td>253071774</td>
</tr>
</tbody>
</table>
3.3. Noun distribution with word frequency

Despite a general trend of distribution for word classes within English in Section 3.1 or across different languages in Section 3.2, the specific feature in different frequency band is still unclear. To elaborate, we further examined the five major word classes as percentage of the first, second and third thousand most frequent lemmas in the BNC. Our prediction is that noun proportion varied with word frequency, specifically, it would be increasing from the most frequent words to the rarest words.

Figure 4 demonstrates that for the first three thousand lemmas in the BNC, compared with a significantly lower percentage of verbs, adjectives, adverbs and propositions, the noun proportion does steadily rise from 37% to 60%. Meanwhile, resembling the general trend in Figure 2, the adjective proportion takes on a similar trend with nouns, which is in opposition to a downward trend of the verb, adverb and proposition proportions.

Meanwhile, what might occur if we extend the first 3000 most frequent lemmas to the first 6000 most frequent? Figure 5 clearly shows a comparable result in that the frequency of noun keeps rising steadily from the first 1000
Noun distribution in natural languages

Figure 4. Distribution of five word classes as percentage of the first, second and third thousand most frequent lemmas in the BNC.

Figure 5. The 6,000 most frequent lemmas in the BNC by frequency band (1,000 lemmas) and word classes.
Figure 6. The 940,000 most frequent lemmas in the BNC by each individual frequency band (1,000 types) and word classes.

lemmas to 2000, while remains quite constant till the first 6000, virtually the same as that in Figure 2. Here, the adjective takes on a similar trend with the noun.

What has been discussed in Section 3.1 is the accumulative word-class proportions in the BNC by frequency band. To further explore the individual word-class proportions within each frequency band, we analyzed the corpus with a much larger size by the same method. In Figure 6, the number on the horizontal axis refers to a specific range, e.g. 3k means the noun proportion within the range of 1000–3000 word types, whereas in Figure 4 or 5, 3k refers to a cumulative of the most frequent 3000 lemmas. The vertical axis refers to the percentage of each word class in a frequency band.

As shown in Figure 6, the noun proportion keeps increasing steadily with the decrease of word frequency and also the increase of word rarity, and reaches a highest point approximately at 300,000, then stabilizes after that. Meanwhile, the adjective proportion is much smaller, but its trend is similar to nouns. As for verbs, the proportion hits the top in the second band (3,000–6,000), and then there follows a declining trend. For other word classes, such as adverbs, prepositions, and pronouns, a general downward trend is witnessed.
In addition to the above, we were also interested in knowing the more refined and detailed distribution features in the least frequent words, also the rarest words in English. To examine the noun proportion and its change in the low-frequency words, we explored the noun distribution in the lowest frequency band, and the results are presented in Figure 7.

Here, the horizontal axis refers to different low frequency bands in the frequency table, for instance, freq100 means the 100 rarest words in the frequency list; and the vertical axis shows the percentage of each word class in the corresponding frequency band.

As displayed in Figure 7, little change occurs in the proportion of low-frequency nouns, adjectives, and verbs, with the noun proportion remaining around 57%. Combined with Figure 2, we can see that, despite a general rising trend in noun proportion, this increase is finite in that when it reaches a certain high point, the proportion will be maintained at a stable level.

To sum up, the results in this section confirm our prediction that the noun proportion increases with word rarity, and remains steady at the band with
the lowest frequency. Among common words, nouns increase as minor word classes decrease, whereas among rare words, the trend for noun proportion remains even and constant. However, this increase is not infinite, for it reaches a peak approximately around 300,000 and then remains stable after that.

3.4. Relation between type frequency and token frequency

In the current research, we have calculated the frequency of word types as well as word tokens. But probably one may ask: what is the relation between the type frequency of a word class and its token frequency? Specifically, is there a regular relationship between type and token frequencies for nouns, or indeed whether there is a more general relationship between the type and token frequency of open- and closed-class words? To answer these questions, we calculated the token and type frequency of the most commonly used 6000 lemmas in the BNC, and the results are displayed in Figure 8.

![Figure 8. Word Classes in the most common 6K lemmas (percentages of types and tokens).](image)
Here, the word “others” refers to the collection of closed-class words other than pronouns and prepositions. As can be seen from Figure 8, if closed-class words are not considered, nouns rank the highest either in the type or token frequency (52% vs. 21%), followed by verbs (20% vs. 19%), adjectives (18% vs. 6%), and adverbs (7% vs. 7%). These data are quite revealing in that, for the open-class words, probably due to their huge number, type frequency is mostly higher than token frequency, which is especially noticeable among nouns and adjectives. Conversely, the closed-class words show an exactly opposite pattern in that their type frequency is quite low and the token frequency is much higher. For instance, the type frequency of pronouns is less than 1%, whereas the token frequency reaches 7%; the type frequency of prepositions is 1.12%, while the token frequency is as high as 13.4%. Further, for “others”, the type frequency totals 1.7%, but the token frequency is high at 26.3%.

