Details Carry Meaning – the potential of orthographic similarities in prediction of recall dynamics

Blake Andreou

blake.a.andreou@vanderbilt.edu

Abstract

Language, whether in its spoken or written form, is a complex, dense transmission of information. The uncovering of the relative salience of each of the features of language (and their subsequent modeling), has been the focus of much of the research of computer science since the inception of the field. We, up until very recently, have been unable to effectively characterize the meaning encapsulated by the orthography, or structure, of words within a language's lexicon. Dr. Cynthia Siew, in a 2018 publication, effectively solved this puzzle by developing a relationship between a word's features within a constructed orthographic network and eventual performance on trials using that word in lexical decision and speeded naming tasks. I elected to recreate this study to see if I was able to develop this same network/results. To extend these findings, I developed a semantic analog of this orthographic network, and combined all network features using a hierarchical regression to uncover if the meaning implicitly carried in orthography is distinct from sematic information. Ultimately, orthographic network variables, but not semantic variables, were able to account for a significant portion of variance in reaction time and accuracy of both lexical tasks, suggesting that orthographic information may be prioritized over semantic in context-less environments.

Background

Language is a multi-faceted, complex form of information. Data about the position, context, pronunciation, and structure of each word are simultaneously transmitted from speaker to listener, and it is the job of the listener to, with the most salient features of this data, develop a most-likely translation of the conveyed information.

Much of the focus of previous psycholinguistic research has developed intricate, well-developed methods to harness various semantic properties of words to uncover likely heuristics individuals take when processing language to uncover its meaning most efficiently. For example, by developing a quantification of the inter-connectedness of a word's immediate sematic network, Hills and colleagues theorized that the more well-connected a word's network will be, the more likely the word will be added to an individual's network permanently (Hills et al. 2010). Likewise, the phonological information encapsulated within spoken language and the relationship between this information and informational processing has been well-characterized. Notably, previous research done by Siew using a phonological network demonstrated that the more phonologically distinct a word is, the more effectively it is processed for lexical tasks (Siew and Vitevitch 2016).

In Siew's 2018 publication, she extends her 2016 findings to an orthographical equivalent of her phonological network (Siew 2018). To develop the network, she first needed to define orthographic similarity. Unlike semantic similarity, which lends itself to a number of well-defined operational definitions of similarness based in reasonable assumptions (i.e., using cosine similarity depends on similar words having similar features; using co-occurrence stats depends on synonyms being interchangeable), there are not satisfying definitions of orthographic similarness that are based in a reasonable assumption. Furthermore, definitions of orthographic similarness, likely for the reason above, vary considerably between studies in psycholinguistics.

Research regarding the effects of orthographic similarness has been largely ineffective and contradictory, likely due to this dissimilarity in the operational definitions of orthographic networks. Even the simple assertion that a word's number of orthographic neighbors inhibits naming speed, a fact long-established in the phonological network literature (Siew and Vitevitch 2016), is a point of contention in the orthographic domain. Meta-analyses of previous research (Andrews 1997) have suggested that the number of orthographic neighbors had had a facilitatory effect on word processing, but this was quickly refuted in a number of publications asserting that quantity of orthographic neighbors have an inhibitory effect on the same process (Perea and Rosa 2000; Davis et al. 2009).

To best operationally define this orthographic network to allow the most generalizability to previous psycholinguistic research, we considered two major operational definitions of orthographic similarity. The first, referred to as Coltheart's N, defines a similarity connection as a set of words that can be interchanged by substituting one letter for another (Coltheart et al.

1977). This was attempted in Siew's current study but was ultimately ignored as the resulting orthographic network had the major flaw of being restricted to words of the same length. A more generalized version of Coltheart's N, Levenshtein or edit distance is defined as the set of words that can be interchanged by either adding, deleted, or substituting one letter. To exemplify this difference, a word such as 'pot' will be connected to the word 'rot' with both a Coltheart and Levenshtein distance of 1, but only 'plot' will be connected to 'pot' using a Levenshtein approach.

