Semantic, Orthographic, and Morphological Biases in Humans' Wordle Gameplay

Gary Liang, Adam Kabbara, Cindy Liu, Ronaldo Luo, Kina Kim, Michael Guerzhoy

University of Toronto

Abstract

We show that human players' gameplay in the game of Wordle is influenced by the semantics, orthography, and morphology of the player's previous guesses. We demonstrate this influence by comparing actual human players' guesses to near-optimal guesses, showing that human players' guesses are biased to be similar to previous guesses semantically, orthographically, and morphologically.

Introduction

Wordle is a daily word-guessing game where players attempt to identify a hidden five-letter word within six attempts (Wardle 2021). Players usually attempt to minimize the number of guesses they use. Players also usually want to maintain a "streak" of having solved the game within at most 6 guesses.

We explore the difference between near-optimal play and human gameplay, which may be influenced by cognitive shortcuts and biases. In order to estimate near-optimal plays, we use the maximum-entropy heurstic. We verify that that heuristic is near-optimal.

In settings where word association is important, humans are known to be influenced by salient past information, a phenomenon known as *priming* in psychology (Schacter and Buckner 1998). We conjecture that a priming effect exists in the game of Wordle as well. Additionally, we conjecture that humans will tend to depart less from previous guesses in order to minimize cognitive load.

Since, as we show, humans' guesses tend to be close semantically to previous guesses, humans' Wordle plays can be seen as being akin to word-association games.

We review the prior work on priming in psychology, and in particular on how priming influences future word choice. We then review the optimal strategy in Wordle, as well as heuristics that approximate it. We introduce our human guess data. We then present our approach to measuring human biases in Wordle gameplay and demonstrate the systematic differences between human plays ane near-optimal play.

Background: Human Cognitive Processes

Priming is a phenomenon in psychology where past experience influences behavior without the person's explicit knowledge of the influence (Schacter and Buckner 1998). Specifically, one aspect of priming is word association. Prior works have demonstrated the grammatical class, semantic meaning and rhyme of the previous (cue) word would influence the later (response) word by humans.

Deese (1962) had done early research on word association exploring the influence of grammatical class of cue words over word association on the next word. De Deyne and Storms (2008) followed up the study and suggested that no matter whether a noun, a verb or an adjective are given as cues, the resulting association is most likely to be nouns. Furthermore, for noun cues, while still being dominant, the effect of paradigmatic association (associating with the same class i.e. noun) would decrease when changing from first to second and third response.

Steyvers and Tenenbaum (2005) demonstrate that an undirected free association network — constructed from data by Nelson (1999) that collects human participants' first responses associated with given cue words — where each word is a node and two words are connected if there exists a cue-response pair consisting of those two words — reveals that, on average, each word is connected to only 0.44% of the overall dataset. This finding underscores the sparseness of the association network where the probability of each word being the response given a cue word is not equally distributed.

Steyvers and Tenenbaum (2005) also use data collected by Miller (1995) and Fellbaum (1998), and found that the word network constructed based on semantics of words exhibits sparseness, connectedness, neighboring clustering and power-law degree distribution, which are same characteristics exhibited in the free association network, just a varying degree.

Bullinaria and Levy (2007) and McDonald and Lowe (2022) observe the connection between information regarding lexical semantics and patterns of word co-occurrence (words appearing together). De Deyne and Storms (2008) also illustrates that the basic semantic features (coded in Wu and Barsalou (2009)): "taxonomic," "entity," and "situation" are influential in terms of association responses, with "situation" being the most prominent.

Nelson, Bajo, and Canas (1987) demonstrate the effect of rhyme on memory and word association. They run an experiment where subjects would initially study (read aloud) the cue-target pair of a given rhyme; then 1.5-2 minutes after they finished studying, a meaning-related cue word and its semantic relation with the target word would be given and the participants would be required to read it aloud and recall the word they studied (Nelson, Bajo, and Canas 1987). In the experiment, cue words that rhyme with many other words would decrease the accuracy of the respondent, regardless of the meaning-related cue word (Nelson, Bajo, and Canas 1987). Through conducting a further experiment that changed all the cue-target pairs studied to be meaning-related and only half to be also rhyme-related, Nelson, Bajo, and Canas (1987) showed that the effect of rhyming appears only if the subjects actively attend to it when studying the word pairs.

