Predicting learning in a cross-situational word learning setting

Master Thesis

Ronny Brouwers

Snr: 2005553

Master of Science in Communication and Information Sciences,
Master Track Data Science: Business and Governance.

Tilburg University

School of Humanities

Thesis committee:

Dr. A. T. Hendrickson

Dr. P. A. Vogt

January 8, 2018

Preface

In front of you is my thesis: ‘Predicting learning in a cross-situational word learning setting’. This thesis completes the Master of Data Science in Business and Governance at Tilburg University. I hope you enjoy reading it.

First and foremost, I would like to express my sincere appreciation to Dr. Hendrickson. I would really like to thank him for his excellent guidance, inspiration, and suggestions. I learned a lot in our meetings and appreciate the time he took to educate me.

Thank you, dad, mom, and sister for your unconditional love and (financial) support.

Finally, I would like to offer my sincere apologies to my friends. I promise to be less boring over the upcoming weekends. The first round is on me.

Ronny Brouwers

Tilburg, January 2018

Abstract

This thesis project used experimental, cross-situational, word-learning data, and the correct combination of pseudo words and novel objects had to be identified in an ambiguous setting. The experiments consisted of training and testing phases; the subjects were required to learn the right combination in the training phase and the testing phase determined whether the subjects learned the combination correctly. The right combinations could be learned through multiple exposures to the same word-object pair, while each screen in the training phase was ambiguous because it consisted of multiple objects.

The first research question of this project aimed to identify which individual features could be used to predict whether subjects would learn a word pair correctly in the testing phase. Based on the literature, five potential features were identified and tested using a logistic regression and random forest algorithm. The first examined feature was the first presentation, which corresponded to the moment that a word-object pair was introduced for the first time in the training phase of the experiment. Both algorithms showed that the first presentation was a poor predictor. The second feature was the final presentation, which corresponded to the last moment that a word-object pair was seen during the experiment. The final presentation feature also showed no predictive power. Next, the feature word distribution was found, which could be used to determine whether the distribution of the presentation of the same word-object pair can be used to predict learning (tight versus loose distributions). The models showed worse predictions than the baseline model that was used. The results also showed that the time subjects needed to select a word-object combination could be used to predict the correctness of their guess. Finally, one context effect was examined in this project to test whether a word-object pair was learned better when no uncertainty was present in the experiment. The results showed that the more frequently a word-object pair was presented without uncertainty, the more likely the pair was to be learned correctly. Combining the non-predictive features in one model did not achieve a predictive model, and combining the predictive features did not result in better classifications than the individual features.

The second research question focused on identifying different types of learners, as the literature showed that different subjects may learn differently. In other words, the individual features could have a different impact on each subject. A Gaussian mixture model and hierarchical clustering were used, and clustering analysis showed that the identified clusters were poorly separated and contained much noise. Therefore, the individual clusters could not be interpreted.

This study evaluated five different features for their potential predictive power on correctly guessed words. New studies could focus on additional testing of features which were not treated in this research. In addition, future research could use the features that are predictive to build a complete model with multiple features that can predict correctly guessed words with high accuracy.

Contents

1. Introduction

When hearing or reading novel words, the meaning of these words is unknown and uncertain. Children and adults learn new words by resolving referential uncertainty and discovering relationships between meaning and words. Learning words involves a large amount of ambiguity, as the correct referent is uncertain for novel words. In 1960, philosopher Quine published his book, Word and Object, and he investigated the concept of meaning in one chapter. Ambiguity and uncertainty play a role when learning new words and may occur when a non-native speaker says a word and points in the direction of an object. A famous example that Quine provides is an anthropologist who observers a speaker with an unknown language while this speaker points his finger in the direction of a rabbit and says the word 'gavagai'. The anthropologist may be confident that the speaker is referring to the English word rabbit, although the speaker could have intended to refer to the word food or animal. Thus, correctly mapping word-object pairs with a single encounter may be problematic.

Yu and Smith (2007) stated that learners use statistics to find the right word-object pairs, as they learn the right combination by mapping the word-object pair through cross-trial statistical relations. Correspondingly, multiple exposures to word-object pairs will result in the correct association between word and referent as they co-occur numerous times. The isolation of word-referent pairs through multiple exposures is called cross-situational word learning. In a study by Yu and Smith (2007), participants had to learn word-object mapping. The study contained a training phase with a series of training screens; several novel objects were displayed on each screen and the same number of pseudo words were spoken. Each pseudo word was combined with a single novel object and the participant had to select the correct combination of spoken word and object. In the beginning, the right combination is highly ambiguous, but the same word-object pair is presented in multiple screens surrounded by different word-object pairs throughout the training phase and combinations can therefore be made.

Most research on cross-situational word learning has been conducted to prove the theory that supports the existence of cross-situational word learning: statistical learning by word-referent mapping (Trueswell, Medina, Hafri & Gleitman, 2013; Yu & Smith, 2007; Smith & Yu, 2008). The studies suggest that learners can learn word-referent pairs in highly ambiguous settings by calculating cross trial statistics. However, the frequency of word occurrences is not proportionally distributed in a real-word setting. In natural language, word frequency occurrences follow Zipfian distribution according to Zipf's law (Zipf, 1949), where the frequency of a word is directly proportional to its ranking in frequency counting. Some studies suggest that learning in a Zipfian distribution is more difficult, as some word-referents only occur with a low frequency (Blythe, Smith, & Smith, 2010; Vogt, 2012). Hendrickson and Perfors (under review) found that learning is usually improved in a Zipfian environment in a more recent study, contrary to the findings of other studies. The authors state that a possible explanation is decreased uncertainty about low frequency words, as high frequency words in the same screen are more likely to have been learned.

The following part of the introduction (section 1.1) describes the problem statement and translates it into two research questions. The next section (1.2) explains the scientific and practical relevance of the topic, and subsection 1.3 outlines the structure of the remainder of this thesis.

1.1. Problem statement and research questions

Until now, little attention has been paid to predicting learning in cross-situational word learning, especially regarding which features predict whether a word will be learned correctly. Most studies focus on complete computational models and simulation of cross-situational, word-learning data, and less on individual features. Identifying which factors and circumstances provide maximum learning results in a cross-situational learning context could provide valuable information about efficient methods of word learning for children, and for adults to learn non-native languages. After investigating which features may predict word learning, it is possible to build a complete model that combines the individual features that show predictive power. This model is of interest because it would help determine whether a combination of features could be used to build a strong predictive model. On the other hand, it would also be interesting to learn whether features that do not individually predict learning can predict learning when they are combined with other features. The following research question (RQ) and its related sub-questions (SQ) address the previous statement:

RQ_1: Which features predict cross-situational word learning?

SQ_1: Which individual features are adequate to predict cross-situational word learning?

SQ_2: Does combining all the non-predictive features into one model lead to a predictive model?

SQ_3: *Does combing all the predictive features into one model lead to a more powerful predictive model than all features separately?*

In addition to identifying features that predict word learning in cross-situational word learning, different types of learners may exist. The first research question examines whether a feature is a predictor of learned words for all subjects together, though it would also be interesting to learn whether features are predictive to learner A without influencing learner B. This issue is addressed extensively in the related work section (section 2). Since the factors that influence learning may differ from person to person, different types of learning may be found. In conclusion, finding clusters of different types of learners could offer important insight into word learning. Research question 2 addresses this topic:

RQ_2: Can clusters that suggest the existence of different types of learners be found?

1.2. Scientific and practical relevance

An important aim of this thesis is to be scientifically and practically relevant to the field and to extend the existing literature. The subsections below explain the scientific and practical relevance of this research project.

1.2.1. Scientific relevance

This thesis contributes to the existing scientific work and literature using multiple machine learning techniques to identify features that predict word learning. The available data for this thesis contain multiple experiments with many participants. Existing comparable studies with equal or more participants were not found, which provides an opportunity to add scientific relevance with the aid of a large sample size. As stated above, some studies have found evidence that learners use statistics to find the right word-object pairs. Exploring and identifying factors that predict word learning could help to explain how optimal memory and mapping of word-object pairs works. Insight into these features could be valuable to science since optimal learning environments could be created.

The second research question about whether different types of learners can be distinguished is highly debated (see related work section), though particular features could have different impacts on different learners. For example, one group of learners may perform better when word-object pairs are repeated quickly and consecutively, and another group could benefit from word-object pairs that are spread out. The controversy over different learning types is addressed in more detailed in section 2.

1.2.2. Practical relevance

In addition to scientific relevance, this thesis contributes on a practical level. By identifying features that positively influence learning, opportunities arise for persons who are willing to learn new words more efficiently and promptly. Moreover, an optimal learning setting can be created by emphasizing the features that contribute to faster learning, which could create opportunities for companies and institutions that are active in language teaching industries. If evidence is found that different types of learners exist, companies could offer a test to learners which would determine the type of learner for a subject. Thus, language education companies and institutions could offer customized word-learning solutions to their users.

It is important to realize that the digitization of learning facilities and content has recently led to a market share of education applications of approximately 16% in the total application market (Global Education Apps Market-Market Study 2015-2019). Further forecast shows a compound annual growth rate of approximately 34.7%. Online education is growing and creating opportunities for companies and institutions that offer customized and effective language-learning solutions.

1.3. Thesis outline

The next chapter, section 2, describes the relevant literature and related work regarding learning in general and cross-situational word learning. In section 3, a detailed description of the datasets is provided, the experimental procedure is explained, and the descriptive statistics of all the features that were tested are presented. Moreover, this section describes which algorithms were used and how the algorithms were evaluated. The results of the tests are provided in section 4, which is split into three subsections. Subsection 4.1 presents the results for each individual feature test as described in the problem statement and research questions. The results of the complete non-predictive and predictive features are then described in the second part of the results section. In subsection 4.3, the results of clustering algorithms are shown, and in section 5, the findings are discussed in relation to the previous literature. Finally, the answers to the research questions are briefly summarized in the conclusion chapter (section 6).

2. Related work

This section discusses relevant literature and related work to establish a broader context for the thesis topic. It reviews related work that has been conducted in the field of machine learning and cross-situational word learning to justify the added value of this thesis. Furthermore, previous work can be used to identify valuable features in optimal cross-situational word learning. A general overview of cross-situational word learning is described in the introduction section (see Section 1), and in this section, features that potentially predict word learning are presented. Subsection 2.1 discusses features that may predict word learning based on the existing literature in a cross-situational context or in a more general learning context. The following subsection (2.2) demonstrates the different possible types of learners, and section 2.3 shows how this thesis adds value to the existing work and literature.

2.1. Exploratory variables

As stated in the introduction (see Section 1), little is known about variables that can predict learning in cross-situational, word-learning experiments. This subsection addresses studies that have examined variables that influence or predict successful word learning. Since few studies focus on individual features in cross-situational word learning, more general sources of features that influence learning are also part of this section.

Frequency

Numerous studies show that frequency significantly influences successful cross-situational word learning. These studies demonstrate that increasing the repetition of word-object pairs increases the likelihood of successful learning (e.g., Kachergis, Yu & Shiffrin, 2015; Kachergis, Shiffrin & Yu, 2009; Frank, Goodman & Tenenbaum, 2009; Yu & Smith, 2012). If the number of repetitions in an experiment varied per pair, the pairs that were more frequent were learned more often. All the previously mentioned studies show evidence that frequency is a predictor of successful word learning in a cross-situational setting.

Serial-position effects

In addition to the presentation frequency of a word referent, the moment of presentation may be important, as a considerable amount of literature has been published about serial-position effects on learning a series of words or items. The serial-position effect is a widely adapted theory by Ebbinghaus (1913) which states that people are likely to recall the last and first word in a series better than the words in between. Although there are relatively few studies in the area of serial-position effects in cross-situational word learning, it seems plausible that the serial-position effect theory also influences cross-situational word learning as learning a series of words is part of this learning technique. It would be interesting to investigate whether word-object pairs that are introduced early in an experiment (primacy effect) are learned more often than word-object pairs that are introduced later. For word objects that are presented at the end of an experiment, these recently (recency effect) seen word objects are memorized better.

Referential uncertainty

Another issue with cross-situational word learning is referential uncertainty. Some studies present evidence that word learning is possible even at high levels of uncertainty. For instance, one study (Smith, Smith & Blythe, 2010) proved that word learning is effective even at high levels of uncertainty, though the authors state that learners learn more successfully and faster when referential uncertainty is low. In cross-situational word learning, referential uncertainty can differ throughout the training phase. In some screens, specific word-object pairs may already be learned, leading to a lower word-object uncertainty for the remaining pairs. When a screen consists of four word-object pairs and the first three are guessed correctly, the last word-object pair is isolated, and its uncertainty is low or zero. In an investigation into the learning of isolated words, Brent and Siskind (2001) found that learners learn words more effectively when exposed in an isolated setting compared to a referent, uncertain setting. A broader perspective was adopted by Lew-Williams, Pelucchi and Saffran (2011) who argued that isolated words should be an addition to presenting non-isolated words. The combination of learning words in uncertain and certain settings enhances word learning according to the authors. The authors of that study examined the descriptive statistics of this context effect, and it would be interesting to examine whether the context effect of isolated words can predict word learning in a cross-situational setting.

Reaction time

Another variable in experiments is reaction time, and whether the reaction time that a subject requires to guess a word-referent pair can be used to predict the correctness of the guess is of interest. A question could be about whether a fast response time indicates that a word-referent pair is learned correctly, or the opposite. No known work in the field of cross-situational word learning has researched whether response time can be used to predict correctly guessed pairs, although several studies have investigated response time and decision making in other experiments. For example, Rubenstein (2013) found that there was a close connection between short response times and wrong decisions. In another study, Schotter and Trevino (2014) demonstrated the predictive power of reaction times in a strategic decision-making situation. Furthermore, in a global game experiment, the authors predicted the future choices of the subjects using various reaction times as a predictor. While a global game experiment differs from a word-learning task, making strategic decisions is a valid part of selecting the correct word-object pair in cross-situational word learning.

