Squibs

An Asymptotic Model for the English Hapax/Vocabulary Ratio

Fan Fengxiang*
Dalian Maritime University

In the known literature, hapax legomena in an English text or a collection of texts roughly account for about 50% of the vocabulary. This sort of constancy is baffling. The 100-million-word British National Corpus was used to study this phenomenon. The result reveals that the hapax/vocabulary ratio follows a U-shaped pattern. Initially, as the size of text increases, the hapax/vocabulary ratio decreases; however, after the text size reaches about 3,000,000 words, the hapax/vocabulary ratio starts to increase steadily. A computer simulation shows that as the text size continues to increase, the hapax/vocabulary ratio would approach 1.

1. Introduction

Words in English texts have a very peculiar distribution. On the one hand, between 50–100 top frequency words typically account for about 50% of the words in any text (Kennedy 1998); on the other, generally, about half of the words of the vocabulary of a text occur only once in the text (Baayen 1996; Kornai 2002). These lexical singletons are referred to as **hapax legomenon** (plural form: hapax legomena), hapax for short.

Hapaxes play a very important role in language studies. For example, the ratio between the number of hapaxes and vocabulary size (hereinafter referred to as *HVR*) is widely used in studies such as vocabulary growth (Tweedie and Baayen 1998), vocabulary richness and author identification (Holmes 1991), language typology (Popescu and Altmann 2008), the degree of analytism (Popescu, Mačutek, and Altmann 2009), and so on.

The high percentage of the top 50–100 words within a text is understandable, as only two of them, *the* and $a(n)$ would account for over 5% of the total word tokens of a text. However, the seemingly constant and high *HVR* of a text or collection of texts is baffling. Intuitively, as the length of a text or collection of texts increases, the *HVR* would decrease; as text length approaches infinity, all the words in the language would have occurred, and the number of hapaxes would approach zero. But the known facts so far do not seem to corroborate this intuition. For example, in Lewis Carroll's 26,505-word *Alice's Adventures in Wonderland*, 44% of the vocabulary are hapaxes (Baayen 2001); in Mark Twain's 71,370-word *The Adventures of Tom Sawyer*, the percentage is 49.8% (Manning and Schütze 2001); in the 43-million-word Merc Corpus, this percentage is

* School of Foreign Languages, Dalian Maritime University, Dalian, China.
E-mail: fanfengxiang@yahoo.com.

Submission received: 5 February 2010; accepted for publication: 2 May 2010.

56.6% (Kornai 2002). There seems to be no explanation for this strange behavior of hapax legomena in the literature.

The high *HVR* even in a mega-corpus poses problems for natural language processing; it would suggest at least sparseness of lexical, semantic, syntactic, discoursal, and pragmatic information on roughly half of the vocabulary. The following questions ensue: What are the factors behind this enigmatic distribution of *HVR*? Is it possible to substantially reduce *HVR* by increasing the size of a corpus? If so, how large should such a corpus be? These questions are the focus of this article.

In this study, the dynamic relationship among the vocabulary size V, number of hapaxes H, and text length N was examined in the 100-million-word British National Corpus (BNC). In the study, the orthographic word concept is adopted as a working definition, that is, a word (also called **word token**) is a string of contiguous alphanumeric characters with a space on either side (Kučera and Francis 1967; Biber et al. 1999). The alphanumeric character set Σ is defined as

$$\Sigma = \{a, b, c \ldots z; A, B, C \ldots Z; 0, 1, 2 \ldots 9\}$$

The word ω is defined as

$$\omega \in \Sigma^{+}$$

There are 62 characters in Σ; however, in this study, all words are case-insensitive (i.e., words such as *Language*, *LANGUAGE*, and *language* are regarded as the same). So there are actually 36 characters in Σ. Another concept used is **lemma**, which refers to the set of words having the same stem, the same major part-of-speech, and the same word-sense (Jurafsky and Martin 2000). In this study, the vocabulary of a text or a corpus is the set of different lemmas within the text or corpus.