By using this edit distance definition, Siew was able to use properties of the resulting orthographic network to uncover that the words with many orthographic neighbors enjoyed a benefit in lexical processing in both a lexical decision and a speed naming task. Extending this, she showed that the relative interconnected-ness of the immediate network of a given word allowed for improved lexical decision performance, but worsened performance in speeded naming.

The present study will attempt to recreate this study to demonstrate the construction of an orthographic similarity network and connect the features of this network to lexical task data. To extend these findings, we will also consider the inclusion of a semantic network, defined by cosine similarity of GloVe vector representations of these words, in conjunction with the orthographic network, to conclude if the meaning implicitly stored in orthography is distinct and differentiable to the information encased in semantic network metrics.

Methods

My strategy for developing the orthographic network was derived from the Siew 2018 study described in brief above. To attempt to keep the relative strengths of the two networks as close as possible with respect to the network statistics eventually derived from each respective network, the general procedures for the two networks as far as processing and building are nearly identical. Therefore, I will first describe the general constructions of the network, and then describe the individual constructions unique to each respective network.

General Network Construction

Our networks were both developed with nodes representing the words within the corpus (described in the *data acquisition* section), and edges representing a connection considered 'similar'. The operational definition of similar will change depending on the orthographic or semantic connection involved and will be defined later.

First, each network was generated by comparing each node in the network to each other node, and creating an edge between the nodes if they were defined as 'similar' to each other. Next, to maximize the strength of the information encased in the networks themselves, we elected to look only at the largest connected components (LCCs) of the generated semantic and orthographic networks. To do this, we first removed all hermits – words without any edges.

We then performed breadth first search with a known “heavily-connected” word from the hermit-less networks to find the LCC of each network.

Once the LCC was generated, we derived the following statistics from each network: degree, clustering coefficient, and closeness centrality. Degree is the number of similar words from each word. This was derived in our networks from the edge count from each node. The clustering coefficient represents the degree to which the similar nodes to a given word are connected to each other. The closeness centrality is the inverse of the average path length to all other connected words in the network. The clustering coefficient and closeness centrality measures were both captured using the *networkX* python module (Hagberg et al. 2008).

Orthographic Network Construction

The orthographic network construction, as stated earlier, was intended to replicate the methods of Siew’s original study regarding this network.

Similarity was defined as an edit distance of 1. This was chosen as opposed to a Coltheart distance of 1 to generalize the network and allow it to include as many nodes as possible by allowing links between words of differing lengths.