Word distribution

In the literature on cross-situational word learning, the experiments consist of several screens in which several word-referent pairs are shown. Depending on the condition of the experiment, the same word-referent pair returns on multiple screens if the occurrence frequency of the word-referent pair is higher than one. Repetition is a key aspect of cross-situational word learning, as pairs cannot be learned if all pairs only occur once. One possible feature could be that the distribution of this repetition influences word learning (e.g. are pairs better learned when they are presented shortly after each other or are pairs learned better when they are spaced out over the learning phase?). No previous study in cross-situational word learning has investigated whether the distribution of the occurrences of a word-object pair influences whether a pair is learned correctly. General learning theories describe two strategies in learning series of words or other concepts, blocking and interleaving learning. In blocking, one word or other skill is learned or practiced at a time ('AAABBBCCC'), whereas the words or skills are mixed together in interleaving ('ABCABCABC').

In 1885, Ebbinghaus was one of the first to find that spaced learning is more effective than blocked learning. Another more recent study by Rohrer (2012) discovered that students are more likely to make mistakes when all exposures to a concept are grouped together. Students made fewer errors when exposures to the same concept were spaced out according to the interleaving strategy. Most researchers investigating interleaving and blocking learning agree that the interleaving strategy is more effective. In contrast with the other studies, Carpenter and Mueller (2013) found that foreign language learners made fewer errors with a blocking strategy. Through multiple experiments, the authors showed that blocking benefited the learning results when native English speakers tried to learn French words. Moreover, this was true regardless of the number of foreign words that had to be learned. The authors stated that one possible explanation for this contrast in the field is that the efficiency of either one of the strategies may depend on the processing requirements of the task. It would be interesting to investigate whether the word distribution can be used to predict if a word will be learned correctly in a cross-situational word learning experiment.

Summary of exploratory variables

The previous studies provide insight into potentially predictive variables in cross-situational word learning, and the present study focused on five potentially predictive variables of correctly guessed words. First, this study investigated whether the moment that a word object is introduced for the first time is a predictor of whether the word is guessed correctly in the testing phase. This test is based on the primacy effect in the serial-position theory. Second, this thesis investigates whether the final time that a word object is presented can be used to predict the correctness in the test phase. This test is also based on the serial-position theory, but in this case, the recency effect was evaluated. The next test of word distribution determines whether the distribution of a word-object pair can be used to predict correctness in the test phase. A predictive model was then built to evaluate whether the response time of subjects in the training phase can be used to predict whether that word-object pair will be guessed correctly in the testing phase. Finally, whether isolated word-object pairs can be used to predict correctness in the testing phase was investigated.

Many studies have found that an increased frequency (number of repetitions) of a word-object pair results in better retention of that word-object pair. Since many studies have proven that frequency is a critical predictor of correctness, this study did not devote an individual test to the feature of frequency. In addition to testing the features individually, the outcome of the individual feature testing can be used to make a complete model with all features that predict whether a word-object pair will be guessed correctly. In the same way, individual features that cannot predict the outcome can be combined to evaluate whether a combination of these features leads to predictive powers.

2.2. Different types of learners

All the literature in the first part of the related work section (see section 2.1) describes learning theories that apply to learning in general, without differentiating between subjects. Examining whether a feature can be used to predict word learning for one subject, though the same feature may have no influence for another subject, is of interest. Likewise, a feature could be predictive for two different types of learners, but for one group, the feature could move in an inverse direction in relation to the outcome variable. Both of these results could suggest the existence of different types of learners. The expression 'different types of learners' refers to the possibility that persons learn and memorize information in different ways. In the general learning literature, the existence of different types of learners is highly debated, and there is no known research in cross-situational word learning that aims to identify different types of learners.

One famous model that demonstrates the differences between individual learning styles is the Dunn and Dunn learning styles model (Dunn & Dunn, 1978). According to this learning style theory, individual learners benefit from different factors when learning. Many of these factors are included in this model such as the structure of the task, the variety of the tasks, and physiological elements. Based on this theory, many web applications use questionnaires to suggest a specific learning style to individual learners. Moving from general learning differences to the specific area of word learning, the exploratory variables section (2.1) introduced the terms interleaving and blocking in learning. Although Bjork and Bjork (2011) found that most of the subjects performed better in interleaving learning, approximately 20% of the subjects made fewer errors with block learning. In other words, the results could indicate differences between learners. In a gender-related learning study, Dye et al. (2013) found evidence that boys and girls store and memorize words differently. Conversely, Pashler, McDaniel, Rohrer, and Bjork (2008) did not find evidence of different types of learners. The authors stated that many published guidebooks and other education materials are built around the concept of different learning styles, despite of the absence of scientific evidence. Kirschner (2017) is another researcher that rejects the individual learning styles theory, stating that the studies that report evidence of different learning styles fail to pass the criteria for scientific validity.

While some studies show evidence for different types of learners, other studies demonstrate that they do not exist. However, no study has attempted to identify different types of learners in a cross-situational, word-learning experiment. Clustering subjects with similar characteristics could provide insight into the identified features and show differences in learning between different subjects.

2.3. Added value of the thesis

In relation to the first research question, the existing work and literature addressed in the sections above show that little is known about features that predict learning in a cross-situational learning setting. The number of subjects in the data used for this thesis exceeds the number of subjects in all the studies mentioned above. Moreover, previous research findings on the existence of different types of learners have been inconsistent and contradictory. Some studies state that different types of learners exist while other studies cannot find evidence for the existence of different types of learners. As stated above, since data was available from many subjects for this thesis project, answering the second formulated research question could contribute to the field of word learning.

3. Methods

This section provides a precise description of how the current study was conducted. The first subsection describes the datasets and the initial experimental design that was employed to obtain the data from the subjects. A detailed description of the data, the features, and the structure are presented in subsection 3.1.2 and subsection 3.1.3. Furthermore, subsection 3.2 shows the descriptive statistics of the dependent variable that was used in the experiments, and then describes how data cleaning and pre-processing has been applied to obtain meaningful features and clean data. In the following sections, the exploratory data analysis (3.3), applied classifiers (3.4), training and test set (3.5), evaluation criteria (3.6), and method of unsupervised clustering (3.7) are described.

3.1. Datasets and data description

The data that was used for this thesis project consists of fifteen separate datasets that contain recorded experimental data from different experimental designs. The data was collected throughout 2016 by Dr. A.T. Hendrickson, professor at Tilburg School of Humanities in the department of Communication and Information Sciences. Adult participants were recruited from Amazon's Mechanical Turk and their ages ranged from 18 to 60 years old. The data used for this thesis project has not been published yet.

3.1.1. Experimental data

The datasets contain recorded data from subjects that participated in a cross-situational word learning experiment. In this experiment, participants were exposed to visual stimuli (i.e. photos of novel but realistic objects) and were required to listen to an audio recording that consisted of pseudo English words. The goal for participants was to combine the novel objects with the English pseudo words, and each object had only one name. In addition, the datasets contain a training phase and a testing phase. In the training phase, several objects were shown on one screen, in which the number of objects per screen depended on the condition. Each experiment consisted of multiple screens and each screen contained several objects depending on the condition. The training phase gave the subjects the opportunity to learn the word-objects pairs and the goal of the testing phase was to record memorized word-object pairs. Depending on the condition, the word-object pair appeared at an equal (uniform condition) or unequal (Zipfian condition) frequency across the training phase. The stimuli which were used for this experiment were obtained from the Novel Object and Unusual Name (NOUN) database (Horst & Hout, 2015), and an overview of the stimuli is shown in appendix A.

3.1.2. Data description

Experimental data was collected from 2690 unique participants throughout 2016. Some of the datasets are replicas of other datasets from a different time in 2016, while other datasets contain deviant conditions and experimental setups. A detailed description is presented for each dataset in appendix B, and a general description of the experimental designs is provided below. One of the most relevant conditions in various datasets is the word distribution:

- Zipfian *word-object pairs are presented with an unequal frequency. Words are distributed according to Zipf's law (Zipf, 1949).*
- Uniform *all word-object pairs are presented with an equal frequency.*

In addition to distribution, there are differences between experiments in the number of training screens, the number of words per screen, the number of unique words, and whether 'check items' are used. These differences are explained below:

- Number of training screens *the number of screens participants are presented with in the training phase (total presented words = number of screens * number of words per screen)*
- Number of words per screen *the number of words in one screen from which the participants must guess*
- Number of unique words *the total number of unique word objects to learn in an experiment*
- Use of check items *some experiments use check words which only appear once to check whether participants are writing words*

The available data for this thesis project contains approximately ten different experimental designs, from which some experiments were not cross-situational, word-learning experiments. Moreover, some experiments did not contain enough subjects for useful predictive implementation. The following four experiments were selected to answer the formulated research questions:

- 28 unique word-object pairs in the uniform condition with 70 training screens
    - *Each screen contained four objects and each pair was shown ten times*
- 40 unique word-object pairs in the uniform condition with 70 training screens
    - *Each screen contained four objects and each pair was shown seven times*
- 28 unique word-object pairs in the Zipfian condition with 70 training screens
    - *Each screen contained four objects and each pair was shown 1 to 65 times*
- 40 unique word-object pairs in the Zipfian condition with 70 training screens
    - *Each screen contained four objects and each pair was shown 1 to 65 times*

The final number of subjects for each experiment is shown in the next subsection (see section 3.2.2) after missing values, duplicates, and outliers are omitted from the data.

3.1.3. Data structure

Each row in the dataset represents the details of a single word-object pair that is presented to a subject in the training or testing phase of the experiment, such as information about the selected word, the correct word, the current screen number, and the accuracy of the current word-object pair. In other words, each row presents a participant's action. Appendix C provides an overview of all columns in the initial datasets before preprocessing.

3.1.4. Software

In this research, the programming language Python 3.0 (Python Software Foundation) was used with the aid of the open-source Jupyter Notebook web application server (notebook 4.2.3). The server ran on a local PC with Anaconda (Anaconda Navigator 1.3.1), and all the data preprocessing, analyzing, and visualizing that occurred in this research project was done with Python. The following packages of Python were used:

• Numpy	(Walt, Colbert & Varoquaux, 2011)
• Json	
• Pandas	(McKinney, 2015)
• Matplotlib	(Hunter, 2007)
• Sklearn	(Pedregosa et al., 2011)
• SciPy	(Walt, Colbert & Varoquaux, 2011)

3.2. Data manipulation

This subsection on data manipulation describes all the steps that were taken to retrieve clean and usable data to perform and implement the desired machine-learning algorithms. To retrieve optimal results, preprocessing is one of the most important aspects of machine learning (Raschka & Mirjalili, 2017). In addition, feature engineering is important for this thesis since relevant indicators had to be created from the available raw experimental data. The steps to address missing values and outliers are also described in the current subsection. Before the preprocessing was conducted, whether the outcome variable was consistent over all experiments was investigated.

3.2.1. Outcome variable

The target variable in all of the following tests is the feature accuracy of a word-object pair in the test phase of the experiment. The feature accuracy is a binary variable, in which a value of one indicates a correctly guessed word-object pair and a value of 0 indicates an incorrect guess. Before proceeding to the data manipulation section, the descriptive statistics of the dependent variable in the uniform condition are shown in Table 1, while the descriptive statistics of the outcome variable in the Zipfian condition are presented in Table 2.

Table 1. Target variable (correctly guessed words) description in the uniform condition.

Feature name	**Feature description**	**28 word-object pairs** ***(N = 6,636)***				**40 word-object pairs** ***(N = 2,960)***			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
Accuracy	Accuracy in test per word-object pair; 1 = correct guess, 0 = incorrect guess.	.37	.50	.00	1.00	.25	.43	.00	1.00

Table 2. Target variable (correctly guessed words) description in the Zipfian condition.

Feature name	**Feature description**	**28 word-object pairs** ***(N = 8,876)***				**40 word-object pairs** ***(N = 3,880)***			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
Accuracy	Accuracy in test per word-object pair; 1 = correct guess, 0 = incorrect guess.	.44	.50	.00	1.00	.27	.45	.00	1.00

The previous tables show that more word-object pairs are guessed correctly in the condition with 28 pairs, and subjects seem to score better in the Zipfian experiments than in the uniform experiments. Figure 1 below shows the proportion of correctly guessed words in the 28 and 40 word-object pair conditions for both the uniform and Zipfian experiments in two boxplots.

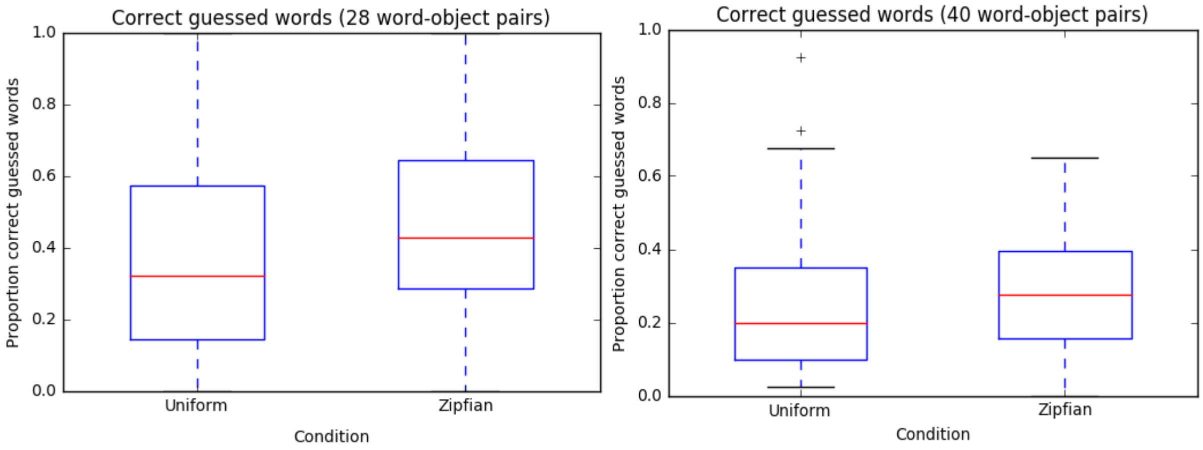


Figure 1. The proportion of correctly guessed words in the test phase per subject in the experiment with 28 unique word-object pairs. The x-axis of the boxplot is split for the uniform and Zipfian conditions.

