Technical Reports on Language Technology
Report No 80, 2022
http://www.njas.helsinki.fi/salama

Optimal collection of words for Wordle[1]

Arvi Hurskainen
Department of Languages
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

Applications of the word game called *Wordle* are booming in the net. Several languages have applications. The original version was based on a grid with five columns and six rows. Also, there was a possibility to guess only one word per day. There are also such versions, where one can add the number of columns to up to eleven, and the game can be played without time restrictions. However, very little has been discussed the selection of words for the game. This report discusses the criteria for selecting the words.

Key Words: *word game, Wordle, morphological analysis.*

1 Introduction

Several applications of the word game *Wordle* have appeared in the net. Because the original code is in open access, it has been relatively easy to apply the game to various languages. In some English applications there are also various alternatives for playing, starting from easy words to more difficult-to-guess words. Also, the number of letters in a word can be selected between four and eleven.

These applications use the lists of base forms of the words, and inflected forms are excluded. In a morphologically very simple language, such as English, this might be a sensible solution, but the case is very different with morphologically complex languages, such as Finnish. For this reason, the Finnish version *Kieluri*[2] has included also the inflected words into the game. In Finnish, words inflect mostly with suffixes. The situation is even more complex in Bantu languages such as Swahili, which inflect to both directions. Therefore, it seems justified to include base forms and inflected forms into the word lists of the game. In this report I discuss the criteria for selecting the words and the methods for doing the selection.

---

[2] edix.fi/kieluri

2 Selection of words using text corpus

In the original *Wordle* implementation, the selection of words was done using an English language dictionary. This method produces a list of words in base form. If our aim is to include also inflected forms of words, we must use another method. The obvious solution would be to use a large text corpus, where words appear in forms appropriate in each context. The collection of words using this method would produce such word forms, which people normally are acquainted with.

But there are questions. Is the list of words obtained using the corpus method representative enough? Does it cover all the words that users consider necessary? It has base form words and inflected words - true - but is there something important missing?
I tested this when compiling word lists for the Swahili version of Wordle. I extracted from Helsinki Corpus of Swahili[3] all such words which had three, four or five alphabetic characters, turning upper case characters into lower case. As a result, there were 22,958 words in the list. It turned out that most of them were either typos or non-Swahili words. Instead of going through them word by word I resorted to another method.

3 Selection of words using morphological analyser

The morphological analyser was constructed using the finite state method, where language forms a tree, and each word-form starts from the root and branches out morpheme by morpheme into a full word-form. Using the morphological lexicon, it was possible to construct lists of such word forms that have three, four or five letters. The list includes base forms and inflected forms.

Then I compared these two lists, the one extracted from the corpus and the list constructed using the morphological analyser. The result was that the corpus list and analyser list had 4,691 common words. This number includes also those 1,378 words that my original analyser list did not contain, but which the analysis system recognised when analysing the list of the not-common words in the corpus list. The analyser list had also 4,317 additional words that were missing in corpus list. This means that about half of the grammatically correct word forms were missing in the corpus. The breakdown of these numbers is in Table 1.

Table 1. Words obtained using extraction from corpus and from morphological lexicon.

Explanation	**Number**	**Correct**	**Wrong**
Raw list from corpus	22.958		13.950
Appears in Corpus + Salama	3.313	3.313	
Appears in Salama alone	4.317	4.317	
Found from corpus with Salama	1.378	1.378	
Total		**9.008**	**13.950**

[3] http://urn.fi/urn:nbn:fi:lb-2014032624

It was expected that the number of words with three, four and five letters is different. This is shown in Table 2.

Table 2. The number of three, four and five letter words.[4]

Number of letters	**Sum**
Three	328
Four	2.510
Five	7.546
Sum total	**10.384**

We see that the number of words increases when the number of letters in the word increases. It can be concluded that the total number of words in the game is close to optimal. It does not have too large numbers of choices, yet it has enough difficult-to-guess words to keep the interest of the player. If words with more than five letters would be included, the number of words would increase tremendously, especially if also inflected word forms are included.

4 Enhancing the word lists

The tests with the selection of words described above show that from the user's perspective it has too many difficult-to-guess word forms. They are either rare, or their form, although morphologically correct, is not semantically natural. A language such as Swahili has a noun class system, and the classes have semantic bearing, although not all nouns fall neatly into a given class.

Verbs pose special problems, because the subject of the finite verb is marked by a special prefix of the particular class. Therefore, if the verb means something that only humans can do, only the subject prefixes denoting humans can be accepted, and all other alternatives must be removed.

Another problem is the transitivity of the verb. The object prefix can be accepted only to transitive verbs.

In reducing the words, I have used two main criteria. First, the word should appear in the corpus at least three times – not necessarily in the same form as the target word. Second, it should be a semantically acceptable form.

Using these criteria, the number of words was reduced considerably, as shown in Table 3.

[4] The sum total of words in this table is bigger than in Table 1, because here also the additional words found from the corpus using the morphological analyser are added.

Table 3. The number of three, four and five letter words in the enhanced lists.

Number of letters	Sum
Three	231
Four	1.458
Five	3.219
Sum total	**4.908**

5 Discussion

A word game, such as *Wordle* and its various applications to other languages, can be constructed in various ways, depending on the purpose of the game. It can be constructed for entertainment only, as was the case with the original implementation. It can also be geared to learning purposes, combined with its entertainment function. Applications such as *Kieluri*[5] for Finnish language and *Maneno*[6] for Swahili language are clearly of the latter type. Because they include also inflected word forms into the word lists, they force the player to think the language from a deeper point of view. The words in language inflect according to the grammatical rules, and they are not simply a sequence of lexical words. Also, when the inflected forms are produced using a morphological analyser, also such forms will be found that very seldom, if ever, occur in written text.

The benefits of including also inflected word forms into the word lists depend very much on the language type. Languages with very little morphology, such as English, may not benefit much by including also inflected forms into the word lists. Languages such as Finnish have complex inflection paradigms, and they certainly would benefit from including also inflected forms. Even more important is to include inflected forms to such languages, which inflect to both directions, using prefixes and suffixes. Swahili and other Bantu languages are examples of this type.

The selection of word length in word lists also raises questions. The original version *Wordle* has only words with five alphabetical characters, and they are in base form. There are also applications that allow selection between words with up to eleven letters. This might work, if the lists contain only basic forms, but certainly not, if also inflected forms in morphologically complex languages would be included. I have calculated that a single verb in Swahili produces hundreds of thousands different word forms. Therefore, the word length must be kept as low as six letters at most. Even this might be too much in part of the languages. Currently *Maneno* and *Kieluri* have lists with three, four and five letters. Both include also inflected word forms.

[5] edix.fi/kieluri
[6] edix.fi/maneno