Artificial Intelligence Technologies and Applications
C. Chen (Ed.)

doi:10.3233/FAIA231295

Simulation and Analysis of Wordle Based on HSW Model

Tong QIU[1], Yu ZHONG[2]
Brunel London School, North China University of Technology, Beijing, China

Abstract Wordle is a popular puzzle currently offered daily by the New York Times. Players try to solve the puzzle by guessing a five-letter word in six tries or less, receiving feedback with every guess. Wordle continues to grow in popularity and versions of the game are now available in over 60 languages.
In this paper, the scientific computing method is used to establish a wordle decision model by combining the statistical data of authoritative English corpus and a wordle solution model based on heuristic algorithm. Based on this, we propose the HSW model to fit and analyze the real game situation of wordle players, so as to make suggestions for designers of wordle.
Based on the previous social research reports and the principle of normal distribution, we summarize and propose a VP model that can describe the vocabulary size of the population. Based on the wordle game rules, we pioneered an heuristic solution strategy with multiple controllable variables. Combined with the previous VP model, we optimized the algorithm used and built a large-scale simulation wordle game model. Through continuous parameter adjustment,we obtained a HSW model that fits the real wordle game situation. Based on the HSW model, we discuss its structural characteristics to explain the different reasons for the results of wordle games in reality. In order to describe the difficulty of a given solution word in wordle, we establish a difficulty evaluation model based on the average number of guess rounds. Finally, we compare the attributes of possible words with fixed irrelevant variables to determine the influence of specific attributes on word difficulty.

Keywords. Wordle, Per Capita Vocabulary, Psychology Simulation, Heuristic Solution.

1. Introduction

Wordle is a popular puzzle. Brooklyn-based software engineer Josh Wardle created the game in October 2021 for his partner, who enjoys word games. In November, 90 people had played the game, a number that ballooned in early January 2022 to 300,000 and has continued to rise, 14% of Americans play Wordle, according to a new Morning Consult poll [1]. As a newly emerging thing, it is worthy of further study, there are some research results about the computational complexity of the solving Wordle [2-5]. The latest study appeared in April 2023[2], which uses the ARIMA time series prediction model to predict future user number and defines the word attribute by combining the word frequency and letter frequency through entropy weight method.

[1] Corresponding Author: Tong QIU, *Brunel London School, North China University of Technology, Beijing, China; email:* ella20010305@163.com
[2] Yu ZHONG, *Brunel London School, North China University of Technology, Beijing, China; email:* zhongyu@ncut.edu.cn

In this paper, we research 5-letter Wordle in Hard Mode. MCM has generated a file of daily results for January 7, 2022 through December 31, 2022. This file includes the date, contest number, word of the day, the number of people reporting scores that day, the number of players on hard mode, and the percentage that guessed the word in one try, two tries, ... , six tries, or could not solve the puzzle. Combining sociology, psychology, linguistics, and computer science, we put forward scientific explanations and predictions for changes in a file of daily results. We First use mathematical models to explain the variability of a file of daily results give predictions for a given vocabulary on a given date, analyze the correct probability of the model's prediction result, give the uncertainty factors that affect the prediction result. Then we establish a classification of vocabulary difficulty and analyze the attributes of each difficulty level of vocabulary in the classification. Finally, we List some other interesting features of this data set.

2.Modeling

At first,we give some reasonable assumptions:

1. Players participating in wordle games conform to statistical rules. There are enough players involved in wordle games, and they are all people who's cultural level is close to the average of the whole population but slightly higher.
2. Players participating in the wordle game will win the game as a behavioral purpose.
3. Players ' vocabulary is related to English word frequency, and players have similar heuristic game thinking.

The following notations shown in table1 will be used in the paper:

Table 1 Notations

Symbol	Definition
W	Lexicon of English
N	Number of all English words
W_1	Lexicon of a virtual player
v	Vocabulary of a virtual player
d	Weight of each word
p	Possibility of a virtual player recognize a fixed word

2.1 Model 1: The Virtual Gamer

According to UPI 's findings, most U.S. adults have a vocabulary of more than 42,000 words[6]. People learn 1-2 words on average every day before middle age. Therefore, in our model, we simulate the age gap in the player population by adjusting the amount of virtual human vocabulary. According to Twitter statistics, Most Twitter users are aged between 25-34, followed by the 35-46 age bracket. The platform is used least by 13-17-year-olds[6, 7].

