Revisiting Orthographic Effects in Spoken Word Recognition: Insights from Pretrained Language Models

Liu, Timothy
tiyliu@ucdavis.edu

Corina, David
dpcorina@ucdavis.edu

Sagae, Kenji
sagae@ucdavis.edu

Abstract

The paired lexical decision task is a common task used for studying online speech processing. Several key priming effects underlying the composition of the mental lexicon have been found using this priming paradigm, non-exhaustively including orthographic priming effects, semantic priming effects, and phonological priming. This study revisits the effects of orthographic priming in an auditory lexical decision task with English heteronymic pairs, word pairs that share the same orthographic form but have distinct phonological codes. Although heteronymic pairs present an ideal condition for orthographic priming effects to surface, heteronyms are often related to each other semantically, making it difficult to isolate possible orthographic effects from semantic priming effects. To this end, we present a novel methodology for using language models to generate semantically matched prime target controls to compare reaction times against. Using these semantically matched controls, we gather reaction time results from a sample of 29 English speaking university student and conduct Bayesian Regression analysis on 153 heteronymic prime target pairs and 343 control pairs. We find no significant difference in reaction times between heteronymic pairs and semantically matched pairs.

1 General Introduction

There is a large body of evidence that suggests that literacy affects the processing of spoken words (Booth et al., 2004; Morais, Cary, Alegria, & Bertelson, 1979; Pattamadilok, Knierim, Kawabata Duncan, & Devlin, 2010; Petersson, Reis, Askelöf, Castro-Caldas, & Ingvar, 2000). There are a growing number of studies that suggest that orthographic effects are robust in on-line spoken word recognition tasks in English (Chéreau, Gaskell, & Dumay, 2007; Jakimik, Cole, & Rudnicky, 1985; Perre, Midgley, & Ziegler, 2009; Seidenberg & Tanenhaus, 1979), in French (Furgoni, Martin, & Stoehr, 2025; Grainger, Diependaele, Spinelli, Ferrand, & Farioli, 2003; Ziegler & Ferrand, 1998), and in Portuguese (Ventura, Morais, & Kolinsky, 2007; Ventura, Morais, Pattamadilok, & Kolinsky, 2004). These investigations generally use a paired lexical decision paradigm, in which a prime word is followed by a target word, to which participants make validity judgments. Robust effects on reaction time have been established for nonce-word targets (Meyer & Schvaneveldt, 1971), semantically related targets (Neely, 1976; Schvaneveldt, Meyer, & Becker, 1976), and phonologically related targets (Slowiaczek & Hamburger, 1992). In this paper, we revisit those effects with the aid of pretrained language models.

Reaction time data for non-orthographic priming effects in paired lexical decision tasks can generally be understood under one of two models: Spreading Activation Semantic Network Theory (SASN) and the cohort model (Collins & Loftus, 1975; Collins & Quillian, 1969; Marslen-Wilson & Welsh, 1978). SASN is a model of cognitive processing that presents the mental lexicon as a collection of semantically interconnected nodes, such that every word is associated with a node. Nodes can activate other nodes through associative connections, such as semantic relatedness, phonological relatedness, or potentially orthographic relatedness. The cohort model, is a model of phonological word processing that proposes a "cohort" of related words is activated as bottom up acoustic phonetic and top down , semantic, pragmatic, and syntactic information becomes available to the listener, and unsuitable candidates are subsequently pruned until one remains. The mechanism for priming in the cohort model stems from a perseveration of the pruned candidates after the target word is recognized. Orthographic effects in spoken word processing are not well specified or well studied under either of these models; the bulk of research has been done with visual word priming, which is suited to SASN, due to the immediate nature of visual word processing. Meanwhile, results from auditory experiments are better understood through the cohort model, which accounts for the temporal spatiality of spoken words. Türk and Domahs (2022) use an adaptation of another model from the domain of visual word processing, the Bimodal interactive Activation Model (BIAM; Grainger and Holcomb 2009), in order to explicitly explain orthographic effects across a variety of languages. The BIAM is one that allows for interactivity between the orthographic and phonological representations of language during recognition, which helps to explain the differences in orthographic effects depending on the depth of the target language's orthography. Of particular note, however, is the difficulty in disentangling effects due to bottom up phonetic processing and effects due to top down semantic, orthographic, and syntactic processing. In examining the effects of orthographic overlap in spoken word processing, Chéreau et al. (2007) compare "brat" and "spat" as the orthographically related condition, and "brat" and "plait" as the orthographically unrelated condition. It is not entirely clear to what extent the semantic similarity between brat and spat (e.g. the brat spat) played a role in priming (see 6.7 for a selection of potentially confounding prime target pairs).