Background: Wordle solving mechanisms

The objective of Wordle depends on the player — it can be maintaining the streak (i.e. try not to lose today's game), win in as few guesses as possible, or even winning the game using funny words.

However, most of the solving mechanisms designed aimed to optimize objectives regarding the number of guesses like , minimizing the average number of guesses, minimizing the number of guesses in the worst case, etc. Those mechanisms can be classified into two classes: the exact optimization approach and heuristic approaches. The best approaches based on heuristics achieve results that are only marginally inferior to exact methods.

Exact Solution for Wordle

Bertsimas and Paskov (2024) found an optimal and efficient solution for Wordle that minimizes the average number of guesses using dynamic programming. They describe the game as a Markov Decision Process and use various pruning techniques to iterate through all the possible actions given a state to keep a value function that represents the optimal average guesses given the state. When choosing an action, the algorithm chooses to transition to the state with smallest value function. Bertsimas and Paskov (2024) show that the word "SALET" is the best starting guess and the minimum average number of guesses required is 3.421. They demonstrate that under this approach the program never loses (i.e. it always completes the game within 6 guesses).

Heuristic approaches

Heuristic-based approaches to Wordle do not guarantee an optimal result but are relatively competitive. For instance, Bonthron (2022) proposes using rank-one approximations with latent semantic indexing and get the average number of guesses of 4.04, Anderson and Meyer (2022) propose using reinforcement learning, but do not report performance results regarding average number of guesses, a direct entropy-based (Shannon 1948) approach is also possible where the Doddle implementation of it has the average number of guesses of 3.432, which is close to the optimal (Liu 2022) (Cross 2022).

Doddle is an open-source Wordle solver, implemented in Python, and have designed two heuristic-based strategies: `minimax` and `entropy` (Cross 2022).

Doddle's `minimax` heuristic aims to minimize the number of guesses for the worst-case scenario with search depth of 1 (for each guess, it is only considering over all the situations after that single guess). For each guess, it iterates through all possible words in the game and chooses the one that minimizes the size of maximum partition (the amount of possible solutions after this guess given the worst case/actual answer occur) as the guess. Given the starting guess as "SALET", it is guaranteed to finish the game in 5 guesses and have the average number of guesses to be 3.482 (Cross 2022).

Doddle's entropy-based heuristic (also with depth 1) revolves around reducing the uncertainty at each step by choosing the guess that decreases (in average) the most number of potential solutions after that guess (Shannon 1948) (Cross 2022). The entropy for a move reflects the distribution of possible future game states after the guess. Let s be the current state that represents all of the previous guesses and coloring (which contains the letters not used by the target word, the letters not in the same location as the target word and the letters in the same locations as the target word). Let g_s be the list of future states given a the current state s and guess g , g_{s_i} be the i th element (state) in the list g_s and n_s be the number of potential solutions given state s, the decrease in Shannon entropy given a guess g is calculated using the formula below

$$E_g = -\Sigma_{g_s} \frac{n_{g_{s_i}}}{n_s} \log_2 \left(\frac{n_{g_{s_i}}}{n_s} \right)$$

The solver aims to choose a guess that, on average, maximizes E_g, which is the guess that decreases the most bits of entropy (where each bit of entropy decrease would cut the potential solutions by half) (Cross 2022). This provides the greatest amount of information and narrows down the possible solutions faster than `minimax`; given the starting guess as "SALET" (the optimal guess), it is also guaranteed to complete the game in 6 guesses and have the average number of guesses of 3.432 (Cross 2022) .

Data

The human guess data was sourced from Reddit. The machine-generated guesses is obtained from Doddle, an open-source Wordle solver introduced earlier. Although an ideal comparison would be with the optimal model, due to computational limitations, this study opted for Doddle as the most practical alternative. It's important to note that the performance difference between the exact dynamic programming solution and the heuristic entropy solver is minimal: the exact solution achieves a minimum average of 3.421 guesses, while the heuristic-based solver has an average of 3.482 guesses for its minimax heuristic and 3.432 guesses for its entropy-based heuristic. Hence from this point forth, the heuristic entropy solver will be referred to as *near-optimal*. Specifically, the near-optimal guesses are generated using Doddle as follows: for each human game-play, the algorithm will first gather the human first guess and generate

the near-optimal second guess; then, it'll get the first two human guesses and generate the near-optimal third accordingly, and repeat this trend until human finishes the game. It is recognized that if the algorithm just followed their own guesses (i.e. generate its near-optimal first, then second based on its first and so on), it would quickly diverge from the human guessing routes, which make the comparison between them meaningless as now both the human and near-optimal algorithm have different previous state and known information.

