Deconstructing Transitional Probabilities: Bigram Frequency and Diversity in Lexical Decision

Russell Turk (russell.turk2012@my.ntu.ac.uk)[1]

Gary Jones (gary.jones@ntu.ac.uk)[1]

Duncan Guest (Duncan.guest@ntu.ac.uk)[1]

Angela Young (angela.young@ntu.ac.uk)[1]

Mark Andrews (mark.andrews@ntu.ac.uk)[1]

[1]Department of Psychology, Nottingham Trent University, 50 Shakespeare Street
Nottingham, NG1 4FQ

Abstract

Statistical learning paradigms traditionally use transitional probabilities as a measure of statistical distribution within a language. The current study suggests that alternative metrics may exist that can account for differences in language processing ability. Two primed lexical decision tasks are used to examine the effects of bigram frequency and diversity on speed and accuracy of word recognition. It is demonstrated that both frequency and diversity contribute to word recognition performance; findings and theoretical implications are discussed.

Keywords: Statistical learning; lexical decision; language

Introduction

Humans are superlative learners capable of identifying and tracking patterns in their environment, both implicitly and explicitly. This ability has been investigated using both implicit and, more recently, statistical learning paradigms (Perruchet & Pacton, 2006) across a number of different domains including shapes (Kirkham, Slemmer, & Johnson, 2002), music (Daikoku, Yatomi, & Yumoto, 2014; Koelsh et al., 2016; Saffran et al., 1999), tactile stimuli (Conway & Christianson, 2005) and, most prominently, language acquisition (Newport & Aslin, 2004; Saffran, Aslin, & Newport, 1996; Thiessen, & Erickson, 2013; Vouloumanos, 2008) highlighting the ability of learners, ranging from infant (Saffran et al., 1996) to adult (Koelsh et al., 2016), to track the transitional probabilities (TPs) within a given set of stimuli.

Over the past two decades a plethora of researchers have investigated this phenomenon and have found transitional probabilities to be a robust indicator of performance across a number of different tasks and languages (e.g. Liu & Kager, 2011; Toro, Sinnett, Soto-Faraco, 2005). This has led to the acceptance of TPs as the standard metric of co-occurrence within natural (and artificial) languages. However, if we consider that the TP of any given stimulus stems from an interaction between the frequency of sequence XY and the number of potential candidates for Y then we are presented with two alternative metrics of statistical distribution. These, in turn, can be used to investigate the types of statistics which learners can attend to.

When applied to words in natural language these metrics can be termed *Bigram Frequency*, which is equal to the total number occurrences for a given sequence of two words within a language or representative selection thereof; and *Bigram Diversity* which can be defined as the number of items that potentially follow word X in the sequence XY.

It is logical to presume that both bigram frequency and diversity would be predictive of performance in language-related tasks. Evidence from Freudenthal et al. (2015) demonstrates that a frequency-based chunking mechanism can successfully reduce output errors in children's speech. This suggests that learners can track not only the TPs of the bigrams but also the frequency with which they occur. No evidence yet exists for a diversity-driven account of language proficiency. Nonetheless, it is recognised that predictability is an important facet of language processing (Bates & MacWhinney, 1987; Glenberg & Gallese, 2012; Goldberg, Casenhiser, & Sethuraman, 2005; Pickering & Garrod, 2004, 2007; Van Berkum et al., 2005); it follows therefore that a larger number of potential competitors for stimulus Y would serve to reduce predictability and thereby prove detrimental to response fluency.

Historically statistical learning paradigms such as those developed by Jusczyk and Aslin (1995, also Saffran, Aslin & Newport, 1996) have exposed learners to artificial languages with carefully built-in TPs. This allows for admirable control of the input at the expense of both diversity and complexity. It has been argued that these languages are too simplistic to assess the extent to which learners are able to process distributional statistics within natural language (Frank et al, 2010; Johnson & Tyler, 2010). To highlight this point, Saffran et al. (1996) reported

inter-syllable TPs of 1.0 in their seminal study whereas naturally occurring TPs are often considerably lower (the bigram *little baby* has a TP of less than 0.002).