As an attempt to disentangle the effects of semantic priming between primes and targets in an auditory paired lexical decision task, we present a heteronymic priming paradigm with semantic controls generating with computational language models. Heteronyms are word pairs that are orthographically identical, but are phonologically distinct, e.g. “object” as a verb, and “object” as a noun. Previous attempts have examined identical rimes (Seidenberg & Tanenhaus, 1979), but as far as we are aware, no study investigating orthographic priming effects has attempted to use heteronym pairs, which ostensibly represent best case pairs for demonstrating effects due orthographic relatedness. Using heteronyms as stimuli presents an ideal environment for orthographic effects to surface. However, finding suitable heteronymic pairs in English is often a challenge due inherent semantic similarity, e.g. between “reject” the verb and “reject” the noun. Such heteronym pairs are called weak heteronyms, and introduce semantic priming confounds (Martin, Jones, Nelson, & Nelson, 1981). In addition to controlling for implicit semantic priming between the target words, our generated controls serve to account for any additional semantic priming risk that heteronymic pairs may present.

2 Prior Work

Orthographic effects are well established in speech perception; the orthographic consistency effect manifests as a tendency for listeners to more quickly identify spoken words with more consistent spellings than words with inconsistent spellings (Ziegler & Ferrand, 1998). In the current literature on priming, there is a strong consensus that orthographic priming effects are present in on-line speech processing, particularly in conjunction with phonological effects. Jakimik et al. (1985), Chéreau et al. (2007), Miller and Swick (2003) all conduct auditory paired lexical decision task experiments and find significant effects of overlapping orthography in environments where there was also phonological overlap. However, the effects of orthographic overlap alone are not significant; Miller and Swick (2003) found no interference or facilitation in a purely orthographically overlapping condition (e.g. deaf - leaf). With respect to other languages, there appears to be an interaction between the depth of a language’s orthography and the extent to which underlying orthographic processes impact online spoken word processing; Türk and Domahs (2022) present evidence from German and English that in a spoken word recognition task, orthographic overlap in languages with shallow orthographies may induce inhibitory effects, while orthographic overlap in languages with deep orthographies may induce facilitatory effects. However, to our knowledge, there has yet to be a study that examines orthographic priming effects in homographic environments.

Condition Name	**Summary**	**Example Prime - Target**
`homograph pair`	heteronymic pairs	bass_fish - bass_music
`semantic control`	non-heteronymic pairs semantically matched to `homograph pair`	salmon - guitar
`unrelated pair`	unrelated prime with heteronymic target	history - bass_music
`filler`	filler rhyming pairs	brat - spat
`nonce-word pair`	nonce-word pairs derived from `homograph pair`	bass_music - basp
`nonce-word filler`	nonce-word pairs derived from filler pairs	yawn - dopsh

Table 1: Reference table for each condition, along with example prime–target pairs.

3 Methods

3.1 Materials

3.1.1 Experimental Condition || *homograph pair*

Semantically unrelated strong heteronyms are selected from Martin et al. (1981). Monosyllabic homophonous heteronyms (e.g. do/do homophonous with dew/dough) and heteronyms with a proper noun variant (e.g. natal/Natal) are dropped from consideration to avoid word confusion confounds and word frequency or proper noun word confounds respectively. Several word pairs contain a heteronymic variant not common in American English (e.g. tarry/tarry). We use the Corpus of Contemporary American English (COCA; Davies 2008) as a reference for establishing whether a heteronymic variant should be considered as ubiquitous. COCA is a text corpus, and returns data for particular orthographic strings, along with frequency counts, common word contexts, and parts of speech. In the case of heteronym pairs that have distinct parts of speech, the heteronym pair is considered ubiquitous if COCA contains a separate entry for both parts of speech (e.g., COCA returns entries for object_noun and object_verb separately). For these pairs, COCA is sufficient for obtaining frequency values for both heteronym variants. For heteronym pairs that do share a part of speech, we turn to the Wikipedia homograph dataset (Gorman, Mazovetskiy, & Nikolaev, 2018). The Wikipedia homograph dataset is a dataset of homographs scraped from Wikipedia; if a heteronym pair does have an entry in this dataset, it is dropped from consideration (e.g. row_verb/row_quarrel). The final resulting heteronym list consists of 13 strong and distinct heteronyms (Appendix 6.7)[1].

[1] Strong heteronymic pair slough/slough was dropped from consideration at the control condition generation stage; BERT does not tokenize slough as a whole word, and cosine

