Study on Visual Word Recognition in Bangla across Different Reader Groups

Manjira Sinha
Conduent Labs India
Bangalore, India

Tirthankar Dasgupta
TCS Innovation Labs
Kolkata, India

Anupam Basu
IIT Kharagpur
Kharagpur, India

{manjira87, iamtirthankar, anupambas}@gmail.com

Abstract

This paper presents a psycholinguistic study of visual word recognition in Bangla. The study examines the relationship among different word attributes and word reading behaviors of the two target user groups, whose native language is Bangla. The different target user groups also offer insights into the subjectivity of written word comprehension based on the readers background. For the purpose of the study, reading in terms of visual stimulus for word comprehension has been considered. To the best of the knowledge of the authors, this study is the first of its kind for a language like Bangla.

1 Introduction

Recognition and understanding of words are basic building blocks and the first step in language comprehension. At this stage, the form (visual representation) joins the meaning (conceptual representation). Therefore, the cognitive load associated with word reading is a significant contributor to the overall text readability. The present study aims to capture the salience effects of different word attributes on the word reading performance in Bangla, the second most spoken (after Hindi) and one of the official languages of India with about 85 million native users in India [1] . The features studied in this work, encompass orthographic properties of a word like length in terms of the number of visual units or akshars; number of unique orthographic shapes i.e, the characteristic strokes and complexity measures based on the familiarity of the akshars and strokes in a word. Phonological properties of a word such as number of syllables and spelling to sound consistency have also been taken into account along with the semantic attributes of a word like number of synonyms and number of senses. Moreover, the feature list also includes word collocation attributes such as orthographic neighborhood size and phonological neighborhood size, which situate the given word with respect to other members in the vocabulary. The effects of the word attributes have been measured in terms of the reaction time and performance accuracy data obtained from empirical user experiments.

The paper is organized as follows: Section 2 presents the relevant literature study, section 3 describes the participants details; section 4 and 5 states data preparation and the psycholinguistic experiment respectively; section 6 presents the feature descriptions and the experimental observations against words and non-words; finally, section 7 concludes the paper.

2 Related Works

Research in word recognition has been central to many areas in cognitive-neuroscience (Frost et al., 2005), educational processes (Seidenberg, 2013), attention (Zevin and Balota, 2000), serial versus parallel processing (Coltheart et al., 1993), connectionism (Plaut et al., 1996) and much more. Typically, two different techniques are used to study visual word recognition: the lexical decision tasks and the naming task (Balota et al., 2004). In lexical decision task, a letter string is presented to a participants are asked to decide whether the given string is a valid word in their language. On the other hand, in the naming task, participants are asked to read allowed a letter string as quickly as possible. The time taken by a subject to complete each task after the visual presentation of the target is defined as the response time (RT). An analysis of the reaction times of the subjects

[1] http://www.ethnologue.com/statistics/size

S Bandyopadhyay, D S Sharma and R Sangal. Proc. of the 14th Intl. Conference on Natural Language Processing, pages 427–434, Kolkata, India. December 2017.

reveals the actual processing of words in brain. The early works in word recognition involves two distinct models: the activation model or the logogen model (Morton, 1969) and the search model (Forster and Bednall, 1976); both of these two models are based on the fundamental premises of the frequency effects in word recognition. The frequency effect in word recognition claims that the high frequency words are recognized more accurately and quickly than the low-frequency words (Murray and Forster, 2004). The logogen model assumes recognition of words in terms of the activation of the constituent linguistic features (called the logogens). Each logogen has got a base activation value (also called the resting activation) that facilitates the recognition process. The resting activation of a given logogen is determined by its frequency of occurrence. That is, high frequency words have higher base activation value than the low frequency words. The search model, on the other hand, assumes that words are organized according to their frequencies and are searched serially. (Taft and Hambly, 1986) have a proposed hybrid model that includes features of both the activation and serial search process. The interactive activation (IA) model (Diependaele et al., 2010) follows the connectionist approach and also incorporates the logogen model. In this framework, a word is initially perceived via the basic orthographic, features which in turn activate the higher level syntactic and semantic features. The IA model also accounts for the word superiority effect that assumes alphabets are recognized more accurately and quickly when they occur in a word as compared to a non-word (Grainger and Jacobs, 1996). An important extension of the IA model is the dual-rout cascaded (DRC) model (Coltheart et al., 2001). This model assumes two parallel process of word recognition: the lexical route and the sub-lexical route. The lexical route accounts for the recognition process through the parallel activation of the orthographic and phonological features of a word. On the other hand, the sub-lexical route possesses a serial processor that converts graphemic representations into phonemic forms. As an alternative to two different processing paths n the DRC model, the parallel distributed processing model (PDP) (Seidenberg and McClelland, 1989) has proposed a single architecture to explain different processing outputs. The model incorporate the distributed nature by assuming that each word is associated with some distinct activation pattern across a common set of features used to recognize the word. The features may include, orthography, phonology, morphology or semantic. Generalizations of the PDP model for non-words and irregular words have been proposed by (Plaut et al., 1996)

