


Language optimality for a game of Scrabble

Optymalność języka do gry w Scrabble

Maciej BOCHENEK

Uniwersytet Warszawski/ University of Warsaw

E-mail: bochenek.maciej94@gmail.com, 

Abstract: This article deals with the problem of measuring a language’s optimality for a game of Scrabble in an objective manner. Firstly, the notion of optimality in the context of the game was defined as “a language that puts maximal strain on a player’s memory”. This notion has two dimensions: a language’s vocabulary size and vocabulary diversity. To measure a language’s optimal vocabulary size, one first needs to calculate the number of all the possible combinations of the language’s letters, up to the longest playable word length, and sum up the results, which are then halved and rounded up or down to the nearest integer. To calculate a language’s vocabulary diversity, the first step is to calculate the optimal probability distribution for every letter in the language, which is equal to the alphabet size inverted. The second step is to calculate the actual probability of each letter in each position in the language’s word list. Each actual probability needs to be subtracted from the optimal, their absolute values summed up, and the result needs to be divided by the alphabet size times the length of the longest word, to extract the mean deviation. Finally, the result needs to be subtracted from 1, which yields us the vocabulary diversity score. To compare the vocabulary diversity of two languages, the smaller alphabet needs to be supplied with placeholder letters with null probability in every position, and the language with shorter words needs to be measured with additional letter positions, where all its letters have null probability, to match the other language.

Keywords: optimality, vocabulary size, vocabulary diversity, Scrabble, Scrabble Towers, Scarabeo, Super Scrabble

Introduction

Scrabble is a word-arrangement game with worldwide popularity. Therefore, not only has the game been translated into many languages, but also has seen many variants and successors to the original. When playing, one might have wondered if a given language is suitable for the game or not, or whether it would be more challenging to play in a different language. This article aims to answer the question of language suitability for a game of *Scrabble* (and certain variants) by developing mathematical methods of measuring the information toll a language takes on a player’s memory.

1. Defining optimality

In a game of *Scrabble*, one does not need to speak a given language to play; what suffices is vocabulary knowledge. Therefore, a language shall henceforth be understood as a list of words. From this point of view, *optimality* of a language for a game of *Scrabble* is tied

to the number of words one needs to memorise to ensure victory. The length of the words has an upper limit, defined by the size of the board. A standard board size is 15×15 ¹, although variations exist, such as *Scarabeo*², *Super Scrabble*³, or *Scrabble Towers*⁴. Thus, it is useful to consider a generalised board $B = d_1 \times \dots \times d_n$. But the list of words need not be only long, the words need to be diverse as well; consider the following sets, which can serve as exemplary mini-languages:

$$L_1 = \{aab, aac, abb, acc\},$$

$$L_2 = \{abc, cba, bac, acb\}.$$

Both L_1 and L_2 contain the same number of words, composed of the same letters, yet one is hopefully keen on agreeing that the words in L_2 are slightly more difficult to memorise. This is due to the fact that predictable phonotactics allow for shortcuts in memorising; L_1 can be shrunk down to a set containing to words: $\{a^2(b+c), a(b+c)^2\}$, which reads *twice "a" followed by either "b" or "c", and once "a" followed by twice "b" or twice "c"*. L_2 cannot be simplified this way. Therefore, to maximally overload a player's memory, the following must be satisfied:

1. *The language possesses the largest vocabulary to memorise;*
2. *The language possesses the richest vocabulary.*

2. Word generation

A list of words can be further generalised as a set of strings S over an alphabet A , much like in Formal Language Theory (Hopcroft et al., 2005). The sizes of the sets S and A (in other words: the number of elements in each set) are denoted as $|S|$ and $|A|$, respectively. This generalisation allows for counting the number W of words of a given length a language can generate, ignoring phonotactics. To do so, one needs to raise the number of letters in an alphabet to the desired word length. For example:

$$|A| = 3,$$

$$|S| = 2,$$

$$W = |A|^{|S|} = 3^2 = 9.$$

But to count the total number T of words a language can generate of all lengths shorter or equal to the largest dimension of a board, the calculation needs to be repeated for every permitted word length:

¹ Hasbro, Scrabble rules (<https://scrabble.hasbro.com/en-us/rules>). [Downloaded 30.05.2024].

² Wikipedia, Scrabble letter distribution (https://en.wikipedia.org/wiki/Scrabble_variants). [Downloaded 17.11.2019].

³ *ibid.*

⁴ Duncasaurus, Scrabble Towers (<https://www.duncasaurus.com/scrabble-towers>). [Downloaded 17.11.2019].

For $B = d_1 \times \dots \times d_n$,

$$T = |A|^1 + \dots + |A|^k = \sum_{i=1}^k |A|^i, \text{ where } k \text{ is the length of the longest playable word on the board } B.$$

For example, if a language has 10 letters in its alphabet, and the board is 3×3 , then $T = 10^1 + 10^2 + 10^3 = 10 + 100 + 1000 = 1110$. In other words, we can generate up to 1110 distinct, playable words.