The previous results show that the proportion of correctly guessed words highly depends on the condition that subjects are in. Word-object pairs are more likely to be learned in the Zipfian experiment in the condition with 28 word-object pairs. Due to the variety of outcome variables that depend on the condition, the machine learning algorithms were performed separately for each group:

- Uniform and 28 word-object pairs
- Uniform and 40 word-object pairs
- Zipfian and 28 word-object pairs
- Zipfian and 40 word-object pairs

3.2.2. Preprocessing and feature engineering

A significant amount of preprocessing and feature engineering was required for this thesis project as the available data was spread over fifteen different datasets, feature engineering was necessary to create meaningful features, and the data included many different experiments. As explained in the previous subsection, each row in the datasets describes an action of a participant. The structure of the raw datasets was not desired since it could not serve as input to the machine-learning models. The appropriate format for most tests was to create a row for each word-object pair. For example, ten participants in an experiment who are shown 28 word-object pairs results in a data frame of 280 rows *(10 * 28 = 280)*. This also created an opportunity to summarize the results per subject by grouping the results based on *subjectID* when desired.

To achieve the desired data format, a data frame was created with the Pandas library, including the following columns: *SubjectID, correct_word, condition, experiment, and seen1 to seen65*. The *seen* columns represent the screen numbers in which a word-object pair is shown in the training phase. The maximum screen number is 65, as that was the maximum frequency that a pair occurred in the training phase in the Zipfian condition.

Word distribution

The word distribution test could only be completed in the uniform condition because some pairs had no distribution in the Zipfian condition if they only occurred once. Previous preprocessing led to a data frame with all index numbers of when a word-object was presented in the training phase. Distances between word-object presentations were determined by simple calculations based on the word-object screen indexes. When a word-object appeared at the screen indexes of *0, 5, 12, 60, 65, and 78,* the distance between the same word-object presentation was calculated by the formula of *distance = N+1 – N* for all indexes in the array. In the previous example, this would result in a distance array of *5, 7, 48, 5, and 13.* Feature engineering was important for the test word distribution, as the raw data was inefficient for further data analysis. Five variables were created to serve as input features: *minimum distance, mean distance, sum of distances, maximum distance,* and *tightness of the word distribution.*

The *minimum distance* feature was calculated by taking the lowest number for each word-object array. *Sum distance,* mean *distance,* and *maximum distance* were calculated similar to the *minimum distance*, but with the sum, mean, and maximum statistic. *Tightness of the word distribution* was created by scoring word-object arrays that had some of the words clustered close together. A tight cluster was defined as at least three word-object pairs presented on less than 15 screens. This number of word-object pairs and screen combinations was chosen by testing different values and taking highest scoring one (result section).

First presentation

After the first steps of preprocessing, the data was formed such that all word-object pairs were indexed for each presentation in the training phase. The previously created column *seen1* contained the screen indexes for all word-object pairs across all subjects for word-objects that appeared for the first time in the training phase. The feature *seen1* is sufficient in the uniform condition and the Zipfian condition, because all word-object pairs appear at least one time in each condition.

Final presentation

In the uniform condition, all words appear at an equal frequency. When words appear ten times (28 pairs condition) in the training phase, the final presentation of a word is found in the *seen10* column. In the condition with 40 pairs, words appear seven times and the index number of the final presentation can be found in the *seen7* column. Unlike the uniform condition, word-object frequency is unequal in the Zipfian condition. The frequency of words varies from a single presentation up to 65 presentations. In this condition, the number of columns that contain the screen index number differ depending on the number of presentations. As most machine learning algorithms cannot deal with data of unequal (column) lengths, a new feature named *last_seen* was created by iterating over each row and by taking the content of the last word-object index column that contained a value.

Reaction time

New variables were created to include the reaction time in the data frame, and the reaction time was measured each time a respondent made a guess. In the 28 unique word-object pair condition, the variables of *reaction time seen 1* to *reaction time seen 10* were created, and the variables of *reaction time seen 1* to *reaction time seen 7* were created in the condition with 40 words. In the Zipfian condition, the frequency of the same word-object pair ranged from a single presentation to 65 presentations. Thus, descriptive features had to be created to obtain valid input of the same length for the machine learning, and the features *mean reaction time, minimum reaction time, and maximum reaction time* were created.

Context effects

New features were built to determine whether subjects are more likely to learn a word if they correctly selected all other word-object pairs presented on the training screen. In all experimental data that was used for this thesis project, each training screen consisted of four word-object pairs. Moreover, a feature called *number of last occurrences* was created which reflected the number of times a specific word-object pair had to be selected as last in a training screen to be correct. A second feature was designed to capture the number of times that the first three word-object pairs on a screen were selected correctly when a specific word-object pair was asked for last in the training screen. The last selected word object is always correct when the previous three are guessed correctly and this feature is called *number of last occurrences while all word-object pairs in screen are guessed correctly.* For example, ‘word-object A’ appears ‘X times’ as the final selected word-object pair (first described feature) and of those ‘X times’, all word-object pairs are correctly selected ‘Z times’ (second described feature).

3.2.3. Missing values, duplicates, and outliers

The provided datasets contained missing instances for some subjects from the experiments, and the missing rows translated into missing actions and user input of the concerned subjects. Subjects that did not complete the experiment were omitted because the actions of these subjects were uncertain, and 68 subjects were deleted. In addition, 55 duplicate instances of subjects were found across the datasets and were deleted.

For the test reaction time in the uniform condition with 40 word-object pairs, one error was found (negative number) and this row was therefore omitted. Another 17 instances were found where respondents took more than 30 seconds to provide an answer, and these rows were also omitted since those outliers highly influenced the corresponding descriptive statistics. In the same test with 28 pairs, 13 errors (negative numbers) and 61 outliers were omitted. Finally, 16 outliers were omitted in the Zipfian condition with 40 word-object pairs, and with 28 pairs 67 instances were deleted. After removing missing values, duplicates, and outliers, the following number of subjects were left for each condition:

28 unique word-object pairs experiment

- Uniform condition
    - *237 unique subjects*
    - *6,636 rows*
- Zipfian condition
    - *317 unique subjects*
    - *8,876 rows*

40 unique word-object pairs experiment

- Uniform condition
    - *74 unique subjects*
    - *2,960 rows*
- Zipfian condition
    - *97 unique subjects*
    - *3,880 rows*

3.3. Exploratory data analysis

The exploratory data analysis offers insight into the descriptive statistics of all the features and the target variable with respect to predicting the different formulated tests. Furthermore, this section is separated according to each test that was formulated in the previous sections. For tests in the Zipfian condition, the *frequency of a word-object pair* was added as a mediating variable, which is important because the variables used in the tests can reflect the approximate frequency of a word-object pair. For example, if the first introduction of a word-object pair in the Zipfian condition is on the last screen of the training phase, the frequency of that pair is also one.

3.3.1. Word distribution

The distribution of words across all screens in the training phase is random, as each word-object pair appears on a random screen across subjects. This test determined whether words that are presented in a tight cluster are learned differently (better or worse) than words that are spaced out over the training phase. This test also investigated if and how the distribution of words predicts the dependent variable accuracy per word-object pair. In other words, this analysis was conducted for each word-object pair rather than all word-object pairs combined for each unique participant.

Uniform condition

In this experiment, the independent features of *mean, minimum, maximum, sum, and tightness* were defined to predict the target feature *accuracy in test per word-object pair*. In this test, only the uniform condition could be examined. Table 3 shows the feature descriptions of the features used for this experiment.

Table 3. Feature descriptions of the uniform design for the experiment word distribution.

Feature name	**Feature description**	**28 word-object pairs (*N = 6,636)***				**40 word-object pairs (*N = 2,960)***			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
Mean	Mean distance between equal word-object pairs	6.48	.81	2.22	7.78	8.88	1.61	3.17	11.50
Minimum	Minimal distance between equal word-object pairs	1.27	.54	1.00	4.00	1.96	1.20	1.00	8.00
Maximum	Maximal distance between equal word-object pairs	17.42	5.17	6.00	45.00	21.03	6.60	6.00	50
Tightness	Integer number for each time at least three equal word-object pairs appear in fewer than fifteen screens	1.29	1.22	.00	7.00	2.13	1.14	.00	4.00
Sum	Sum of all the distances per word-object pair.	58.34	7.31	20.00	70.00	53.29	9.68	19	69

Note. The experiments with 28 and 40 word-object pairs had the same number of training screens. Each word-object pair was presented ten times in the first condition (28 words) and seven times in the later condition (40 words).

Table 3 shows that the mean distance in terms of screens between a word-object and the next presentation of the same word-object pair is 6.48. In other words, the average space between a word-object to reoccur is around 6.48 screens. The sum of the distances shows the total distance for all presentations of a word-object pair; the minimum sum in the datasets is 20 screens and the highest sum is 70 screens. The minimum of 20 shows a tight word-object pair cluster where all ten presentations of a word-object pair are achieved on 20 screens. The highest sum is 70 screens, which mean that the presentations of that word-object pair are spaced out over all 70 screens.

3.3.2. First presentation

As explained in previous sections, the training phase of the cross-situational word learning experiment is divided between several screens. This section determines the impact of the first presentation of a word-object introduction on memorizing the word to learn whether early introduced word-object pairs have an advantage over later introduced word-object pairs, or vice versa. An analysis was conducted to estimate the influence of first word-object presentations.

Uniform condition

For this test, the feature of *first word introduction* was used to predict *accuracy in test per word-object pair*. The feature *first word introduction* is an integer which reflects the screen number in which a word-object pair is introduced. The table below offers insight into the descriptive statistics of the previously mentioned variables.

Table 4. Feature descriptions of the uniform design for the experiment’s first presentation

Feature name	**Feature description**	**28 word-object pairs** *(N = 6,636)*				**40 word-object pairs** *(N = 2,960)*			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
First word introduction	Screen number in training phase of first introduced word-object pairs.	5.46	5.44	0	46	7.84	7.29	0	45

In the uniform condition, the first word-object presentation of all pairs range from 0 to 46 screens. Figure 2 shows the accuracy per screen number in the test phase for word-object pairs that were introduced in the first ten screens of the training phase. The percentage of correct words is highest for the first screen of the experiment with 28 word-object pairs, but this does not apply to the experiment with 40 pairs. The figure does not show any other specific trends from the first screen to the last screen.

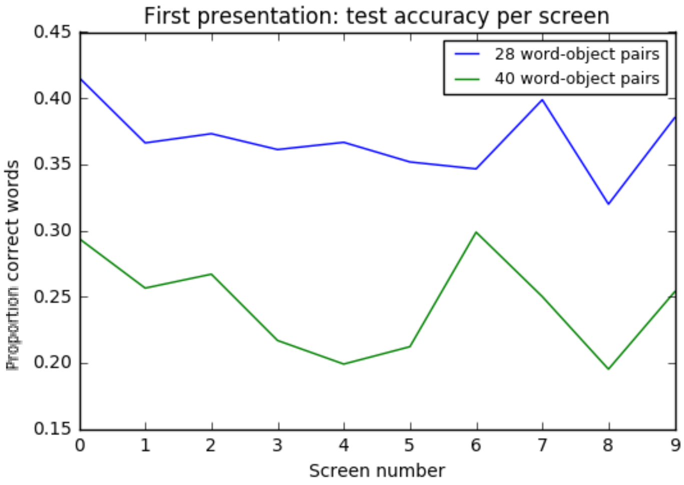


Figure 2. Percentage of correctly guessed word-object pairs in the test phase of the experiment for the first ten training screens based on the first presentation of a word-object pair.

Zipfian condition

In the Zipfian condition a word-object pair could be introduced in the last screen of the training phase as a pair may only be presented once. When a word-object pair is introduced in the last few screens of the training phase it may also reflect the frequency that a word-object will appear. The results show a significant negative correlation between *first word introduction* and *frequency of word-object pair (*r = -.38) in the 28 unique word-objects condition. In the condition with 40 words, a significant negative correlation was also found between the same variables (r = -.37). The higher the initial screen number of a word introduction, the lower the possible frequency. Therefore, frequency was added as a mediator in the Zipfian condition for the first presentation test. The importance of word introduction as a predictor can be measured by assembling two models: one model with word introduction and frequency as predictors, and one model with only word frequency as a predictor. The differences between scores can then be calculated. Table 5 presents an overview of the descriptive statistics for the features used in this Zipfian condition test.

Table 5. Feature descriptions of the Zipfian design for the first presentation experiment

Feature name	**Feature description**	**28 word-object pairs (*N* = 8,876)**				**40 word-object pairs (*N* = 3,880)**			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
First word introduction	Screen number in training phase for first introduced word-object pairs.	11.68	13.96	.00	70.00	16.05	16.32	.00	69.00
Frequency	The number of presentations of a word-object pair.	10	12.83	1.00	69.00	7.00	10.85	1.00	67.00

Table 5 shows that words in the Zipfian condition were introduced starting from the first screen (screen 0) in the training phase to the last screen (screen 69 or screen 70). The mean screen introduction in the 28-pair condition is lower than in the 40-pair introduction, since the same word-object pairs were presented more often in the 28-word condition. This is also reflected in the descriptive statistics of the *word frequency.*

Note. A line plot with the proportion of correctly guessed words is un-informative in the Zipfian condition as a word-object pair may be introduced from the first screen to the last screen.

3.3.3. Final presentation

The final presentation experiment was similar to the first presentation experiment, but it described the influence of the final presentation of word-object pairs instead of evaluating the influence of first presentations. In other words, it determined whether the final presentation of a word-object influenced the accuracy of the test.

Uniform condition

The *word last seen* feature was used to predict *accuracy in test per word-object pair*. The integer feature reflects the screen number on which a word-object pair is last seen in the training phase. Table 6 shows an overview of the features and descriptive statistics.

Table 6. Feature descriptions of the uniform design for the final presentation experiment

Feature name	**Feature description**	**28 word-object pairs** ***(N = 6,636)***				**40 word-object pairs** ***(N = 2,960)***			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
Word last seen	Screen number in training phase for last presentation of word-object pairs.	63.87	5.45	25	70	61.13	7.40	25	69

The table above shows that the last presentation of a word-object ranges from screen 25 to screen 70. Figure 3 shows the accuracy per screen number for the last ten screens, for word-object pairs their last presentation. The line graph does not show a clear trend of accuracy based on words and their last presentation.

