Involuntary Language Processing and Lexical Features' Effects

Eda Eylul Yagci
STUDENT NUMBER: 2019662

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:
Dr. Giovanni Cassani
Dr. Bruno Nicenbaum

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
June 2022

Preface

Writing this thesis was certainly challenging, yet rewarding . I got to conduct and supervise an experiment and learned to utilize software I wasn't very familiar with before and I believe that it will definitely come in handy in the direction I want to proceed in. I would first like to thank my supervisor Dr Giovanni Cassani for their valuable insight and support, and Dr Bruno Nicenbaum for their help in building the experiment. While writing this thesis, I was once again reminded that the problems that appear intimidating, usually are fairly easy to solve once you can get down to starting them. And for helping me feel motivated enough to handle the issues I faced while writing this thesis, I would like to thank the unconditional love and support of my parents and friends.

Eda Eylul Yagci
June 2022

Involuntary Language Processing and Lexical Features' Effects

Eda Eylul Yagci

While the lexicon's role in reading is indisputable, the effects of its lexical factors aren't explored completely. Thus, this thesis aimed to investigate the effects of lexicality and other lexical features by conducting a simple reaction time experiment. To do so, the feature values of the experiment were extracted from a mega-scale lexical study called the Dutch Lexicon Project, where the feature values of 14.000 words and 14.000 pseudowords were present. The experiment was conducted on 6 Tilburg University students and they were asked to react to the stimulus as fast as possible without making any lexical judgements. The results revealed that while the participants did not have a major difference in their reaction times between words and pseudowords, some of the lexical features of both of them still played a role even when told not to make lexical judgments. The effects of certain lexical features, how they interact with each other, and the effects of their interactions are further discussed.

1. Introduction

Language processing is the cognitive function of understanding the context of audible or visual information that is being perceived. For adults, besides imagery, one of the most common ways of perceiving visual information is achieved via reading, which, for the larger part of the literate adult population, does not require a lot of effort. However, this process that can feel effortless can be further examined to reveal insight in regards to psycholinguistic questions and problems.

Computational linguistics is an interdisciplinary field that applies techniques of computer science to analyze and synthesize language and speech. Applying computational linguistics to the aforementioned psycholinguistic issues can aid in drawing up or confirming conclusions and relations that have not been made previously.

This study focuses on whether visual language processing can occur involuntarily by conducting a psycholinguistic simple reaction time experiment. It builds on a previously conducted lexicon project called the Dutch Lexicon Project where reaction times for accurately differentiating Dutch words from Dutch pseudowords (strings with no particular meaning which resemble the words of the language they are constructed in) are explored **(Keuleers, E., Diependaele, K., & Brysbaert, M, 2010).** The simple reaction time experiment, however, is interested in the reaction times to any visual stimuli, including words and pseudowords,

regardless of their lexicality. While this simple reaction time experiment and the project conducted by Keuleers and Brysbaert differ on the approach of reacting to stimuli, both are interested in the effects of the lexical features.

These lexical features can consist of how often a word is found in a corpus or daily language, how many contexts the word can be used in, the length of the word, how many operations the word requires to transform into another word and so on.

There have been several studies that explore lexical features, some of these studies have been examining the effects of these lexical features on pseudowords as well. While finding how common a pseudoword is in daily language or how many different contexts it is spoken in can be challenging, the fact that pseudowords still inhabit certain lexical features like length and edit distance(operations need to be done on a string for it to reach its target string) makes it possible to explore linguistic relations in regards to pseudowords themselves or actual words.

Psycholinguistic experiments that involve visual text stimuli are generally conducted in a way that requires their participants to make lexical decisions based on the stimuli they are reacting to. However, in this simple reaction time experiment, by telling participants to disregard lexicality while observing their reaction times to stimuli, the experiment eliminates the participants' need to make active linguistic judgments. By applying this approach, the experiment aims to not only look for an overall difference in involuntary reactions to words and pseudowords but also aims to investigate more automated reactions concerning the lexical features of the text stimuli.

1.2 Research questions

This study therefore focuses on answering the following research questions:

1. Is there a systematic difference in reaction times when participants are asked to react to words and non-words, disregarding their lexicality?
2. Do the lexical features of words and non-words, such as length, OLD20, bigram frequency, length or others affect the reaction times of participants when participants are asked to react to words and non-words, disregarding the lexicality of the stimuli?
3. Do the lexical features of words such as frequency, length and contextual diversity affect the reaction times of participants for words only, when participants are asked to react to words and non-words, disregarding the lexicality of the stimuli?

1.3 Findings

A systematic difference between reaction times to words and pseudowords was not observed in the scope of this experiment. However, lexical features still proved to have an impact on the reaction times. Lexical features more concerned with the physical structure of the string were more impactful, with string length leading. Unlike most lexical decision tasks, lexical features composed of values assigned from their use in language, like contextual diversity and word frequency did not play a very distinct role, showing that the participants were able to react to the stimuli without reading them.

Related work

The usage of pseudowords in psycholinguistics is not necessarily a recent addition. In the context of psycholinguistic experiments, pseudowords can sometimes also be referred to as wug words. This alias for pseudowords traces back to a psycholinguistic experiment conducted in 1958, in which psycholinguist Jean Berko Gleason presented children with pseudowords like “wug” and asked the children to perform linguistic decisions, like pluralising or transforming the stimuli into past tense **(Gleason, 1958)**. The results revealed that even younger children were capable of correctly producing past tenses, (although better performance for simpler past tense transformations, like -ed, were observed), plurals and other linguistic interactions, demonstrating an internalization of the linguistic rules. "However, it is important to note that the process of generating these pseudowords plays a significant role in the research due to the inherent differences among languages." For example, languages can differ in their syllabic structure and general typology, which could require different rule systems of creating pseudowords (**Klafehn, T. 2011)**.

The pseudowords used in the simple reaction time experiment are taken from a much larger study, the Dutch Lexicon Project, where reaction times to accurately differentiate between 14.000 words and 14.000 pseudowords were measured. The pseudowords were generated using Wuggy, which is a multilingual pseudoword generator **(Keuleers, E., & Brysbaert, M., 2010)**. The pseudoword generator was in fact built by Keuleers and Braysbert to use in their studies such as the Dutch Lexicon Project, as by then, they found the available pseudoword generators to be limiting. The Wuggy algorithm was constructed in a way that allowed for easier generation of polysyllabic strings(strings with more than one syllable) by implementing built in restrictions. The algorithm, which is still available today, thus allowed them to generate the 14.000 pseudowords differing in characteristics(length, syllable formation, with mostly disyllabic strings) that were corresponding to the 14.000 words in their structure.

Lexical decision tasks are procedures that allow the analysis of lexical items such as words and pseudowords by evaluating their lexical formation and access. A basic lexical decision task is identifying words from pseudowords, which has been a point of interest in lexicon projects. These lexical decision tasks are usually followed by a more in-depth analysis of the lexical features of the strings. Although strings can have many lexical features, the more commonly investigated lexical features in analysis for lexical decision tasks have been word frequency, contextual diversity, ngrams, word length, and edit distance.