4. Discussion

The present study investigates noun distribution across different languages in terms of word frequency. The results show that the proportion of nouns follows an explicit regularity in that it increases with word rarity. Specifically, among words with high frequency, nouns increase as minor word classes decrease; whereas among least frequent words or rare words, the trend for noun proportion remains at a stable level.

In what follows, we will first discuss how the findings in the current paper expand and expound previous research of word class distribution in English and across languages, and then discuss the connection between noun proportion regularity and the noun bias debate.

4.1. Noun proportion within English and across languages

The present study explores noun proportion across different languages based on large corpora with a magnitude as large as or even larger than 100 million word tokens. Due to this huge magnitude, importantly, the current findings expand and enhance the validity of previous research on noun distribution we present here.
Our finding that noun distribution follows a regularity and increases with word rarity is echoed with the research on noun proportion in Chinese. Yin (1986) made an initial attempt to explore the distribution of word classes. The research compared the word classes on the one hand in narrative, argumentative, and expository texts, and on the other hand in written and spoken texts. The results showed that the noun proportion in all the corresponding written texts reaches approximately 30%, while that for the spoken texts is 23%. Importantly, a comparative analysis of word class distribution in ancient, medieval, early modern and modern Chinese demonstrated an important fact that, diachronically, the distribution of word classes remains a “constant”. In a language, there exist many “invariants” or “constants”, and the distribution of word classes is an important grammatical constant.

Largely in line with Yin’s research, Guo’s statistics were based on a 20,000-word corpus of Modern Chinese with written and spoken data included (Guo 2002: 275). The results similarly demonstrated that the noun proportion is 34% for the written data and 13% for the spoken data. In spite of a significant difference in noun proportion existing between Yin’s diachronic research and Guo’s synchronic research in spoken Chinese, the similar noun distribution pattern for written Chinese is clear: 30% in Yin’s research and 34% in Guo’s data.

Finally, a recent study (Zhao 2012) explored the dynamic growth patterns of words and content words in English for Specific Purposes (ESP) and general English at 4000-word intervals. The results showed that, with an increase of cumulative tokens, two different corpora exhibit a similar pattern of overall vocabulary increase, and the vocabulary increase curves for content words are also quite similar, with nouns increasing more rapidly than other content words.

The current finding is also quite consistent with Uhlířová (2000). Her results demonstrated that the total frequency of nouns, verbs and adjectives, which are open classes, makes more than three quarters of running texts in human language, and these word classes have a lower repetition rate than the grammatical (or functional) ones, because the inventory of the latter is more or less closed. When it comes to the open- vs. closed-class words, our results suggest that there exists a general relationship between the type and token frequency of a word class. That is, as for the open-class words, their type frequency is mostly higher than their token frequency, whereas for the closed-class words, their type frequency is much lower than their token frequency.
4.2. Regularity of noun proportion across languages

The present study indicates that a regularity of noun distribution exists across languages, regardless of different language features. This serves as a stride forward for research on word class distribution in natural languages by proposing a fairly definite noun proportion of 30–45%. Indeed, previous research on linguistic typology has recognized that all human tongues have such great similarities. Thom (1970) proposed that the small number of elementary sentence types corresponds to a small number of typological structures underlying events in the exterior world. Quite in line with this proposal, Bybee (2006, 2010) claims that language is one of the most systematic and complex forms of human behavior, as such the linguistic structures can be derived from the application of domain-general processes. It is the repetitive use of these processes that has an impact on the cognitive representation of language and thus on language as it is manifested overtly. Compared with the preexisting speculations and theories on language similarity, the current research offers an explicit regularity of noun proportion ubiquitous in natural languages.

To put it more specifically, the regularity of noun proportion extends both to isolating and inflectional languages. With regard to language differences, the present study explores seven large representative corpora in English, Russian, Chinese, French, German, Italian and Japanese. On the one hand, modern Chinese and modern English are typical “isolating” languages, which build sentences by rearranging immutable word-sized units and adopting functional words, as in The woman saw the boy vs. The boy saw the woman (Comrie 1989; Pinker 1994; Dryer 1997; Comrie et. al. 2005). On the other hand, languages such as Russian express who did what to whom by modifying nouns with case affixes, or by modifying the verb with affixes that agree with its role-players in number, gender, and person (Bivon 1971; Pinker 1994). However, this clear differentiation does not technically change the noun proportion between these two different types of languages, suggesting that the noun proportion shows a regularity that is shared across languages.