Thus, the true test of statistical learning theories is their application to a more naturalistic dataset, one which retains the complexity and diversity of natural language whilst allowing for the accurate tracking of distributional cues; natural language corpora represent such datasets. The British National Corpus (BNC) is a collection of contemporary natural language which comprises approximately 100-million words of written and spoken British English drawn from a variety of sources ranging from telephone calls to academic journals. By analysing the distributional statistics within the BNC it is possible to present learners with verisimilar but also quantifiable samples of natural language.

This raises another issue however, in that learners already have a great deal of experience interacting with natural languages. This makes traditional methods of testing such as those used by Saffran et al. (1996, also Frank et al., 2010; Jusczyk & Aslin, 1995) unsuitable for natural language stimuli. Thus, two solutions are immediately apparent; the use of unfamiliar or non-native languages or an alternate method of assessment. Non-native languages would seem to be the ideal solution except that the complexity of these languages means that learners require either long periods of familiarisation or simplified samples in order to obtain actionable data. It is therefore favourable to introduce an alternate measure of language proficiency whilst retaining the complexity of the language and avoiding a lengthy familiarisation process.

The current study seeks to address this issue by assessing language proficiency using a primed lexical decision task (LDT) where the first word of a bigram acts as the prime and the second word the target. It is predicted that, using bigram frequency and diversity as statistical primes, response time for stimuli Y will be predicted by the strength of its association with prime X. Based on this prediction two hypotheses are proposed:

H1: Response times on a LDT will be quicker when primed with high frequency bigrams compared to low frequency or non-bigrams, and

H2: Response times will also be quicker when primed with low diversity bigrams compared to high diversity or non-legal bigrams

Method

Participants

Thirty-one participants (25 females) aged between 18 and 41 years (M= 20.77, SD= 4.17) were recruited from Nottingham, UK. All participants reported English as their first language and were screened for language difficulties. Participants took part in both experiments and received research credits in exchange for their participation where applicable. An *a priori* power analysis showed that a sample of at least twenty-four participants was necessary to achieve statistical power of above .8.

Experiment One

Design

Experiment one used a LDT to assess the extent to which bigram frequency affects word recognition. The aim of the experiment was to identify any statistical priming effect that may result from high frequency word pairs within natural language.

Materials

Three 30-item lists were generated using bigrams found within the BNC in addition to one 90-item non-word list which was created using entries from the ARC Non-word database (Rastle, Harrington, & Coltheart, 2002). The BNC contains only samples of British English which increases its validity as a natural language representation for a UK sample.

Bigrams were extracted from the BNC by using a python script to parse the .xml version of the corpus into word pairs before writing them to a database and tallying the number of occurrences. This resulted in a list of 12,293,349 *unique* bigrams. A further script was used to remove any bigrams with a frequency of less than 0.1 per million since these were considered too infrequent to provide meaningful data. The remaining corpus was then filtered to exclude any bigrams containing acronyms, initialisations, contractions, hyphenations, non-standard or non-English words, names, numbers expressed as digits, or words with fewer than three letters.

Table 1: Diagnostic means and standard deviations for target words

	Bigram Type	Log(Frequency)	Concreteness	Letters	Phonemes
Experiment one	High Frequency	3.20 (3.86)	3.10 (1.09)	5.01 (1.48)	4.10 (1.24)
	Low Frequency	3.21 (3.87)	3.11 (1.41)	4.92 (1.41)	3.97 (1.18)
	Non-Bigrams	3.20 (3.86)	3.10 (1.08)	5.00 (1.48)	4.05 (1.27)
Experiment two	High Diversity	2.08 (0.27)	2.85 (0.98)	5.43 (1.17)	4.33 (1.15)
	Low Diversity	2.01 (0.49)	3.96 (1.00)	5.00 (0.88)	4.00 (0.96)
	No Diversity	2.18 (0.02)	3.21 (1.03)	5.06 (1.26)	4.20 (1.19)

Stimuli lists were organised according to the frequency with which the bigrams occur within the BNC; the three lists contained bigrams of high (>100 occurrences) or low frequency (<20 occurrences), or bigrams consisting of words that do not appear together in the BNC. A number of metrics were obtained for each of the bigrams including word frequency (http://ucrel.lancs.ac.uk/bncfreq/flists.html), concreteness (Brysbaert, Warriner, & Kuperman, 2014), number of letters, and number of phonemes. Due to the nature of the sample exact matching across conditions was impossible without compromising the number of available bigrams, however word lists were balanced so as to not differ significantly on any of these characteristics (each $p > 0.05$); list diagnostics are presented in table 1. Individual word frequencies were log-transformed. Examples of stimuli can be seen in Table 2.