3.1.2 Control Condition || *semantic control*

Now that we have a robust list of heteronymic pairs for our experimental condition, we turn to BERT (Devlin, Chang, Lee, & Toutanova, 2019) and return to the Wikipedia homograph dataset (Gorman et al., 2018) to establish our `semantic control` condition. As previously stated, a reliable control condition is necessary to disentangle non-semantic priming effects from semantic priming effects. BERT is a pretrained language model that learns semantic relations between words, a property of embedding spaces within language models(Mikolov, Chen, Corrado, & Dean, 2013). An embedding space is a higher dimensional space in which words are mapped to unique vectors tuned during the pretrained language model training process to implicitly capture semantic information as measured by cosine similarity; words with a cosine similarity near 1 are more closely related. A pre-trained BERT model has its own frozen weights; to "train" the model to distinguish between heteronym pairs, we take a naive but effective approach of averaging embeddings across various contexts. We use 10 sentences for each heteronymic variant from Gorman et al. (2018) and calculate average embeddings for each word[2]. To find semantically related prime target pairs to use as semantic controls, contextual embeddings are generated for all words[3] in the Leipzig 2018 English 1 Million Word Corpus (Quasthoff, Goldhahn, & Eckart, 2015) and indexed by cosine similarity to the averaged heteronymic word embeddings in a k-d tree, following the methodology set by Dale (2020). Each heteronym pair is matched to a similar pair of words measured by cosine similarity, controlling for syllabic counts within the generated control words (e.g. bass_fish/ bass_music was matched to salmon/guitar). The final list of control pairs is listed in Appendix 6.7.

3.1.3 Unrelated Condition || *unrelated pair*

To account for possible lexical decision effects specific to a target heteronymic variant, we include an unrelated prime word condition, where heteronymic targets are preceded by a prime selected at random from the `semantic control` conditions controlling for syllable counts between the new prime and target, and ensuring that the randomly selected prime was not a control for the selected heteronym. Participant groups C and D are created by replacing `homograph pair`

similarity measures for slough_verb resulted in incomprehensible token matches, likely due to infrequency of slough in its verb form in the original training dataset. BERT uses tokens generated by the WordPiece algorithm Wu et al. (2016). WordPiece is a statistical algorithm that relies on training from a tokenization corpus.

[2] For heteronym pairs not represented in the Wikipedia homograph dataset but already established to be ubiquitous through COCA, 10 sentences were generated with ChatGPT (OpenAI, 2023) using the prompt: "Provide 10 sentences using the [disambiguator] [HV] without inflecting the morphology.", where [disambiguator] was replaced with the part of speech of the target Heteronymic variant (HV) or some disambiguator in the case that the HV pair shared part of speech, and [HV] was replaced by the target HV.

[3] Technically, contextual embeddings are generated for each token in the Leipzig dataset, and those token embeddings are compared to the contextual word embeddings generated by BERT for the heteronym variants (which we call words on the basis of establishing that each heteronymic variant is stored as a whole word token).

conditions from the A and B participant groups with `unrelated pair` conditions (see Counterbalancing).

3.1.4 Filler Condition || *filler*

To obfuscate the intention of the experiment, we include filler trials selected at random from (Chéreau et al., 2007).

3.1.5 Nonce-word conditions || *nonce−word pair*, *nonce-word filler*

For the `homograph pair` and filler condition, we generate nonce-words conditions `nonce-word pair` and `nonce-word filler`. We generate nonce-words by replacing the final consonant with a random consonant such that the resulting string is not an English word[4].

3.2 Counterbalancing

To address prime-target order confounding attributable to frequency effects, we establish A and B groups. A participant assigned to group A would respond to heteronym variant one (e.g. object_noun) preceded by prime heteronym variant two (e.g. object_verb). A participant assigned to group B would respond to the same trial, but with the order of the heteronyms swapped (e.g. object_noun followed by object_verb). To address any possible long-term repetition priming effects caused by heteronymic primes in the nonce-word pair impacting targets in the homograph pair condition, we partition X and Y groups within each of the A, B, C and D participant groups. Participants in the Y condition have the order of the `homograph pair` and `nonce-word pair` trials with matching heteronym pairs swapped. Participants were randomly assigned to positively respond to the target word with either their left or right hand to account for any difference in reaction time due to handedness. 17 participants were asked to respond positively to target words with their right hand, and 12 participants were asked to respond positively with their left hand. In total, a participant from Group A or Group B would be exposed to 13 `semantic control` trials, 13 `homograph pair` trials, 13 `nonce-word pair` trials, 13 `filler` trials, and 26 `nonce-word filler` trials. A participant from Group C or Group D would be exposed to 13 `semantic control` trials, 2 or 1 `homograph pair` trials, 13 `nonce-word pair` trials, 13 `filler` trials, 26 `nonce-word filler` trials, and 12 or 11 `unrelated pair` trials.

3.3 Recording

Auditory experimental stimuli was produced by a college-aged Caucasian male who was born and raised in the Bay Area. All recorded stimuli were delivered in a carrier phrase. All stimuli recordings were recorded in sessions of 50 words. Each session was normalized with the Normalize function in Praat, and Praat

[4] In the case of re.su.me, we replaced the final vowel and appended a random consonant.

was used to extract the stimuli by hand (Boersma & Weenink, 2001). After all stimuli were extracted, all stimuli were collectively normalized to 75 db in Praat.