2. Data and Analysis

To study the dynamic relationship between N, V, and H, the entire BNC was divided into equi-sized text blocks automatically by the computer. The size of each of the text blocks was initially set to 4,500 words, a size suitable for this study because in computing the growth curves of V, H, and *HVR* of a large corpus, text blocks much smaller than this (such as the average size of the text blocks of the one-million-word Brown and LOB corpora, which is about 2,000 words) would considerably increase the number of text blocks and would therefore result in a longer computing time. However, the actual size of each of the text blocks is $N \approx 4,200$ words because of the removal of textual structure codes in the text blocks such as *sn="10"*, */head*, */p*, and so on; and punctuation tags such as *cPUN*. The total number of such text blocks is 23,709. These text chunks were subsequently tokenized and lemmatized; characters that are not included in Σ were ignored, except for word-linking hyphens, which were replaced with white spaces. These processed text blocks were then recombined into 10 test sets, each having 2,371 text blocks, totaling about 10,000,000 word tokens, with the exception of the tenth set, which has 2,370 text blocks. The formation of each set was done by random sampling without replacement from the 23,709 text blocks. During the formation of each set, as the text blocks were continuously sampled and pooled one by one to form a set, the growth of V, H, and the corresponding *HVR* were computed along the way. The V, H, and *HVR* of each of the ten sets are close to the means, which are 103,588.9, 42,384.6, and 0.4091, respectively.

However, when the *HVR* curves were plotted, an interesting pattern was revealed. (See the right panel, Figure 1.) The *HVR* curves drop sharply initially, then stop drop-

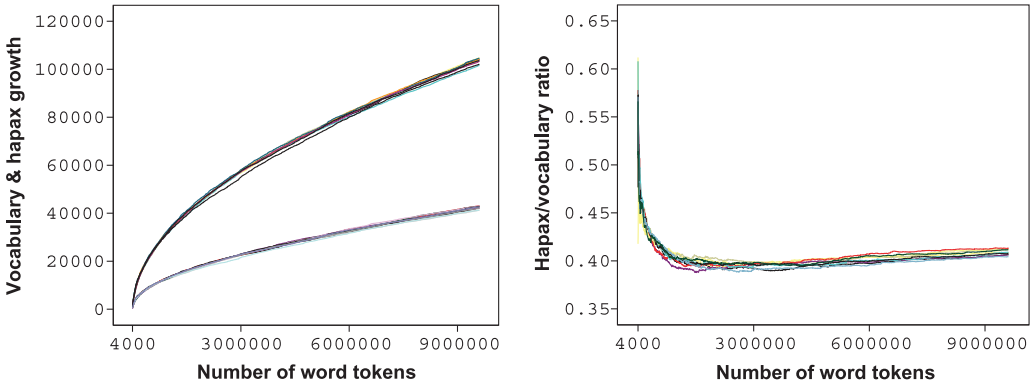

Figure 1
Growth curves of V and H (the left panel), and the HVR curves (the right panel), along with the increase in set sizes. The V growth curves are in the upper curve cluster, and the H growth curves are in the lower curve cluster.

ping at some point far from the end, and start to rise, slowly but persistently, all the way to the end. The number of word tokens at which HVR is the lowest and from which it displays a general upward trend until the end is referred to as POR (Point Of Return). Table 1 shows the initial HVRs ($N \approx 4,200$), the minimum HVRs, the final HVRs, V and H at minimum HVRs, and PORs. The minimum HVRs of the ten sets are fairly close, around 0.3928, and so are the final HVRs, around 0.4091, although the PORs have a wide dispersion, from 1,515,629 to 4,382,688, averaging 2,957,179.5. It seems to defy common sense that PORs should exist in all the ten sets, and are much smaller than the sizes of the sets.

To see what would happen to HVR and POR when N is much larger than 10,000,000, the growth of V, H, and HVR of the entire BNC were computed at an interval of 21 text blocks, about 89,000 word tokens. There are 1,129 such intervals, each formed by random sampling without replacement from the 23,709 text blocks. The purpose of reforming the 23,709 text blocks into 1,129 larger text chunks was to reduce computing time. The result is shown in Figure 2. The vocabulary size of the entire BNC is 346,578, and the number of hapaxes is 154,403. The initial HVR is 0.4583, the minimum HVR is

Table 1
The initial HVRs (HVR_i), the minimum HVRs (HVR_m), the final HVRs (HVR_f), H at HVR_m (H_{hvrm}), V at HVR_m (V_{hvrm}), and the PORs.