Table 2: Example stimuli for experiment one

Bigram Type	Example Stimuli (prime target)
High Frequency	recent times; last night; other hand
Low Frequency	craggy face; local access; time across
Non-bigrams	oval hipster; meet gone; chilli call

Procedure

Participants were presented with letter strings and were asked to indicate whether the string constituted a real English word by pressing either 'z' or 'm' on a standard QWERTY keyboard; key mapping was systematically varied so that half of all participants used 'z' to indicate a word and 'm' to indicate a non-word whilst half responded with 'm' for words and 'z' for non-words. Strings were presented for a maximum of 1500ms and were immediately preceded by a 75ms prime. All prime-target pairs mapped exactly onto bigrams from the stimuli lists whereby the first word of the bigram acted as a prime and the second word as the target. A fixation point was presented in the centre of the screen for 500ms prior to each trial. Prime-Target pairs were presented in two blocks each containing fifteen low-frequency bigrams, fifteen high-frequency bigrams, fifteen non-bigrams, and forty-five non-word trials. The order of presentation for both blocks and trials was randomised for each participant.

Analysis and Results

All participants scored more than 80% on the LDT. Data was then trimmed to exclude incorrect responses as well as those made faster than 200ms, slower than 1500ms (Perea et al., 2016), or more extreme than three standard deviations from the participant's mean (Madan et al., 2016), following this procedure 2.29% of correct trials were removed across participants.

All response time data were log-transformed; data was then analysed categorically using a repeated-measures analysis of variance to identify any differences in response time between the high and low frequency bigrams (M = 6.310, SD = .080), non-bigrams (M 6.507, SD = .101) and non-words (M = 6.548, SD = .130).

Bigram frequency had a significant effect on response time, $F(3,28) = 53.759$, $p < .001$, $\eta_p^2 = .852$. Post hoc pairwise comparison using Bonferroni correction show that words in the non-bigram condition were recognised more slowly than those in both the high ($p < .001$) and low ($p < .001$) bigram frequency conditions. There was no difference between high and low frequency bigrams ($p = .305$). Non-words were recognised more slowly than words in the high frequency ($p < .001$), low frequency ($p < 0.001$), and non-bigram conditions ($p < .038$). Figure 1 illustrates these differences.

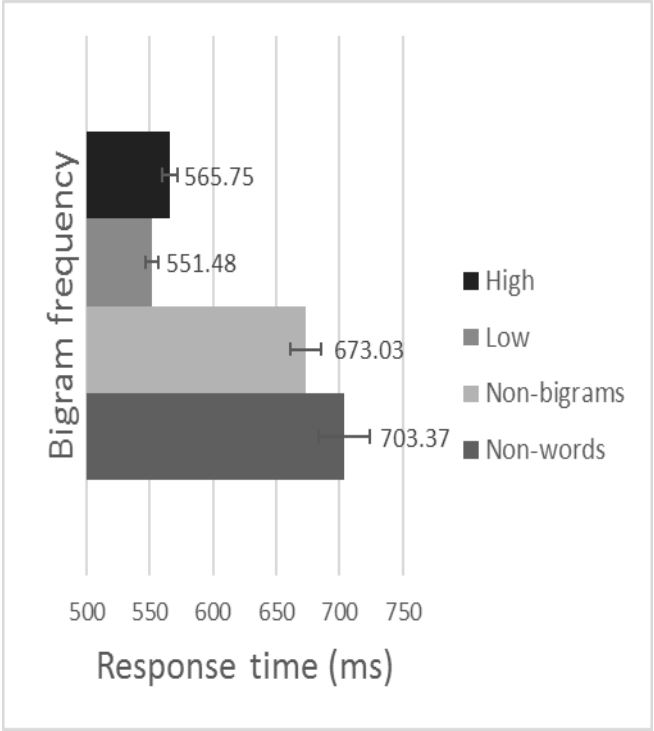

Figure 1: Non-transformed group means for bigram frequency, bars depict standard error