3 Participants

In order to understand how the different cognitive processes vary across different user groups, two categories of users have been considered for each user study. Group 1 consists of 25 native users of Bangla in the age range 21-25 years, who are pursuing college level education and group 2 consists of 25 native users in the age range 13 to 17 years (refer to figure 1). In this paper, the variations in age and years of education have been taken into account. Moreover, we have considered a distribution over medium to low socio-economic sections with monthly household income ranges INR 4500 to INR 15000. The Socio-Economic Classification (SEC) has been performed according to the guidelines by the Market Research Society of India (MRSI) [2]. MRSI has defined 12 socio-economic strata: A1 to E3 in the decreasing order. The containment of the socio-economic range was necessary as it directly affects education, literacy and thus the state of comprehension skills of a reader. In addition, to capture the first-language skill, each native speaker was asked to rate his/her proficiency in Bangla on a 1-5 scale (1: very poor and 5: very strong), see figure 2.

Type	Background	Mean age (Standard deviation)
Group 1 (adult): 25 native speakers of Bangla	Education: pursuing graduation	22.8 (1.74)
	Socio Economic Classification: B2-D2	
Group 2 (minors): 25 native speakers of Bangla	Education: pursuing school education	15 (1.24)
	Socio Economic Classification: B2-D2	

Figure 1: Participants' details

[2] http://imrbint.com/research/The-New-SEC-system-3rdMay2011.pdf

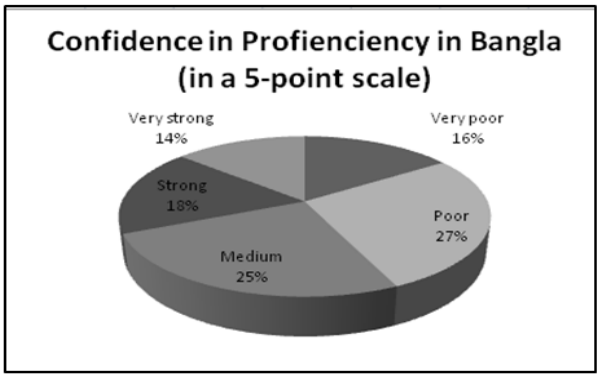

Figure 2: Proficiency in the mother tongue

4 Data preparation

From a Bangla corpus [3] of about 400,000 unique words, we have sampled 3500 words for the study. The words were selected in such a way that they represent the 'average' words over the corpus. The median values of word frequency distribution and length distribution lie at 368 and 5 respectively (refer to figur 3 for some sample words used in experiment). In a psycholinguistics, to preserve the experimental standard, it is essential to restrict the participants from making any strategic guess about the input stimuli. This has been achieved by randomly introducing non-words in between the valid words during the experiment. However, designing non-words are a non-trivial process, and often the reader's response to the different types of non-words opens up new insights into the process of word comprehension. Some examples of non-words are provided in figure 4.

5 Experimental Procedure

We have conducted lexical decision task (LDT) experiment (Meyer and Schvaneveldt, 1971) to study the visual recognition of Bangla words by native speakers. In this experiment, a participant is presented with a visual input, generally, a string of letters that can be words, non-words or pseudo words. Their task is to indicate, whether the presented stimulus is a valid Bangla word or not. The reaction time against each participant and the accuracy against each experimental stimulus across all the participants are recorded for further analysis. The time window for a user to submit any response has been set at 4 seconds, failing that a No Response is recorded. In either cases it is followed by hash signs (####) followed by the next letter string with 2.5 second delay. No response against a stimulus is automatically recorded as wrong response by the user.