3.4 Procedure

The experiment was performed on a Dell Latitude 7480 laptop running Windows 10 using Presentation® software (Version 24.1, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). Participants wore a set of beyerdynamic DT 770 headphones to receive language stimuli and used the keyboard to record their responses. Participants were tested individually in a quiet room at a comfortable distance from the laptop and wore a pair of headphones for 1 session of 15 minutes, under the supervision of a proctor in the same room. After consenting, participants were instructed to use the Windows 10 built-in sliding volume adjustment to set the volume level to a setting between 0 and 100 that was comfortable for them (mean=55.17). Participants were then shown an instruction screen, and played a practice block of 9 words (Appendix 6.7), during which time they were informed that they could adjust the volume. Participants were asked to use the A and L keys to respond either positively or negatively to the target word (see Counterbalancing). Participants were asked to keep their hands on the keyboard for the duration of the experiment. During the experiment, primes were followed by targets after a delay. An interstimulus interval of 500ms was selected within the range of established interstimulus intervals within the literature of auditory priming ranging from 20ms (Chéreau et al., 2007) to >800ms (Miller & Swick, 2003). Immediately after the playing of the target word, an image response prompt was shown to participants to indicate that they should respond. Reaction times were measured from the start of the target stimulus. Participants were given a maximum window of 3 seconds to respond; if they timed out, the trial was considered incorrect and the experiment moved on.

3.5 Participants

29 participants are included in the final study. Participant demographics consist of 28 university undergraduate students and 1 university graduate student, representing a strongly literate population. 8 male and 6 female participants were recruited through word of mouth, and participated without compensation. An additional 3 male and 12 female participants were recruited through SONA, and participated for course credit. All participants were at least 18 years old, and gave informed consent to participate in the experiment through overview of possible risks and benefits before beginning the experiment. 5 participants were left-handed and 24 participants were right-handed. All participants had normal hearing and were native English speakers. Participant distributions across counterbalancing groups are shown in Table 2.

Group	X	Y
A	5	4
B	5	5
C	3	2
D	3	2

Table 2: Distribution of participants in each of the participant groups (N=29).

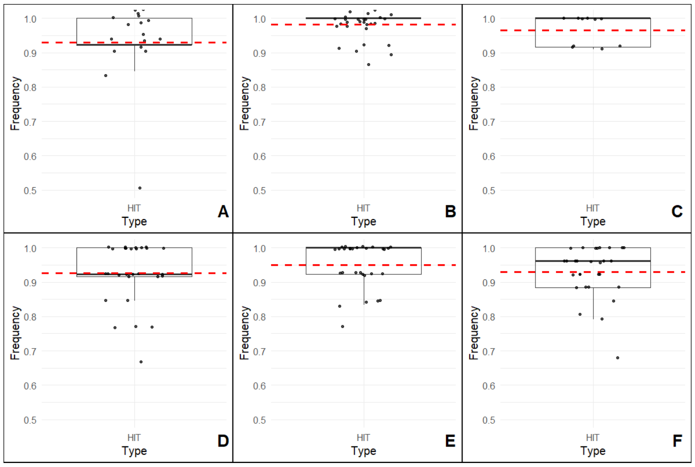

Figure 1: Hit rates for each participant across conditions. Each subplot represents a different condition: (A) `homograph pair`, (B) `semantic control`, (C) `unrelated pair`, (D) `nonce-word pair`, (E) `filler`, and (F) `nonce-word filler`. The red dotted line indicates the mean hit rate for each condition.

4 Results

4.1 Participant Task Accuracy

Participants generally did well on all tasks in terms of accuracy (Fig. 1). Mean accuracy in the `homograph pair` condition was 92.99%. Mean accuracy in the `semantic control` condition was 98.14%. Overall accuracy in the `unrelated pair` condition was 94.85%. Mean accuracy in the `nonce-word pair` condition was 92.68%. Mean accuracy in the `filler` condition was 94.92%. Mean accuracy in the `nonce-word filler` condition was 92.99%[5]. Trials where participants timed out and did not respond were not included in the final analysis. Boxplot distributions for each condition are shown in Figure 1. Table 4.1 shows the exact number of missed trials for each condition.

[5] One participant in the C condition was only exposed to two homograph pair trials, and answered one incorrectly. This results in an apparent outlier with 50% accuracy, but their data was included because their performance on the remaining tasks was normal.

Condition	Analyzed Trials	Correct Trials	Incorrect Trials	Miss Trials	Total Trials
homograph pair	134	153	10	0	163
semantic control	347	369	7	1	377
unrelated pair	101	110	4	0	114
filler	305	357	19	1	377
nonce-word pair	339	343	27	7	377
nonce-word filler	690	695	52	7	754

Table 3: Total number of trials per condition included in the final analysis. Trials were excluded if participants responded faster than 300 ms. A "miss" indicates no response within three seconds of stimulus onset.

4.2 Nonce-Word Sanity Check

To verify the integrity of our data, we first conducted an analysis comparing the *nonce-word* conditions against the *real-word* conditions. Our goal was to replicate the well-established effect of delayed reaction times to nonce-word targets.

