

**Ring-Writer™ Development Note**  
**Re: Dictionaries**  
**Date: January 4, 2006**  
**From: Tim Scanlon, Ring-Writer Inventor**

We were able to generate an ordered list of English words quite easily. We analyzed a large body of text data taken from the Internet, and organized the words in the order of their number of occurrences in the data sample. To give an idea of the scope of the search, there were more than 7 million occurrences of the word “the” (twice as many as the runner-up, which was “a”).

Starting from this original list of words, we can create Ring-Writer dictionaries of any size that we want, simply by cutting off the list at the appropriate point. The dictionaries (beyond a reasonable minimum size) contain the same most common words, which will also be presented in the right order by Ring-Writer. Some cleaning up was necessary, to remove control strings and such, as well as some proper names arising from short-lived news articles. Also, whenever possible, words sharing the same root were grouped together in order to be proposed together by Ring-Writer. The frequency element is only of real importance for short words (five letters or less).

A dictionary of 25,000 words includes most words that one is likely to use in writing emails and SMS messages (apart from proper names of course). There is some argument for starting off with a small dictionary, even something like 3000 words, in order to reduce the memory requirements of the loaded Ring-Writer dictionary, which uses a lot of indexing in order to make alternates generation acceptably fast. With a small dictionary, the user will of course spend more time adding words, but this should not be excessively inconvenient.

Using small dictionaries will also allow us to have several dictionaries active at the same time (we did this on the PC prototype), and not to switch between re-loaded dictionaries as we are forced to do in the current version.

In Version 2, we shall offer the user another alternative. The user would start with the large dictionary, and then after a reasonable period of utilization (probably several months) instruct Ring-Writer to build a customized dictionary by simply throwing away all the words that the user never wrote.

The idea of a customized, or “stripped down”, dictionary, is particularly attractive for highly inflected languages, where verbs are conjugated, nouns have cases, and adjectives have to agree in number and gender with their nouns. Consider, for instance, French verbs.

Generating a list of French words based on frequency gives the same reasonable technical result as in English, but creates noticeably anomalous situations as far as verbs are concerned. For example, a given dictionary cut-off might include “parlez” (as in you speak), but not “parlons” (as in we speak). Even though this will may correctly reflect the relative importance of the two forms, the user will quite naturally find it odd that one is present and not the other.

One might decide to automatically add all the verb forms in the present tense, say, but that only shifts the problem to the future tense, or the imperfect, or conditional, and leads on a slippery path to the imperfect subjunctive!

Similar considerations apply for nouns in languages where nouns have cases, and adjectives (or past participles) which must agree in gender and number with their qualified nouns. While it may seem odd to have a singular form of a noun, but not the plural form, or an adjective in its masculine version, but not in the feminine, the cost of including other forms explicitly (that is, not as determined by the frequency count) is either the expansion in memory size of the dictionary, or the exclusion of certain other words in all their forms. A simple example in English, used in the User Guide, would be the inclusion of the plural “quintets”. Would this be a good thing, if it resulted, for instance, in the exclusion of the word “quartet”?

All of this means that a-priori judgments are hard to make, and so a user-customized dictionary, based initially on a fairly large word list, makes a lot of sense. Also, in this case, the user knows that if a given word form is missing, it’s probably because he or she had never used it over a fairly long period. This then is the approach that we shall take for Version 2.

In the meantime, we need to improve and extend our base dictionaries. We currently have word lists for Danish, English, French, German, Italian, Portuguese, Spanish, and Swedish. The English word list is good, and French and German are not too bad, but we have not been able to put much work into the other languages. Please send us the words that you think should be added to your dictionary, at [team@ring-writer.com](mailto:team@ring-writer.com), using as Subject the name of your language, and we’ll include your additions in the Version 2 product. If anyone wants to offer us a new language word list (free of strings of course), with the words ordered according to frequency, we shall include it in an intermediate release of the V1 Trial Product, and receive additions from other users.