Fifty users from the two target user groups participated in the LDT experiment. The 5000 experimental words (2500 words and 2500 non-words) were distributed randomly among 67 equal sized 75-word sets. Each user was presented maximum of three sets a day with at least one hour gap between two sets. Before recording experimental data, a sample set made up of 20 words was presented to the users to make them accustomed with the experiment.

6 Observations

All the incorrect responses and extreme reaction times (RT: the time taken to respond to a stimuli) have been discarded. Participants and experimental words having less than 70% accuracy have also been discarded. Finally, 440 words with RT of 42 (22 from group 1 and 20 from group 2) participants have been used for further study.

The RTs of each user have been normalized by z-transformation (Balota et al., 2007). The mean z-score over all users for a word has been computed. Negative z-scores indicate shorter response latencies. Paired t-test has been performed between results of the two user groups and $p < 0.05$ has been found signifying the difference between reading characteristics of the two user groups. Next, we have studied the influence of different word features on the outcome of the lexical decision task. The word features studied in this paper have been selected based on their prominence in the literature (Yarkoni et al., 2008) and their relevance with respect to Bangla. The features are:

•**Morphological Family size:** The morphological family size of a word **w** comprises of all the inflected, derived and compound paradigms that contains the word *w* (De Jong IV et al., 2000).

•**Word length (linear):** The length is measured in terms of the number of visual units or akshars; as Bangla belongs to the abugida group, mere alphabetic word length does not reflect the difficulty encountered in reading (Sinha et al., 2012b).

•**Number of complex characters in a word:** Complex characters are the consonant conjuncts or *jukta-akshars* present in a word.

•**Number of unique shapes in a word:**

[3]The Unicode corpus of Bangla was developed by the authors as a part of a broader study, the details are not in scope of this paper.

Category	Word (frequency, length in akshars, number of unique features)
High frequency long words (HL)	পরিবার (*paribAra*, family) (13403, 4, 6); আন্তর্জাতিক (*AntarjAtika*, international)(6936, 5, 13); তথ্যপ্রযুক্তি(*tathyaprayukti*, information technology) (2332, 5, 13), উল্লেখযোগ্য (*ullekhayogya*, worth mentioning)(1143, 5, 13)
High frequency short words (HS):	লঙ্ঘন (*la.nghana*, to cross) (541, 3, 7); অঞ্চল (*a~ncala*, region)(2163, 3, 8); বাড়ি (*bA.di*, house) (37984, 2, 6); কেন্দ্রীয় (*kendriya*, central) (22465, 3, 11) ;
Low frequency long words (LL):	নির্বাচকমণ্ডলী (*nirbAchakamandali*, selection committee) (74, 7, 13); স্বাতন্ত্র্যবোধ (*sbAtantrabodha*, feeling of freedom)(3, 5, 10);
Low frequency short words (LS):	দপ্তর (*daptara*, office) (10, 3, 8); পীড়িত (*pI.dita*, sufferer) (95, 3, 9); শিকার (*shikara*, hunt) (73,3, 7)

Figure 3: Examples of valid-words for experiment [4]

Type	Example
Proper non-word	ণজতথী (*NajatathI)*
Non-words from valid words	
Character deletion	রাজকীয় (*rAjakiYa*, royal) > রাজকী
Character insertion	রন্ধনশালা(*randhanshAlA*, kitchen) > রন্ধনশালাম
Jumbled	সংবাদপত্র (*sa.NbAdapatra*, newspaper) > সংদপবাত্র
Substitution by similar orthographic or phonological unit	তারিফ (*tAripha*, praise)> তারিপ (*tAripa*), চালান (*chAlAna*, transaction)> টালাণ (*TAlANa*)

Figure 4: Construction of non-words for experiment

Bangla script uses the space in a non-linear way and the akshars hangs from a distinct horizontal head-stroke called mAtrA. The letters are made up of combinations of different shapes or strokes. All together 57 unique strokes have been identified and indexed accordingly. The initial hypothesis is that more the number of distinct shapes in a word; the more difficult it is to comprehend.