3. The relationship between board size and vocabulary size

One would think that the more words can be generated, the better a language is for a game of *Scrabble*. However, this is the case only up to a certain point; if the number of playable words is equal to T , a player does not need to memorise any word – they can use any combination of letters. Therefore, the information strain put on a player is equal to 0. However, if any one string from a language that generates T words is considered illegal, a player must memorise it in order to avoid playing it. Similarly, if we accept one letter combination as legal into an empty language (a language with no words), this one word needs to be memorised to play. If both operations, of subtracting from T and of adding to an empty language, are continuously repeated, the eventual conclusion is that the largest strain on memory is put at a vocabulary size half of T . Consider: at the optimal vocabulary size $V_{opt} = \frac{T}{2}$, a player needs to memorise either the playable half or the unplayable one. For example:

For $|A| = 10$, $B = 3 \times 3$,

$$T = 10^1 + 10^2 + 10^3 = 10 + 100 + 1000 = 1110,$$

$$V_{opt} = \frac{T}{2} = \frac{1110}{2} = 555.$$

The closer a language using 10 letters is to generating 555 words of the length 3 or less, the closer it is to its optimality regarding vocabulary size. Conversely, if a language generated, say, 1000 words, it would be more efficient to memorise the 100 words that were not allowed. If V_{opt} is not an integer, the result can be rounded up or down to the closest one – this can be expressed as $V_{opt} = \left\lfloor \frac{T}{2} \right\rfloor$ (Graham et al., 1989). If convenient, one can include both steps (of calculating T and V_{opt}) in the single equation $V_{opt} = \left\lfloor \frac{\sum_{i=1}^k |A|^i}{2} \right\rfloor$.

4. Vocabulary diversity measured

The idea of vocabulary diversity has been so far left undefined. Let us go back to L_1 and L_2 – the issue of a more predictable phonotactics stems from probability distribution. The probability of a in L_1 was $P_a = \frac{6}{12} = 0.5$, since out of 12 letters, half of them was a . As for the other letters, their probabilities were $P_b = P_c = \frac{3}{12} = 0.25$. In L_2 , the probability of each letter was $\frac{4}{12} = 0.33$. This gives us a hint that the more equal the distribution of probabilities, the more diverse a vocabulary. From this reasoning, the optimal probability for greatest vocabulary diversity can be deduced and expressed by the following equation:

$$\Omega = \frac{1}{|A|}.$$

Let us look at an example to better grasp the idea:

For $A = \{a, b, c, d\}$,

$$|A| = 4,$$

$$\Omega = \frac{1}{4} = 0.25.$$

Now that it is known how to calculate the optimal probability of every letter in every position, we need to know how distant a language is from the optimum. To do so, firstly we need to use the following equation, akin to calculating standard deviation:

$$\sigma = \frac{|\Omega - P_{a_1}| + \dots + |\Omega - P_{a_m}| + \dots + |\Omega - P_{n_1}| + \dots + |\Omega - P_{n_m}|}{|A| \times |S_m|} = \frac{\sum_{i=1}^{|A|} (\sum_{j=1}^{|S_m|} |\Omega - P_{ij}|)}{|A| \times |S_m|}.$$

Let us break this down: $|\Omega - P_{a_1}|$ calculates the modulus of the difference between the optimal probability and the actual probability of the first letter on the first position. Then, we repeat the calculation for the first letter on every position, up to the last position m in the longest playable word, S_m . Next, we do the same calculations for every other letter, counting their actual probabilities in every position. The modula are then summed up. Lastly, the sum is divided by the number of letters in the alphabet $|A|$ times the length of the longest playable word S_m , to calculate the arithmetic mean. To better comprehend, please look at the following example:

For $L_1 = \{aab, aac, abb, acc\}$,

$$P_{a_1} = 1, P_{a_2} = 0.5, P_{a_3} = 0,$$

$$P_{b_1} = 0, P_{b_2} = 0.25, P_{b_3} = 0.5,$$

$$P_{c_1} = 0, P_{c_2} = 0.25, P_{c_3} = 0.5,$$

Therefore,

$$\Omega = 0. (3),$$

$$\sigma = 0.06(418).$$

The last step is to locate the σ on the scale $\Gamma = 1 - \sigma$. If the deviation from the optimal probability distribution is null, the language in question receives the maximal diversity of vocabulary score $\Gamma = 1$. Let us see how L_1 scores:

For $\sigma = 0.06(418)$,

$$\Gamma = 1 - \sigma = 1 - 0.06(418) = 0.93(518).$$

This last step is not necessary; one can stop the measurements at calculating the mean deviation from the optimum, and the result will show the same information, but from a different perspective – rather than showing how *diverse* a given vocabulary is, it would show how *predictable* it is.