In addition, another important parameter of virtual players is the probability of remembering a word. An important factor affecting this probability is word frequency. We count the frequency of English words according to the UK National Corpus Word

Frequency Order in ECDICT[8]. The statistical results are shown in Figure 1.

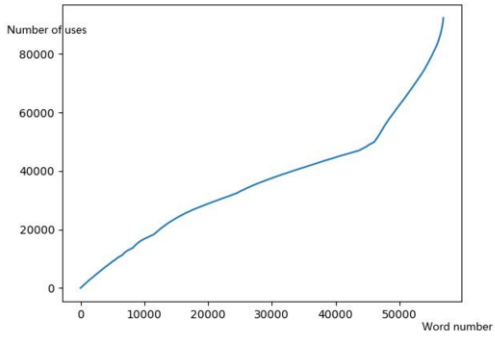

Figure 1 The relationship between a word and the number of times it is used

The curve in figure 1 has an inflection point, it is obvious that the vocabulary usage rate is extremely high after the inflection point, so we can believe that almost all players know these high-frequency vocabularies. For the low-frequ- ency words, because of the fact that players engaged in different industries will master a large number of low-frequency words related to the industry, special groups of population may learn special low-frequency vocabulary. In this model, people engaged in different occupations in society are simulated by changing the field of low-frequency vocabulary mastered by virtual people. In the establishment of virtual human, according to the research of Enghin Atalay et al., shown as Figure 2, the overall situation of a large number of special virtual humans is consistent with the American social structure[6]. In addition, fixed letter combination frequency also affect the probability of remembering a word.

Job Title		6-Digit SOC Occupations		4-Digit SOC Occupations	
Description	Count	Description	Count	Description	Count
Secretary	117.1	436012: Legal Secretary	178.5	4360: Secretary	374.4
Sales	56.0	439022: Typist	177.8	4390: Other Admin.	284.2
Assistant	53.7	414012: Sales Rep.	135.2	1320: Accountant	204.0
Accounting	50.9	132011: Accountant	128.0	4330: Financial Clerks	191.7
Clerk	49.9	412031: Retail Sales	110.2	4140: Sales Rep.	158.4
Typist	49.7	436014: Secretary	94.8	1511: Computer Sci.	133.9
Salesperson	42.2	436011: Exec. Secretary	88.6	4120: Retail Sales	131.3
Engineer	41.8	291141: Nurse	81.3	1720: Engineers	107.8
Manager	41.1	433021: Bookkeeper	80.7	2911: Nurse	107.1
Bookkeeper	40.5	411011: Sales Supervisors	70.7	4340: Record Clerks	107.0

Notes: This table lists the top ten job titles (columns 1-2), the top 10 6-digit SOC codes (columns 3-4), and the top 10 4-digit SOC codes (columns 5-6) in the *Boston Globe*, *New York Times*, and *Wall Street Journal* data. The counts are given in thousands of newspaper job ads.

Figure 2 Common occupations in US [9].

However, it is difficult to explain a variety of word attributes based on limited data when analyzing the difficulty of words in wordle games from the perspective of guessed words in games. We established the virtual player model, combined with the Heuristic solution model, and constructed the HSW model to study the decision-making mode of real individuals participating in wordle games, so as to analyze how a specific word attribute affects the difficulty of wordle games with higher interpret ability.

Let W is the set of all English words, a total of N English words, of which each English word is w_i, and its corresponding word frequency is f_i. We may as well arrange f_i from large to small to construct the sequence F_i. The process of constructing a v-vocabulary W_I is:

$$W_{I0} = \{w_i | w_i \in W\ ,\ F_i = f_i\ , i \in N^*, i \leq 5000\ \}.$$

For a certain individual I, the probability of cognitive w_i is $P_i = \left\{\frac{f_i}{\sum_{j=1}^{N} f_i}\right\}$.

Denote the random variable ξ, for a certain ξ_i , it's possible result $w_\xi \in W - W_{In}$, for each possible w_ξ ,ts probability of occurrence is P_ξ, $P_\xi$is proportional to P_i, and the sum of all possible P_ξ is 1. Hereinafter, ξ_n , which indicates the result of the random variable based on the probability, then.

$$W_{I(n+1)} = W_{In} \cup \{\xi_n\}.$$

In the end, $W_I = W_{Iv}$.