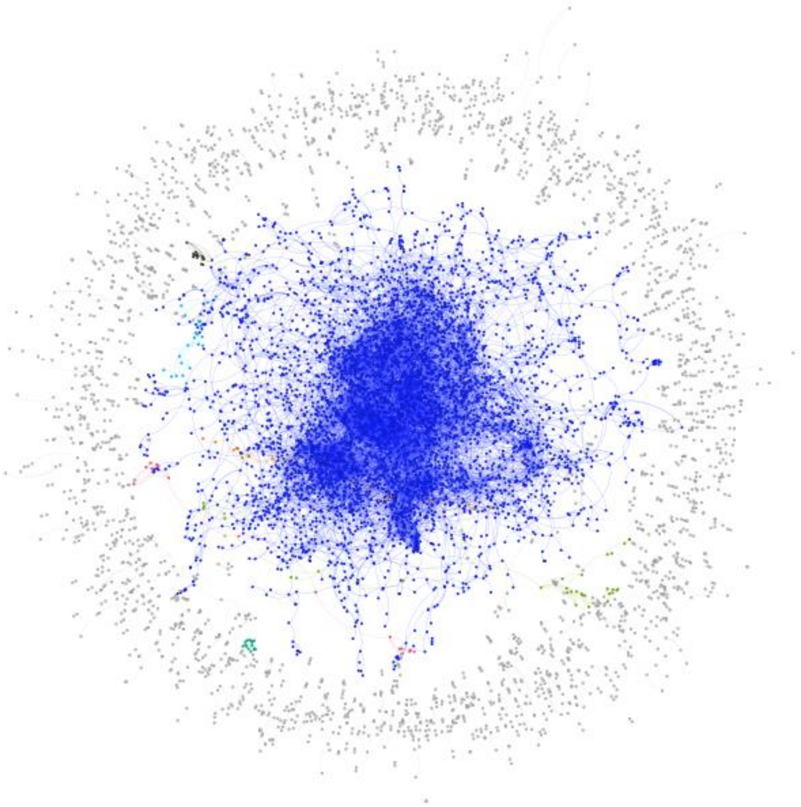

Figure 1 Visualization of Orthographic Similarity Network with edit-distance of 1. Adapted from Siew 2018.

Semantic Network Construction

The semantic network construction was designed with Siew's methods in mind and was intended to match the network dynamics of the orthographic version in as many dimensions as possible.

Similarity was defined by first converting all words in the corpus to a 64-dimensional GloVe vector (Pennington et al. 2014). Then, all vectorized words were compared to all other words, and if the cosine similarity of these vectors were above a threshold *k*, we created an edge of weight 1. If the similarity was at or below the threshold, we did not create an edge. We chose to use a discrete binary definition of similarity, rather than a continuous distribution, to prevent the semantic network from functioning as a weighted network, as we felt that this could implicitly include data features that the orthographic network was not able to capture. K was defined over several trials as 0.6, to keep the hermit proportion as close to the orthographic variant while simultaneously minimizing the difference between the network statistic of the semantic network and the orthographic variant.

Data Acquisition

The data, like Siew's study, was taken from the English Lexicon Project (ELP) (Balota et al. 2007). This project collected a substantial number of data points regarding two lexical tasks: a lexical decision task (*do the letters you see spell a real word?*), and a speeded naming task (*say the real word you see as fast as you can.*). Both tasks were performed across multiple sites with a corpus of approximately 40,000 common English words. We were able to utilize both the behavioral statistics as well as the word pool itself from this project.

The behavioral statistics consisted of the normalized reaction time for each word in the task, as well as the raw accuracy of each word in the task, averaging across participants. We elected to use the normalized reaction time in lieu of the raw reaction time to prevent outliers from affecting the measures of central tendency.

Statistical Modeling

We developed a series of hierarchical models to uncover the potential distinction between these two models in their contribution in explaining the behavioral statistics in the two ELP tasks.

This was done by first capturing all the known first-level regressors for each word in the respective LCC of each network. For the current study, these were defined as the number of letters, number of syllables, number of phonemes, and log likelihood of occurrence. These are all known to strongly affect recognition dynamics of words and were therefore regressed out prior to the inclusion of any network statistics.

To see the unique effects of each model, after we performed a regression including these first level regressors, we performed a second linear regression by including the first-level variables along with the second-level network variables of degree, clustering coefficient, and closeness centrality of one of the networks. This allowed us to directly compare the variance captured by the second-level variables that is not already captured by the first-level variables by utilizing the R^2 measurement.

Results

Network Statistics – Comparing Orthographic and Semantic Networks

When generating the orthographic network, we successfully replicated the dynamics of Siew's manufactured orthographic network. This network has a hermit proportion of 40.7%, which we replicated closely in the semantic network – 40.4%.

When removing hermits and developing the LCC of each respective network, we see a general trend of higher inter-connectedness in the semantic network (Table 1). The degree and the clustering coefficient are substantially higher in the semantic network than the orthographic network. The closeness centrality in both networks, however, is similar, with orthographic networks having a measure of 0.11 compared to semantic network's 0.12.

	Degree	**Clustering Coefficient**	**Closeness Centrality**
Orthographic	5.77	0.27	0.11
Semantic	8.92	0.43	0.12

Table 1: Network Statistics of Orthographic v. Semantic Networks.