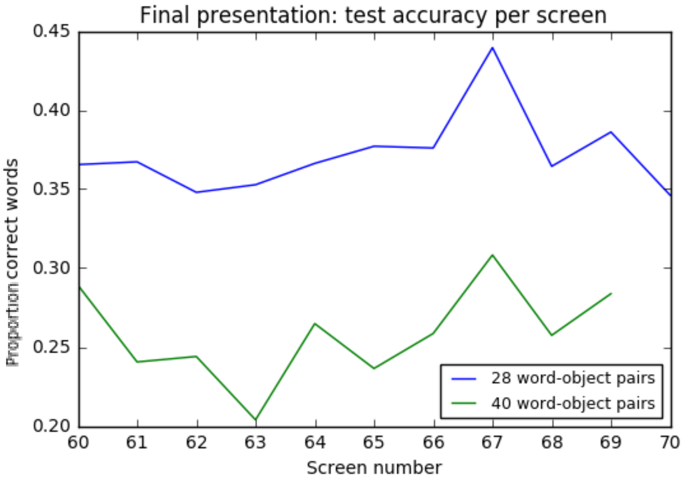


Figure 3. Percentage of correctly guessed word-object pairs in the test phase of the experiment for the last ten training screens based on the final presentation of a word-object pair. Part of the 28 word-object experiments also contained four 'check items' in which each check item appeared once. Therefore, there was one extra screen in the training phase.

Zipfian condition

The final presentation test in the Zipfian condition is similar to the first presentation test in the same condition. In this case, a word-object's final presentation could be in the first screen of the training phase, and this number could reflect the frequency of the word-object pair. The correlation results show a significant negative relation between *final word introduction* and *frequency of word-object pair* (r = .38) in the 28 word-object pair condition and (r = .37) in the condition with 40 word-objects. Since the frequency increases with the number of the last presentation, frequency was added as a mediator in the Zipfian condition. The importance of final presentation as a predictor can be measured by assembling two models: one with word introduction and frequency as predictors and one with only word frequency as a predictor. The difference between scores can then be calculated. The descriptive statistics of both the predictors are shown in Table 7.

Table 7. Feature descriptions of the Zipfian design for the final presentation experiment

Feature name	**Feature description**	**28 word-object pairs (*N* = 8,876)**				**40 word-object pairs (*N* = 3,880)**			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
First word introduction	Screen number in training phase for first introduced word-object pairs.	57.74	14.14	.00	70.00	52.62	16.71	.00	69.00
Frequency	The number of presentations of a word-object pair.	10	12.83	1.00	69.00	7.00	10.85	1.00	67.00

As the table suggests, the last time a word-object pair was presented ranges from screen 0 to screen 70. In contrast to the uniform condition, some words only occur once in the Zipfian condition, and the last (and first) presentation can be shown on screen 0.

Note. A line plot with the proportion of correctly guessed words is un-informative in the Zipfian condition as the final presentation of a word-object pair may be presented from the first screen to the last screen.

3.3.4. Reaction time

The goal of the reaction time test was to predict whether respondents gave the correct word-object combinations in the test phase, with reaction time as the independent feature. Reaction time was measured during the training phase of the experiment and tested against the correctness of the selected word-object pair in the testing phase of the experiment.

Uniform condition

The features of *accuracy seen 1* to *accuracy seen 7* were used (i.e. seven same word-object presentations) to predict *accuracy in test per word-object pair* in the uniform condition with 28 word-object pairs. In the uniform condition with 40 object pairs, the features of *accuracy seen 1* to *accuracy seen 10* are used (10 same word-object presentations) to predict correct guessed words. Table 8 offers an overview of the features' descriptive statistics in the uniform condition (all statistics are in seconds), and shows that the average reaction time descends as the number of presentations of the same word-object increases.

Table 8. Feature descriptions of the uniform design for the reaction time experiment (RT).

Feature name	**Feature description**	**28 word-object pairs (*N = 6,562)***				**40 word-object pairs (*N = 2,942)***			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
RT screen 1	In seconds	1.95	1.59	.01	23.29	1.85	1.40	.03	21.67
RT screen 2	In seconds	1.81	1.32	.01	24.98	1.79	1.26	.01	23.59
RT screen 3	In seconds	1.78	1.40	.01	28.33	1.70	1.16	.07	26.68
RT screen 4	In seconds	1.74	1.40	.01	28.97	1.70	1.30	.01	29.24
RT screen 5	In seconds	1.68	1.46	.01	26.51	1.73	1.40	.01	22.18
RT screen 6	In seconds	1.68	1.50	.01	27.04	1.66	1.35	.01	29.96
RT screen 7	In seconds	1.63	1.62	.01	29.65	1.60	1.14	.02	17.96
RT screen 8	In seconds	1.64	1.53	.01	28.92	-	-	-	-
RT screen 9	In seconds	1.59	1.45	.01	28.53	-	-	-	-
RT screen 10	In seconds	1.56	1.45	.02	27.11	-	-	-	-

Note. In the experiment with 40 word-object pairs, the same pairs only occur seven times.

Zipfian condition

As explained in the preprocessing section (subsection 3.2.1.), the reaction times per screen could not be used in the reaction time experiment in the Zipfian condition. The descriptive statistics for the features of *mean reaction time, minimum reaction time, and maximum reaction time* in the Zipfian condition are shown in Table 9.

Table 9. Feature descriptions of the Zipfian design for the experiment reaction time (RT).

Feature name	**Feature description**	**28 word-object pairs** ***(N = 8,809)***				**40 word-object pairs** ***(N = 3,864)***			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
Mean RT	Average RT (in seconds)	1.64	.77	.01	14.41	1.61	.91	.01	29.00
Min RT	Minimum RT (in seconds)	1.00	.53	.00	14.31	1.10	.79	.00	28.99
Max RT	Maximum RT (in seconds)	3.10	2.61	.01	29.33	2.58	2.13	.01	29.61

The descriptive reaction times are similar to each other in the 28 word-object and 40 word-object pair conditions. Moreover, the reaction times are faster on average in the Zipfian condition than in the uniform condition (Table 8).

3.3.5. Context effects

There are various potential context effects that could predict word learning, and one possible context effect was tested for this thesis: are subjects more likely to learn a word if they correctly selected the object for all other words presented on the screen? The same features were used for the tests in the uniform and Zipfian conditions.

Uniform condition

Table 10 presents an overview of the descriptive statistics for the features used in the context effects test. Statistics for the features of *last correct pair* and *number of screens without uncertainty* are shown in the table below. In the 28 unique pairs condition, the average number of screens in which a word-object is presented lastly while all words are guessed correctly is .68. In the condition with 40 words, the average number of screens is .36. This means that pairs are represented without uncertainty more often in the 28 pairs condition.

Table 10. Feature descriptions of the uniform design for the context effects experiment

Feature name	**Feature description**	**28 word-object pairs (*N* = 6,636)**				**40 word-object pairs (*N* = 2,960)**			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
Last correct pair	number of last occurrences in training phase	2.50	1.35	.00	8.00	1.75	1.13	.00	6.00
Number of screens without uncertainty	number of last occurrences while all word-object pairs in screen are guessed correctly in training phase	.68	.93	.00	6.00	.36	.64	.00	4.00

The figure below (Figure 4) shows the percentage of correctly guessed words for the different values of number of screens without uncertainty in the uniform condition. The graph shows a clear pattern; the more often a word-object pair is shown while all other word-object pairs are guessed correctly, the higher the average accuracy is in the test phase.

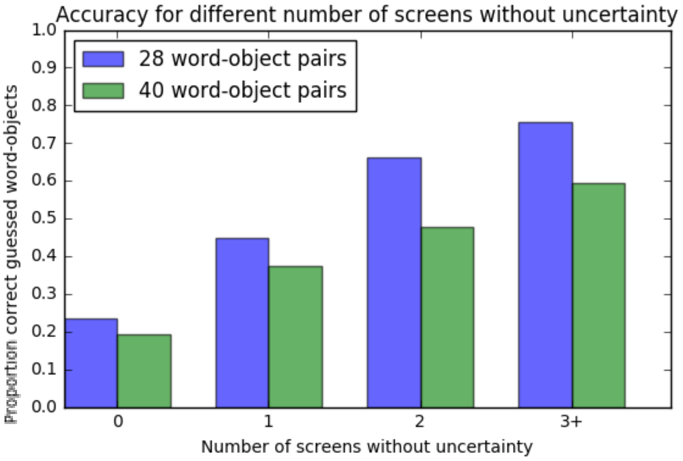


Figure 4. The proportion of correctly guessed word-objects based on the number of screens without uncertainty. The number on the x-axis represents the number of times that a word-object pair is on a screen where no uncertainty is present because all other word-object pairs on that screen have been selected correctly.

Zipfian condition

The features used in the uniform condition were also used in the Zipfian condition, and Table 11 offers insight into the descriptive statistics that were used in this current test. As in the uniform condition, the condition with fewer unique words on average (1.53) has more screens on which a word-object must be selected last when all other pairs are guessed correctly, compared to the average of the condition with 40 pairs (1.02).

Table 11. Feature descriptions of the Zipfian design for the context effects experiment

Feature name	**Feature description**	**28 word-object pairs**				**40 word-object pairs**			
		(N = 8,876)				***(N = 3,880)***			
		Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
Last correct pair	number of last occurrences in training phase	2.50	3.47	.00	29.00	1.75	2.92	.00	28.00
Number of screens without uncertainty	number of last occurrences while all word-object pairs in screen are guessed correctly in training phase	1.53	2.52	.00	24.00	1.02	2.05	.00	19.00

Figure 5 presents the percentage of correctly guessed words for the different number of screens without uncertainty in the Zipfian condition. The graph shows a clear pattern, where the more often a word-object pair is shown while all other word-object pairs are guessed correctly, the higher the average accuracy is at the test phase.

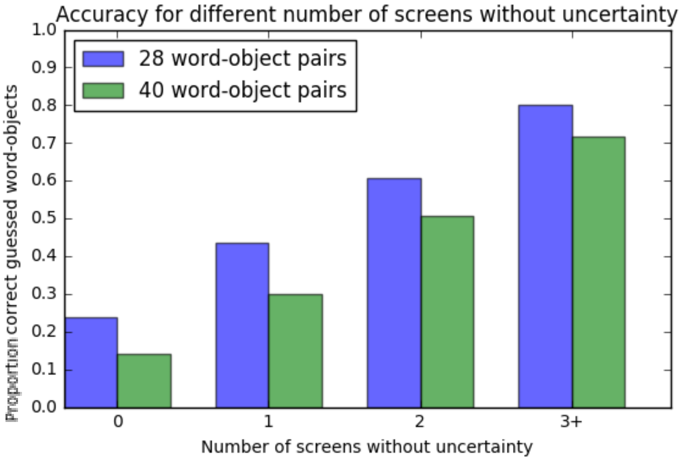


Figure 5. The proportion of correctly guessed word-objects based on the number of screens without uncertainty. The number on the x-axis represents the number of times that a word-object pair is on a screen where no uncertainty is present because all other word-object pairs on that screen have been selected correctly. Note: part of the variance in the graph is based on the frequency of a word-object pair.

3.4. Applied classifiers

In this subsection, the algorithms that were used for the classification tasks are reviewed. An important aspect of machine learning is that there is no universal algorithm that fits all problems (Caruana & Niculescu-Mizil, 2006). Since there are relatively few studies that have built predictive models in the domain of cross-situational word learning, the selection of machine-learning algorithms for this thesis was not based on previous research, but on the known advantages of specific algorithms.

In this thesis project, two distinct machine learning algorithms were used to build predictive models: a logistic regression algorithm and a random forest algorithm. Both algorithms were implemented for each of the individual feature experiments and for the complete predictive models. Using two different models provided an opportunity to find different patterns and to compare the models to each other. The algorithms used in the clustering analysis are described in the unsupervised clustering section (see Section 3.7). In addition, a justification for the selected models and a description of the chosen algorithms are presented below.

3.4.1. Logistic Regression

The logistic regression algorithm is a technique that is used for binary prediction in which an observation belongs to a category based on a probability estimate. One strength of the algorithm is that the coefficients used for decision making are informative, and they show the strength of the relationship between the predictors and the outcome variable. In a similar manner, the direction (positive versus negative) between the features and outcome variable are given (Schein & Ungar, 2007; Raschka & Mirjalili, 2017; Caruana & Niculescu-Mizil, 2006).

3.4.2. Random Forest

The strengths of the random forest algorithm are the ability to handle unbalanced data and non-linear features, and the ability to extract feature importances (Breiman, 2001; Raschka & Mirjalili, 2017; Biau, Devroye & Lugosi, 2008). The ability to deal with non-linear features is a strong asset of the classifier. Another great strength of the random forest classifier is the ability to access the feature importance for all of the features used.

3.5. Training and test set

For the algorithms, the datasets were divided into training and testing sets. The training set was a randomly selected partition of 80 percent, and the remaining 20 percent was used as the testing set. To generalize the results and evaluate the performance on unseen data and avoid underfitting and overfitting (high bias versus high variance), a ten-fold cross validation was used for the training set. Cross validation can be used to fine-tune the models' parameters. In the k-fold cross validation method, the training set was randomly split into *k* folds without replacement. To train the model, *k – 1* folds were used and one was kept separately to measure performance. This procedure was repeated *k* times using ten folds (Raschka & Mirjalili, 2017), and the outcome was the mean *k* score.

3.6. Evaluation criteria

Two different metrics were used to evaluate the outcome of the tests: accuracy and area under the receiver operating characteristic curve (AUROC). Since evaluating with accuracy as the only metric is not desired in tasks with imbalanced classification problems, the models were also evaluated with the AUROC curve.

3.6.1. ZeroR Classifier

To compare and evaluate the accuracy scores of the machine learning algorithms, the ZeroR (or Zero Rule) classifier was used. This classifier finds the majority class and predicts that class for all instances (Beckham, 2015). In this thesis project, the outcome class is imbalanced and the ZeroR classifier is more suitable than a random guessing baseline.