Reaction time (RT) data were log-transformed to approximate a normal distribution. We then fit a multilevel Bayesian regression model to data from the `homograph pair` and `nonce-word pair` conditions. The model included fixed effects of experimental condition (`Code`) and trial index (N), random participant-level effects of `Code` and N, and random intercepts for the target word. The full model specification is given in Equation 1:

$$\mathrm{RT} \sim 1 + \texttt{Code} \times N + (\texttt{Code} \times N \mid P) + (1 \mid \mathrm{target}) \tag{1}$$

Reaction times were measured from the onset of the auditory stimulus. Responses faster than 300 milliseconds were excluded as implausibly fast.

We specified a Student-t prior on the intercept with a mean of 6.9[6], standard deviation of 1.7, and 3 degrees of freedom. For all other effects (fixed and random), we used Student-t priors with mean 0, standard deviation 1.7, and 3 degrees of freedom.

[6] $6.9 = \log(1000)$. We assume a baseline reaction time of 1000ms.

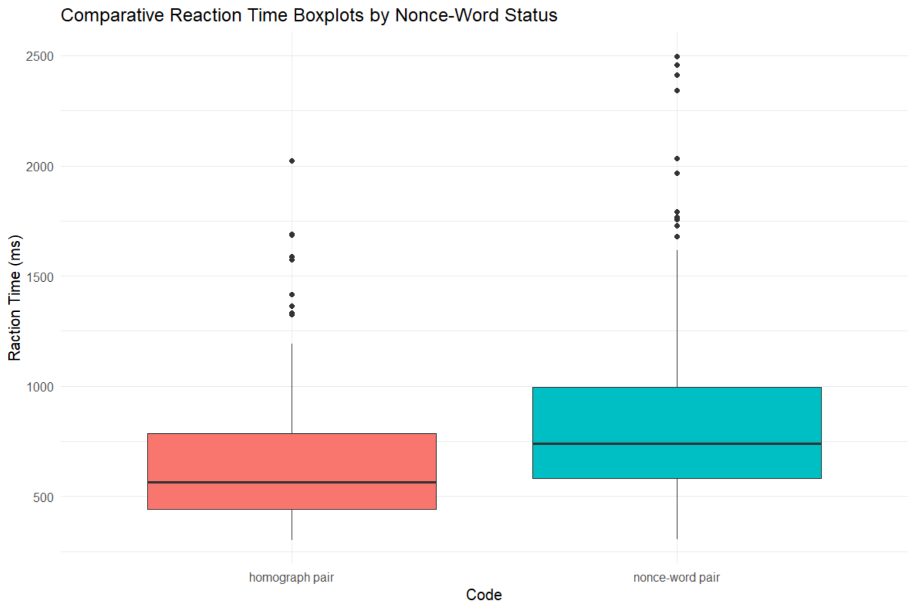

Figure 2: `homograph pair` (N=153) and `nonce-word pair` (N=343) reaction time (ms) comparisons.

Level	Estimated Effect	Estimated Effect Re-exponentiated	Estimation Error	95% CI
Intercept	6.57	713.37	0.07	6.43 – 6.72
Nonce-word	0.18	140.79	0.05	0.08 – 0.28

Table 4: Summary of regression coefficients for the model comparing real word and nonce-word reaction times. Re-exponentiated effects are displayed in milliseconds[8].

Level	Estimated Effect	Estimation Error	95% CI
Intercept	0.30	0.05	0.21 – 0.42
Nonce-word	0.10	0.04	0.01 – 0.19

Table 5: Summary of participant group level effects for the real word/nonce-word comparison model; standard deviation of participant-dependent values for *Level*, respectively. Results remain log-transformed.

[8] Since we use log transformed data to conduct analysis, re-exponentiated values are calculated by taking the sum of log transformed values and re-exponentiating the difference of those values. We follow this convention in all following analyses.

Trial index effects are not included in Table 5; there was no significant effect of trial index, and will henceforth no longer be included in the results. We find a main effect of the nonce-word condition on reaction time; on average, participant reaction time increased by 140.79 milliseconds when encountering a nonce-word condition (C.I. = [0.01, 0.19]). This result validates the integrity of the experiment by replicating the well-established increased reaction time to nonce-word targets (Fig. 2).

4.2.1 homograph pair and unrelated pair Comparison

We examine if there is any significant difference in reaction times between the homograph pair and unrelated pair conditions. We fit a multilevel Bayesian regression model with the same model function in Equation 1 on log transformed reaction times for the homograph pair and unrelated pair condition. Our null hypothesis assumes no orthographic effects on priming.

Level	**Estimated Effect**	**Estimated Effect Re-exponentiated**	**Estimation Error**	**95% CI**
Intercept	6.43	620.17	0.08	6.27 – 6.59
unrelated pair	0.00	0	0.07	-0.14 – 0.15

Table 6: Summary of regression coefficients for the model comparing the homograph pair and unrelated pair conditions. Re-exponentiated effects are displayed in milliseconds.