Set No.	HVR_i	HVR_m	HVR_f	H_{hvrm}	V_{hvrm}	POR
Set 1	0.5605	0.3896	0.4064	23,727	60,900	3,509,249
Set 2	0.4950	0.3927	0.4086	18,056	45,978	2,015,669
Set 3	0.6078	0.3964	0.4115	24,936	62,902	3,734,768
Set 4	0.5420	0.3880	0.4063	15,088	38,882	1,515,629
Set 5	0.6119	0.3962	0.4124	21,906	55,295	2,834,215
Set 6	0.5780	0.3937	0.4134	16,211	41,174	1,624,036
Set 7	0.6080	0.3930	0.4089	22,627	57,568	3,031,637
Set 8	0.5732	0.3948	0.4078	26,998	68,379	4,382,688
Set 9	0.5710	0.3885	0.4044	19,476	50,130	2,591,759
Set 10	0.5662	0.3949	0.4117	26,618	67,396	4,332,145
Average	0.5714	0.3928	0.4091	21,564.3	54,860.4	2,957,179.5

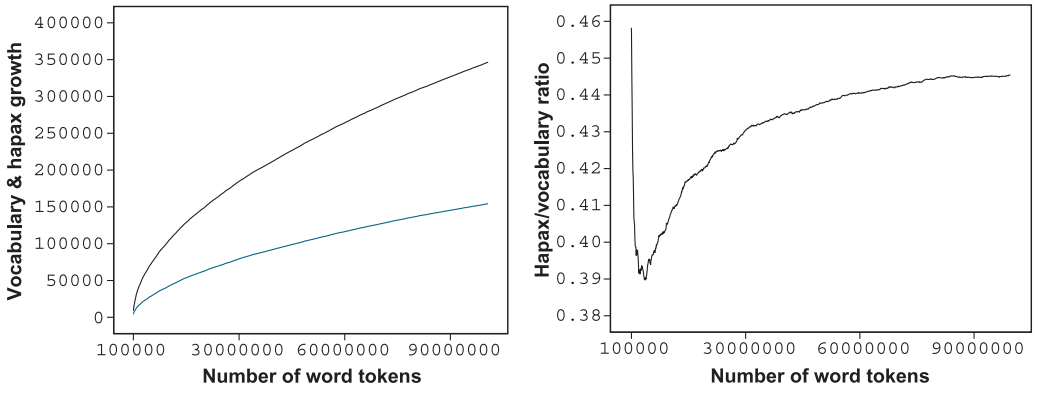

Figure 2
Growth curves of *V* and *H* (left panel), and the *HVR* curve (right panel), along with the increase in set size.

0.3899, and the final *HVR* is 0.4455. The *POR* is 3,820,340. The *HVR* curve quickly reaches *POR*, and then goes up steeply. The rise gradually slows down but still continues till the end.

To find out the general composition of the 154,403 BNC hapaxes, 1,000 of these hapaxes were randomly sampled and examined. These sampled hapaxes can be roughly divided into three major types: within-dictionary words, out-of-dictionary words, and typos. The total number of within-dictionary words is 232. They consist of 137 general words (*niffy*, *superreader*, *theonomous*, etc.); 93 techno-scientific words (*metavolcanosedimentary*, *bitstream*, etc.); and two Early-modern English words (*auncyent*, *howshold*). The total number of out-of-dictionary words is 718. They consist of 416 personal, organizational, and place and brand names of various national origins (*Khachaturyan*, *Nimmanhemin*, *Canham*, etc.); 161 alphanumeric strings (*hhfd*, *dlgs*, *90kyr*, etc.); 123 Arabic numerals (*0431*, *2562*, *817200*, etc.); 16 foreign words (*mkukumkuku*, *gevald*, *gesprach*, etc.); and two interjections (*aagghhh*, *khrrrrr*). There are 50 typos (*appication*, *disolv*, *challeneger*, etc.).