3.6.2. Area Under the Receiver Operating Characteristic Curve (AUROC)

The AUROC curve is plotted with the true positive rate as a function of the false positive rate with varying thresholds. Moreover, the area under the curve (AUC) explains the predictive power of the model. A score of 0.5 means that the model performs very poorly, has no predictive power, and scores similar to random guessing. The highest score is 1.0, which indicates a perfect classifier (Hanley & McNeil, 1982; Bradley, 1997).

3.7. Unsupervised clustering

The second research question of this thesis project focuses on finding different types of learners. With an unsupervised clustering approach, the aim was to find distinctions between different subjects. If different clusters are found, each individual cluster can be investigated to locate the characteristics of that cluster.

This current subsection describes all the steps taken to conduct the unsupervised clustering learning method. The first part of this subsection (3.7.1.) illustrates which data is used to serve as input data for the clustering algorithms. Next, subsection 3.7.2. describes the steps to improving pattern recognition for the potential clusters. The last subsection (3.7.3.) of this section deals with the evaluation methods of the chosen clustering algorithms.

3.7.1. Clustering data

In the clustering analysis, the individual features described in the previous section were combined to identify different types of learners. All four conditions used in the previous sections were combined for one clustering analysis across all experiments. The features frequency and word distribution cannot be used in the clustering analysis, as these features were only tested in one of the two experiments (uniform or Zipfian).

In the individual feature experiments, each sample consisted of one word-object pair, and no connection between word-object pairs and subjects were made. To find possible different types of learners, the data had to be reshaped such that it contained information per subject, instead of information per word-object pair. There are several methods for retrieving features for each subject, instead of each word-object pair. One logical way to solve this problem is to use estimated regression coefficients that serve as input to the clustering algorithms (Tarpey, 2007). The main benefit of using the coefficients from the model above and using a scoring metric such as accuracy is that coefficients also show the direction of the decision making (i.e. negative versus positive coefficients). Moreover, two methods for using regression coefficients as input to the clustering analysis were found. Both the methods use the logistic regression algorithm to build a model for each subject, as the logistic regression offers the opportunity to extract the coefficients that are used in the decision function.

The first method is to build a logistic regression model and repeat this model for each subject with all the individual features together. The intercept and the coefficients per feature that are obtained for each subject can then be used as input features for the clustering analysis. The second potential method is to build a logistic regression model and repeat this model for each subject and for all the individual features separately. In this case, four features are used which will result in four predictive models per subject. The advantage of the second option is that each individual feature has its own intercept, while only one intercept is available in the first method because all the features are combined. To provide more clarity, an illustration of the two methods is shown in Figure 6 and Figure 7. The second method was selected for the clustering analysis, as each feature has its own intercept and more information is available to serve as input.

To build reliable machine-learning models for each subject, some subjects had to be omitted in the clustering analysis. Some subjects incorrectly guessed almost all word-object pairs, and some guessed almost all word-object pairs correctly, which leads to highly imbalanced classes in the target variable and can cause problems if few instances are positive or negative classes. Moreover, modeling subjects that guessed everything correctly or incorrectly does not add information. Therefore, subjects that answered four or fewer questions correctly and four or fewer questions incorrectly were omitted from further analysis, and 491 subjects remained.

Method 1 – One predictive model per subject

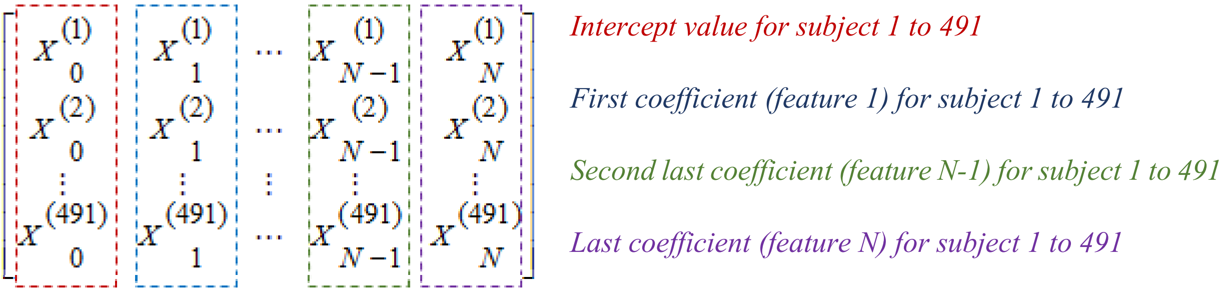

$X_0^{(1)}$ $X_1^{(1)}$ … $X_{N-1}^{(1)}$ $X_N^{(1)}$
$X_0^{(2)}$ $X_1^{(2)}$ … $X_{N-1}^{(2)}$ $X_N^{(2)}$
$X_0^{(491)}$ $X_1^{(491)}$ … $X_{N-1}^{(491)}$ $X_N^{(491)}$
Intercept value for subject 1 to 491
First coefficient (feature 1) for subject 1 to 491
Second last coefficient (feature N-1) for subject 1 to 491
Last coefficient (feature N) for subject 1 to 491

Figure 6. *Method 1, an example matrix of the results of a logistic regression model for all of the individual tests combined per subject.*

Method 2 – Predictive model per feature and per subject

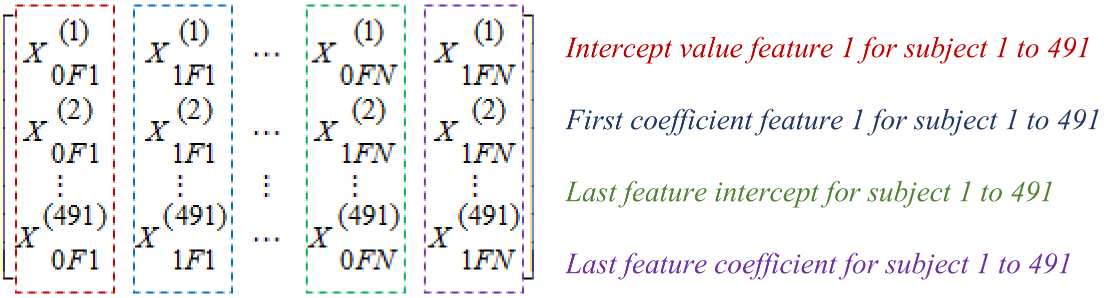

$X_{0F1}^{(1)}$ $X_{1F1}^{(1)}$ … $X_{0FN}^{(1)}$ $X_{1FN}^{(1)}$
$X_{0F1}^{(2)}$ $X_{1F1}^{(2)}$ … $X_{1FN}^{(2)}$ $X_{1FN}^{(2)}$
$X_{0F1}^{(491)}$ $X_{1F1}^{(491)}$ … $X_{0FN}^{(491)}$ $X_{1FN}^{(491)}$
Intercept value feature 1 for subject 1 to 491
First coefficient feature 1 for subject 1 to 491
Last feature intercept for subject 1 to 491
Last feature coefficient for subject 1 to 491

Figure 7. *Method 2, an example matrix of the results of a logistic regression model for each individual test per subject.*

The final matrix, used for the unsupervised clustering contains 11 features (dimensions) before dimensionality reduction, which include the following elements:

- First presentation
  - Intercept first presentation
  - One coefficient
- Last presentation
  - Intercept last presentation
  - One coefficient
- Intercept reaction time
  - Intercept reaction time
  - Three coefficients: minimum reaction time, mean reaction time, and maximum reaction time
- Context effect
  - Intercept context effect
  - Two coefficients: last correct pair and number of screens without uncertainty

3.7.2. Dimensionality reduction

Dimensionality reduction is desired for the unsupervised clustering analysis as distance measures do not work well in high-dimensional spaces (Raschka & Mirjalili, 2017). The most widely used technique for dimensionality reduction is the principal component analysis (PCA). Before the PCA was implemented, the data was normalized to maximize the variance.

Figure 8 shows the cumulative proportion of the variance explained for different numbers of dimensions. All 11 dimensions are needed to capture all the variance of the data. To reduce the highly dimensional data, however, a lower number of dimensions was selected. The figure shows that eight dimensions capture more than 90% of the variance. Therefore, eight components were selected that serve as input to the clustering algorithms. The negative side of eight-dimensional data is that visualizing the clusters becomes harder. Since the captured variance is below .6 when only three dimensions are selected, a significant part of the variance would be lost.

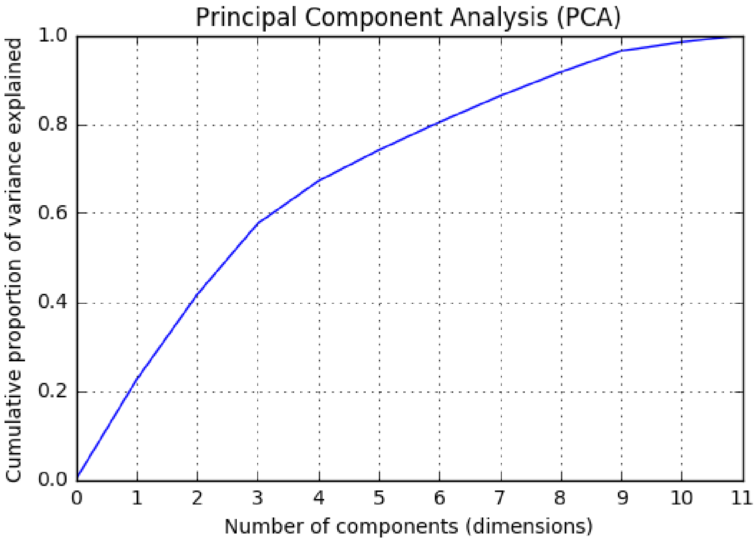


Figure 8. Number of components in the principal component analysis and the cumulative proportion of the variance that it explains.

3.7.3. Clustering algorithms

Each clustering algorithm has strengths and weaknesses, and accessing more than one clustering algorithm provides an opportunity to compare methods and identify the best performing method. In this thesis project, two distinct clustering algorithms were used: a Gaussian mixture model and an agglomerative hierarchical clustering approach. The advantages of both algorithms are stated below.

Gaussian mixture model

An important advantage of the Gaussian mixture model is that a cluster does not need to follow a specific structure, which enhances flexibility compared to a k-means clustering algorithm. Moreover, research has shown that the Gaussian mixture model can effectively be used for high-dimensional data (Azizyan, Singh & Wasserman, 2014). The Gaussian mixture model seems to be a valid choice as the input data is high dimensional and the model does not follow a specific structure.

Hierarchical clustering (agglomerative)

Each observation starts as its own cluster in hierarchical clustering and from there, observations are merged into larger clusters. A strong feature of hierarchical clustering is that the structure of high dimensional data can be visualized with the aid of a dendrogram. The dendrogram displays the hierarchical relationship among the observations. Another asset of hierarchical clustering is the ability to see relationships in the data on different levels of the dendrogram.

3.7.4. Unsupervised clustering evaluation

To evaluate the quality of the unsupervised clustering analysis the silhouette coefficient was used (Rousseeuw, 1987). The silhouette analysis was used to validate the consistency of the clusters that were found. This metric measures how tightly grouped the instances in the clusters are (Raschka & Mirjalili, 2017), and it ranges between -1 and 1. A score near 1 indicates a perfect separated cluster, while a score near 0 indicates no identified structure in the data. Furthermore, a negative score indicates the possibility that a sample is assigned to the wrong cluster.

4. Results

The first part of this section provides the results of the five individual tests that are defined in the related work section, and the second part offers the results of the complete models. One complete model was made containing all non-predictive features together and one complete model was made that contains all the features that were predictors of correctly guessed words. Finally, the last part presents the results of the clustering approach for different types of learners. The parameters of the models that were used for each test, the complete models, and the clustering analysis are shown in the appendices (appendix D).

4.1. Part I – Feature testing

Subsection 4.1 presents the results of the following features: the word distribution test, first presentation test, final presentation test, reaction time test, and context effect test. For experiments where both the uniform and Zipfian condition were tested, the two conditions are split into two subsections.

4.1.1. Word distribution

The word distribution test investigated whether the distribution of words (tight versus spaced out distributions) can predict word learning. The table below (Table 12) shows the accuracy per model including the baseline model. Both the logistic regression and the random forest classifier algorithm scored below the ZeroR classifier baseline model, which predicts the majority class. In other words, a model which ignores all predictors had higher accuracy than the models that were used.

Table 12. Performance of models in the uniform condition for the word distribution test

Model	**28 word-object pairs**		**40 word-object pairs**	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.63	-	.75
Logistic Regression	.53	.53	.48	.46
Random Forest Classifier	.54	.56	.65	.69

Note. This table shows the accuracy scores for the ZeroR baseline model, Logistic Regression model and Random Forest Classifier model for both the 28 and 40 word-object pair word condition. The cross-validation score is the over 10-folds averaged accuracy score on the training set of the data while the test set accuracy is the retrieved accuracy of the models on the test set. The format of this table is used throughout the result section.

As explained in the methods section, the true positive rate as a function of the false positive rate with varying thresholds is plotted to obtain the ROC curve. Figure 9 shows the ROC curve for the logistic regression and random forest algorithm, and for the random guessing baseline model. The left figure shows the results of the machine learning experiments for the experiment in which 28 unique word-object pairs had to be learned, while the right figure contains the information from the experiment with 40 unique word-object pairs.

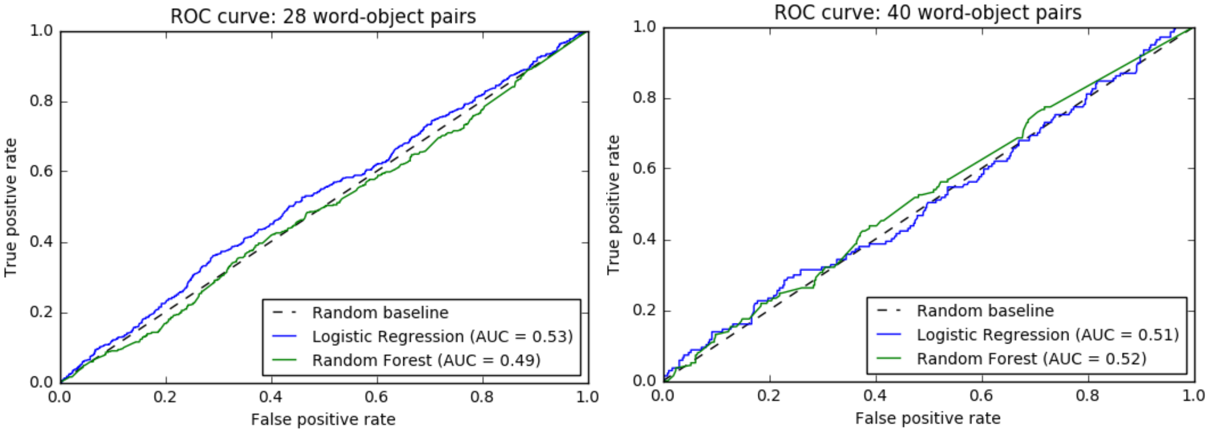


Figure 9. Receiver operating characteristic curve for the uniform word distribution test in the experiment with 28 word-object pairs(left) and in the experiment with 40 word-object pairs (right).