Word frequency is how often a word is found in a corpus, and while the sizes of corpora can differ and the sizes of some corpora might be unknown, the lexical feature also has ways of standardizing its measurement. Contextual diversity is a measurement of the number of contexts a word has been used in. A high correlation is often found between the two features, as words that occur more often have a higher likelihood of occurring in more contexts **(Plummer, P., Perea, M., & Rayner, K. 2014)**.

Word frequency has empirically been accepted amongst the strongest influencers in lexical decision and word naming tasks. However, more recent studies have proposed that word frequency is confounded with contextual diversity. Despite the high correlation between the two features, contextual diversity has been shown to have better predictive leverage over word frequency in isolated word identification tasks as well as more predictive leverage over reaction times in word naming and lexical decision tasks **(Adelman, J. S., Brown, G. D. A., & Quesada, J. F. 2006)**. While contextual diversity might perform better as a predictor in certain tasks, word frequency's effect is not completely disregarded. This is due to the predictive power of word frequency being determined consistently larger in lexical decisions than in word naming, yet still present in both tasks **(Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. 2004).**

While it may be more difficult to measure the word frequency and contextual diversity of pseudowords as they can not be found in a corpus, correlations between reaction times to the pseudowords and the word frequency of the word it was generated from have been determined by researchers **(Perea, M., Rosa, E., & Gómez, C. 2005)**. The generation of the pseudoword was found to still play a role in reaction times; pseudowords that were generated by transposing two adjacent internal letters showed a larger effect in comparison to pseudowords generated by replacing one internal letter of the base word.

However, pseudowords do have lexical features that can be measured independently from their base word, such as their length and edit distance. Edit distance is a measure that compares sequences and counts how many edits(insertion, deletion, and/or substitution) are required for one sequence to reach the other sequence. While there have been discussions of correctly measuring edit distance, as it isn't applicable to some languages, and different tasks can require different methods of measurement, edit distance continues to be utilized in psycholinguistics and can provide insight about strings. This can include the strings' base word frequency, the base word syllable counts as well as how orthographically distinct the strings are (**Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J., 2015)**.

While lexical decision tasks can offer many advantages and insights regarding linguistics, how they are constructed can alter the results of the experiment. With an increase of large-scale lexical decision studies being carried out **(see, e.g., Balota et al., 1999, 2004, 2007; Keuleers et al., 2010, 2012; Ferrand et al 2010)** it is possible to track how the properties of the lexical features differ from each other in time, such as earlier experimenters having to use less sophisticated methods of pseudoword generation and alternate measures for word frequency being implemented in more recent studies. The difference in these properties could lead to the results of the experiment being affected substantially **(Lieber, R., Štekauer, P., & Baayen, H. 2014)**. As lexical decision tasks ask their participants to react to isolated words, it isn't possible to analyze how reaction times to the word could differentiate in certain sequences and contexts.

Another limitation stems from the fact that the participants are asked to make metalinguistic judgments, meaning that as they are actively focused on the

linguistic task, the lexical judgements they make throughout the experiment can differ from the lexical judgment they might make in their daily lives. The simple reaction time experiment attempts to overcome this limitation by asking the participants to not engage in any lexical decision making; participants are told to react as soon as they see the stimuli. The simple reaction time experiment can thus investigate beyond the aspects that are limited by lexical decision tasks: It can explore whether or not participants are able to actually react without making lexical judgements by exploring the reactions between pseudowords and words, and then can look into the effects of the mentioned lexical features when no metalinguistic judgment is required.

Methods

3.1 Initial Data and Variable Explanations

The simple reaction time experiment was carried out in these 3 main steps: Gathering and pre-processing the initial data, constructing data frames from the pre-processed initial data to build the experiment, and pre-processing and analyzing the data that the completed experiment generated. The unprocessed initial data was gathered online from the publicly available Dutch Lexicon Project **(Keuleers, E., & Brysbaert, M., 2010)**.The Dutch Lexicon Project was a megastudy conducted with 39 participants(to completion), that investigated the reaction times for accurately identifying 14.000 words from 14.000 pseudowords in 50 blocks. The data index of DLP consisted of the project's items, stimuli and trials. As the Dutch Lexicon Project was a mega-scale lexicon study only the features necessary for constructing the data frame and analyzing the results were kept.

The pre-processing and combining of the data was done using the programming Language R **(R Core Team, 2020)**.The variables retained from the items file consisted of the spelling and the lexicality of 28.000+ strings. The variables retained from the trials file were the spelling of the strings, participant numbers and trial counts. Finally, from the stimuli file, the variables retained were again the spelling of the strings, alongside their "OLD20", "subtlex.frequency", "subtlex.cd", "summed.bigram", and "nchar" values. The retained variables from all three files were then merged into a joint dataset in R, using spelling as the merging item.

The "OLD20" variable stands for Orthographic Levenshtein Distance 20 and measures a string's average Levenshtein distance of its 20 nearest neighbors. Levenshtein distance is a string metric for measuring the minimum number of character edits- such as inserting, deleting or substituting a letter- to change the string at hand to the target string. "subtlex.frequency" and "subtlex.cd" both start with "subtlex" as they were measured using the SUBTLEX-NL which is a large database containing multiple word frequencies of 44 million dutch words, gathered from the Dutch film and television subtitles. "subtlex.frequency" measures a word's frequency, which is how often the word can be found in the database. In the meantime, "subtlex.cd", with cd standing for contextual diversity, measures how many different contexts the word can be found in. "summed.bigram" is the total measurement count of the frequency distribution of every bigram in a string. A bigram is every adjoining 2-letter unit of the string. Finally, "nchar" signifies the length of the string.

While measures like OLD20, bigram frequency and length can apply to both words and pseudowords, the word frequency and contextual diversity counts of pseudowords can not be measured. Thus, two different datasets were created for research questions 2 and 3.

Table 1 : Statistical Values of Variables Used in the Simple Reaction Time Experiment for Research Question 3

	Min	**Max**	**Mean**	**SD**
General Fields				
Length	3	9	6.2	1.44
Contextual Diversity	1	7729	702.794	1,715.07
Word Frequency	1	45,392	2,213.27	7,744.63
Logged Contextual Diversity	0.30	3.88	1.96	0.93
Logged Word Frequency	0.31	4.56	2.10	1.02
Orthographic Neighborhood				
OLD20	1.0	3.6	1.9	0.55
Bigram Frequency				
Summed Bigram Frequency	31.989	528	212.574	114.87
Logged Bigram Frequency	1.50	2.72	2.25	0.26
Experiment Variables				
Reaction Times	104.182	803.91	207.108	87.604
Trial	3	239	3.587	70.07

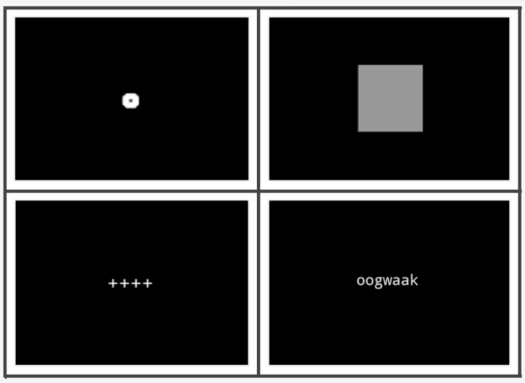
3.2 **Building the Experiment**

The construction of the simple reaction time experiment was carried out using OpenSesame **(Mathôt, S., Schreij, D., & Theeuwes, J. 2012),** which is a program used in the development of behavioral experiments for psychology. The experiment consisted of 2 practice blocks, followed by 6 trial blocks. The practice blocks consisted of 15 elements each, while the trial blocks contained 60 elements each. The first practice block, alongside the first and last trial blocks contained squares of differing intensities to control for the reaction times of the participants. The intensities of these squares were set using shades of black and white, in expectancy of faster reaction times to whiter squares as the experiment had a black background. The remaining blocks contained the text stimuli. For text stimuli, the string counts were divided evenly between words, pseudowords and strings of plus signs with matching lengths. Out of the resulting strings, 5 each were used for the practice blocks and 20 each were used for the trial blocks. A representation of all the different stimuli types can be seen in **Table 2**.

