W(ha/il)t's the (wo/ba)rd?

A cross-sectional analysis of English word pairs that satisfy the properties of the Split Decisions word puzzle

Oliver Daniel

April 6, 2022

Abstract

From three English wordlist corpora, pairs of words are generated which might appear in the Split Decisions word puzzle as published by the New York Times, i.e., at least 5 characters in length and differing by exactly two consecutive characters. The corpora are contrasted among themselves, and then the lexical and phonological properties of those words which form pairs – and those which don't – are analyzed for causative insights. Such insights will be alluded to here once discovered.

Introduction

A few times a year, the New York Times games department releases a new issue of Split Decisions, a crossword-like word puzzle. Similar to crosswords, the game board consists of boxes arranged in sequence, one for each letter of a word running either left-to-right or top-to-bottom, intersecting with other words at right angles. However, instead of cryptic descriptions of correct answers, Split Decisions solvers are only given pairs of two-letter substrings for each word, either of which serves to complete a valid English word. For example, given the following diagram:

$$\binom{PL}{GR}----$$

A correct fill for the remaining boxes might be ___EASE, creating both PLEASE and GREASE. For the remainder of this paper, such a pair of words will be referred to as a *split pair*. Conversely, an incorrect fill might be ___IERS, which produces the valid word PLIERS and the invalid word *GRIERS[1]. Note that each combination of clues (hereon a *split*) might accept several different fills for the common substring, e.g., ___UNGE (PLUNGE, GRUNGE); however, as with crosswords, only one solution will also satisfy all the crossing constraints produced by the rest of the puzzle.

Note that, although a letter may appear on both sides of a split, say, (IT/TA), the letter will *never* appear at the same index in both words, e.g., (IT/AT), as such a pair of words would only differ by a single letter instead of two. Also, although such a puzzle would be fascinating to both solve and analyze, crossing words do *not* go through split pairs, only common letters.

Upon understanding the rules of the puzzle, one may start to wonder: what sorts of words form split pairs? Are some splits more common than others? How do the properties of English, as both a spoken and written language, affect the guessability of a given split pair? In this paper, we perform preliminary statistical analysis to investigate these questions, using a variety of computational methods.

Data

I used (R Core Team 2021), (Wickham et al. 2019), and (Wickham 2016). The wordlists come from GNU/Linux (specifically `/usr/share/dict/words` on an Ubuntu distribution), (*Collins Official Scrabble®*

[1] The asterisk prefix will be used throughout this paper to indicate incorrect/invalid words.

Words 2019), and (NASPA Dictionary Committee 2020).

Results

Discussion

Appendix

References

Collins Official Scrabble® Words. 2019. Harper Collins.

NASPA Dictionary Committee. 2020. *NASPA Word List, 2020 Edition*. NASPA.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.