•**Orthographic word complexity:** During visual word recognition, the reader has to recognize the orthographic patterns (Selfridge, 1958). Word level representations interact with the letter level representations i.e, the characteristic shapes or strokes (refer to). As no standard dataset on shape combinations in Bangla letters is available, the unique shapes or strokes have been identified intuitively across all the Bangla letters including the consonant conjuncts. The Bangla Akademi font has been considered as standard Bangla orthography. All together 57 unique strokes have been identified and numbered. Every Bangla letter has been represented as a combination of the constituent shapes. To capture the interactive nature

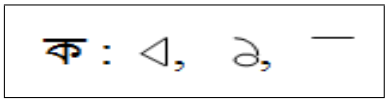

Figure 5: characteristics strokes of Bangla akshars

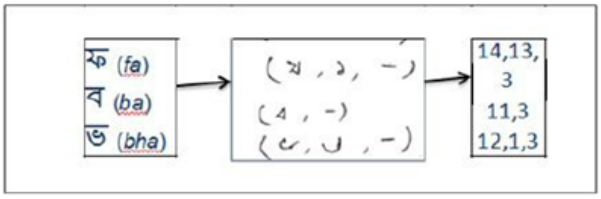

Figure 6: Mapping of Bangla akshars to characteristics shapes

of visual complexity, an orthographic complexity

model has been derived in the following way:

(a) The difficulty (d(s)) of a characteristic shape (i) or stroke is inversely proportional to its familiarity or frequency (f(s)). The frequency of the shapes has been calculated from the unique word list of the Bangla corpus without considering the frequency of each word.

$$d(s) = 1/f(s). \quad (1)$$

(b) The difficulty (d(a)) of an akshar (a) depends on the sum of the complexity of its shapes normalized by the number of shapes (n)

$$d(a) = 1/n \sum_i d(s_i) \quad (2)$$

(c) Finally, the difficulty (d(w)) of a word (w) is the sum of the complexity of its constituent akshars normalized by word length (l) and multiplied by the inverse of the word frequency (f(w))

$$d(w) = 1/f(w)\Big[1/l\big(\sum_j d(a_j)\big)\Big] \quad (3)$$

•**Orthographic & phonological neighborhood:** We have constructed akshar based, orthographic shape based and phonological pattern based neighborhood structure. The akshar based distance measure treats all akshars as of same visual complexity regardless of their orthographic properties, this is the reason distance among words based on orthographic strokes has been treated separately. At each level of orthographic information, the neighbors have been categorized into three groups based on their distance from the given word.

•**Number of syllables:** The syllabification of the Bangla words has been performed using a Bangla Grapheme to Phoneme conversion tool, developed inhouse.

•**Semantic neighborhood:** This measure represents the number of semantic neighbors of a word within the lexical organization of the language. This is computed from the semantic lexicon described in (Sinha et al., 2012a).

The mean and standard deviation values of the word features described above have been presented in figure 7. We have analyzed the RT corresponding to the above features using Spearman's correlation coefficient. The coefficient values between each word attribute and word recognition performance for the two user groups have been presented in figure 8.

From 8 we can observe that the correlation coefficients values for lexical decision latencies and decision accuracies are always less than 0.5, though they are different for different groups. The difference in the coefficient values may be attributed to the different reading patterns of the two groups. Number of syllables has similar correlation coefficients as word length because most often the akshars boundaries match the phonological syllable boundaries. The measure of orthographic word complexity possess low correlation coefficients with reaction times and accuracies, this can be an outcome of considering only the orthographic attributes of a word, isolating it from the phonological or semantic dimensions. In future, the measure needs to be augmented with those word features.