Data collection

The data used is based on separate dump files for the top 40,000 subreddits. Each subreddit has separate files for comments and submissions. Specifically, the data for this research project is collected from the r/Wordle subreddit, where people share their guesses online contributing to a total of 83,000 data entries (Watchful1 2023). Regex is used to identify lines in Wordle posts where users have displayed both their square results and their guesses. Regex searches for the combination of colored squares and five-letter guesses enclosed in special HTML-like tags (`gt;!WORD!lt;`), ensuring that only complete guess lines are extracted. For instance, given text:

>!STALE!<
>!SLUMS!<

Regex will:

- match the first line >!STALE!< and extract `STALE`.
- match the second line >!SLUMS!< and extract `SLUMS`.

Data cleaning is described in Appendix .

Methods

Measuring Human Biases

To quantitatively assess the influence of human cognitive biases in Wordle games, human plays are compared to their entropy-based near-optimal counterpart, where five different metrics described below are utilized in an attempt to reveal different aspects of human biases (semantic, orthographic, and morphological). For each guess in the data, the metrics below are computed through comparing that guess with the previous one (instead of comparing with all prior ones) unless otherwise stated.

Levenshtein Distance quantifies the minimum number of edits—insertions, deletions, or substitutions—needed to transform one word into another (Levenshtein 1966). This feature captures how closely a player's subsequent guesses align with their previous ones in terms of structural similarity. A smaller Levenshtein distance indicates that the player is selecting guesses that are more similar to their prior attempts, potentially reflecting a reluctance to explore novel letter combinations or a preference for minimizing cognitive effort.

Semantic distance The Word2vec distance is computed using negative cosine similarity between word embedding pairs. Words are represented as vectors using Word2Vec (Mikolov et al. 2013), a neural network model that converts words into continuous vector representations. Word embeddings such as Word2Vec being closer together in Euclidean space is more likely to the words corresponding to those embeddings being semantically related than for two embeddings far apart. This property measures semantic relatedness of guesses, allowing a quantitative measurement to the extent that players' guesses are influenced by the meanings and associations of previous guesses. A smaller Word2Vec distance indicates a tendency to rely more on semantically related words, suggesting a bias toward guessing conceptually or contextually similar words.

The GloVe distance is computed using negative cosine similarity between word embedding pairs as well. Words in this case are represented as vectors using an unsupervised learning algorithm that uses word co-occurance statistics, similarly to word2vec (Pennington, Socher, and Manning 2014).

Character-level difference measures the extent to which players deviate from their initial guesses, quantified by the number of differing characters between subsequent guesses. More character-level difference suggests a greater willingness to explore alternative solutions. Conversely, minimal deviation indicates an over-reliance on early guesses.

Shared Tokens (Syllables) This metric measures the frequency with which players reuse previously employed subwords in their guesses. These subwords are identified using the `nltk SyllableTokenizer`, which operates by breaking words down into their constituent syllables based on phonetic patterns. `SyllableTokenizer` syllabifies words based in the Sonority Sequencing Principle (SSP) (Selkirk 1984), a language-agnostic algorithm proposed by Otto Jepersen in 1904. The SSP determines syllable breaks based on the sonorous quality of phonemes, which is influenced by the openness of the lips during articulation. The `SyllableTokenizer` begins by assigning a sonority value to each phoneme according to a predefined hierarchy. By default this hierarchy categorizes English phonemes into vowels, nasals, fricatives, and stops, with vowels receiving the highest sonority values (Hench and Estes 2024). When a word is tokenized, the algorithm analyzes the sequence of phonemes, identifying potential syllable breaks based on their sonority levels (Hench and Estes 2024). This way, `SyllableTokenizer` breaks the 5-letter Wordle words into smaller patterns that symbolize patterns within words. This feature is viewed as an indicator of cognitive resource expenditure where a higher rate of shared tokens may reflect a cognitive bias towards familiar patterns rather than exploring novel word combinations.

Rhyme To determine whether two words rhyme or not, their phonic transcription was used. This was achieved with the help of the `pronouncing` library, which provides a phonetic transcription based in the CMU Pronouncing Dictionary (CMU 2015) . Two words are considered to have a

perfect rhyme if they have matching phonetic endings which include stressed vowels (Per 2019). We assess if the guess rhymes with the previous one. If the guess rhymes with the previous one, it may suggest a bias toward the phonologically related words.