A further repeated-measures analysis of variance was also conducted to assess any differences in response accuracy between the four conditions. Response accuracy also shows an effect of bigram frequency, $F(3,28) = 6.796$, $p = .001$, $\eta_p^2 = .421$. Post hoc analyses using Bonferroni correction show that participants responded less accurately to words from the non-bigram condition than those in the high ($p = .005$) or low ($p = .002$) frequency conditions. All other comparisons were non-significant (each $p > .062$). Figure 2 shows means and standard error for accuracy.

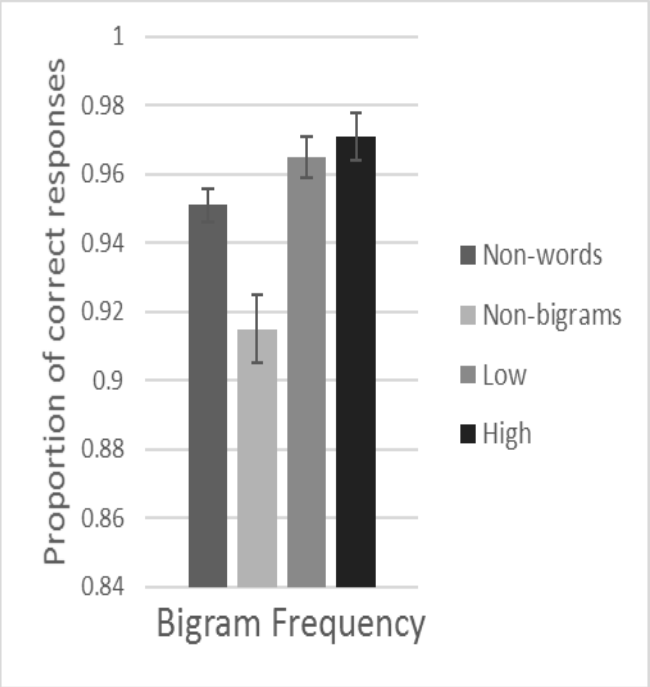

Figure 2: Proportion of correct responses by group, bars depict standard error.

Experiment Two

Design

Experiment two used a LDT to assess the extent to which bigram diversity affects word recognition. The aim of the experiment was to identify any statistical priming effect that may result from the predictability of the second word in a bigram given the diversity of the first.

Materials

Stimuli were obtained and processed using an identical procedure to experiment one with the exception that the word lists were organised according to high (>100 potential followers) or low (<2 potential followers) diversity or bigrams consisting of primes that do not have followers within the BNC. Word lists were balanced in the same way as the first experiment, each p >0.06 with the exception that the low diversity list differed significantly from both the high and no diversity list on concreteness (high: p <0.01, no: p <0.01); this is due to the relative scarcity of low diversity bigrams within the BNC and the theoretical decision to prioritise controlling individual word frequency since this represents the largest predictor of word recognition performance (Brysbaert & New, 2009; Ferrand et al., 2010; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers et al., 2012; Yap & Balota, 2009). List diagnostics are presented in table 1. Individual word frequencies were log-transformed. Example stimuli can be seen in Table 3; none of the bigrams were repeated across the two experiments.

Table 3: Example stimuli for experiment two

Bigram Type	Example Stimuli
High Diversity	that place; with number; this ancient
Low Diversity	revolve around; beady eyes; gilded cage
No-Diversity	yonder month; ribbed final; orate red

Procedure

The experimental procedure was identical to that used in the first experiment.

Analysis and Results

All participants scored more than 80% on the LDT. Data was trimmed in the same way as the first experiment and a total of 2.04% of correct trials were removed. Response time data was log-transformed.

Data was analysed categorically using a repeated-measures analysis of variance to identify any differences in response time between the high (M = 6.375, SD = .054), low (M = 6.395, SD = .059) and no diversity (M = 6.422, SD = .581) bigrams as well as non-words (M = 6.548, SD = .130); means and standard error can be seen in Figure 3.