Level	**Estimated Effect**	**Estimation Error**	**95% CI**
Intercept	0.22	0.08	0.05 – 0.37
unrelated pair	0.10	0.08	0.00 – 0.28

Table 7: Summary of participant group level effects for the homograph pair/unrelated pair comparison model; standard deviation of participant-dependent values for *Level*, respectively. Results remain log-transformed.

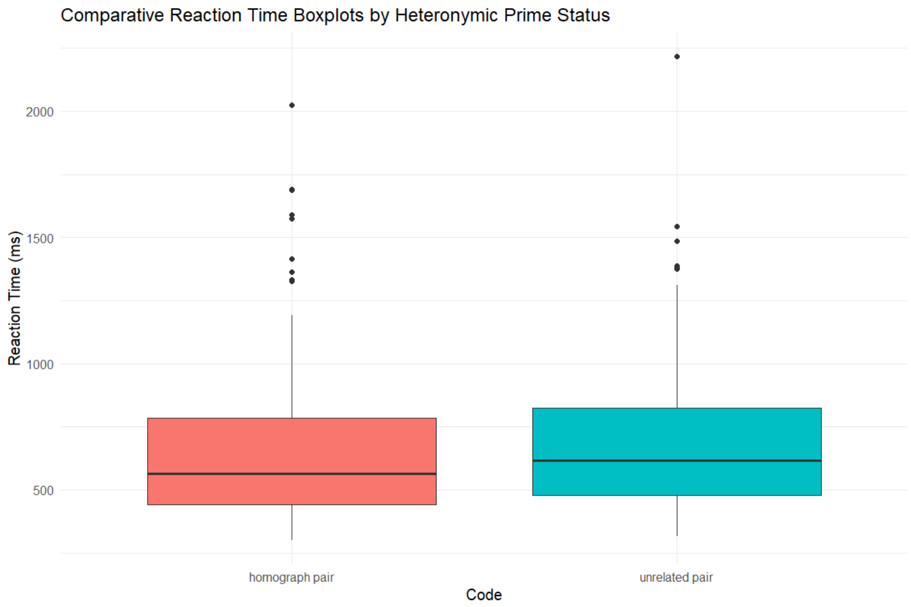

Figure 3: homograph pair (N=153) and unrelated pair (N=110) reaction time (ms) comparisons.

As expected, we see a very small main effect of the unrelated target word prime (mean=0.02, C.I.=[-0.12, 0.14]). The potential main effect overlaps with 0, giving us strong confidence that there was no significant effect of heteronymic primes when compared to non-heteronymic primes.

4.2.2 homograph pair and semantic control Comparison

Finally, we analyze the reaction times between the homograph pair and semantic control conditions. We use the same model (1) to examine the effects.

Level	Estimated Effect	Estimated Effect Re-exponentiated	Estimation Error	95% CI
Intercept	6.36	578.25	0.06	6.25 – 6.47
semantic control	0.01	5.811	0.04	-0.07 – 0.04

Table 8: Summary of regression coefficients for the model comparing the homograph pair and semantic control conditions. Re-exponentiated effects are displayed in milliseconds.

Level	Estimated Effect	Estimation Error	95% CI
Intercept	0.22	0.04	0.15 – 0.31
semantic control	0.03	0.02	0.00 – 0.09

Table 9: Summary of participant group level effects for the homograph pair/semantic control comparison model; standard deviation of participant-dependent values for *Level*, respectively. Results remain log-transformed.

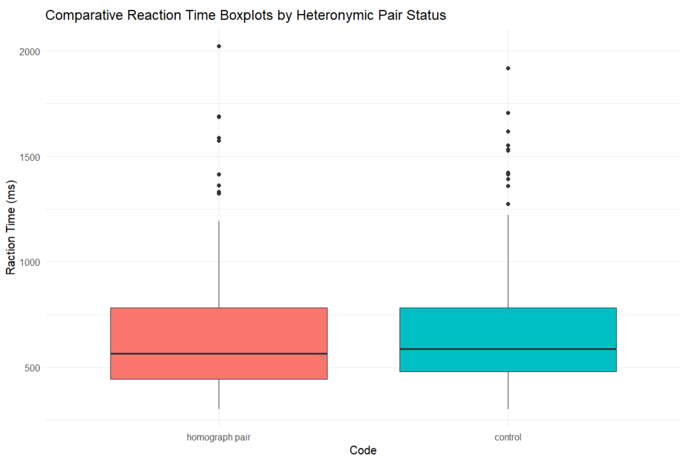

Figure 4: homograph pair and semantic control reaction time (ms) comparisons.

We find no significant effect of heteronymic priming when compared against semantically controlled non-heteronym prime target pairs (mean=0.01, C.I.=[-0.07, 0.04]).