As with V_{opt} , if one so prefers, the three steps necessary to calculate a language's vocabulary diversity (optimal probability distribution, mean deviation, and the final score) can be expressed with the single equation:

$$\Gamma = 1 - \frac{\sum_{i=1}^{|A|} (\sum_{j=1}^{|S_m|} \frac{1}{|A|} P_{ij})}{|A| \times |S_m|}.$$

Please note that the steps discussed to a language's vocabulary diversity measure it relative to the language's possibilities. In other words, the score tells us only how diverse this particular language is in relation to its own maximal possibilities. Therefore, one cannot compare the scores from two languages and state that one is more diverse than the other or not, unless their alphabets are of equal size. To compare any two languages, extra two bits of mathematics are required – one would need to supplement the smaller alphabet with *phantom letters*, that is, with placeholder letters that do not exist in it, to even out the sizes, and assign them null probability in every position. Additionally, if one language does not have words beyond some length, which the words from the other language achieve, similar phantom positions need to be taken into account, which all the letters from the first language have null probability of appearing in. For example:

For $L_1 = \{aab, aac, abb, acc\}$, $L_3 = \{xxyd, xxzd, xyyd, xzzd\}$,

Let $\Omega(L_1|L_3)$, $\sigma(L_1|L_3)$, and $\Gamma(L_1|L_3)$ be the optimal probability distribution, standard deviation, and vocabulary diversity of L_1 compared to L_3 , and

$\Omega(L_3|L_1)$, $\sigma(L_3|L_1)$, and $\Gamma(L_3|L_1)$ be the optimal probability distribution, standard deviation, and vocabulary diversity of L_3 compared to L_1 .

Therefore, $\Omega(L_1|L_3) = 0.25$, $\sigma(L_1|L_3) = 0.25$, and $\Gamma(L_1|L_3) = 0.75$, and

$\Omega(L_3|L_1) = 0.25$, $\sigma(L_3|L_1) = 0.28125$, and $\Gamma(L_3|L_1) = 0.71875$.

In conclusion, the vocabulary of L_1 is only slightly more diverse (0.75 vs. 0.71875) than that of L_3 , despite L_1 having a much higher score (0.93(518) vs. 0.75) prior to the alphabetical and positional expansions necessary for the comparison.

Summary

In search of answering the question which language is better for a game of *Scrabble*, the rules of the game allow us to reduce any language to a list of words, or combinations of letters. With this in mind, we can define *optimal* in the context of the paper as *maximally overloading a player's memory*. There are two dimensions to understanding this: (a) a language requires most words to memorise, and (b) the language's vocabulary is most diverse.

The relation between the size of a board used for a game of *Scrabble* is not straightforward; if a language permits all combinations of its letters, a player does not need to memorise any word – they can just place whatever combination they wish. Thus, to calculate the optimal vocabulary size, one needs to half the total number of

combinations, following the equation $V_{opt} = \frac{\sum_{i=1}^k |A|^i}{2}$, where V_{opt} is the optimal vocabulary size for a given board size, $|A|$ is the number of letters in a language's alphabet, and k is the longest playable word.

To calculate the diversity of a vocabulary, one needs to follow the equation $\Gamma = 1 - \frac{\sum_{i=1}^{|A|} (\sum_{j=1}^{|S_m|} \frac{1}{|A| - P_{ij}})}{|A| \times |S_m|}$, where Γ is the diversity of a given language's vocabulary, $|A|$ is the number of the letters in a language's alphabet, $|S_m|$ is the length of the longest playable words, and P_{ij} is the probability of a given letter i occurring in a given position j in a word within the language's word list. This, however, only calculates how diverse a language is given its own limitations – to compare any two languages, they must be equated regarding their alphabets' sizes and word lengths. To do so, the language with the smaller alphabet needs to be assigned placeholder letters with null probability of appearing in every position, and the languages whose longest word is shorter than the longest word of the other needs to be assigned similar supplementary positions, where each letter of its alphabet has a null probability of appearing in.

Bibliography

- Duncasaurus, Scrabble Towers. (URL <https://www.duncasaurus.com/scrabble-towers>). [Downloaded 17.11.2019].
- Graham, R.L., / D.E. Knuth/ O. Patashnik (1989), *Concrete Mathematics, second edition*, Addison-Wesley Publishing Company, Massachusetts.
- Hasbro, Scrabble rules. (URL <https://scrabble.hasbro.com/en-us/rules>). [Downloaded 30.05.2024].
- Hopcroft, J.E. /R. Motwani/ J.D. Ullman (2005), *Wprowadzenie do teorii automatów, języków i obliczeń*, Warsaw.
- Wikipedia, Scrabble letter distribution. (URL https://en.wikipedia.org/wiki/Scrabble_variants). [Downloaded 17.11.2019].