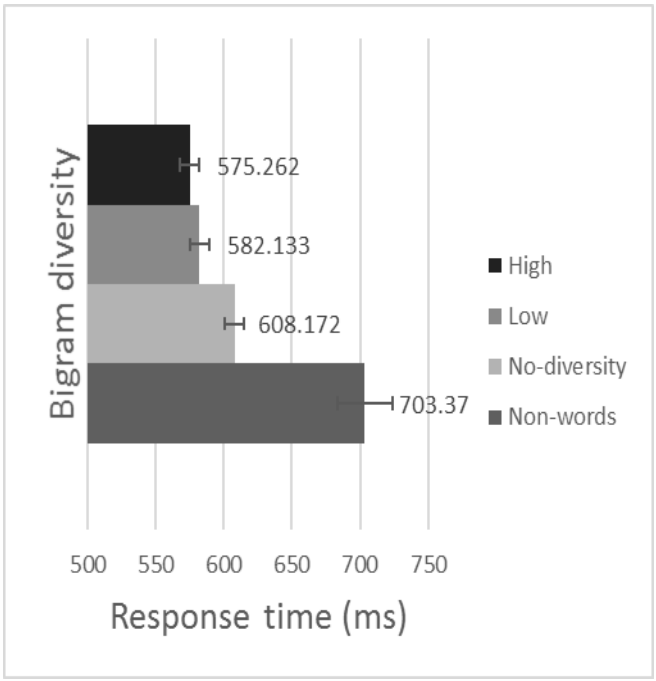

Figure 3: Non-transformed group means for bigram diversity, bars depict standard error

Bigram diversity had a significant effect on response time, $F(3,28) = 35.932$, $p < .001$, $\eta_p^2 = .794$. Post hoc pairwise comparison using Bonferroni correction shows that non-words were recognised more slowly than those in high ($p < .001$), low ($p < .001$) and no diversity ($p < .001$) conditions. Words in the no diversity condition were also recognised more slowly than those in both the high ($p = .007$) and low ($p = .011$) diversity conditions; there was no significant difference between high and low diversity bigrams ($p = .261$).

Bigram diversity had no effect on response accuracy, $F(3,28) = 1.486$, $p = .208$.

Comparison with Transitional Probability

Transitional probabilities were calculated for all bigrams using the formula:

$$P(Y|X) = \frac{P(Y, X)}{P(X)}$$

Where Y is the target stimulus and X is the initial word of a given bigram pair.

An item analysis was then run using a multiple linear regression with data from both experiments to assess the relationship between response time (log-transformed) on a LDT and the three key variables bigram frequency, bigram diversity, and transitional probability, $F(3, 168) = 2.937$, $p = .035$. Individual coefficients (see Table 4.) indicate that bigram frequency represents the strongest predictor of word recognition performance; neither bigram diversity or TP were significant predictors of response time.

Table 4: Coefficients and *p*-values

	Beta	Stand. Beta	P
Bigram frequency	-2.51e-5	-.189	< .016
Bigram diversity	-1.21e-5	-.060	.456
Trans. Probability	-.038	-.107	< .169

Discussion

The current study aimed to assess whether bigram frequency and bigram diversity would have an effect when used as primes in a LDT. Findings from the categorical analyses suggest a binary interaction between bigram frequency and response time where naturally occurring bigrams are recognised significantly more quickly than illegal bigrams or non-words. The same is also true for bigram diversity.

This suggests that any amount of exposure to a language is beneficial regardless of the frequency or diversity of individual structures within the input. This is an interesting effect which may have been overlooked by previous studies that have focussed on TPs since the methodologies employed tend to focus on recognition of familiar versus unfamiliar strings. It could be argued however that the bigram frequencies presented in the current study, although highly infrequent, do not accurately represent the extremes of low frequency within the BNC. It is therefore suggested that further investigation needs to access frequencies of less than 0.1 per million in order to identify the absolute minimum amount of exposure required to elicit statistical priming effects.