5 Discussion

In this paper, we investigate orthographic priming effects in heteronymic word pairs, which arguably provide the strongest test of orthographic effects during auditory word processing. A key attribute of this experimental design was to provide appropriate semantic controls for our experimental pairs. We used pre-trained language models in order to generate these semantic controls, utilizing their ability to generate semantic space to search for suitable control pairs. We find that reaction times to audially presented heteronymic pairs are not significantly different to reaction times to these control pairs. These findings provide support for a model of on-line speech processing that does not necessarily invoke orthographic representations in literate adults, similar to the model of orthographic processing in languages with shallow orthographies proposed by Türk and Domahs (2022). Although this methodology addresses a potential confound in the investigation of orthographic priming effects, there are limitations. We did not conduct an analysis investigating the extent to which pretrained model semantic spaces accurately predict reaction times accounted for by semantic priming. Although the control pairs pass an intuitive check of semantic similarity to the experimental pairs, the exact semantic priming effects that they mitigate might vary between words. Pretrained language models have also improved significantly in recent years; the results may vary with a newer model. We acknowledge the potential for phonological priming effects to confound the results, due to the phonological similarities between heteronymic pairs. Newer language models that utilize ipa tokenization might be useful for investigating the impact of phonological priming with phonologically generated controls.

6 Declarations

6.1 Funding

This work did not receive external funding.

6.2 Conflicts of Interest/Competing Interests

We are not aware of any conflicts of interest in the production of this work.

6.3 Ethics approval

All work was conducted with approval from the Institutional Review Board at the University of California, Davis, protocol 433927.

6.4 Consent to Participate

All participants gave informed consent to participate in the experiment.

6.5 Consent for publication

. All participants gave informed consent for their data to be anonymized and published.

6.6 Availability of Data and Materials

Anonymized data is available at https://github.com/timo-liu/RevisitingOrthographicEffectsWithLLMs.

6.7 Code Availability

Code associated with this project is available at https://github.com/timo-liu/RevisitingOrthographicEffectsWithLLMs.

References

Boersma, P., & Weenink, D. (2001, January). PRAAT, a system for doing phonetics by computer. *Glot international*, *5*, 341–345.

Booth, J. R., Burman, D. D., Meyer, J. R., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2004, September). Development of Brain Mechanisms for Processing Orthographic and Phonologic Representations. *Journal of cognitive neuroscience*, *16*(7), 1234–1249. doi: 10.1162/0898929041920496

Chéreau, C., Gaskell, M. G., & Dumay, N. (2007, March). Reading spoken words: Orthographic effects in auditory priming. *Cognition*, *102*(3), 341–360. doi: 10.1016/j.cognition.2006.01.001

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. doi: 10.1037/0033-295X.82.6.407

Collins, A. M., & Quillian, M. R. (1969, April). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. doi: 10.1016/S0022-5371(69)80069-1

Dale, D. (2020). *Learn to extract embeddings from BERT.*

Davies, M. (2008). *The corpus of contemporary american english (COCA).*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805). arXiv. doi: 10.48550/arXiv.1810.04805

Furgoni, A., Martin, C. D., & Stoehr, A. (2025, March). A cross linguistic study on orthographic influence during auditory word recognition. *Scientific Reports*, *15*(1), 8374. doi: 10.1038/s41598-025-92885-x

Gorman, K., Mazovetskiy, G., & Nikolaev, V. (2018, May). Improving homograph disambiguation with supervised machine learning. In N. Calzolari et al. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA).

Grainger, J., Diependaele, K., Spinelli, E., Ferrand, L., & Farioli, F. (2003, November). Masked repetition and phonological priming within and across modalities. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(6), 1256–1269. doi: 10.1037/0278-7393.29.6.1256

Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processes in visual word recognition. *Language and Linguistics Compass*, *3*(1), 128–156. doi: 10.1111/j.1749-818X.2008.00121.x

Jakimik, J., Cole, R. A., & Rudnicky, A. I. (1985, April). Sound and spelling in spoken word recognition. *Journal of Memory and Language*, *24*(2), 165–178. doi: 10.1016/0749-596X(85)90022-1

Marslen-Wilson, W. D., & Welsh, A. (1978, January). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63. doi: 10.1016/0010-0285(78)90018-X

Martin, M., Jones, G. V., Nelson, D. L., & Nelson, L. (1981, May). Heteronyms and polyphones: Categories of words with multiple phonemic representations. *Behavior Research Methods & Instrumentation*, *13*(3), 299–307. doi: 10.3758/BF03202018

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. doi: 10.1037/h0031564

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, September). *Efficient Estimation of Word Representations in Vector Space* (No. arXiv:1301.3781). arXiv. doi: 10.48550/arXiv.1301.3781

Miller, K. M., & Swick, D. (2003). Orthography Influences the Perception of Speech in Alexic Patients. *Journal of Cognitive Neuroscience*, *15*(7), 981–990. doi: 10.1162/089892903770007371

Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, *7*(4), 323–331. doi: 10.1016/0010-0277(79)90020-9

Neely, J. H. (1976, September). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*(5), 648–654. doi: 10.3758/BF03213230

Pattamadilok, C., Knierim, I. N., Kawabata Duncan, K. J., & Devlin, J. T. (2010, June). How does learning to read affect speech perception? *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(25), 8435–8444. doi: 10.1523/JNEUROSCI.5791-09.2010