The graphs above demonstrate that the AUC is around .50 for both the conditions, and these scores indicate a result similar to random guessing. Since the logistic regression and random forest model score below the ZeroR classifier and random guessing models, *word distribution* is a poor predictor of *accuracy at test*.

4.1.2. First presentation

In this subsection, the following question is answered: Does the moment that a word-object pair is introduced have an impact on correctly predicted words in the test phase? The first part of this experiment addresses the uniform condition while the second part deals with the Zipfian condition.

Uniform condition

The table below (Table 13) shows the accuracy per model including the baseline model. Both the logistic regression and the random forest classifier algorithm score below the ZeroR classifier baseline model, which predicts the majority class. In other words, a model which ignores all predictors scores a higher accuracy than the models that were used.

Table 13. Performance of models in the uniform condition for the first presentation test.

Model	28 word-object pairs		40 word-object pairs	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.63	-	.75
Logistic Regression	.48	.47	.60	.58
Random Forest Classifier	.50	.54	.74	.74

The results in the previous table show that the random forest classifier algorithm achieved the highest accuracy of the two selected models. The random forest classifier's accuracy score is .01 below the baseline model. The difference between the two models is considerable because the random forest classifier classifies almost all examples as incorrectly guessed words, which indicates that first presentation is an insufficient predictor; Figure 10 (ROC curve) supports that statement. In both the experiments, the selected models perform poorly in the uniform condition, and the AUC scores are around .50 which indicates results similar to random guessing.

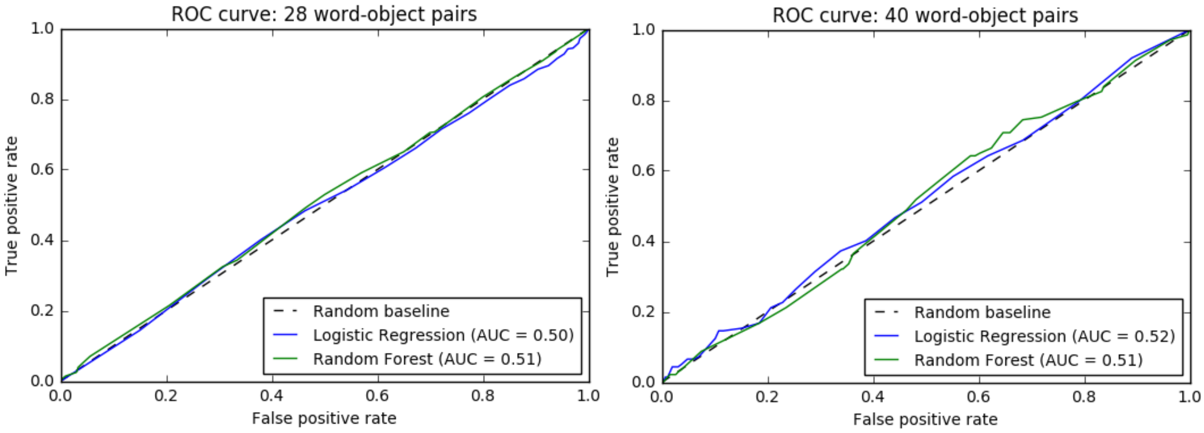


Figure 10. Receiver operating characteristic curve for the first presentation test in the experiment with 28 word-object pairs (left) and in the experiment with 40 word-object pairs (right).

Zipfian condition

The following table (Table 14) presents the results of four models, including a logistic regression and random forest classifier model for the features of *first presentation* and *frequency*, and a logistic regression and random forest classifier model with only *frequency* as a predictor.

Table 14. Performances of models in the Zipfian condition for the first presentation test.

Model	28 word-object pairs		40 word-object pairs	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.56	-	.73
LR frequency only	.67	.66	.77	.76
RF frequency only	.66	.66	.77	.75
Logistic Regression	.65	.66	.74	.75
Random Forest Classifier	.63	.62	.72	.72

Note. This table is extended with a Logistic Regression (LR) and Random Forest (RF) model with frequency as the only predictor. Section 3.3 (exploratory data analysis) described the reason for adding frequency only models.

The previous table shows that the two models with *frequency* as the only predictor, reach the same or higher accuracies as a model with *first presentation* included in the models. The feature of *first presentation* does not contribute to a higher accuracy in the tested models. The ROC curves in Figure 11 support these findings; in both conditions (28 and 40 pairs), the models without *first presentation* scored better than or similar to the models with *first presentation* added. To conclude, the accuracy metric and ROC curves show that *first presentation* does not add any predictive power to predicting correctly guessed words in the Zipfian condition.

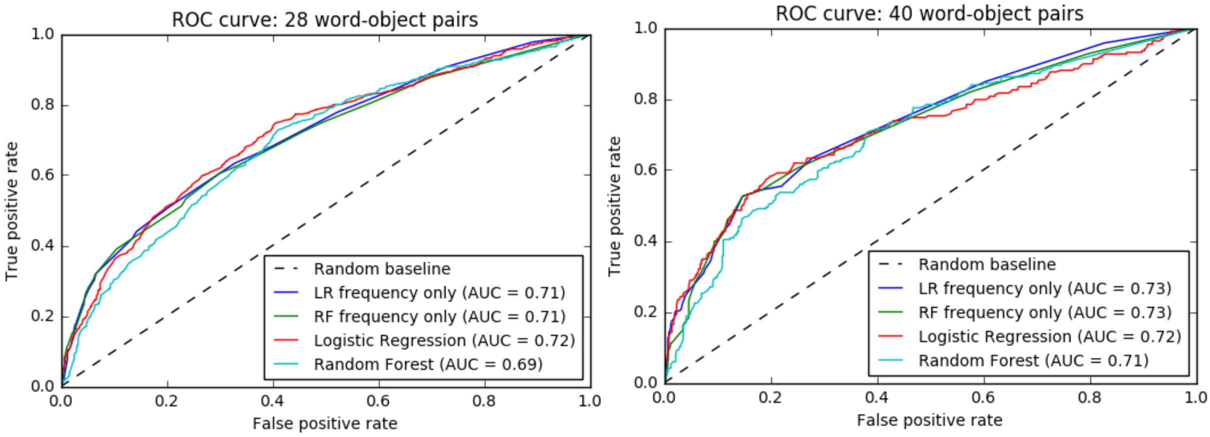


Figure 11. Receiver operating characteristic curve in the Zipfian condition for the first presentation test in the experiment with 28 word-object pairs(left) and in the experiment with 40 word-object pairs (right).

4.1.3. Final presentation

The current experiment determines whether the correct learning of a word-object pair can be predicted by the moment that a word-object pair is presented for the last time. Both the logistic regression model and the random forest classifier scored lower than the ZeroR model in the uniform condition (Table 15). Thus, the majority class predictor outperforms the models that include final presentation as a predictor.

Table 15. Performance of models in the uniform condition for the final presentation test

Model	**28 word-object pairs**		**40 word-object pairs**	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.63	-	.75
Logistic Regression	.48	.48	.49	.47
Random Forest Classifier	.62	.61	.71	.70

The random forest classifier scored lower than the ZeroR model, but was superior to the logistic regression model. This result can be explained by the fact that the random forest classifier model classified nearly all examples as incorrectly guessed words, which also explains why the model almost reached the baseline model. Furthermore, since the ROC curve shows (Figure 12) no effect for both experiments (28 vs 40 word-object pairs), final presentation is a poor predictor of correctly guessed words in the uniform condition.

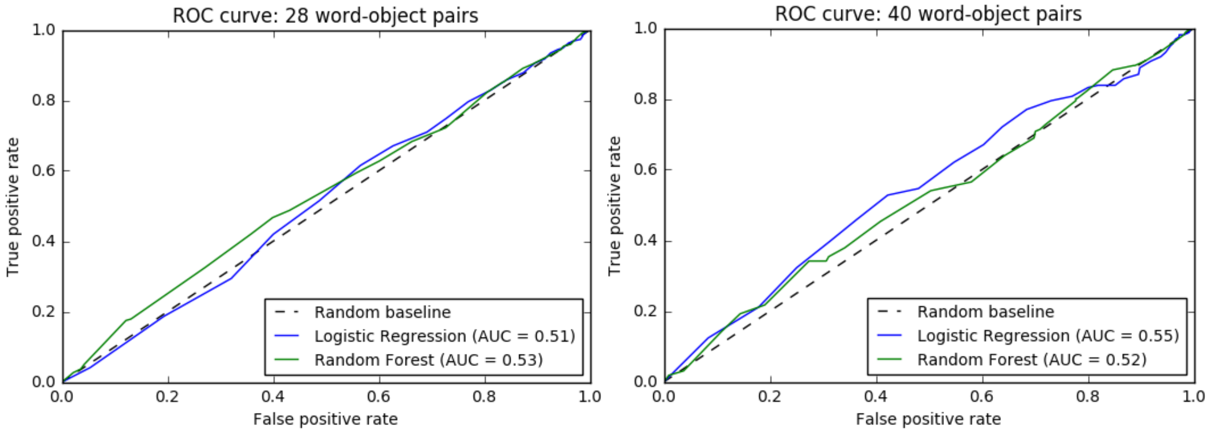


Figure 12. Receiver operating characteristic curve for the uniform final presentation test in the experiment with 28 word-object pairs(left) and in the experiment with 40 word-object pairs (right).

Zipfian condition

In the Zipfian condition, *frequency* is added to capture a baseline and compare whether the *final presentation* feature improves the baseline model with *frequency* only. Table 16. presents the results of the four models. The first two models describe the results of the logistic regression and random forest classifier model for the features of *final presentation* and *frequency*. The next two models describe the results of the logistic regression and random forest classifier model with only *frequency* as a predictor.

Table 16. Performance of models in the Zipfian condition for the final presentation test.

Model	**28 word-object pairs**		**40 word-object pairs**	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.56	-	.73
LR frequency only	.67	.66	.77	.76
RF frequency only	.66	.66	.77	.75
Logistic Regression	.66	.65	.70	.69
Random Forest Classifier	.66	.66	.74	.73

Table 16 above shows that the models with *first presentation* and *frequency* as predictors score lower or the same on accuracy as the models with *frequency* as the only predictor. Adding *first presentation* to the model does not result in a better predictive model. The ROC curves in Figure 13 support these findings; the models without *final presentation* score better or similar to the models with *final presentation* added for both the conditions (28 and 40 pairs). To summarize, the accuracy metric and ROC curves show that *final presentation* does not add predictive value to the model. Thus, it is a poor predictor of correctly guessed words in the Zipfian condition.

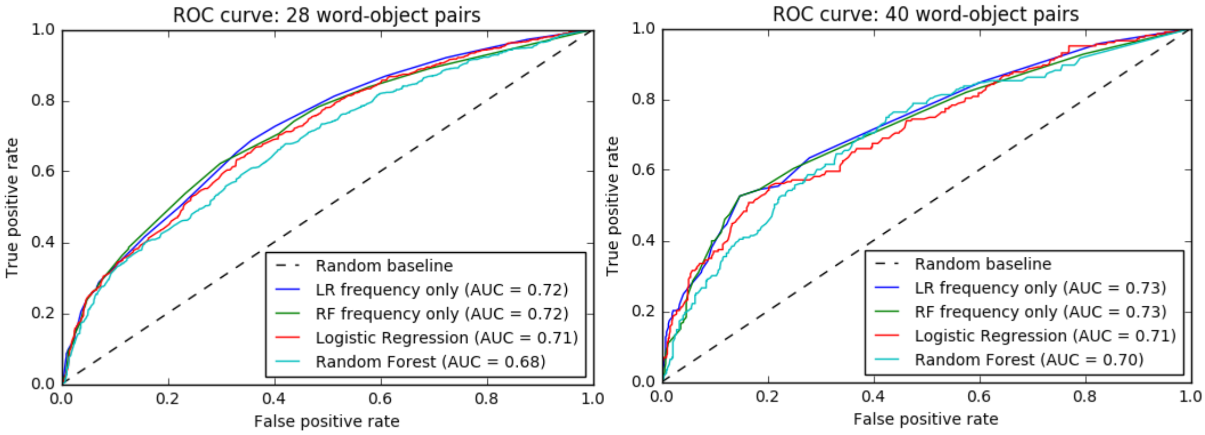


Figure 13. Receiver operating characteristic curve in the Zipfian condition for the final presentation test in the experiment with 28 word-object pairs (left) and in the experiment with 40 word-object pairs (right).

4.1.4. Reaction time

The reaction time experiment determined whether the time that subjects need to make a guess in the training phase can be used to predict whether they correctly give the right word-object combination in the test phase. Table 17 presents the outcome of the two selected models and the ZeroR baseline model. These models show no sign of overfitting, as the accuracy results of the test sets are higher than the average cross validated accuracy. In the 40-pair condition, the accuracy of the logistic regression model scores far below the ZeroR baseline model while the logistic regression algorithm scores just below the baseline model. Furthermore, the accuracy outcome of the random forest classifier model in the 28 pairs condition shows that *reaction time* is a better predictor of *correctly guessed words* than the ZeroR classifier.

Table 17. Performance of models in the uniform condition for the reaction time (RT) test.