With regard to the regularity of word classes across languages, recent years has witnessed a number of efforts in research on relative frequencies of nouns, pronouns, and verbs. Seifart (2011) pointed out that the noun–verb

---

2 [http://www.eva.mpg.de/lingua/conference/2013_NTVR/files/program.html]. We thank one of the anonymous PSiCL reviewers for providing us with this information.
ratio correlates with the argument marking system of the language. Polinsky (2012), however, demonstrated a correlation between the noun–verb ratio and the headedness of the language (OV/VO). Recently, this correlation has been also found in Basque, Spanish, Catalan and Galician (Pastor et al. 2013). Given this evidence, it is obvious that the correlation between noun–verb ratio and linguistic typology is still controversial and calls for more research; meanwhile, it has its great theoretical significance. Admittedly, the main aim of our current research is only to verify that the noun proportion follows a regularity in natural languages, and the correlation between noun proportion and linguistic typology is beyond our current scope; still, we need to ask whether there is any regularity of noun proportion and noun–verb ratio in natural languages, and this is in need of further research.

4.3. Noun proportion regularity and the noun bias debate

So, how does this regularity of noun proportion emerge within one language or even across different languages? Popescu et al. (2007) hypothesized that the frequency distribution of words is semantically structured; and even individual frequency classes have their particular semantic spectrum, changing from class to class. Further, they examined the distribution of nouns, verbs, adjectives and adverbs, and found that the generating mechanism is the valency of the word classes.3 This is quite consistent with our findings. The appearance of a class in text leads to the appearance of another class which lies in the domain of its valency.4 Though the realization of a special class from this domain is controlled probabilistically (e.g. a noun admits an adjective but does not need to have it in all cases, but on the contrary, an adjective requires a noun; a verb presupposes a noun or a pronoun but the pronoun can be an inflection, e.g. Latin *vocamus*, etc.; this is realized differently in different languages), the associated classes follow the distribution of the main class.

Apart from taking the valency as the generating mechanism, we further argue that our findings provide solid evidence to untangle the long-held debate of noun bias (Gentner 1978, 1982; Gentner and Boroditsky 2001), that is, whether the preponderance of nouns in children’s vocabulary is language

---

3 Liu (2011) provides a quantitative analysis of English valency.
4 The passive valency of a noun well abides by the modified right-truncated Zipf-Alekseev distribution (Liu 2009b: 270).
universal or language specific. On the one hand, some developmental psy-
cholinguists believed that the noun bias is a linguistic phenomenon universal
to all children across cultures and languages. Gentner (1982) studied the ear-
ly vocabularies of children in six different languages: English, German, Ja-
apane, Kaluli, Mandarin Chinese, and Turkish, and concluded that the noun
bias in children’s vocabularies is universal, and children acquire nouns be-
fore verbs because the concepts encoded by nouns are cognitively developed
earlier than the concepts encoded by verbs (Gentner 1978, 1982; Gentner and
Boroditsky 2001). The noun bias has been confirmed in other studies of chil-
dren learning English (Goldfield 1993; Shatz 1994), Italian (Caselli et al.
1995), French (Bassano 2000; Bassano et al. 2005; Bassano et al. 1998), and
Hebrew (Dromi 1987). On the other hand, however, some other scholars
have questioned the universality of the noun bias. For example, children
learning Mandarin (Tardif 1996; Tardif et al. 1999), Korean (Gopnik and
Choi 1995; Gopnik et al. 1996) and Philipino (Lucas and Bernado 2008) use
more verbs than nouns. They argue that the predominance of nouns relative
to verbs in some languages may be due to the predominance of nouns in the
linguistic inputs that children hear from adults. Thus, the noun bias may not
be a universal feature of vocabulary development for all young children. In-
stead, it may be produced by features of parental input and speaking context
(Tardif et al. 1999).

Our findings demonstrate that both the language universal and the lan-
guage specific hypotheses may be held true. As shown in Sections 3.1 and
3.3, both in American or British English, and across seven different lan-
guages, the noun proportion is evidently about or larger than 37%. This indi-
cates that the two hypotheses are equally rational. Statistically, the significant
difference between the noun proportion and the proportion of other word
classes exists among virtually all languages, and hence this distribution pat-
tern accordingly suggests that the noun bias is language universal. Plus, as
claimed by Tardif et al. (1999), the predominance of nouns can be attributed
to the parental speech patterns or speaking context, and these kinds of factors
could also be integrated into the language factor, that is, the language specif-
ic is also language universal.

5. Conclusions

The current study has demonstrated that, consistent with Hudson’s observa-
tion, noun distribution in natural languages follows a regularity, and noun.
proportion is evidently about or larger than 37%. Also, noun proportion increases with word rarity and then remains constant after the most frequent 300,000 lemma.

The data analysis based on very large language corpora in the current study has provided an easy-to-operate approach to explore the regularity of noun proportion in natural languages, and the results have proved our hypothesis and also Hudson’s observation, that is, that noun distribution is regularity-abiding. However, it is evident that this conclusion needs to be tested on more languages and also more diversified language data. Meanwhile, the underlying mechanism of this regularity still remains a mystery. Further research is necessary to uncover this regularity with behavioral and neurological experiments.

REFERENCES


Noun distribution in natural languages


Noun distribution in natural languages


Address correspondence to:
Haitao Liu
School of International Studies
Zhejiang University
866 Yuhangtang Road
310058, Hangzhou, China.
lhtzju@gmail.com