Cohen's d   Cohen's d is a measure of effect size that quantifies the standardized difference between two means, in this case, human and model performance (Sullivan and Feinn 2012). Cohen's d transforms the absolute difference between means into standard deviation units, enabling a direct comparison of the magnitude of this difference across various metrics. Effect sizes are traditionally classified as small ($d = 0.2$), medium ($d = 0.5$), and large ($d \geq 0.8$) (Carson 2012). A small effect indicates only a subtle difference between the groups, whereas a large effect suggests a more substantial divergence that is visually apparent to even casual observation.

Cohen's d also provides an intuitive interpretation in terms of percentile overlap between two distributions. For example, a d of 0 indicates complete overlap, with no discernible difference between groups, while a d of 0.8 means that the mean of one group corresponds to the 79th percentile of the other group, with an overlap of just 53%.

Experiments

We compare how human guesses/plays differ systematically from near-optimal play. We obtain distributions of human plays and near-optimal plays, and compare them. We assess the effect size using Cohen's d, and we computed the p-values based on the t-statistics for the difference between the two distributions.

We analyze separately games starting from different positions. We use the notation $c_g g c_y y c_b b$, where (c_g, c_y, c_b), where the number of "green" guesses (correct letter in the correct place) is denoted with c_g, the number of "yellow" guesses (correct letter in the incorrect place) is denoted with c_y, and the number of letter guesses that are incorrect is c_b.

Below, we present some observations on the results of our comparison of human play with near-optimal play.

Full results are presented in the Appendix.

Levenshtein Distance   Studying the Entropy-based play, in states such as *1g0y4b*, *2g0y3b*, and *1g1y3b*, large average Cohen's d values can be observed with -1.652, -1.135, -1.178, respectively. Large cohen's d indicates significant differences between human and model play. These results suggest that humans are more likely to guess words that are structurally closer (smaller Levenshtein distance) to their previous attempts compared to the model. This behavior could stem from cognitive biases, such as a reluctance to make drastic changes after receiving certain types of feedback. When a few letters have been confirmed as correct, humans tend to stick closely to their previous guesses. For example, after receiving feedback confirming just one correct letter, players could often become more conservative in their subsequent guesses, whereas the model continues to explore more broadly.

In contrast, states such as *0g1y4b* and *0g0y5b* exhibit smaller Cohen's d values, suggesting less of a gap between human and model behavior. These states provide minimal confirmation (with no or very few correct letters), encouraging players to make more exploratory guesses, which aligns their behavior more closely with the model.

Interestingly, states like *3g2y0b*, *1g4y0b*, *2g3y0b*, and *1g3y1b* also show smaller Cohen's d values, even though these states offer much more feedback (many correct letters). In these cases, the solution space is significantly constrained, leading both the human and model to adopt more conservative strategies, as there is little left to explore. Here, human cognitive biases toward conservative guessing persist, but now this conservatism aligns with the model's behavior due to the narrowed solution space.

Additionally, extremely low p-values and high t-statistics are demonstrated across most states. This shows that the differences in Levenshtein distances between human guesses and the model's guesses are statistically significant.

In summary, when minimal confirmation is provided (such as states with 0 correct letters), human guesses more closely resemble those of the model, as reflected by smaller effect sizes. As more partial confirmation is given (a few correct or almost-correct letters), humans exhibit a stronger bias toward sticking with familiar patterns, diverging from the model's exploratory strategy. Finally, when a lot of feedback is provided, both the human and model behave conservatively, leading to more similar guess patterns.

A similar pattern emerges when evaluating minimax-based play. States such as *1g0y4b*, *2g0y3b*, and *1g1y3b* again show large Cohen's d values (-1.39, -1.06, and -1.03, respectively), confirming that humans tend to stick closely to previous guesses when they have some feedback, while the model continues to explore more freely. The divergence between human and model play in these states is consistent with the cognitive biases observed in entropy-based play: after receiving partial confirmation, humans become more conservative, while the model remains exploratory.

On the flip side, states like *0g1y4b* and *1g4y0b* show smaller Cohen's d values (0.0039 and 0.137, respectively), reflecting a greater alignment between human and model guesses. In these cases, where minimal or complete feedback is available, humans behave more like the model, either because they are exploring due to lack of confirmation or because both human and model strategies converge in the face of constrained possibilities.