2.2 Model 2: Heuristic Solutions[10]

Now we established a feasible heuristic model. The model calculates the decision value d of each word in W_I to determine the words filled in the wordle. The decision of d for each word follows the steps:

1. The higher the word frequency of a word is, the higher the initial word decision value d will be;

2. Word contains a yellow letter, and when the letter is different from the previous position, the decision value d of the word will increase;

3. A word contains a gray letter, the decision value of the word d will be reduced;

4. In the hard mode, a word contains a green letter, and when the letter is different from the previous position, the decision value of the word is regarded as negative infinity;

5. The more unknown vowels a word contains, the higher the decision value d of the word will be.

6. The more the unknown letters contained in a word repeat each other, the smaller the decision value d of the word will be;

7. The decision of d for the previously guessed word will be regarded as negative infinity;

8. When there are words in the lexicon that conform to the existing yellow, green, and gray prompt information, the word with the largest current decision value d is selected as the current round of guessing words.

9. When any word in the lexicon cannot meet the existing yellow, green and gray prompt information, the qualified words are randomly selected from the set W.

Finally, a simulation model can be built by combining the virtual player 's lexicon and a simulated wordle game: the heuristic solution model is based on the virtual player 's lexicon, making different choices in each round, and getting the feedback information of the guessed vocabulary from the simulated wordle game, so as to make more correct decisions in the next round. The calculation process of our heuristic solution of wordle model (HSW model) can be summarized by Figure 3.

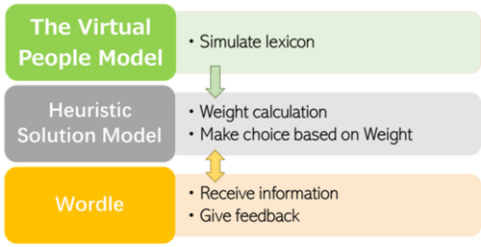

Figure 3 The HSW model

3.Model Application

By adjusting the parameters in HSW model, we obtain a simulation model close to the reality. The result shows consistency with the reality in different difficulty words, which proves the model is effective. The final adjustment values of all influence weights are shown in table 2.

Table 2 Optimal fitting weights

Variable name	Weight
Word frequency	0.40
Letter frequency	0.06
Vowel letter	0.03
Repeat letters	-0.10
Yellow hint	1.00
Grey hint	-2.00

The final fitting results show that word frequency is the most influential factor except for yellow letter cues, which indicates that the idea of ' guessing familiar words first ' is statistically used by most players. In the simulation model, when there are alternative words in the virtual lexicon, players tend to give priority to guess words with higher word frequency. Therefore, those high-frequency words in the wordle game are more likely to be selected in the first few rounds, which reduces the average number of guessed rounds of high-frequency words.

Based on the HSW model, 25,000 virtual players are created, and the weights of each variable of the fitted heuristic algorithm are calculated. The wordle using EERIE as the solution word is simulated. The results are as table 3:

Table 3 The results of "EERIE" simulation

Tries	Number	Rate
1	0	0.00
2	255	0.01
3	1749	0.07
4	2712	0.11
5	4017	0.16
6	4753	0.19
x	11514	0.46

According to the model hypothesis and the analysis, the model we proposed conforms to the game thinking of most wordle players in statistics. As long as the

group of wordle game players does not change greatly in a short time, the prediction of this model will be accurate. However, considering that MCM has published a possible wordle answer in advance, if enough people know this fact in advance, on March 1,2023, they will guess the word EERIE for the first time, which may lead to an increase in the number of first-time guesses, but will not have a significant impact on the 2-7 guesses. At the same time, considering that the MCM game has a certain topicality, it may lead to changes in the wordle player group on March 1,2023, which in turn affects its average vocabulary.

During the test, we adjusted the weight of the influence factor of part of speech. We found that within the weight range of 0.01-0.3, part of speech will not affect the final result of the test, so the model is not sensitive to part of speech. In addition, our lexicon is basically unchanged. When a low-frequency word becomes a popular term on the Internet, this uncertainty factor will also make the model unable to make correct predictions.

We use the HSW model to simulate a more comprehensive word-difficulty table to summarize and improve the accuracy of the analysis, shown in Figure 4.

