

## Re: Word count of minimum vocabulary

---

*Source:* <http://sci.tech-archive.net/Archive/sci.lang/2006-07/msg00170.html>

---

- *From:* Mok-Kong Shen <[mok-kong.shen@xxxxxxxxxxx](mailto:mok-kong.shen@xxxxxxxxxxx)>
  - *Date:* Mon, 03 Jul 2006 22:31:37 +0200
- 

Mok-Kong Shen wrote:

[snip]

On the other hand, I like to remark that in my original post in this thread my intention was rather to inquire whether it is possible to "compose" a text to express any given thought (this may eventually have to be suitably restricted to certain domains of discourse, e.g. private and business correspondence) in terms of a common (preferably standardized) vocabulary of minimum size or, equivalently, to "paraphrase" an arbitrary "given" text (with eventual restrictions as mentioned above) in such a way that all the words are to be found in such a common minimum-sized vocabulary. In that topic comparatively little has been discussed to date in this thread in my view. However, for Chinese an estimate of the vocabulary size of 1000 words has been given in Lee's post. Since Chinese ideographs don't have "derivatives", a vocabulary of size  $2^{10}=1024$  would thus serve the purpose. I conjecture that, for some special domains of discourse, it may be conceivable and practicable to employ even somewhat smaller vocabularies. On the other hand, Chinese is a rather special language (in particular, it doesn't have an alphabet and is deemed by quite many people to be rather hard to learn), even though it is one of the major natural languages of the world. English, on the other hand, is without doubt to be considered the most important natural language for a large number of application fields today and in the future. What is the size of a vocabulary (with and without counting the "derivatives") for English that corresponds in functionality to the one for Chinese of size  $2^{10}$ ? And how should one proceed to determine

## Re: Word count of minimum vocabulary

its content? I should appreciate it very much, if there could be some further discussions on such questions in this thread.

I should mention (even though possibly unnecessarily) one of the applications of the minimum-sized vocabularies. In learning a foreign language, it frequently is the case that one doesn't want to master that language perfectly (including in-depth studies of literature classics etc.) but only to be able to communicate one's own thought (in a certain domain of discourse) to the foreigners freely and perfectly. An example is a businessman going abroad to do business. It's then usually not a problem, when his discourse partner uses a word that he doesn't know, for he could normally ask for explanation. On the other hand, if he couldn't find the right words to express his own thought, then he would have a problem. Therefore, the vocabulary to be learned by him should be complete in the sense that he could with it express correctly all his thoughts in the domain of discourse of his interest. That the size of the vocabulary should be minimum is certainly trivial: the learner could then master it with the least amount of time and effort, which is certainly desirable.

There are several follow-ups to my posts in other branches of the tree of this thread. I don't have much to reply to them, excepting the following which I write below instead of separately and individually in the other branches of the tree so as to minimize the number of posts that a general reader of this thread would have to open.

Daniels:

Chinese does not use ideographs.

Herring:

If you call them "characters" it's no harder to type, and you won't get people popping up every other post to point out that the Chinese writing system isn't ideographic.

Answer:

From Collins Concise Dictionary:

ideogram or ideograph: a sign or symbol, used in a writing system such as that of China, that directly represents a concept or thing, rather than a word for it.

Herring:

Re: Word count of minimum vocabulary

Re: Word count of minimum vocabulary

Just guessing here, but I doubt if those 323 million words are all different.

Answer:

From the envelope of Collins Concise Dictionary:

The Bank of English is a computerized collection of more than 323 million words of current English created by Collins Dictionaries and the University of Birmingham to record how language is used and is changing.

M. K. Shen