This consistency between the entropy-based and minimax-based evaluations highlights the persistent nature of human cognitive biases. In both models, humans tend to adhere to structural similarities between their guesses, particularly when feedback is only partial, causing their behavior to deviate from the model's more exploratory tendencies.

Semantic Distance   The Cohen's d values for Word2Vec semantic distances of the Entropy play hover around zero, indicating very small effect sizes across most states. This suggests that humans and the model exhibit similar patterns when selecting semantically related words during gameplay. States like *3g2y0b*, *0g5y0b*, and *3g1y1b* show minimal differences in behavior between humans and the model.

Figure 1: Word2Vec Distance Histogram for State 3g2y0b - three completely right letters, and two letters are correct but misplaced. The small Cohen's distance and p-value indicate that the semantic choices of humans and the model are quite similar in this state. Both rely on semantic context when substantial feedback is provided.

Figure 2: GloVe Distance Histogram for State 0g0y5b where no correct letters are found. The larger Cohen's distance value highlights a greater divergence between human and model behaviors, with humans leaning more towards semantically related words compared to the more exploratory nature of the model's guesses.

In states with minimal feedback (such as *0g0y5b*, where none of the guessed letters are correct), relatively higher Cohen's d values can be observed. This may be due to the model exploring more randomly, while human players are guided by cognitive biases, leading to less random guesses. As a result, human and model behavior differ more in these states. Conversely, when players receive more informative feedback (e.g., when some letters are correct but not in the right positions, such as in states like *3g0y2b* or *3g1y1b*), Cohen's d values decrease. This indicates that human guesses become more semantically aligned with the model's, likely because with some letters fixed, there are fewer possible solutions. Although Cohen's d values remain small overall, these patterns reveal important trends in how feedback influences the alignment between human and model behavior. Additionally, the low p-values in many states emphasize that even small deviations in guessing patterns are statistically detectable, despite the overall alignment between human and model play.

When semantic distance for Entropy play is computed using GloVe embeddings, the overall trends remain similar, but the results are more pronounced. For instance, the Cohen's d values for states with little feedback (i.e., more black tiles indicating incorrect guesses) are larger compared to Word2Vec. For example, the semantic distance for state *0g0y5b* with GloVe has a Cohen's d of -0.437, significantly greater than the -0.099 observed for Word2Vec. This indicates that GloVe is better at capturing the differences between human and model play in these uncertain states, making the trend more apparent than in the Word2Vec analysis.

A similar behavior is observed for the minimax strategy. Semantic distances analyzed with Word2Vec embeddings show Cohen's d values close to zero, indicating small ef-

fect sizes. However, the same trend holds: in high-feedback states such as *3g2y0b*, Cohen's d is relatively low at 0.002, while in low-feedback states like *0g2y3b* or *0g0y5b*, Cohen's d values are relatively higher, at -0.059 and -0.099, respectively. This again suggests that human players tend to deviate more from the machine's strategy when less feedback is provided to guide their guesses.

The low t-statistics associated with many states further highlight that while the effect sizes are small, the differences between human and model behavior in terms of semantic distance are consistently detectable. This suggests that although humans and the minimax model display similar semantic strategies in word guessing, subtle yet significant deviations remain, particularly in situations where human biases lead to less optimal word selection.

Repeating the same analysis for Minimax strategy with GloVe embeddings yields results similar to those observed with Word2Vec embeddings. The general trends remain consistent, reinforcing the notion that GloVe better captures the semantic distinctions between human and model play in states with minimal feedback.

Overall, the semantic distance feature demonstrates a general alignment between human and model strategies, with subtle deviations in human responses to limited feedback, possibly hinting at cognitive processes like semantic associations that differ slightly from the model's approach. However, compared to other properties studied, such as Levenshtein distance, the Cohen's d values for semantic distance are much smaller.

To extend the analysis, the semantic distances were calculated between each human guess and its previous guess, between each near-optimal guess and its previous guess, and pairwise semantic distances between all possible solution

candidates in Wordle. The semantic distance for all possible solution candidates was studied to establish a benchmark that represents the natural semantic relationships within the vocabulary.