Number of unique shapes and complex characters also do not show significant correlation. Spelling to sound consistency also has a moderate correlation with the groups. This shows that speakers are not much sensitive towards the minor inconsistencies in spelling to sound mapping. The correlation coefficients of distant orthographic and phonological neighbors, immediate orthographic neighbors at shape level and semantic neighborhood are not significant for both groups. These indicate that after a threshold distance, the similarity or dissimilar-ity of the given word with other words in vocabulary does not affect the readers decisions. In addition, at shape level, the number of immediate orthographic neighbors may be unimportant due to the fact that often an akshar is constituted with more than 2 characteristic orthographic shapes and therefore, while reading, such minor changes in orthographic properties may go unnoticed.

Finally, the present calculation of semantic neighborhoods has been based on exhaustive language information (Sinha et al., 2012c), but the actual users may not possess such deep language knowledge and therefore are less affected by the semantic neighborhood structure. On the other hand, the number of senses or meaning of a word does not have inhibitory effect on the decision making process as the no ambiguity had to be resolved here, instead the use of a word in different

No.	Feature	Mean	Standard Deviation
1	Word frequency (log base 10)	1.81	1.09
2	Morphological family size	2.36	1.27
3	Word length (in akshar)	4.48	1.29
4	Number of complex characters	0.71	1.35
6	Number of unique shapes	10	3
7	Orthographic word complexity	-4.04	0.29
8	Immediate orthographic neighbors (akshar level)	2.11	0.16
9	Close orthographic neighbors (akshar level)	3.65	0.27
10	Distant orthographic neighbors (akshar level)	4.23	0.34
11	Immediate orthographic neighbors (shape level)	1.96	0.17
12	Close orthographic neighbors (shape level)	3.47	0.25
13	Distant orthographic neighbors (shape level)	4.71	0.41
14	Number of syllables	3.60	1.10
15	Immediate phonological neighbors	2.32	0.17
16	Close phonological neighbors	3.71	0.31
17	Distant phonological neighbors	4.33	0.36
18	Spelling to sound consistency	0.67	0.21
19	Number of senses of a word	0.13	0.06
20	Semantic neighborhood	18	6

Figure 7: Properties of valid words for experiment

contexts have increased its chance of encountering with the native readers of Bangla more often.

Moreover, the decisions against non-words are equally interesting to the decisions against the valid words. Non-words such as *kakShataNa* [correct: (katakShaNa, time duration)], *AkampIta* [correct: (akampita, steady)] and *TAlAN* [correct: (cAlAna, transaction)] have almost always been perceived as correct words by the readers due to their orthographic and phonological proximity to the correct words. On the other hand, proper non-word i.e, an arbitrary letter string such as *NajatathI* has been accurately classified as invalid. This indicates that the cognitive processes of reading are sensitive to the probability of what akshar pattern can occur in a valid Bangla word.

7 Conclusion

In this paper, we have presented a study on the comprehension difficulty of visual word recognition in Bangla stored as a lexical decision database. Number of interesting observations has been made from the experimental data and the observations have been complemented with rational inferences based on them. The correlation coefficients among word attributes and reaction time data has revealed that individually no feature has a large covariance factor, but the collective effect of all of them determines the cognitive load for comprehension. Moreover, using a reference language corpus based only on text from printed sources has proven to be a short-coming for drawing meaningful inferences. Some initial insights on the decisions corresponding to the non-words have also been presented.

References

David A Balota, Michael J Cortese, Susan D Sergent-Marshall, Daniel H Spieler, and MelvinJ Yap. 2004. Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133(2):283.

D.A. Balota, M.J. Yap, K.A. Hutchison, M.J. Cortese, B. Kessler, B. Loftis, J.H. Neely, D.L. Nelson, G.B. Simpson, and R. Treiman. 2007. The english lexicon project. *Behavior Research Methods*, 39(3):445–459.

M. Coltheart, B. Curtis, P. Atkins, and M. Haller. 1993. Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review; Psychological Review*, 100(4):589.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. Drc: a dual