The generation of the text stimuli was achieved with Lexops **(Taylor, J. E., Beith, A., & Sereno, S. C. 2020),** which is an R package used in generating controlled stimuli for psychology experiments. While the joint dataset contained the variables required for the entirety of the study, the necessary features for constructing the data frames for the string stimuli only required the strings' spelling, lexicality, nchar and OLD20 values. A total of 170 strings from the joint dataset were required for the experiment, 10 for the practice block, 160 for the trial blocks. With Lexops, these 170 strings were generated from the 28.000 strings of the DLP, with an even split by lexicality, controlled for matching

lengths(nchar) and controlled for matching Levenshtein edit distances**(OLD20)** with a minute(0.1) constraint. After generating the strings, 85 strings formed out of plus signs with matching lengths were manually added to the data frame. The resulting data frame which consisted of 255 strings(15 for practice blocks and 240 for trial blocks) and their mentioned feature values was then shuffled once, while keeping the same feature values of the strings, to create another data frame as a way of controlling for an effect regarding the order of the strings. This control variable was dubbed as "subject parity" as every odd subject saw one version of the experiment with every even subject seeing the other version.

Table 2: Representations of the stimuli and controls as seen by participants. First square showing the fixation dot, Second square showcasing squares of differing intensities, Third square showcasing strings consisting of varying plus signs and the Fourth square showcasing words/pseudowords.



3.3 Experiment Design

A total of 6 participants, 4 female and 2 male, completed the simple reaction time experiment. The participants were recruited using the Human Subject Pool of Tilburg University and received course credits upon completion. The experiment, which consisted of 8 blocks, 2 practice and 6 trials, lasted 25 minutes on average. During these blocks, the task was the same regardless of the stimuli presented: participants were sat in front of a computer and were told to react to the stimuli as fast as they could, without making any judgements.

A block's procedure was as follows: The participant was presented with a fixation dot which was then followed by a stimulus. The durations between the fixation dots and the stimuli were randomized and created in R using exponential distribution so that the participants could not anticipate the duration of the following stimulus. If the participant reacted before the stimulus appeared they would receive a message warning them that they were too fast. If the participant took longer than 2000ms to respond after the stimulus had appeared, they would be met with a warning message telling them that they were too slow. The trial blocks took around 3 minutes each, but as the participants were told to simply react as fast as possible and not make any judgments, the blocks were separated with 2 minute intervals to avoid overwhelming the participants with seemingly endless stimuli.

3.4 Cleaning and Evaluating the Experiment Responses

For each participant, OpenSesame outputted a dataframe with 221 columns. Out of these 221 columns, the ones required for the experiment were "spelling", "colorhex", "subject_nr", "condition", "lexicality", "RT", and "subject_parity", so a large portion of the data frame was dropped using R. “colorhex” showed the color of the squares and was used to control reaction times to squares of varying intensities. “condition” was a three-way separation in between words, pseudowords and strings of plus signs and was used to control reaction times to plus signs. “Subject_parity " was used to control reaction times regarding the order of the strings. As OpenSesame did not output trial counts, trial counts were generated and added to the data frame in R. The resulting data frame was then merged with the joint dataset constructed from the initial data, with “spelling” as the merger variable, to analyze the response times of the simple reaction time experiment and the effects of the kept lexical features. This operation was carried out for all of the resulting data frames from OpenSesame and were added onto each other to create the final data frame to be evaluated.

The final data frame consisted of 16 columns. This data frame was then further analyzed using the lme4 R package **(Bates D, Mächler M, Bolker B, Walker S, 2015)** which is a package that provides functions for fitting and analyzing mixed models. This way, the stimulus response times were able to be investigated while controlling for factors that might have an effect outside of the lexical features, such as subject parity or the participant itself.

3.5 Statistical Approach

The final dataset, which was constructed by combining the variable values from the Dutch Lexicon Project and the participant results from the simple reaction time experiment, was then further analyzed using the R programming language **(R Core Team, 2020)**. Variables with larger values - bigram frequency scores, SUBTLEX-NL contextual diversity and SUBTLEX-NL frequency - were log transformed to avoid scaling issues while constructing models. The data was then analyzed using a linear mixed model approach instead of a linear model approach to account for the random effects that resulted from the experiment. The R package Lme4 was utilized to fit the linear mixed-effects models (**Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker, 2015)**. Random effects were then tested by looking into their correlation components and variances as output by the model summary. The models were then compared by taking out random effects with very high correlations, and if an improvement was found, the random effect was disregarded. Finally, the Buildmer package was utilized to automatically perform backward stepwise elimination to identify the maximal model that could still converge **(Cesko C. Voeten, 2022)**.

After finding the base maximal converging model structure, the variables were fed separately as predictors, in their relating groups and/or interactions, and then all together for model comparisons. An ANOVA test was then performed for each research question and their models, which resulted in the model's AIC and BIC scores, alongside their chi-squared statistics. The AIC scores and the p-value significance of the chi-squared statistics were the main selection criteria during model comparisons.

4. Results

In this section, the results from the experiments will be further analyzed based on their individual and interaction effects via linear-mixed models in R. The results aimed to investigate three main questions. First, the effects of lexicality concerning reaction times for words and pseudowords were explored. Second, the effects of the mutually possessed lexical features for words and pseudowords were compared. And finally, the effects of the entirety of the lexical features taken from the DLP data were investigated for words only.

4.1 Effect of Lexicality on Reaction Times

Table 3: Anova Test Results for Model Comparisons Regarding Research Question 1. Base: lmer(RT ~ (1|subject_nr) + (1|spelling), data = all.no.pluses)

Predictor	*AIC*	Pr(>Chisq)	R-squared
Base	11059	-	0.264
Lexicality Effect	11061	0.947	0.265
Lexicality and Subject Parity Interaction Effect	11064	1.000	0.265
Lexicality and Subject Parity Interaction and Trial Effect	11034	2.435e-08***	0.290
Lexicality and Trial Effects	11032	2.433e-08***	0.291

Table 3 demonstrates Anova Test results for model comparisons regarding the first research question. Models of multiple varying predictors and their interactions were added to the baseline model. The model with the lowest AIC score, as well as the most significant Chi-squared statistic was the model with lexicality and trial effects as predictors(AIC = 11032, Pr(>Chisq) = 2.433e-08, R-Squared = 0.291). As leaving subject parity effect showed an improvement in the AIC score and Chi-Squared statistic of the models, the control variable was tested again and left out of the research questions during model comparisons(AIC = 11034, Pr(>Chisq) = 2.435e-08, R-Squared = 0.290). The preferred model for Research Question 1 thus had Lexicality and Trial effects as predictors with no interactions.