Lexical Decision Task – RT:

When observing the lexical decision task RT with just the first level regressors (number of letters, number of phonemes, number of syllables, log likelihood of occurrence), the regressors were able to account for 43.7% of the variance.

When including the orthographic network variables along with the first level regressors, the complete set of regressors were able to now account for an additional 0.66% of the variance, $\Delta R^2 = .0066$, $F(3, 11{,}350) = 45.1$, $p < .001$. Specifically, all three network variables were shown to have a significant effect on RT in the lexical decision task.

When including the semantic network variables along with the first level regressors, the complete set of regressors were able to now account for no additional variance. All three network variables were found to have no significant effects on RT in the lexical decision task.

Lexical Decision Task – Accuracy

When observing the lexical decision task accuracy with just the first level regressors (number of letters, number of phonemes, number of syllables, log likelihood of occurrence), the regressors were able to account for 33.9% of the variance.

When including the orthographic network variables along with the first level regressors, the complete set of regressors were able to now account for an additional 0.39% of the variance, $\Delta R^2 = .0039$, $F(3, 11{,}350) = 22.4$, $p < .001$. Degree was the only network variable found to have a significant effect on RT ($\beta = 0.0028$, $t = 6.82$, $p < .001$).

When including the semantic network variables along with the first level regressors, the complete set of regressors were able to now account for no additional variance. All three network variables were found to have no significant effects on accuracy in the lexical decision task.

Lexical Naming Task – RT

When observing the lexical naming task RT with just the first level regressors (number of letters, number of phonemes, number of syllables, log likelihood of occurrence), the regressors were able to account for 25.7% of the variance.

When including the orthographic network variables along with the first level regressors, the complete set of regressors were able to now account for an additional 0.99% of the variance, $\Delta R^2 = .0099$, $F(3, 11{,}350) = 49.6$, $p < .001$. Degree and closeness centrality were able to account for a significant proportion of this variance.

When including the semantic network variables along with the first level regressors, the complete set of regressors were able to now account for no additional variance. All three network variables were found to have no significant effects on RT in the lexical naming task.

Lexical Naming Task – Accuracy

When observing the lexical naming task accuracy with just the first level regressors (number of letters, number of phonemes, number of syllables, log likelihood of occurrence), the regressors were able to account for 15.4% of the variance.

When including the orthographic network variables along with the first level regressors, the complete set of regressors were able to now account for an additional 0.55% of the variance, $\Delta R^2 = .0055$, $F(3, 11{,}350) = 24.4$, $p < .001$. Like the RT for this task, degree and closeness centrality were able to account for a significant proportion of this variance.

When including the semantic network variables along with the first level regressors, the complete set of regressors were able to now account for no additional variance. All three network variables were found to have no significant effects on RT in the lexical naming task.

Discussion

Orthographic Network Effects v. Semantic Effects

The most glaring conclusion from the regressions run is that the addition of semantic network variables had virtually no role in explaining the variance in either RT or accuracy for both tasks in the English Lexicon Project.

This is a particularly interesting finding, as research dating back to the earliest stages of psycholinguistic study has found a benefit from semantic information on tasks such as lexical decision making (Fischler 1977). In fact, the initial derivation of this very task was utilized to help introduce the concept of semantic priming in the early 1970s (Meyer and Schvaneveldt 1971).

However, it is important to note that in all experiments utilizing this task, those that find a semantic benefit are manipulating the presentation of words in some capacity to look for a potential effect of semantic relatedness between consecutive words specifically. For example, Dr. Fischler's 1977 procedure explicitly tests certain presentation order-combinations to test for priming given that the previous word is/is not related to the current word. In the present study, we had a naïve presentation order so as not to directly test for semantic effects, and it appears that this lack of semantic 'direction' prevents the utilization of semantic priors to make lexical decisions.