The mean semantic distance for the candidate words is 0.971, which is similar to both human (0.958) and near-optimal (0.974) distances. This suggests that neither humans nor the model deviate significantly from the inherent semantic structure of the vocabulary. However, the slight differences in means highlight subtle distinctions in behavior. Human guesses tend to have closer semantic connections, possibly reflecting a reliance on familiar word patterns. In contrast, the machine's slightly higher semantic distance may indicate a broader exploration strategy, prioritizing uncertainty reduction over strict semantic proximity. The results are shown in Fig. 3

Figure 3: Comparison of Semantic Distances Using GloVe Embeddings. Human and near-optimal guesses are compared to the previous guesses. Candidate words are compared to all $2,309$ Wordle candidate. Human guesses tend to be more similar to the previous guess than Optimal guesses are. Candidate words are about as far away from each other on average as consecutive Optimal guesses, with more optimal guesses at the far end of the distribution, indicating the need to sometimes completely change the guess's letters, thereby making its semantics slightly further away on average.

It's important to note that the overall distribution of semantic distances is narrow across all three categories (human, near-optimal, and candidate words), as shown in Figure 3. This suggests that most guesses, whether made by humans or machines, fall within a similar range of semantic closeness. This indicates that Wordle's guessing process is constrained by the natural structure of the vocabulary, regardless of whether it's driven by human intuition or machine strategy.

In summary, while humans show a slight tendency to rely on semantic relationships between words during gameplay, this does not result in significant divergence from near-

optimal gameplay. Both human and model guesses remain closely aligned, constrained by the natural semantic relationships inherent in the vocabulary.

Shared Tokens (Syllables) The Cohen's d values for entropy play across different states are mostly close to zero, suggesting that humans and the model behave similarly when it comes to reusing tokens in their guesses. The small effect sizes imply that this feature is less influenced by cognitive biases compared to Levenshtein distance.

However, as shown in Figure 14, certain states like *3g0y2b* and *2g0y3b* exhibit relatively larger effect sizes, with Cohen's d values above 0.07. In these states, which confirm the presence of some correct letters although not necessarily in correct position, human players tend to behave somewhat differently from the model, reusing syllables more frequently. In contrast, states such as *0g1y4b* and *0g0y5b* which provide less confirmation, appears to prompt similar exploratory behavior in both humans and the model, thereby reducing the influence of cognitive biases.

A similar behavior is observed for the minimax strategy. The Cohen's d values are close to zero, indicating small effect sizes. However, the same trend holds: in partial confirmation states such as *3g0y2b* and *2g0y3b* , Cohen's d is relatively high at 0.070 and 0.084 respectively, while in little confirmation states like *0g1y4b* or *0g0y5b*, Cohen's d values are relatively smaller, at 0.015 and 0.017, respectively. This again suggests that human players tend to deviate more from the machine's strategy when partial confirmation are provided to aid the guessing process.

This behavior aligns with observations from Levenshtein distance, reflecting a common human tendency to hold onto partial information (e.g., some correct letters although that may not be in correct location) and reuse similar syllables in subsequent guesses. Overall, token reuse shows relatively small effect sizes across most states, indicating that human and model play are largely similar in this regard.

The reason that the common syllables metric are having less Cohen's d values is suspect to be that syllables are a bigger unit than both shared characters and Levenshein distance (the maximum number of syllables change is 2 while for shared characters and Levenshtein it's 5), so the number of changes are greatly suppressed, thus making the two appearing similar.

Shared Characters The Cohen's d values for entropy play reveal notable differences between human and model behavior at the character level. For instance, states like *1g0y4b*, *2g0y3b*, *0g1y4b*, and *1g1y3b* exhibit high Cohen's d values (0.85 and above), indicating a significant divergence between human and model strategies. This suggests that humans are more likely to retain specific characters (whether correct or incorrect) in their guesses after receiving feedback about character matches. The fact that these states confirm the presence of one or two correct characters (either green or yellow) prompts humans to make guesses that are more character-consistent than the model's guesses, potentially due to cognitive biases like fixation on previously confirmed partial information.

For states such as *2g2y1b*, *4g0y1b* and *1g3y1b*, where

Figure 4: Shared chars histogram for 1g0y4b — Humans tend to retain more characters from their previous guesses, unlike the model which exhibits a more exploratory approach.

moderate Cohen's d values (between 0.2 and 0.5) are observed, we see that humans still deviate somewhat from the model's behavior but to a lesser extent. These situations provide too much confirmation where it would make both human and model act conservatively due to constrained solution space as described above.