Table 4: Summary Statistics of the Preferred Model for Research Question One.

Predictor	*β*	Confidence Intervals	t-value	p-value	se
Intercept	177.94	135.89 – 219.98	-	-	-
Lexicality Effect [W]	-0.84	-10.96	-0.164	0.870	5.129

Trial Effect	0.22	0.15-0.30	5.825	<0.001***	0.038

Table 4 demonstrates the fixed effect estimates, confidence intervals, t-values, p-values and standard errors for the final model of the first research question. Participants were estimated to respond around 0.84 ms faster to words than they would do to pseudowords. A significant effect for trial was found (β = 0.22, t = 5.825, p = <0.001, se = 0.038). However, lexicality did not seem to significantly affect reaction times (β = -0.84, t = -0.164, p = 0.870, se = 5.129).

4.2 Effects of Multiple Variables on Reaction Times for Words and Pseudowords

Table 5: Anova Test Results for Research Question 2 Model Comparisons

Predictor	*AIC*	Pr(>Chisq)	R-squared
Base	11059	-	0.264
Lexicality Effect	11061	0.947	0.265
OLD20 Effect	11061	0.422	0.265
Length Effect	11057	0.035	0.265
Lexicality and Length Interaction	11061	1.000	0.266
Lexicality, OLD20 and Length Three-way Interaction	11066	1.000	0.269
OLD20 and Length Interaction, and Lexicality, Bigram Frequency and Trial Effects	11033	0.107	0.267
Lexicality, OLD20 and Length Interaction, and Bigram Frequency and Trial Effect	11035	0.8032	0.267
All Variables, No Interaction Effects	11032	5.024e-07***	0.269

Table 5 demonstrates Anova Test results for model comparisons regarding the second research question. The model with the lowest AIC score, as well as the most significant Chi-squared statistic was the model with lexicality, OLD20, length and trial effects as predictors(AIC = 11030, Pr(>Chisq) = 1.071e-08, R-Squared = 0.269). The second model with the lowest AIC score was the model with a two-way interaction of OLD20 and length, alongside the lexicality and trial effects (AIC = 11033, Pr(>Chisq) = 0.110, R-Squared = 0.268). The model with the second most significant Chi-squared statistic was the model with a three-way interaction between lexicality, OLD20 and length, alongside the trial effect (AIC = 11035, Pr(>Chisq) = 8.698e-07, R-Squared = 0.268). The preferred model for Research Question 2, which had the lowest AIC score and the most significant

Chi-squared statistic together, thus had Lexicality, OLD20, Length and Trial effects as predictors with no interactions in-between them.

Table 6: Summary Statistics of the preferred model for Research Question Two.
Base: lmer(RT ~ (1|subject_nr) + (1|spelling), data = all.stirngs.no.pluses)

Predictor	*β*	Confidence Intervals	t-value	p-value	se
Intercept	205.52	125.19 – 302.09	-	-	-
Lexicality [W]	-0.83	-8.20 – 9.56	-0.164	0.870	5.054
OLD20	1.18	-22.20 – 24.06	0.198	0.843	5.954
Length	-4.81	-10.75 – 9.09	-2.003	0.045**	2.399
Bigram Frequency	-1.09	-27.82 – 25.64	-0.080	0.936	13.620
Trial	0.22	0.15 – 0.30	5.914	<0.001***	0.037

Table 6 demonstrates the fixed effect estimates, confidence intervals, t-values, p-values and standard errors for the final model of the second research question. The intercept was estimated to be 205.52 ms. Like the first model, Trial appeared to have a significant effect, with a 0.22 ms slower response being expected per unit of trial ($\beta = 0.22$, $t = 5.914$, $p = <0.001$, $se = 0.037$). Alongside the trial effect, the length of the string also showed an impactful change in reaction times, with an estimated 4.81 ms faster response time as the string increased in character count($\beta = -4.81$, $t = -2.003$, $p = 0.045$, $se = 2.399$). Lexicality for words and OLD20 did not appear to affect the reaction times as impactfully, much higher p-values t-values much closer 0 ($\beta = -4.81$, $t = -0.164$, $p = 0.870$, $se = 5.054$; $\beta = 1.18$, $t = 0.198$, $p = 0.843$, $se = 5.954$).

4.3 Effects of Multiple Variables on Reaction Times for Words Only

For the final research question, two sets of model comparisons and two sets of preferred model estimates were reported. This was due to the high correlation between the variables' contextual diversity and word frequency. First, the model comparisons and estimates for contextual diversity, then the model comparisons and estimates for word frequency can be found.

Table 7: Anova Test Results for Research Question 3 Model Comparisons Concerning Contextual Diversity

Predictor	*AIC*	Pr(>Chisq)	R-squared
Base	5098.4	-	0.238
Contextual Diversity Effect	5099.9	0.4898	0.240
OLD20 Effect	5100.0	0.5214	0.240
Length Effect	5099.1	0.2521	0.240
Bigram Frequency Effect	5099.5	0.3547	0.240
OLD20 and Length Interaction	5101.4	0.3558	0.243
OLD20 and Length Interaction, and CD Effect	5103.2	0.6488	0.245
OLD20, Length and CD Three-way Interaction	5106.8	0.5834	0.249
OLD20 and Length Interaction, Bigram, Trial and CD Effects	5094.8	0.2935	0.247
OLD20, Length and CD Three-way Interaction, and Bigram and Trial Effects	5099.3	0.0031**	0.251
All Variables, No Interaction Effects	5093.9	0.0007***	0.250

Table 7 demonstrates Anova Test results for model comparisons regarding the third research question concerning contextual diversity. The model with the lowest AIC score, as well as the most significant Chi-squared statistic was the model with contextual diversity, OLD20, length, bigram frequency and trial effects as predictors(AIC = 5093.9, Pr(>Chisq) = 0.0007, R-Squared = 0.250). The model with the second lowest AIC score and second most significant Chi-squared statistic was the model with a three-way interaction of OLD20 and length and contextual diversity, alongside the bigram frequency and trial effects (AIC = 5099.3, Pr(>Chisq) = 0.0031, R-Squared = 0.251). The preferred model for Research Question 3 regarding contextual diversity thus had contextual diversity, OLD20, length, bigram frequency and trial effects as predictors with no interactions in-between them.