Comparison of the key predictors also suggests that bigram frequency outperforms TPs as a predictor of response time in a statistically primed LDT. This can be attributed to the lower computational costs associated with tracking bigram frequency compared to the calculation of TPs. To the authors knowledge, the current study is the first to assess statistical learning using a LDT. These findings should therefore be interpreted with caution until they can be demonstrated in alternative paradigms.

It is proposed that the findings presented are evidence for the use of metrics other than TP in statistical learning paradigms, particularly when applied to natural language where TPs tend to be very small. A case can also be made that LDTs are a viable paradigm for the investigation of statistical effects in natural language where traditional recognition tasks may not be appropriate.

Crucially, they suggest that theories of statistical learning can deal with the scale-up in variety and complexity that comes from moving between artificial and natural languages. This begins to address one of the most fundamental criticisms of statistical learning theory.

When interpreting the data presented herein it would be prudent to consider that, although the BNC constitutes a multifarious selection of British English it does not encapsulate the entirety of written and spoken language. It is therefore posited that any findings presented be considered as representative rather than absolute in their accuracy. Future investigation should include the analysis of alternate corpora in order to ensure that any results are not artefactual in nature.

It is recognised that neither bigram frequency or diversity represent a complete account of statistic learning, nor is it suggested that learners utilise these metrics in place of TPs. Rather, it is posited that bigram frequency and, to a lesser extent, diversity constitute ‘another brick in the wall’ which may one day lead to a comprehensive understanding of how humans process language.

In conclusion, the current study demonstrates that individuals are capable of using bigram frequency and diversity to respond to statistical primes in a lexical decision task and that these metrics may be comparable to transitional probabilities when applied to natural language.

Acknowledgments

Data cited herein have been extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

References

Bates, E., & MacWhinney, B. (1987). Competition, variation, and learning. In B. MacWhinney (Ed.), Mechanisms of language acquisition. Mahwah, NJ: Erlbaum.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods, 46*, 904-911.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of experimental psychology. Learning, memory, and cognition, 31*, 24-39.

Daikoku, T., Yatomi, Y., & Yumoto, M. (2015). Statistical learning of music- and language-like sequences and tolerance for spectral shifts. *Neurobiology of Learning and Memory, 118*, 8-19

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Meot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*, 488-496.

Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2016). Developmentally plausible learning of word categories from distributional statistics. In Proceedings of the 38th *Annual Conference of the Cognitive Science Society*, 674-679.

Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex, 48*, 905–922.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2005). The role of prediction in construction-learning. *Journal of Child Language, 32*, 407-426.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*, 1-23.

Keuleers, E., Diependaele, K. & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology* 1:174.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*, 287-304.

Kirkham, N., Slemmer, J., & Johnson, S. (2002). Visual statistical learning in infancy: Evidence for a domain general mechanism, *Cognition, 83,* B35-B42

Koelsh, S., Busch, T., Jentschke, S., & Rohrmeier, M. (2016). Under the hood of statistical learning: A statistical MMN reflects the magnitude of transitional probabilities in auditory sequences, *Scientific Reports, 6.*

Liu, L., & Kager, R. (2011). How do statistical learning and perceptual reorganization alter dutch infants perception to lexical tones. In Proceedings of the 17th International Congress of Phonetic Sciences, 1270-1273.

Madan, C. R., Shafer, A. T., Chan, M., & Singhal, A. (2016). Shock and awe: Distinct effects of taboo words on lexical decision and free recall, *The Quarterly Journal of Experimental Psychology*, 1-18

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*, 127-162.

Perea, M., Marcet, A., Vergara-Martínez, M., & Gomez, P. (2016). On the Dissociation of Word/Nonword Repetition Effects in Lexical Decision: An Evidence Accumulation Account. *Frontiers in psychology*, 7.

Perruchet, P., Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches, *Trends in Cognitive Sciences, 10*, 233-238.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? Trends in Cognitive Sciences, 11, 105-110.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology, 55A*, 1339-1362.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* (New York, N.Y.).

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27-52.

Thiessen, E. D., & Erickson, L. C. (2013). Discovering words in fluent speech: The contribution of two kinds of statistical information. *Frontiers in Psychology*, 3

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97*.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 443.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory & Language, 60*, 502-529.