The word length distribution of the entire BNC hapaxes was also analyzed. The average length of the hapaxes is 7.74 characters, compared with the average length of 7.29 characters of the non-hapaxes. The longest hapax has 104 characters (it is the longest word in the BNC as well), and the shortest has two. The following are some examples:

1. krumfettmahanamanahulamoranosecanosemahanaritiraoosalalarama lapalapapopapopobaalacataltootoochepereelong (104 characters)
2. kahahahahahahahahahahahahehehehehehehehehehehehehehehehahahahahaha hahah (63 characters)
3. llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch (58 characters)
4. 5cggaggggccctagagggccctagagggccccccaaaaaccccccaaaaaccccccc3 (57 characters)
5. 0100110101010111010101001011010111001O (38 characters)
6. ywtfghikmccowlxpmtxwkirlmhjdb (29 characters)

7. antidisestablishmentarianism (28 characters)
8. tetramethylindocarbocyanine (27 characters)
9. whothinkstorideanangeldown (26 characters)
10. verfassungsschutzbericht (24 characters)
11. yyyyyyyyyyyyyyyyyyy (20 characters)
12. actggaagggttagtttg (18 characters)
13. yslrwqieiifkvwksl (17 characters)
14. 5ddddddddddddd6 (16 characters)
15. chippingdale (12 characters)
16. reporeted (9 characters)
17. kwaouvi (7 characters)
18. baaba (5 characters)
19. akfu (4 characters)
20. xw (2 characters)

As seen in this list, many of the hapaxes are pure random alphanumeric strings; many of the names, personal or non-personal, and foreign words appear to be random strings to the reader because of their diversified linguistic origins. Assuming the longest possible word in the English language has 104 characters, because there are 36 characters in Σ (ignoring case), the maximum possible vocabulary size V_{max} of the English language would be

$$V_{max} = \sum_{l=1}^{104} 36^l$$

V_{max} tends to infinity. The number of within-dictionary words of the English language plus those formed by compounding, derivation, and higher-probability alphanumeric strings would account for only a very tiny fraction of V_{max}. As more and more text blocks were sampled from the BNC, many of these words would have occurred more than once, but extremely low probability within-dictionary words and out-of-dictionary alphanumeric strings would accumulate. When N reached POR, the accumulation was large enough to reverse the downward trend of HVR, and the HVR curve started to go upwards. This suggests that as N approaches infinity, almost all the within-dictionary words would have appeared more than once, and the vocabulary growth would mainly be the growth of hapaxes formed by random alphanumeric strings, and the HVR curve would gradually approach its horizontal asymptote 1. That is,

$$\lim_{n\to\infty} \frac{H}{V} = 1$$

3. Computer Simulation

A computer simulation was designed and performed to test this hypothesis. In the simulation, Lewis Carroll's *Through the Looking-Glass and What Alice Found There* was

used as a miniature language source that contains all the within-dictionary-words of the miniature language. The novel has 30,566 word tokens and 2,754 word types. A simulation corpus was built by repeated sampling with replacement from the miniature language source. The size of the sample was 100, each consisting of 94 words randomly drawn from the source, with six low-probability random alphanumeric strings added. These strings were automatically generated by the computer with lengths between six and nine characters, and would roughly simulate the number of low-probability hapaxes in a natural English text of about 4,000 words, as there are about six low-probability hapaxes in a 4,200-word text chunk of the BNC. The following are some of these computer-generated random alphanumeric strings: *sygtxue*, *ungwba*, *aruvfyr9*, *layyieubk*. As the size of the simulation corpus increased, V, H, and *HVR* were computed. The simulation corpus finally consisted of 639,506 such samples, with N = 63,950,600, V = 3,838,940, H = 3,835,368, the initial *HVR* = 0.8554, *POR* = 20,300, the *HVR* at *POR* = 0.6004, and the final *HVR* = 0.9991. The result is shown in Figure 3. At N = 192,000 the *HVR* curve looks similar to that of the BNC; after the *HVR* curve reaches *POR*, it starts to go up and gradually approaches its asymptote. If we were able to build a corpus with N hundreds of times larger than the BNC, its *HVR* curve would be similar to that of the simulation corpus.