Table 8: Summary Statistics of the preferred model for Research Question 3 Regarding Contextual Diversity

Base: lmer(RT ~ (1|subject_nr) + (1|spelling), data = words.only)

Predictor	*β*	Confidence Intervals	t-value	p-value	se
Intercept	213.64	125.19 – 302.09	-	-	-
Contextual Diversity	0.68	-8.20 – 9.56	0.151	0.880	4.518
OLD20	0.93	-22.20 – 24.06	0.079	0.937	11.766
Length	-2.97	-12.78 – 6.85	-0.594	0.454	4.992
Bigram Frequency	-7.37	-47.15 – 32.41	-0.364	0.716	10.120
Trial	0.20	0.09 – 0.31	3.587	0.001	0.056

Table 8 demonstrates the fixed effect estimates and the confidence intervals for the final model of the third research question concerned with contextual diversity. The intercept was estimated to be 213.64 ms. Like the model results of research question 1 and 2, trial appeared to have a significant effect, with a 0.22 ms slower response being expected per unit of trial ($\beta = 0.22$, $t = 3.587$, $p = 0.001$, $se = 0.056$). While no other variables seemed to affect the intercept as impactful as the trial effect, when compared within themselves, length appeared to have the highest impact with an estimated 2.97 ms faster response time as the word increased in character count ($\beta = -2.97$, $t = -0.494$, $p = 0.554$, $se = 4.992$).

Table 9: Anova Test Results for Research Question 3 Model Comparisons Concerning Word Frequency

Predictor	*AIC*	Pr(>Chisq)	R-squared
Base	5098.4	-	0.238
Word Frequency Effect	5099.7	0.3994	0.240
OLD20 Effect	5100.0	0.5214	0.240
Length Effect	5099.1	0.2521	0.240
Bigram Frequency Effect	5099.5	0.3547	0.240
OLD20 and Length Interaction	5101.4	0.3558	0.243
OLD20 and Length Interaction, and Word Frequency Effect	5103.1	0.4664	0.245

OLD20, Length and Word Frequency Three-way Interaction	5106.3	1.0000	0.249
OLD20 and Length Interaction, Bigram, Trial and Word Frequency Effects	5094.7	0.3001	0.247
OLD20, Length and Word Frequency Three-way Interaction, and Bigram and Trial Effects	5098.8	0.0033**	0.251
All Variables, No Interaction Effects	5093.8	0.0007***	0.250

Table 9 demonstrates Anova Test results for model comparisons regarding the third research question concerning word frequency. The model with the lowest AIC score, as well as the most significant Chi-squared statistic was the model with word frequency, OLD20, length, bigram frequency and trial effects as predictors(AIC = 5093.8, Pr(>Chisq) = 0.0007, R-Squared = 0.250). The model with the second lowest AIC score and second most significant Chi-squared statistic was the model with a three-way interaction of OLD20 and length and word frequency alongside the bigram frequency and trial effects (AIC = 5098.8, Pr(>Chisq) = 0.0033, R-Squared = 0.251). The preferred model for Research Question 3 regarding word frequency thus had word frequency, OLD20, length, bigram frequency and trial effects as predictors with no interactions in-between them.

Table 10: Summary Statistics of the preferred model for Research Question 3 Regarding Word Frequency

Base: lmer(RT ~ (1|subject_nr) + (1|spelling), data = words.only)

Predictor	***β***	**Confidence Intervals**	**t-value**	**p-value**	**se**
Intercept	211.48	122.77 – 300.20	-	-	-
Word Frequency	1.31	-6.72 – 9.34	0.332	0.748	4.085
OLD20	1.35	-22.20 – 24.06	0.116	0.908	11.668
Length	-2.96	-12.78 – 6.85	-0.594	0.453	4.983
Bigram Frequency	-7.36	-47.12 – 32.39	-0.364	0.716	10.120
Trial	0.20	0.09 – 0.31	3.578	0.001	0.056

Table 10 demonstrates the fixed effect estimates and the confidence intervals for the final model of the third research question concerned with word frequency. The intercept was estimated to be 211.48 ms. Like the model results of research question 1,2 and the first part of research question 3, trial appeared to have a significant effect, with a 0.20 ms slower response being expected per unit of trial

(β = 0.20, t = 3.578, p = 0.001, se = 0.056). While no other variables seemed to affect the intercept as impactful as the trial effect, when compared within themselves, length again appeared to have the highest impact with an estimated 2.96 ms faster response time as the word increased in character count (β = -2.96, t = -0.594, p = 0.453, se = 4.983).

4.4.1 Control Variables: Strings Composed of Plus Signs with Differing Lengths

Table 11: Control Variable Statistics for Strings Composed of Plus Signs with Differing Lengths

Predictor	*β*	Confidence Intervals	t-value	p-value
Intercept(+++++++++)	203.009	143.290 – 262.727	-	-
+++	54.676	-3.314 – 112.665	1.853	0.065
++++	-.043	-58.033 – 57.946	-0.001	0.999
+++++	40.456	-6.435 – 87.347	1.695	0.091
++++++	29.302	-14.862 – 73.465	1.304	0.193
+++++++	15.327	-29.592 – 60.246	0.671	0.503
++++++++	27.764	-20.754 – 76.281	1.124	0.261

Table 11 demonstrates the variable statistics for plus signs of differing lengths which were introduced as a control variable. There is a somewhat linear decrease in reaction times as the string gets longer, with strings composed of 7, 8 and 9 plus signs showing the overall lowest reaction times. Although the string composed of 4 plus seems to show the lowest reaction time based on estimates, when looked into its t and p values it is possible to observe that this effect isn't regarded as significant (β = -.043, t = -0.001, p = 0.999).

4.4.2 Control Variables: Square of Differing Intensities

Table 12: Control Variable Statistics for Squares of Differing Intensities

Predictor	*β*	Confidence Intervals	t-value	p-value
Intercept(*Lightest Intensity Square*)	216.492	147.550 – 285.434		
Second Lightest Intensity Square	-2.968	-27.011 – 21.075	-0.242	0.809

Middle Intensity Square	14.610	-9.433 – 38.653	1.193	0.233
Second Darkest Intensity Square	0.623	-23.420 – 24.666	0.051	0.959
Darkest Intensity Square	29.726	5.684 – 53.769	2.427	0.015

Table 12 demonstrates the variable statistics for squares of differing intensities which were introduced as a control variable. While the square with the slowest reaction times was the darkest intensity square with an estimated 246 ms, the square with the fastest square was the lightest intensity square with an estimated 216 ms. The same effect observed in table 11 is once again observed, with the model estimating an out of ordinary, faster reaction time for the second darkest square. Although again, when looked into its t and p values it is possible to observe that this effect isn't regarded as significant ($\beta = 0.623$, $t = 0.051$, $p = 0.959$).

5. Discussion

Lexical decision tasks are procedures used in psycholinguistic experiments. The basic procedure aims to measure how fast participants can differentiate words from pseudowords. This distinction requires the participants to make active linguistic judgments. However, a limitation of lexical decision tasks is that the participants are required to make metalinguistic judgements, which could differ from their normal comprehension **(Lieber, R., Štekauer, P., & Baayen, H. 2014)**.

This study aimed to take a different approach to lexical decision tasks by conducting a simple reaction time experiment in which the participants were asked to simply react to the stimuli as fast as possible, without making lexical judgements. By doing so, it attempted to investigate three points of interest; whether the participants were still involuntarily reading the stimulus, the lexical features' effects on both words and pseudowords when participants were told to not make lexical judgements, and finally, the lexical features' effects on just words, when participants were told to not make lexical judgements.