It seems that, due to this naïve presentation, we may be forced to 'build' the word from the ground up to generate the proper word representation to either declare the word as a valid word, or to name the word. Due to this from-scratch generation, we may be forced to utilize heuristics, such as degree of orthographic similar-ness, to scaffold our decision-making in the absence of any pragmatic or contextual information. This may allow for the most efficient generalization in this contextless environment, either for most likely pronunciation rules (for pronunciation purposes for the naming task), or for likelihood of seeing the presented letters as a single word.

Regarding the networks themselves, the semantic network was approximately the same size as the orthographic network with a similar closeness centrality, but there was far more interconnectivity in the semantic network. This implies that there were more edges in the semantic network, but these edges were not able to allow for better traversal of the network. This in turn implies that many of the edges are devoted to heavily inter-connected subregions within the network, which may impair potential facilitatory effects of the usage of semantic network data in these predictive tasks.

Orthographic Effects on Lexical Tasks

Diving deeper into the orthographic network effects on the lexical tasks, we see an interesting pattern emerge within the degree and closeness centrality variables specifically. To reiterate, degree represents the number of words connected to it in the network, or, in our example, the number of words that have an edit distance of one from the current word. Closeness centrality is a measure of how centralized our node is in the network, represented in our network as the inverse of average path length from our given node to all other network nodes.

For the variable of degree, we saw a similar facilitatory effect for words with a high degree as Coltheart initially did when running the first variants of orthographic similarity studies (Coltheart 1977). This makes sense, as our words with more neighbors are more able to be generalized towards and can be more reliably characterized/named due to this increased confidence due to the degree.

Closeness centrality is a more interesting variable, having reverse effects on the reaction times of lexical naming tasks and decision tasks. Specifically, there was a facilitatory effect on RT of lexical decisions, and an inhibitory on lexical naming due to closeness centrality. The facilitatory effect is in line with previous research suggesting that more prototypical ‘wordy’ words enjoy a benefit in these decision tasks (cite here). However, this prototypicality seems to hinder these same words from receiving a corresponding benefit in a lexical naming task. This may be due to the orthographic relatives of the word, which served as scaffolds in lexical decision tasks, now serve as competitive lures in a naming task.

References

Andrews S (1997) The effect of orthographic similarity on lexical retrieval: resolving neighborhood conflicts. *Psychon Bull Rev* 4(4):439–461

Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart (2008) Exploring network structure, dynamics, and function using NetworkX, *in Proceedings of the 7th Python in Science Conference (SciPy2008)*. 11–15.

Balota DA, Yap MJ, Hutchison KA, Cortese MJ, Kessler B, Loftis B et al (2007) The English lexicon project. *Behav Res Methods* 39(3):445–459.

Coltheart M, Davelaar E, Jonasson T, Besner D (1977) Access to the internal lexicon. *In: Dornic S (ed) Attention and performance VI. Academic Press, New York*. pp. 535-555

Davis CJ, Perea M, Acha J (2009) Re(de)fining the orthographic neighborhood: the role of addition and deletion neighbors in lexical decision and reading. *J Exp Psychol Hum Percept Perform* 35(5):1550

Fischler, I (1977) Semantic facilitation without association in a lexical decision task. *Memory & Cognition*(5) 335–339.

Hills TT, Maouene J, Riordan B, Smith LB (2010) The associative structure of language: contextual diversity in early word learning. *J Mem Lang* 63(3):259–273.

Meyer, D.E.; Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*. 90 (2): 227–234.

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *EMNLP* 1532—1543.

Perea M, Rosa E (2000) The effects of orthographic neighborhood in reading and laboratory word identification tasks: a review. *Psicológica* 21(2):327-340

Siew CSQ (2018) The orthographic similarity structure of English words: Insights from network science. *Appl Netw Sci* (3), 13

Siew CSQ, Vitevitch MS (2016) Spoken word recognition and serial recall of words from components in the phonological network. *J Exp Psychol Learn Mem Cogn* 42(3):394–410.