States like *0g0y5b* show smaller effect sizes (between 0.1 and 0.2), suggesting that humans and the model behave more similarly when feedback has no confirmation. This is because humans and the model may both engage in more exploratory or random guessing behavior, reducing the impact of cognitive biases.

Finally, for states like *3g2y0b* and *0g5y0b*, which show Cohen's d values close to zero or slightly negative, humans and the model are virtually indistinguishable in their play. This makes logical sense as in those states all the correct letters are given and there's barely any reason to explore other letters, which would make both human and model to act similarly (i.e. mainly keeping all the letters known).

The extremely small p-values (almost all close to zero) and large t-statistics further validate the statistical significance of these differences, confirming that the observed deviations between human and model performance at the character level are not due to chance.

A similar trend is observed for the minimax strategy even though the results are less pronounced. In states where one to two correct letters are given such as *0g1y4b* and *2g0y3b* , Cohen's d is relatively high at 0.669 and 0.964 respectively (which is less than the 0.940 and 1.033 in entropy play), in states where little confirmation are given like *0g0y5b*, Cohen's d values are relatively smaller, at 0.167 (which greater than the 0.157 in entropy play). In states where all correct letters are given like *3g2y0b*, the Cohen'd value is close to zero (-0.012 and is the same with entropy play). This again suggests that human players tend to deviate more from the

machine's strategy when partial confirmation are provided to aid the guessing process.

To extend the analysis, the shared chars between all the human guesses and that between all the near-optimal guesses per game are generated (Figure 15). The result shows that similar amount of human and near-optimal guesses share 5 characters, which makes logical sense as those would generally only occur when the previous guess already has confirmed all the five correct letters.

It also shows that near-optimal guesses have much more cases where 0 or 1 tokens are shared yet human guesses have more cases when 2 or 3 or 4 tokens are shared. This is congruent with our above observation that when there's partial confirmation (i.e. some letters being correct), human tends to act more conservative thus having more shared tokens yet the model would be more explorative thus having less tokens shared.

Conclusion

Human Wordle gameplay is influenced by semantic, orthographic, morphological biases, which lead to deviations from near-optimal play. While entropy-based strategies, such as those implemented by Doddle, follow a more rational, information-maximizing approach, human players tend to rely on familiar word patterns, semantic associations, and partial confirmations from previous guesses. These tendencies result in suboptimal strategies that often prioritize ease and familiarity over strategic exploration.

Comparing human guesses to near-optimal model guesses, our experiments demonstrate several key patterns. Humans are more likely to guess words that are phonologically similar to their previous attempts, as reflected by smaller Levenshtein distances and a greater reuse of syllables. They also exhibit biases toward semantically related words, especially in states where minimal feedback is provided. Phonological biases, such as favoring words that rhyme, also emerge in the data, albeit to a lesser extent.

Using a large dataset of over 65,000 human games, we demonstrate that priming effects, including semantic and phonological associations, are in effect. Using metrics like Levenshtein distance, GloVe, Word2Vec similarity, and shared syllables, we quantified these biases and showed statistically significant differences between human and model guesses through measures such as Cohen's d and t-tests. The findings suggest that, in the face of uncertainty, humans tend to stick with partially confirmed patterns rather than exploring broader possibilities.

References

2015. The CMU Pronouncing Dictionary. http://www. speech.cs.cmu.edu/cgi-bin/cmudict. Accessed: October 13, 2024.

2019. Exact Rhyme. https://literarydevices.net/exact-rhyme/. Literary Devices.

Anderson, B. J.; and Meyer, J. G. 2022. Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning. *arXiv preprint arXiv:2202.00557.*

Bertsimas, D.; and Paskov, A. 2024. An Exact Solution to Wordle. *Operations Research*.

Bonthron, M. 2022. Rank one approximation as a strategy for Wordle. *arXiv preprint arXiv:2204.06324*.

Bullinaria, J. A.; and Levy, J. P. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39: 510–526.

Carson, C. 2012. The effective use of effect size indices in institutional research. In *31st Annual Conference of the North East Association for Institutional Research*, volume 41, 41–48.

Cross, A. 2022. Doddle. *https://github.com/CatchemAL/Doddle*.

De Deyne, S.; and Storms, G. 2008. Word associations: Network and semantic properties. *Behavior research methods*, 40(1): 213–231.