Regarding the first point of interest, whether or not participants were still actively reading the stimuli before reacting, we can look at the results of the first research question. In most lexical decision tasks , there are clear differences in reaction times to words and pseudowords. The Dutch Lexicon Project, which is a mega-scaled lexical study, found a 20ms~ difference in reaction times within Dutch words and pseudowords carried out in around all 50 blocks of the experiment **(Keuleers, E., & Brysbaert, M., 2010)**. Another study focusing on the effect of lexicality between Italian words and pseudowords compared high frequency words with high frequency pseudowords, and low frequency words with low frequency pseudowords, to find that the effect of lexicality was applicable in both cases, with participants again having a faster reaction to words than they did to pseudowords (18ms~) **(Pagliuca, Giovanni & Arduino, Lisa & Barca, Laura & Burani, Cristina, 2008)**.

However, lexicality did not seem to significantly affect reaction times during the simple reaction time experiment ($\beta = -0.84$, $t = -0.164$, $p = 0.870$, $se = 5.129$, Model Comparison 1). Based on the prior knowledge acquired by lexical decision tasks, which is that lexicality does have an effect in reaction times, the effect not occurring during the simple reaction time experiment could indeed prove that the participants were able to react to the stimuli without making lexical judgements.

After the effect of lexicality as a sole predictor, the effects of the shared lexical features of words and pseudowords were investigated. These effects were the strings' Levenshtein edit distances, lengths and bigram frequency. While studies dating earlier than the 2000's usually suggest the bigram frequency has an effect in reaction times, later studies, including the mega-scaled lexicon projects, seem to argue that bigram frequency does not significantly affect reaction times **(Schmalz, Xenia & Mulatti, Claudio. 2017)**. Subsequently, the effect with the least impact on reaction times for the simple reaction time experiment regarding words and pseudowords was also bigram frequency ($\beta = -1.09$, $t = -0.080$, $p = 0.936$, $se = 13.620$). The effect of word length is slightly trickier. Older and newer studies seem to both confirm that the effect of string length is not linear, however

there is a mutual agreement that as the length of the string increases so does reaction times **(New, B., ferrand, L., pallier, C. *et al. ,2006;* Ginestet, Phénix, Diard, Valdois, 2019)**. However, this was not the case for the simple reaction time experiment. String length was the only other significant effect for research question 2, with participants reacting 3ms~ faster as the words got longer. This held true for the strings composed of plus signs introduced as a control method. The contradicting results could be regarded as yet another proof that the participants were reacting without making any lexical judgements and were reacting to the general shapes of the strings instead; A conclusion can be drawn that as the strings got longer they got bigger and thus easier to react to. String length and OLD20 measures are usually correlated to a certain extent; Although not necessary, as the string gets longer, it usually gets more difficult to reach its 20 nearest neighbors. Despite this, as can be seen in the Dutch Lexicon Project as well, OLD20 and string length does not necessarily affect reaction times to the same extent. While an increase in reaction times was found with the increase of string length(although this trend did not persist for all blocks) the effects of OLD20 appeared to be small throughout the experiment (**Keuleers, E., & Brysbaert, M., 2010)**. This was seen in the simple reaction time experiment as well, with OLD20 having much less of an impact than string length. The complexity of the string not affecting the reaction time in the same way string length has can also be attributed to the participants truly reacting to the stimuli while disregarding its letter characteristics, just as it has occurred with bigram frequency.

Finally, the effects of lexical features regarding only words were investigated. Besides OLD20, bigram frequency, and length, contextual diversity and word frequency were introduced. Two different sets of model comparisons and predictor estimates were conducted by separating contextual diversity from word frequency as predictors from the models, because as mentioned in the theoretical framework, the two variables are often highly correlated **(Plummer, P., Perea, M., & Rayner, K. 2014)**. As expected, the results gathered from the two sets were quite similar.

Both word frequency and contextual diversity have been proven to play a significant role in lexical decision tasks **(Adelman, J. S., Brown, G. D. A., & Quesada, J. F. 2006 ; Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. 2004).** However, this was not the case for the simple reaction time experiment. This makes sense as the changes in reaction times for both word frequency and contextual diversity can be explained by participants recognizing the word and reacting faster or slower accordingly. In the simple reaction time experiment, as the participants were told to withhold making lexical judgements and reading, it is feasible that they did not recognize a more common or uncommon word, thus yielding no significant change in reaction times. The trends with OLD20 and bigram frequency continued with research question 3, with both effects not significantly affecting reaction times. Although string length had a much higher p-value and a t-value much closer to 0 compared with string length effect from research question 2, from all of the effects(besides trial) string

length still had the most consistent impact on reaction times for both research question 2 and 3.

Trial effect, which measured the change in reaction times as participants carried on with the experiment, was consistently significant for all research questions. Participants reacted 0.20~ ms slower per unit of trial for 240 trials in total. It is possible to base this change on participants getting tired or more automated and thus less agile in their reactions as the experiment carried on.

5.2 Limitations and Possible solutions

The simple reaction time experiment was planned to be carried out on 60 participants, however only 6 participants completed the experiment. This was plausibly due to the experiment timeslots clashing with the final exam dates of the Tilburg University students, which was the only pool the participants were selected from. This limitation's effects became quite observable: Subject parity which was introduced as a way of controlling the effect of stimulus order on reaction times quickly lost its purpose as it became more of a divide between individuals than the entire group. Demographic information, which would have been looked into for gender and handedness(dominant hand), were also ignored as they would have output results that resembled individuals more than the demographic itself. Although it is applicable to most experiments, increasing stimulus count would provide more accurate results as well. Compared with most mega-scale lexical studies, the simple reaction time tested on about 1/100'th of the stimulus count. While this was expected concerning the nature of the study, the study is likely to produce different and more accurate results if carried out on a larger scale.

Conclusion

This thesis aimed to investigate these relations: a systematic difference in reaction times between words and pseudowords, the effects of their mutual lexical features, and the effects of the lexical features of just words. It was unable to prove a systematic difference in reaction times concerning lexicality, however with this and the lack of impact concerning other lexical features such as contextual diversity and word frequency in mind, it proved that the participants were able to react to stimuli without reading it.

String length, the feature most concerned with the physical structure of the string, had the most impactful effect on reaction times besides the effect of practice. However, unlike other lexical decision tasks, instead of showing an increase in reaction times, it actually caused reaction times to get faster as the string length increased. This once again proved that the participants were treating the stimuli as imagery instead of strings, as they reacted faster as the “image” got bigger.

It should be once again reminded that this study was conducted with a relatively low stimuli count and with a quite low sample size due to the reasons mentioned in limitations during discussion. For future research, if the study were to be conducted on a larger scale, it could be possible to further investigate the effects of lexical features in stages that were previously inaccessible due to the nature of lexical decision tasks.

Acknowledgements

This research was made possible with the freely accessible online data provided by Keuleers, E., Diependaele, K., & Brysbaert, M, 2010.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. In Psychological Science (Vol. 17, Issue 9, pp. 814–823). SAGE Publications. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual Word Recognition of Single-Syllable Words. In Journal of Experimental Psychology: General (Vol. 133, Issue 2, pp. 283–316). American Psychological Association (APA). https://doi.org/10.1037/0096-3445.133.2.283

Balota, D.A., Cortese, M.J., & Pilotti, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. In Abstracts of the 40th Annual Meeting of the Psychonomics Society (p. 44). Los Angeles, CA: Psychonomic Society.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. In Behavior Research Methods (Vol. 39, Issue 3, pp. 445–459). Springer Science and Business Media LLC. https://doi.org/10.3758/bf03193014

Bates D, Mächler M, Bolker B, Walker S (2015). “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, **67**(1), 1–48. doi: 10.18637/jss.v067.i01.