Perre, L., Midgley, K., & Ziegler, J. C. (2009). When beef primes reef more than leaf: Orthographic information affects phonological priming in spoken word recognition. *Psychophysiology*, *46*(4), 739–746. doi: 10.1111/j.1469-8986.2009.00813.x

Petersson, K. M., Reis, A., Askelöf, S., Castro-Caldas, A., & Ingvar, M. (2000). Language processing modulated by literacy: A network analysis of verbal repetition in literate and illiterate subjects. *Journal of Cognitive Neuroscience*, *12*(3), 364–382. doi: 10.1162/089892900562147

Quasthoff, U., Goldhahn, D., & Eckart, T. (2015, January). Building large resources for text mining: The leipzig corpora collection. In (pp. 3–24). doi: 10.1007/978-3-319-12655-5_1

Schvaneveldt, R. W., Meyer, D. E., & Becker, C. A. (1976). Lexical ambiguity, semantic context, and visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(2), 243–256. doi: 10.1037/0096-1523.2.2.243

Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 546–554. doi: 10.1037/0278-7393.5.6.546

Slowiaczek, L. M., & Hamburger, M. (1992). Prelexical facilitation and lexical interference in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(6), 1239–1250. doi: 10.1037/0278-7393.18.6.1239

Türk, S., & Domahs, U. (2022, December). Orthographic influences on spoken word recognition in bilinguals are dependent on the orthographic depth of the target language not the native language. *Brain and Language*, *235*, 105186. doi: 10.1016/j.bandl.2022.105186

Ventura, P., Morais, J., & Kolinsky, R. (2007, December). The development of the orthographic consistency effect in speech recognition: From sublexical to lexical involvement. *Cognition*, *105*(3), 547–576. doi: 10.1016/j.cognition.2006.12.005

Ventura, P., Morais, J., Pattamadilok, C., & Kolinsky, R. (2004, February). The locus of the orthographic consistency effect in auditory word recognition. *Language and Cognitive Processes*, *19*(1), 57–95. doi: 10.1080/01690960344000134

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016, October). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation* (No. arXiv:1609.08144). arXiv. doi: 10.48550/arXiv.1609.08144

Ziegler, J. C., & Ferrand, L. (1998, December). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review*, *5*(4), 683–689. doi: 10.3758/BF03208845

Appendix A. Practice Block of Word Pairs Given to All Participants

Prime	**Target**	**Expected Response**
trouble	bubble	real word
hurt	flirt	real word
punch	cralt	nonce-word
cream	streep	nonce-word
re'bel	rebek	nonce-word
triple	insane	real word
jump	jul	nonce-word
apple	berry	real word
tear_verb	tear_eye	real word

Table 10: Practice block of word pairs presented to all participants.

Appendix B. All Possible Pairs in the `homograph pair` Condition

Prime	Target
bass_music	bass_fish
bass_fish	bass_music
bow_verb	bow_noun
bow_noun	bow_verb
buffet_noun	buffet_verb
buffet_verb	buffet_noun
dove_verb	dove_noun
dove_noun	dove_verb
entrance_verb	entrance_noun
entrance_noun	entrance_verb
invalid_noun	invalid_adj
invalid_adj	invalid_noun
minute_noun	minute_adj
minute_adj	minute_noun
polish_noun	polish_verb
polish_verb	polish_noun
resume_noun	resume_verb
resume_verb	resume_noun
wind_verb	wind_noun
wind_noun	wind_verb
wound_noun	wound_verb
wound_verb	wound_noun
object_verb	object_noun
object_noun	object_verb
present_verb	present_noun
present_noun	present_verb

Table 11: All possible word pairs in the `homograph pair` condition (heteronym primes).

Appendix C. All Possible Control Pairs Matched to Their Respective `homograph pair` Heteronym

Prime	**Target**	**Matched Heteronym**
history	continue	resume
continue	history	resume
guitar	salmon	bass
salmon	guitar	bass
boat	hat	bow
hat	boat	bow
disrupt	dining	buffet
dining	disrupt	buffet
bird	slipped	dove
slipped	bird	dove
entry	engage	entrance
engage	entry	entrance
paralyzed	fraudulent	invalid
fraudulent	paralyzed	invalid
large	hour	minute
hour	large	minute
oil	Poland	polish
Poland	oil	polish
breeze	grow	wind
grow	breeze	wind
trauma	ended	wound
ended	trauma	wound
item	reply	object
reply	item	object
now	show	present
show	now	present

Table 12: All matched control word pairs and their associated heteronyms from the `homograph pair` condition.

Appendix D. Selection of Potentially Confounding Pairs

Target	**Prime**
brat	spat
gleam	dream
fume	plume
womb	tomb
pork	fork
hoop	loop
yawn	dawn

Table 13: Word pairs selected as potentially confounding due to phonological or semantic similarity.