Deese, J. 1962. Form class and the determinants of association. *Journal of verbal learning and verbal behavior*, 1(2): 79–84.

Fellbaum, C. 1998. WordNet: An electronic lexical database. *MIT Press*, 2: 678–686.

Hench, C.; and Estes, A. 2024. nltk.tokenize.sonority_sequencing.

Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*.

Liu, C.-L. 2022. Using wordle for learning to design and compare strategies. In *2022 IEEE Conference on Games (CoG)*, 465–472. IEEE.

McDonald, S.; and Lowe, W. 2022. Modelling functional priming and the associative boost. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 675–680. Routledge.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Miller, G. 1995. WordNet: An on-line lexical database [Special issue]. *International Journal of Lexicography*, 3(4).

Nelson, D. 1999. The University of South Florida word association norms. *http://w3. usf. edu/FreeAssociation*.

Nelson, D. L.; Bajo, M. T.; and Canas, J. J. 1987. Prior knowledge and memory: the episodic encoding of implicitly activated associates and rhymes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1): 54.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Schacter, D. L.; and Buckner, R. L. 1998. Priming and the brain. *Neuron*, 20(2): 185–195.

Selkirk, E. 1984. On the major class features and syllable theory. *Language Sound Structure: Studies in Phonology/MIT Press*, 107136.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Steyvers, M.; and Tenenbaum, J. B. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1): 41–78.

Sullivan, G. M.; and Feinn, R. 2012. Using effect size—or why the P value is not enough. *Journal of graduate medical education*, 4(3): 279–282.

Wardle, J. 2021. Wordle. Web. Online game.

Watchful1. 2023. Subreddit comments/submissions 2005-06 to 2023-12.

Wu, L.-l.; and Barsalou, L. W. 2009. Perceptual simulation in conceptual combination: Evidence from property generation. *Acta psychologica*, 132(2): 173–189.

Appendix: Sample Data

This appendix contains a series of visualizations that support the main findings of this study by illustrating key differences between human gameplay and near-optimal model guesses in Wordle. The figures present histograms for several metrics used in our analysis, including Levenshtein distance, semantic distance (Word2Vec and GloVe), shared syllables, shared characters, and rhyme occurrence. Each figure compares human guesses against the optimal model guesses under various game states, such as those with partial or no feedback (e.g., 0g0y5b, 2g0y3b). Metrics such as Cohen's d and p-values are provided to indicate the magnitude and statistical significance of the observed differences.

Ultimately, these visualizations highlight the extent to which human players rely on structural and semantic similarities in their guesses, favoring familiarity over exploration, particularly when faced with partial confirmation of correct letters.

The full data we used is in a further Appendix.

Figure 5: Levenshtein distance histogram for 2g0y3b

Figure 6: Levenshtein distance histogram for 0g0y5b

Figure 7: Levenshtein distance histogram for 2g3y0b

Figure 8: Word2vec distance histogram for 0g0y5b

Figure 9: Glove distance histogram for 2g3y0b

Figure 10: Common syllables histogram for 0g0y5b

Figure 11: Shared chars histogram for 2g2y1b

Figure 12: Shared chars histogram for 0g0y5b

Figure 13: Shared chars histogram for 0g5y0b

Figure 14: Common syllables histogram for 3g0y2b

Figure 15: Total shared chars histogram

Figure 16: Proportion of rhyming guesses

Data cleaning

For each guess, the unnecessary parts such as the special symbols (`>!`, `&!lt;`) are removed. To ensure the integrity of the data provided by Reddit users, a cross-referencing process was conducted between the dataset and a Wordle answers database. This approach verified the accuracy of the Wordle IDs submitted and ensured that the answers had not been altered. If a Wordle ID was not provided, the corresponding game was considered illegible, as there was no way to confirm the authenticity of the data. In cases where a Wordle ID was provided without an answer, the last guess was cross-referenced with the Wordle answers dataset. If the last guess matched the correct answer, it was recorded; otherwise, the entry was removed. To maintain consistency, all guesses were converted to lowercase. The data cleaning process eliminated entries where users did not include their guesses or submitted answers for non-Wordle games. Additionally, any unsolved Wordle games were removed from the dataset. As a result of these cleaning efforts, the dataset was reduced from 83,000 entries to a more manageable 65,000 entries. Ultimately, information about the player, words guessed, and the number of guesses each user made are obtained.