Model	28 word-object pairs		40 word-object pairs	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.63	-	.75
Logistic Regression	.54	.54	.63	.62
Random Forest Classifier	.64	.67	.72	.74

The ROC curve shows (Figure 14) that reaction time predicts correctly guessed word-object pairs better than chance performance. In the experiment with 28 word-object pairs, the random forest model (AUC = .69) outperformed the logistic regression model (AUC = .65), and both models performed similarly in the experiment with 40 word-object pairs. The results indicate that *reaction time* is a reasonable predictor of *correctly guessed words.* The coefficients of the logistic regression show that in both the 28 and 40 pairs condition, the longer participants take to guess the first four times a word-object pair is repeated, the higher the odds are that they will correctly guess that word-object pair. After the fourth repetition of the same word-object pair, the word-object pair is less likely to be learned when reaction time increases.

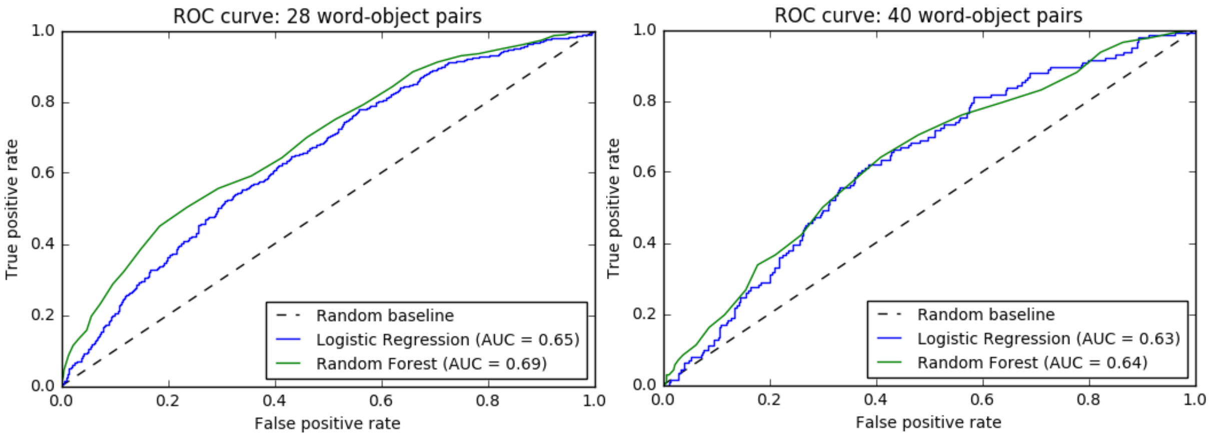


Figure 14. *Receiver operating characteristic curve in the uniform condition for the reaction time test in the experiment with 28 word-object pairs(left) and in the experiment with 40 word-object pairs (right).*

Zipfian condition

Table 18 below shows the result of the accuracy metric across the different models in the Zipfian condition. In both experiments and models, the accuracy scores were above the majority predictor level. The logistic regression and random forest model scores were similar in both experiments.

Table 18. Performance of models in the Zipfian condition for the reaction time (RT) test.

Model	**28 word-object pairs**		**40 word-object pairs**	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.56	-	.73
Logistic Regression	.61	.60	.74	.75
Random Forest Classifier	.62	.59	.76	.75

Figure 15 (ROC curve) shows decent results for both the 28-pair and 40-pair conditions, and demonstrates that *reaction time* is a predictor of *correctly guessed words* as the feature scores above chance levels. The best scoring model according to the AUC metric is the logistic regression in both the 28-pair and 40-pair conditions. In the 28 pairs condition, the area under the curve is .68, while the AUC is .70 in the 40-pair condition. In conclusion, the accuracy of both models and the ROC curves below show *that reaction time* is a predictor of *correctly guessed words*. These findings show that reaction time is a predictor of correctly guessed words in both conditions. Moreover, the coefficients of the logistic regression model were negative, indicating that longer reaction times decrease the chance of guessing the word-object pair correctly, as in the 28-pair and 40-pair condition.

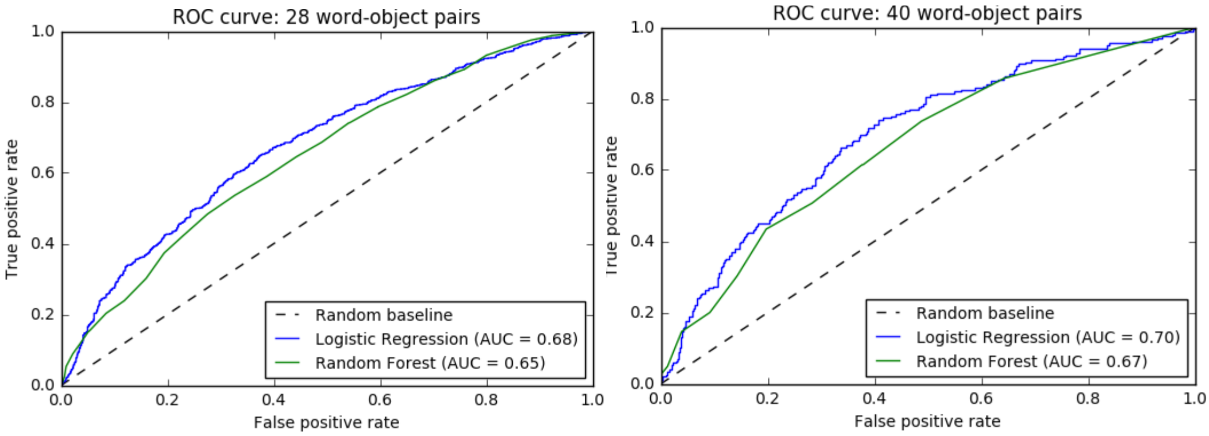


Figure 15. Receiver operating characteristic curve in the Zipfian condition for the reaction time test in the experiment with 28 word-object pairs(left) and in the experiment with 40 word-object pairs (right).

4.1.5. Context effects

This test answers determined whether one can predict if a word will be learned correctly based on how often it is selected, when there is no referential uncertainty because all previous word-objects on a screen have been selected correctly.

Uniform condition

In the uniform condition, the results show that the accuracy metric for the best scoring model in both the 28-pair and 40-pair conditions is higher than the ZeroR majority predictor model (Table 19). In the 40 pairs condition, the random forest classifier model outperformed the logistic regression model. In the condition with 28 word-objects, both the models scored an accuracy of .69 on the test set, which is .06 higher than the baseline model. The accuracy differences between the training and test sets are low, indicating that the models do not tend to overfit the data.

Table 19. Performance of models in the uniform condition for the context effects test.

Model	**28 word-object pairs**		**40 word-object pairs**	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.63	-	.75
Logistic Regression	.69	.69	.74	.74
Random Forest Classifier	.70	.69	.78	.77

The ROC curves in Figure 16 show that in both the 28 and 40 word-object pair conditions, the features can predict *correctly guessed words* better than the baseline. This context effect is strong compared to the previous tests. Both the selected machine learning algorithms score equivalent to each other for true positive rate versus false positive rates. The results also show a higher area under the curve in the condition with 28 word-object pairs compared to the 40-pair condition. The outcome coefficients of the logistic regression show that for each additional screen where a word-object pair is shown without uncertainty, the odds of guessing that word-object pair correctly increase by 1.56 times in the 28-pair condition and by 1.52 times in the 40-pair condition.

In summary, the exploratory data analysis showed that word-objects are more likely to be learned correctly when there is a high number of screens with less referential uncertainty because all word-objects in the same screen are guessed correctly. The logistic regression coefficients support those results. The examined context effect is a predictor of correctly guessed words in the uniform design. Moreover, the selected models appear robust to new data.

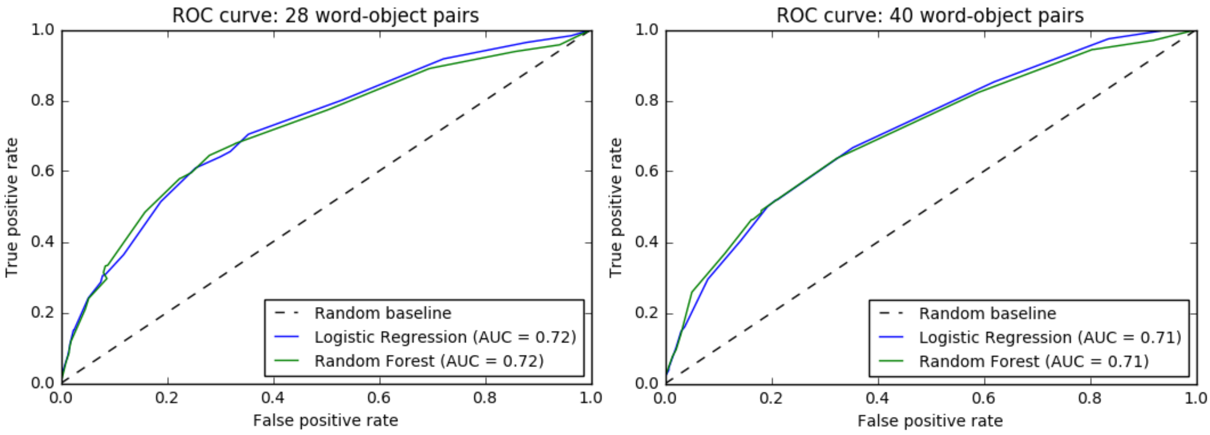


Figure 16. Receiver operating characteristic curve in the uniform condition for the context effect test in the experiment with 28 word-object pairs(left) and in the experiment with 40 word-object pairs (right).

Zipfian condition

In the Zipfian condition, *frequency* is added to capture a baseline and to compare whether the features of *last correct pair* and *number of screens without uncertainty* improve a model with *frequency* as the only predictor. Table 20. shows that adding the features that are defined for this test provide higher accuracy results than the frequency-only models. In both the 28-pair and 40-pair conditions, the logistic regression model outperformed the random forest classifier. The results also show no sign of overfitting as the accuracy score of the ten-fold cross validation results are equal to or lower than the accuracy results of the final test set. The effect of the model is the highest in the 28 word-object pair condition. The best performing setting of the logistic regression model scored .11 higher than the ZeroR random baseline model and .08 higher than the model with frequency only.

Table 20. Performance of models in the Zipfian condition for the context effects test.

Model	**28 word-object pairs**		**40 word-object pairs**	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.56	-	.73
LR frequency only	.67	.66	.77	.76
RF frequency only	.66	.66	.77	.75
Logistic Regression	.74	.74	.79	.79
Random Forest Classifier	.71	.73	.78	.78

Figure 17 clearly shows that the AUROC is higher for the models with context-effect features added. In the case of 28 word-objects, both the logistic regression and the random forest model scored higher than the models with frequency only. Moreover, the results of the ROC curve in the experiment with 40 word-object pairs show that the random forest model performs worse than the logistic regression model, and scores similar to models with *frequency* as the only predictor. The results of the accuracy metric and the ROC curves show that the selected context effect is a decent predictor of *correctly guessed words* in the Zipfian condition. Moreover, the coefficients of the logistic regression model show that for each screen where a pair is presented without uncertainty, the odds of correctly guessing that pair increase by 1.91 in the 28-pair condition and by 1.86 in the 40-pair condition. Thus, adding screens without uncertainty is associated with increased odds of correctly guessing the specific word pair. Finally, the models were robust towards new data.

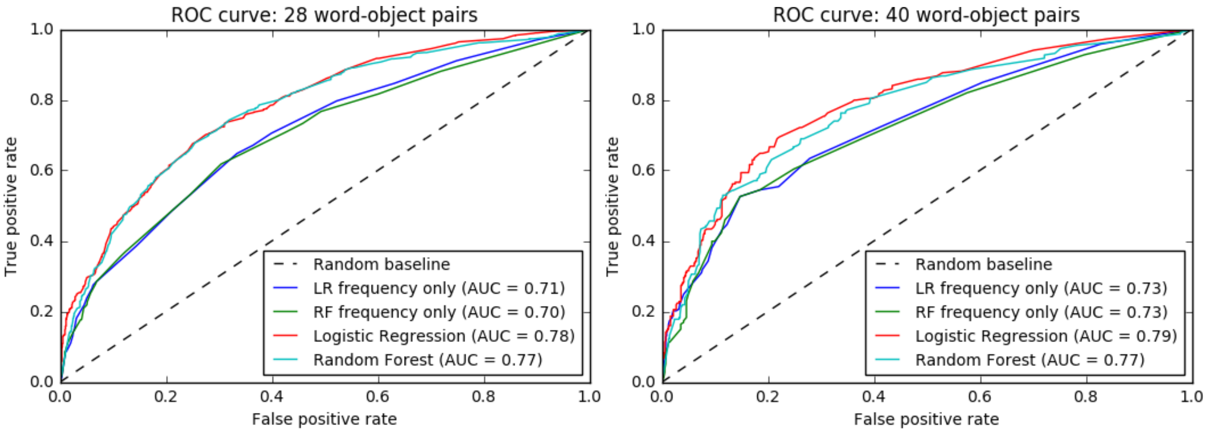


Figure 17. Receiver operating characteristic curve in the Zipfian condition for the context effect test in the experiment with 28 word-object pairs(left) and in the experiment with 40 word-object pairs (right).

4.2. Part II – Complete models

The first part of the results section showed individual features could predict the correctly guessed word-object pairs and individual features that were not capable of predicting the outcome. In this part of the results section, two different models that were built are described and evaluated: one with all individual non-predictive features and one with all individual predictive features.

Individual non-predictive features

- *Word distribution (Zipfian only)*
- *First presentation*
- *Final presentation*

Individual predictive features

- *Reaction time*
- *Context effect*

4.2.1. Model of individual non-predictive features

The individual features that were non-predictive individually were combined to create one model and investigate whether combining non-predictive features can lead to a predictive model. Consequently, the features of the tests for *word distribution*, *first presentation* and *final presentation* were used for this model. Word distribution was not part of the feature testing in the Zipfian experiment; therefore, only *first presentation* and *final presentation* are included in this model in the Zipfian condition.

Uniform condition

The results of the tests in the uniform condition are presented in Table 21, including a recap of the results from the individual tests and the current model. It is evident that combining the features in one model does not show predictive power in terms of the accuracy metric. The accuracies in both the logistic regression and random forest model are still below the majority predictor (ZeroR).

Table 21. Performance of models in the uniform condition for combining the individual non-predictors.