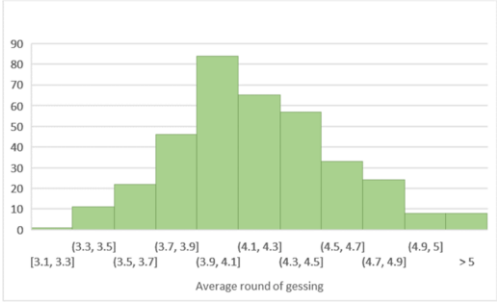

Figure 4 Statistic data of average round of Guessing

We classify wordle vocabulary as: 'hard', 'normal' and 'easy'. The specific classification is as table 4:

Table 4 The classification of solution words by difficulty

Average round of guessing	Classification
<3.5	Easy
3.5 ~ 4.5	Normal
>4.5	Hard

The proportions of the three levels words are shown in Figure 5.

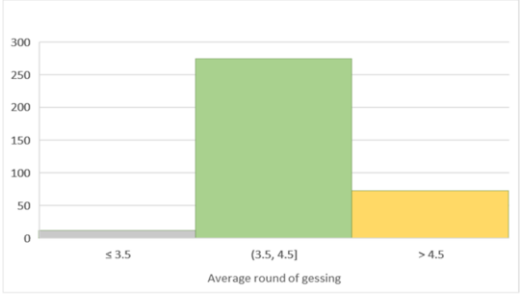

Figure 5 Statistic data of different levels of words

It can be seen that in the given wordle vocabulary, the words with normal difficulty account for the vast majority. The analysis of the HSW model suggests that the word frequency of a word, the letter frequency containing letters, and some specific letter combinations are the main factors that affect the average number of rounds a word is guessed that is, the difficulty of the word. Therefore, we analyze the influence of these factors on the difficulty of words in a statistical way.

At the same time, based on the combination of all letter frequencies of each letter in the word, the relationship with its average round of guessing is shown in the figure, and its Pearson correlation index is -0.65, shown in Figure 6, suggesting that there is a significant linear negative correlation between the two, which verifies the correctness of the HSW model from a statistical perspective.

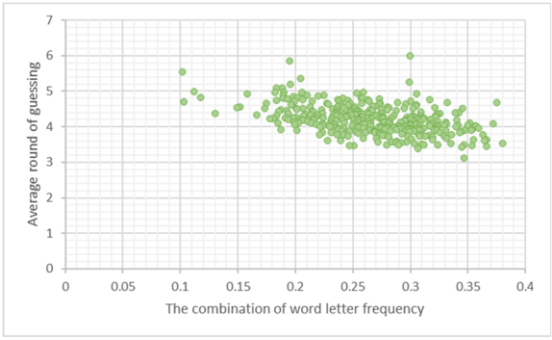

Figure 6 The relation of word letter frequency and average round of guessing

Finally, common consonant letter combinations, such as -tch, -ck, wh-, wr-, etc., are not supported by enough similar words in the given data. We use the

HSW model to simulate, and the results show that in 30 groups of words with similar word frequency, the average number of guessing rounds of words with common consonant letter combinations is always higher than that of words without any consonant letter combinations. Although there are differences between the two in statistics such as median combined average, the signific- ance of the rank sum test is 0.387, which does not indicate that words containing common consonant letter combinations are statistically less significant than words that do not contain these letter combinations, shown in Figure 7.

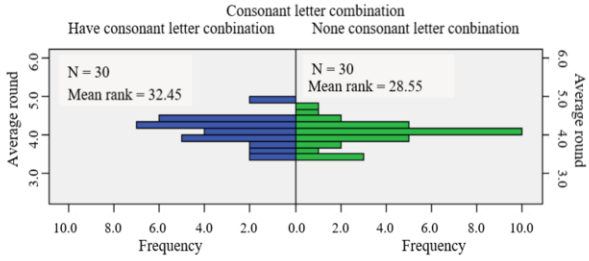

Figure 7 The result of independent-samples mann-whitney U test

In the end, for the word EERIE, because its average round of guessing is greater than 3.5, it is considered to be hard, which is mainly due to its low word frequency.