Feature		Correlation coefficients (r)[significance p-value<0.05]			
No.	Name	Group 1		Group 2	
		RT-z	Accuracy	RT-z	Accuracy
1	Word frequency (log base 10)	-0.43	0.16	-0.39	0.19
2	Morphological family size	0.04	0.03	0.04	0.02
3	Word length (in akshar)	0.35	-0.09	0.39	-0.07
4	Number of complex characters	0.27	-0.16	0.35	-0.23
6	Number of unique shapes	0.28	-0.05	0.37	-0.18
7	Orthographic word complexity	0.19	-0.13	0.21	-0.17
8	Immediate orthographic neighbors (akshar level)	0.13	-0.14	0.17	-0.09
9	Close orthographic neighbors (akshar level)	0.04	0.02	-0.15	0.13
10	Distant orthographic neighbors (akshar level)	0.03#	0.02#	-0.04#	-0.13#
11	Immediate orthographic neighbors (shape level)	0.04#	0.01#	-0.11#	-0.07#
12	Close orthographic neighbors (shape level)	0.03	0.01	-0.14	-0.09
13	Distant orthographic neighbors (shape level)	0.03	0.02	-0.12	-0.06
14	Number of syllables	0.36	-0.11	0.37	-0.18
15	Immediate phonological neighbors	0.22	-0.11	0.19	-0.12
16	Close phonological neighbors	0.04	0.03	0.03	-0.07
17	Distant phonological neighbors	0.06#	0.02#	0.11#	0.08#
18	Spelling to sound consistency	-0.23	0.09	-0.34	0.13
19	Number of senses of a word	-.37	0.21	-0.41	0.23
20	Semantic neighborhood	-0.23#	0.09#	-0.32#	0.07#

Figure 8: Correlation analysis between word attributes and data from LDT (correlation coefficients marked with # are not significant (p-value $>$ 0.05)

route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

Nivja H De Jong IV, Robert Schreuder, and R Harald Baayen. 2000. The morphological family size effect and morphology. *Language and cognitive processes*, 15(4-5):329–365.

K. Diependaele, J.C. Ziegler, and J. Grainger. 2010. Fast phonology and the bimodal interactive activation model. *European Journal of Cognitive Psychology*, 22(5):764–778.

Kenneth I Forster and Elizabeth S Bednall. 1976. Terminating and exhaustive search in lexical access. *Memory & Cognition*, 4(1):53–61.

Stephen J Frost, W Einar Mencl, Rebecca Sandak, Dina L Moore, Jay G Rueckl, Leonard Katz, Robert K Fulbright, and Kenneth R Pugh. 2005. A functional magnetic resonance imaging study of the tradeoff between semantics and phonology in reading aloud. *NeuroReport*, 16(6):621–624.

Jonathan Grainger and Arthur M Jacobs. 1996. Orthographic processing in visual word recognition: a multiple read-out model. *Psychological review*, 103(3):518.

David E Meyer and Roger W Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227.

John Morton. 1969. Interaction of information in word recognition. *Psychological review*, 76(2):165.

Wayne S Murray and Kenneth I Forster. 2004. Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, 111(3):721.

David C Plaut, James L McClelland, Mark S Seidenberg, and Karalyn Patterson. 1996. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological review*, 103(1):56.

Mark S Seidenberg and James L McClelland. 1989. A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523.

Mark S Seidenberg. 2013. The science of reading and its educational implications. *Language learning and development*, 9(4):331–360.

Oliver G Selfridge. 1958. Pandemonium: a paradigm for learning in mechanisation of thought processes.

M. Sinha, T. Dasgupta, and Basu A. 2012a. A complex network analysis of syllables in bangla through syllablenet. In Sobha L Girish Nath Jha, Kalika Bali, editor, *Workshop on Indian Language and Data: Resources and Evaluation, LREC*, pages 131–138, May.

M. Sinha, S. Sharma, T. Dasgupta, and Basu A. 2012b. New readability measures for bangla and hindi texts. Communicated in the 24th International Conference on Computational Linguistics,2012, IIT Bombay, August.

Manjira Sinha, Abhik Jana, Tirthankar Dasgupta, and Anupam Basu. 2012c. A new semantic lexicon and similarity measure in bangla. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 171–182, Mumbai, India, December. The COLING 2012 Organizing Committee.

Marcus Taft and Gail Hambly. 1986. Exploring the cohort model of spoken word recognition. *Cognition*, 22(3):259–282.

T. Yarkoni, D. Balota, and M. Yap. 2008. Moving beyond coltheart?s n: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.

Jason D Zevin and David A Balota. 2000. Priming and attentional control of lexical and sublexical pathways during naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(1):121.