4. Conclusion

This study reveals that *HVR* is not constant at all; it follows a U-shaped pattern. In the case of the BNC, initially, as N increases, *HVR* decreases; after N reaches *POR*, *HVR* starts to increase. This *HVR* distribution is due to the reoccurrence of hapaxes consisting of within-dictionary words and high probability alphanumeric strings, and the accumulation of extremely low probability within-dictionary words and out-of-dictionary alphanumeric strings. If N approaches infinity, *HVR* would approach its horizontal asymptote 1.

The asymptotic property of *HVR* indicates that, at least for a general balanced corpus like the BNC, if it does not have a distinct *POR*, this would suggest that it has not reached the stage where hapaxes of extremely low probability within-dictionary words and out-of-dictionary alphanumeric strings substantially contribute to the vocabulary growth, and this may imply insufficient coverage of the core vocabulary of the text

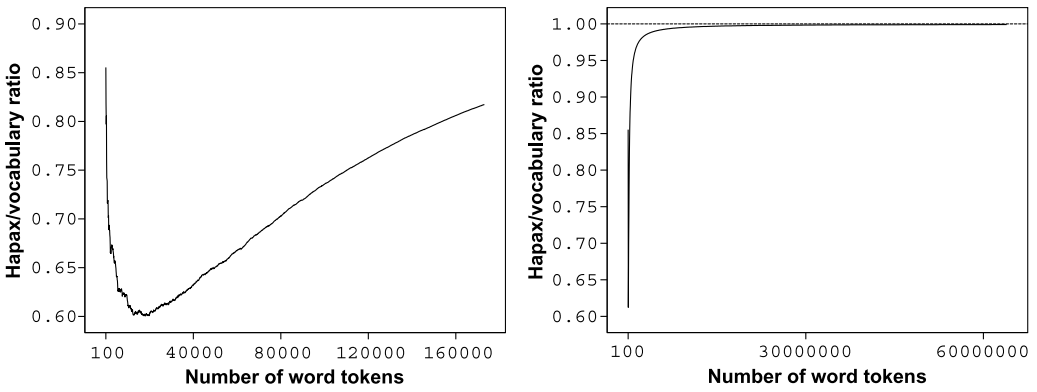

Figure 3
The *HVR* curve of the simulation corpus. Left panel: *HVR* between N = 100–192,000. Right panel: *HVR* between 100–63,950,600; the dotted line is the horizontal asymptote.

population from which it is built. To display a clear *POR*, *N* should be at least 10,000,000 for such a corpus, as shown in Section 2. On the other hand, a corpus with a size much larger than that of the BNC would have higher *HVR*; such a corpus would have more lexical noise, with the within-dictionary words far outnumbered by out-of-dictionary alphanumeric strings, leaving sparseness of lexical, semantic, syntactic, discoursal, and pragmatic information on more than half of the vocabulary of the corpus practically unsolved.

References

Baayen, Harald. 1996. The effects of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics*, 22(4):455–480.

Baayen, Harald. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education Limited, Harlow, Essex.

Holmes, David. 1991. Vocabulary richness and the prophetic voice. *Literary and Linguistic Computing*, 6(4):259–268.

Jurafsky, Daniel and James Martin. 2000. *Speech and Language Processing, an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, Upper Saddle River.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. Addison Wesley, London.

Kornai, A. 2002. How many words are there? *Glottometrics*, 4:61–86.

Kučera, H. and W. Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Manning, Christopher and Hinrich Schütze. 2001. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

Popescu, I. I. and G. Altmann. 2008. Hapax legomena and language typology. *Journal of Quantitative Linguistics*, 15(4):370–378.

Popescu, I. I., J. Mačutek, and G. Altmann. 2009. *Aspects of word frequencies*. RAM, Lüdenscheid.

Tweedie, Fiona and Harald Baayen. 1998. How variable may a constant be? Measure of lexical richness in perspective. *Journal of Quantitative Linguistics*, 32(5):323–352.