Berko, J. (1958). The Child’s Learning of English Morphology. In <i>WORD</i> (Vol. 14, Issues 2–3, pp. 150–177). Informa UK Limited. https://doi.org/10.1080/00437956.1958.11659661

Cesko C. Voeten (2022). buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression. R package version 2.6. https://CRAN.R-project.org/package=buildmer

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. In Behavior Research Methods (Vol. 42, Issue 2, pp. 488–496). Springer Science and Business Media LLC. https://doi.org/10.3758/brm.42.2.488

https://doi.org/10.3758/s13428-020-01356-w

Ginestet, E., Phénix, T., Diard, J., & Valdois, S. (2019). Modeling the length effect for words in lexical decision: The role of visual attention. In Vision Research (Vol. 159, pp. 10–20). Elsevier BV. https://doi.org/10.1016/j.visres.2019.03.003

Johns, B. T. (2021). Disentangling contextual diversity: Communicative need as a lexical organizer. In Psychological Review (Vol. 128, Issue 3, pp. 525–557). American Psychological Association (APA). https://doi.org/10.1037/rev0000265

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. In Behavior Research Methods (Vol. 42, Issue 3, pp. 627–633). Springer Science and Business Media LLC. https://doi.org/10.3758/brm.42.3.627

Klafehn, T. (2011). Myth of the wug test: Japanese speakers can’t pass it and English speaking children can’t pass it either. In Annual Meeting of the Berkeley Linguistics Society (Vol. 37, Issue 1, p. 170). Linguistic Society of America. https://doi.org/10.3765/bls.v37i1.841

Lieber, R., Štekauer, P., & Baayen, H. (2014). Experimental and Psycholinguistic Approaches. In The Oxford Handbook of Derivational Morphology. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641642.013.0007

New, B., ferrand, L., pallier, C., & brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. In Psychonomic Bulletin & Review (Vol. 13, Issue 1, pp. 45–52). Springer Science and Business Media LLC. https://doi.org/10.3758/bf03193811

Pagliuca, G., Arduino, L. S., Barca, L., & Burani, C. (2008). Fully transparent orthography, yet lexical reading aloud: The lexicality effect in Italian. In Language and Cognitive Processes (Vol. 23, Issue 3, pp. 422–433). Informa UK Limited. https://doi.org/10.1080/01690960701626036

Perea, M., Rosa, E., & Gómez, C. (2005). The frequency effect for pseudowords in the lexical decision task. In Perception & Psychophysics (Vol. 67, Issue 2, pp. 301–314). Springer Science and Business Media LLC. https://doi.org/10.3758/bf03206493

Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. In Journal of Experimental Psychology: Learning, Memory, and Cognition (Vol. 40, Issue 1, pp. 275–283). American Psychological Association (APA). https://doi.org/10.1037/a0034058

R Core Team (2020). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reynolds, B. L. (2018). The effects of nonce words, frequency, contextual richness, and L2 vocabulary knowledge on the incidental acquisition of vocabulary through reading: more than a replication of Zahar et al. (2001) & Tekmen and Daloğlu (2006). In International Review of Applied Linguistics in Language Teaching (Vol. 58, Issue 1, pp. 75–102). Walter de Gruyter GmbH. https://doi.org/10.1515/iral-2015-0115

Schmalz, X., & Mulatti, C. (2017). Busting a myth with the Bayes Factor. In The Mental Lexicon (Vol. 12, Issue 2, pp. 263–282). John Benjamins Publishing Company. https://doi.org/10.1075/ml.17009.sch

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). LexOPS: An R Package and User Interface for the Controlled Generation of Word Stimuli. *Behaviour Research Methods*, *52*, 2372–2382. http://doi.org/10.3758/s13428-020-01389-1

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the English Lexicon Project. In Journal of Experimental Psychology: Learning, Memory, and Cognition (Vol. 41, Issue 3, pp. 597–613). American Psychological Association (APA). https://doi.org/10.1037/xlm0000064

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314-324. Opensesame doi:10.3758/s13428-011-0168-7

Appendices and Supplementary Materials

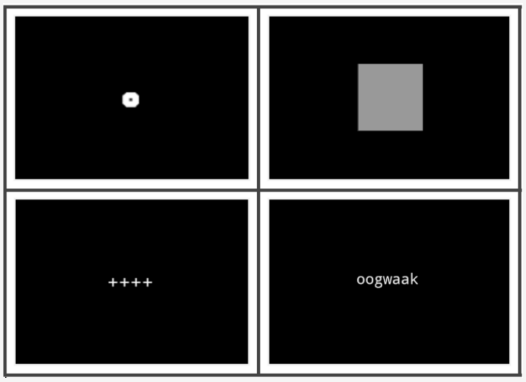
3.2 Statistical Values Of Variables

Table 1 : Statistical Values of Variables Used in the Simple Reaction Time Experiment for Research Question 3

	Min	**Max**	**Mean**	**SD**
General Fields				
Length	3	9	6.2	1.44
Contextual Diversity	1	7729	702.794	1,715.07
Word Frequency	1	45,392	2,213.27	7,744.63
Logged Contextual Diversity	0.30	3.88	1.96	0.93
Logged Word Frequency	0.31	4.56	2.10	1.02
Orthographic Neighborhood				
OLD20	1.0	3.6	1.9	0.55
Bigram Frequency				
Summed Bigram Frequency	31.989	528	212.574	114.87
Logged Bigram Frequency	1.50	2.72	2.25	0.26
Experiment Variables				
Reaction Times	104.182	803.91	207.108	87.604
Trial	3	239	3.587	70.07

3.3 Stimuli Representation

Table 2: Representations of the stimuli and controls as seen by participants. First square showing the fixation dot, Second square showcasing squares of differing intensities, Third square showcasing strings consisting of varying plus signs and the Fourth square showcasing words/pseudowords.



3.4 Stimulus Characteristics

Colorhex: the hex code for the color of the squares.
Condition: A1 for words, A2 for pseudowords, A3 for strings made off of plus signs and A4 for squares.
Lexicality: W for words, N for pseudowords.
Nchar: Number of characters of the stimulus.
OLD20: Levenshtein edit distance measure, range: 1-4
RT: Reaction times to the stimuli
Spelling: String names.
Subject_parity: Determines which version of the data frames the participant sees, can be odd or even.
Subtlex.cd: Raw contextual diversity counts taken from the SUBTLEX-NL database.
Subtlex.frequency: Raw frequency counts taken from the SUBTLEX-NL database.
Subtlex.log10.cd: Log10 of the raw contextual diversity counts taken from the SUBTLEX-NL database.
Subtlex.log10.frequency: Log10 of the raw frequency counts taken from the SUBTLEX-NL database.
Summed.bigram: Sum of non-positional bigram frequencies.
Trial: The identification number of the trials.