4. Additional Interesting Features

Since the week will affect the number and structure of the players, we analyse the impact of the week on the game results which are shown in table 5, and find that the average round of Friday players completing wordle is at least 3.994 times, followed by the average round of Saturday of 3.997 times. On Monday, the average round of players completing wordle was up to 4.194. We speculate that the incumbent as a group of players in the vocabulary of the more abundant groups, their game level is relatively high, they will reduce the completion of the wordle average rounds. After Friday 's work and Saturday 's work, these people are in a state of leisure and entertainment. More time spent playing wordle leads to fewer average rounds of wordle completion. On Sundays and other working days, people with high life pressure and less time to play wordle in busy work lead to higher average rounds of wordle completion. Perhaps to some extent, the average round of wordle can reflect the leisure degree of social groups.

Table 5 The relationship between week and average number of tries

Week	Average number of tries
Friday	3.994
Saturday	3.997
Thursday	4.096
Wednesday	4.122
Sunday	4.136
Tuesday	4.137

With the increase of wordle release time, the number of people playing wordle increases, but the average round of players completing wordle does not decrease. The correlation coefficient between Contest number and average round is 0.0065, shown in Figure 8, it shows that players have not found a more effective solution.

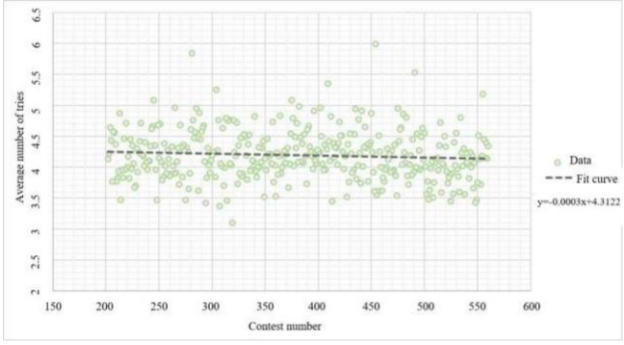

Figure 8 The relationship between contest number and Average number of tries

5. Conclusion

As we know that the difficulty of its wordle game is mainly affected by its word frequency, word letter frequency, vowel and letter combination. The weight of each

factor has been given in Table 1. In addition, although the correlation analysis shows that the influence of common letter combinations on the difficulty of a word is not statistically significant, in a five-letter word, words containing too common letter combinations, or more than three letters, are still a factor that makes solution word more difficult.

Finally, our study shows that an easy solution word usually has the following characteristics: the word frequency is extremely high or belongs to the daily life language or the words contain extremely high letter frequency, such as the letter ' t', the words contain multiple vowel letters; a normal solution word usually involve the word who's frequency is low or out of daily life or the word in which there are high-frequency letters, but the letter does not appear in its most likely position; a hard solution word usually shows that the frequency of words is very low or the words belong to professional vocabulary.

References

[1] Dellatto, Marisa (2022) Millennials Are Driving Force Behind Wordle's Success, Poll Suggests. Forbes.com. 1/20/2022, pN.PAG-N.PAG. 1p.
[2] Rosenbaum, W. (2022) Finding a Winning Strategy for Wordle is NP-complete. arXiv: 2204 .04104.
[3] Lokshtanov D. , Subercaseaux B.(2022) Wordle is NP-hard, arXiv: 2203.16173.
[4] Zhirui Min (2023) A study of the number of Wordle users and experience predictions，Academic Journal of Mathematical Sciences. 4(2): 60-65. DOI:10.25236/ AJMS. 2023. 040209.
[5] Short M. B. (2022) Winning Wordle Wisely--or How to Ruin a Fun Little Internet Game with Math. The Mathematical Intelligencer. 44(3), 227-237. DOI:10.1007/ S00283-022- 10202-0.
[6] Brysbaert M., Stevens M., Mandera P., Keuleers E. (2016) How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. [J]. Frontiers in psychology, 7:1116. DOI: 10. 3389 / fpsyg. 2016. 01116
[7] https://www.websiterating.com/research/twitter-statistics
[8] https://github.com/skywind3000/ECDICT
[9] Atalay E., Phongthiengtham P., Sotelo S., et al. (2017) The evolving US occupational structure [J]. Washington Center for Equitable Growth Working Paper, C, 12052017.
[10] Lishang J. (2005) Mathematical Modeling and Methods of Option Pricing. World Scientific Publishing Company, Singapore. DOI:10.1142/5855.