4.1 Effect of Lexicality on Reaction Times

Table 3: Anova Test Results for Model Comparisons Regarding Research Question 1.
Base: lmer(RT ~ (1|subject_nr) + (1|spelling), data = all.no.pluses)

Predictor	*AIC*	Pr(>Chisq)	R-squared
Base	11059	-	0.264
Lexicality Effect	11061	0.947	0.265
Lexicality and Subject Parity Interaction Effect	11064	1.000	0.265
Lexicality and Subject Parity Interaction and Trial Effect	11034	2.435e-08***	0.290
Lexicality and Trial Effects	11032	2.433e-08***	0.291

Table 4: Summary Statistics of the Preferred Model for Research Question One.

Predictor	*β*	Confidence Intervals	t-value	p-value	se
Intercept	177.94	135.89 – 219.98	-	-	-
Lexicality Effect [W]	-0.84	-10.96	-0.164	0.870	5.129
Trial Effect	0.22	0.15-0.30	5.825	<0.001***	0.038

4.2 Effects of Multiple Variables on Reaction Times for Words and Pseudowords

Table 5: Anova Test Results for Research Question 2 Model Comparisons

Predictor	*AIC*	Pr(>Chisq)	R-squared
Base	11059	-	0.264
Lexicality Effect	11061	0.947	0.265
OLD20 Effect	11061	0.422	0.265
Length Effect	11057	0.035	0.265
Lexicality and Length Interaction	11061	1.000	0.266
Lexicality, OLD20 and Length Three-way Interaction	11066	1.000	0.269
OLD20 and Length Interaction, and Lexicality, Bigram Frequency and Trial Effects	11033	0.107	0.267
Lexicality, OLD20 and Length Interaction, and Bigram Frequency and Trial Effect	11035	0.8032	0.267
All Variables, No Interaction Effects	11032	5.024e-07***	0.269

Table 6: Summary Statistics of the preferred model for Research Question Two.

Predictor	*β*	Confidence Intervals	t-value	p-value	se
Intercept	205.52	125.19 – 302.09	-	-	-
Lexicality [W]	-0.83	-8.20 – 9.56	-0.164	0.870	5.054
OLD20	1.18	-22.20 – 24.06	0.198	0.843	5.954
Length	-4.81	-10.75 – 9.09	-2.003	0.045**	2.399
Bigram Frequency	-1.09	-27.82 – 25.64	-0.080	0.936	13.620
Trial	0.22	0.15 – 0.30	5.914	<0.001***	0.037

4.3 Effects of Multiple Variables on Reaction Times for Words Only

Table 7: Anova Test Results for Research Question 3 Model Comparisons Concerning Contextual Diversity

Predictor	*AIC*	Pr(>Chisq)	R-squared
Base	5098.4	-	0.238
Contextual Diversity Effect	5099.9	0.4898	0.240
OLD20 Effect	5100.0	0.5214	0.240
Length Effect	5099.1	0.2521	0.240
Bigram Frequency Effect	5099.5	0.3547	0.240
OLD20 and Length Interaction	5101.4	0.3558	0.243
OLD20 and Length Interaction, and CD Effect	5103.2	0.6488	0.245
OLD20, Length and CD Three-way Interaction	5106.8	0.5834	0.249
OLD20 and Length Interaction, Bigram, Trial and CD Effects	5094.8	0.2935	0.247
OLD20, Length and CD Three-way Interaction, and Bigram and Trial Effects	5099.3	0.0031**	0.251
All Variables, No Interaction Effects	5093.9	0.0007***	0.250

Table 8: Summary Statistics of the preferred model for Research Question 3 Regarding Contextual Diversity

Predictor	***β***	**Confidence Intervals**	**t-value**	**p-value**	**se**
Intercept	213.64	125.19 – 302.09	-	-	-
Contextual Diversity	0.68	-8.20 – 9.56	0.151	0.880	4.518
OLD20	0.93	-22.20 – 24.06	0.079	0.937	11.766
Length	-2.97	-12.78 – 6.85	-0.594	0.454	4.992
Bigram Frequency	-7.37	-47.15 – 32.41	-0.364	0.716	10.120
Trial	0.20	0.09 – 0.31	3.587	0.001	0.056

Table 9: Anova Test Results for Research Question 3 Model Comparisons Concerning Word Frequency

Predictor	***AIC***	**Pr(>Chisq)**	**R-squared**
Base	5098.4	-	0.238
Word Frequency Effect	5099.7	0.3994	0.240
OLD20 Effect	5100.0	0.5214	0.240
Length Effect	5099.1	0.2521	0.240
Bigram Frequency Effect	5099.5	0.3547	0.240
OLD20 and Length Interaction	5101.4	0.3558	0.243
OLD20 and Length Interaction, and Word Frequency Effect	5103.1	0.4664	0.245
OLD20, Length and Word Frequency Three-way Interaction	5106.3	1.0000	0.249
OLD20 and Length Interaction, Bigram, Trial and Word Frequency Effects	5094.7	0.3001	0.247
OLD20, Length and Word Frequency Three-way Interaction, and Bigram and Trial Effects	5098.8	0.0033**	0.251
All Variables, No Interaction Effects	5093.8	0.0007***	0.250

Table 10: Summary Statistics of the preferred model for Research Question 3 Regarding Word Frequency

Predictor	*β*	Confidence Intervals	t-value	p-value	se
Intercept	211.48	122.77 – 300.20	-	-	-
Word Frequency	1.31	-6.72 – 9.34	0.332	0.748	4.085
OLD20	1.35	-22.20 – 24.06	0.116	0.908	11.668
Length	-2.96	-12.78 – 6.85	-0.594	0.453	4.983
Bigram Frequency	-7.36	-47.12 – 32.39	-0.364	0.716	10.120
Trial	0.20	0.09 – 0.31	3.578	0.001	0.056

4.1 Control Variables: Strings Composed of Plus Signs with Differing Lengths

Table 11: Control Variable Statistics for Strings Composed of Plus Signs with Differing Lengths

Predictor	*β*	Confidence Intervals	t-value	p-value
Intercept(+++++++++)	203.009	143.290 262.727	-	-
+++	54.676	-3.314 112.665	1.853	.065
++++	-.043	-58.033 57.946	-.001	.999
+++++	40.456	-6.435 87.347	1.695	.091
++++++	29.302	-14.862 73.465	1.304	.193
+++++++	15.327	-29.592 60.246	.671	.503
++++++++	27.764	-20.754 76.281	1.124	.261

4.1 Control Variables: Square of Differing Intensities

Table 12: Control Variable Statistics for Squares of Differing Intensities

Predictor	***β***	**Confidence Intervals**	**t-value**	**p-value**
Intercept(*Lightest Intensity Square*)	216.492	147.550 – 285.434		
Second Lightest Intensity Square	-2.968	-27.011 – 21.075	-242	.809
Middle Intensity Square	14.610	-9.433 – 38.653	1.193	.233
Second Darkest Intensity Square	0.623	-23.420 – 24.666	.0.51	.959
Darkest Intensity Square	29.726	5.684 – 53.769	2.427	.015