Model	**28 word-object pairs**		**40 word-object pairs**	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.63	-	.75
LR word distribution	.53	.53	.48	.46
RF word distribution	.54	.56	.65	.69
LR first presentation	.48	.47	.60	.58
RF first presentation	.50	.54	.74	.74
LR final presentation	.48	.48	.49	.47
RF final presentation	.62	.61	.71	.70
LR combined model	**.51**	**.54**	**.50**	**.50**
RF combined model	**.56**	**.57**	**.70**	**.68**

Note. This table shows the accuracy scores of the individual non-predictors found in the previous results section and presents the accuracy results of the combined model of the individual features of word distribution, first presentation and final presentation. A logistic regression (LR) and random forest (RF) algorithm were used, and the scores of the complete combined feature models are in bold text.

The ROC curves in Figure 18 show poorly performing models in both conditions. The area under the curves ranges from .49 to .55 for both models and conditions, which indicates the absence of valid predictors. The model does not perform better than the individual tests (section 4.1.1., 4.1.2. & 4.1.3.) Therefore, based on the accuracy scores and the ROC curves, combining the individual non-predictive features does not result in a model that can predict correctly guessed words.

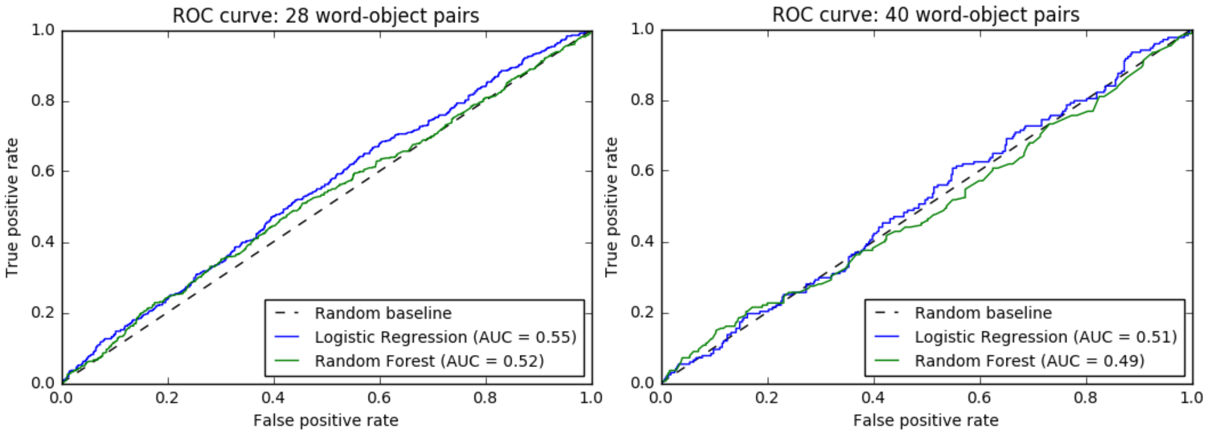


Figure 18. Receiver operating characteristic curve in the Uniform condition in the experiment with 28 word-object pairs (left) and in the experiment with 40 word-object pairs (right). The figure shows the results of the combined test from the tests that were not predictive individually: word distribution, first presentation, and final presentation.

Zipfian condition

Table 22 shows the 10-fold cross validated and test set scores of the accuracy metric for the different models. To compare the current results, the table also includes the results of the individual tests found in the previous part (section 4.1.2. & 4.1.3.). The new models are in bold text and include *first presentation, last presentation,* and *frequency*, and the models still score lower than the *frequency*-only models for both conditions. Furthermore, the accuracy metric shows that even when combined, the features cannot predict correctly guessed words.

Table 22. Performance of models in the Zipfian condition for the individual non-predictors

Model	28 word-object pairs		40 word-object pairs	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.56	-	.73
LR frequency only	.67	.66	.77	.76
RF frequency only	.66	.66	.77	.75
LR freq. + first presentation	.65	.66	.74	.75
RF freq. + first presentation	.63	.62	.72	.72
LR freq. + final presentation	.66	.65	.70	.69
RF freq. + final presentation	.66	.66	.74	.73
LR combined model	**.66**	**.64**	**.68**	**.69**
RF combined model	**.62**	**.61**	**.72**	**.72**

Note. This table shows the accuracy scores of the individual non-predictors from the previous results section and presents the accuracy results for the combined model of the individual features of first presentation and final presentation. Frequency was added as a mediator again, and a logistic regression (LR) and a random forest (RF) algorithm were used. The scores of the complete combined feature models are in bold text.

In addition to the accuracy metric, the ROC curve (Figure 19) also shows no improvement in the models compared to a *frequency*-only model. Correspondingly, the area under the curve is lower in both models and experiments. Since the results show no improvements, it can be concluded that *first presentation* and *final presentation* combined are poor predictors of correctly guessed words.

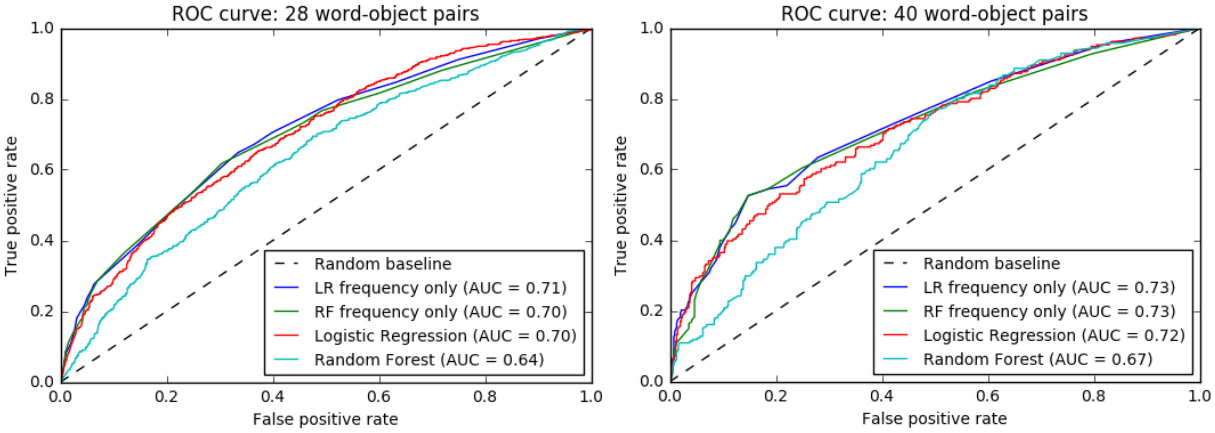


Figure 19. Receiver operating characteristic curve in the Zipfian condition in the experiment with 28 word-object pairs (left) and in the experiment with 40 word-object pairs (right). The figure shows the results of the combined test from the tests that were not predictive individually: first presentation and final presentation. The models with only frequency as a predictor are also included in the graphs.

4.2.2. Model of individual predictive features

In this subsection, the individual predictive features found in the first part of the results section are combined to build a complete model. *Reaction time* and *context effects* were significant predictors of correctly guessed words in both the uniform and Zipfian condition (section 4.1.4. & 4.1.5.). In this section, these predictors are combined into one final model.

Uniform condition

Table 23 below shows the results of the previous individual *reaction time* and *context effects* tests. In addition, the results in bold text show the accuracy results of those tests combined into one model. The accuracy metric indicates that the combined model performs better than a model with *reaction time* as the only predictor in both conditions. Compared to the *context-effect*-only model, the combined model only scored better in the 28-pair condition. In the condition with 40 pairs, the models with both features added performed worse than the models with *context effects* only. Both the logistic regression and random forest combined models are robust towards new data, as the accuracy on the test sets in all cases are higher than the average cross-validated training scores.

Table 23. Performance of models in the uniform condition for combining the individual predictors

Model	**28 word-object pairs**		**40 word-object pairs**	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.63	-	.75
LR reaction time	.54	.54	.63	.62
RF reaction time	.64	.67	.72	.74
LR context effect	.69	.69	.74	.74
RF context effect	.70	.69	.78	.77
LR combined model	**.69**	**.70**	**.62**	**.69**
RF combined model	**.71**	**.73**	**.75**	**.76**

In the previous individual test, the AUC for the *context effect* (section 4.1.5.) was .72 in the 28 word-object pair condition and .72 in the 40-pair condition. Moreover, the individual *reaction time* (section 4.1.5) test showed an AUC of .69 in the first condition and .64 in the condition with more unique words. The results in Figure 20 show that the current model with both features scores better in the 28 word-object pair condition. Conversely, the models perform worse than the individual tests for the condition with 40 pairs. The results of the ROC curves support the findings in the accuracy table above. It can be concluded that the combined model only scores better in the 28 word-object pair condition.

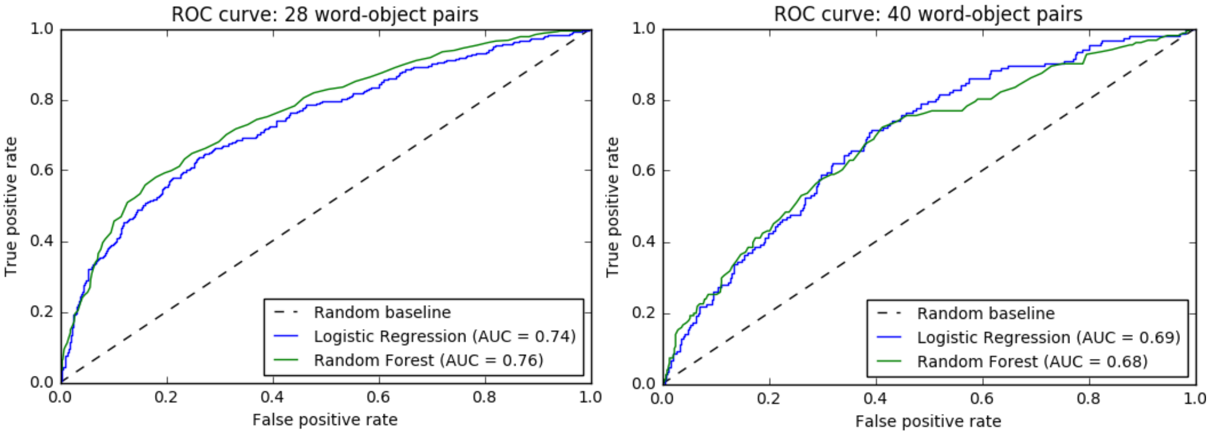


Figure 20. Receiver operating characteristic curve in the Uniform condition in the experiment with 28 word-object pairs (left) and in the experiment with 40 word-object pairs (right). The figure shows the results of the combined test from the tests that were predictive individually: reaction time and context effects

To determine which feature contributes most to the final model in the uniform condition, a feature importance analysis was conducted and is presented in Figure 21. The context variable *number of screens without uncertainty* (CE 2) is the most important variable in the 28-pair condition. Whereas this is not the case in the 40-pairs condition, were the variable is one of the least important features. The feature importance between reaction times shows little variation, they score similarly in both distributions.

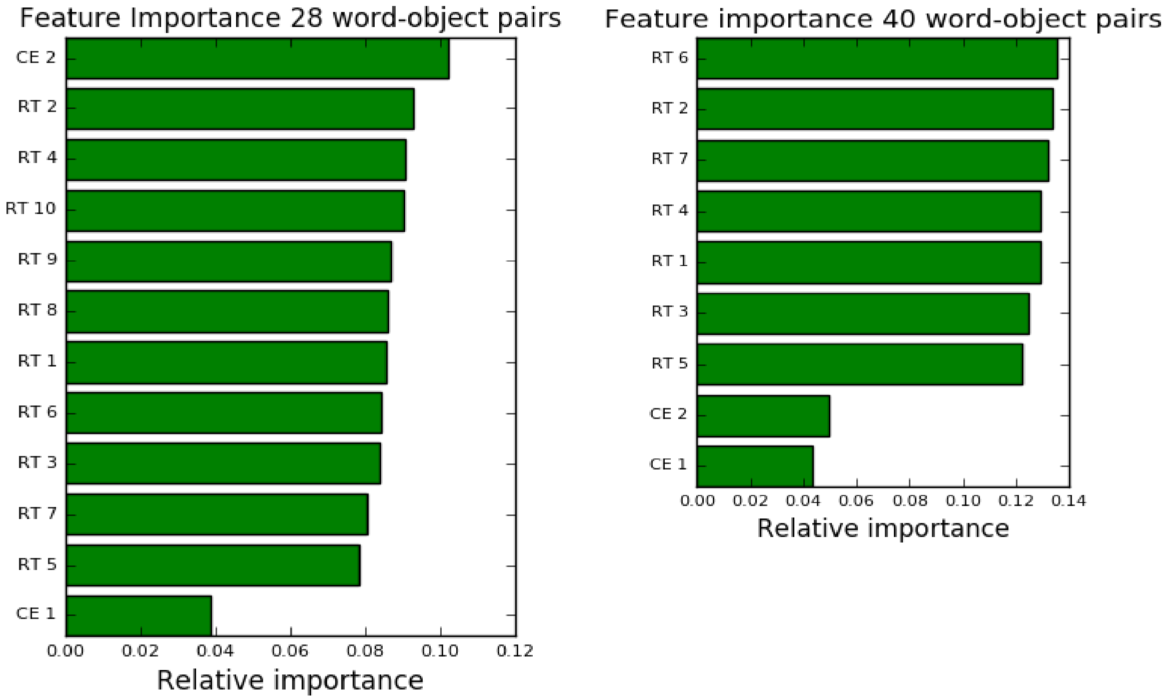


Figure 21. Feature importance in the uniform condition for the 28-pair condition (left) and 40-pair condition (right). The figure includes reaction time 1 (RT 1) to reaction time 10 (RT 10) in the first condition and reaction time 1 (RT 1) to reaction time 7 (RT 7) in the second condition. Both conditions contain context effect 1 (CE 1) and context effect 2 (CE 2), which represent the two variables used in the individual context effect test: last correct pair and number of screens without uncertainty.

Zipfian condition

Table 24 shows the accuracy scores of the previous individual feature tests and the current combined model. First, the accuracy scores of the *frequency*-only model are outlined in the table below, including the *reaction time* test (section 4.1.4.), and the *context effect* test where *frequency* was added as a mediator (section 4.1.5.). The two rows in bold text show the accuracy score of the models that combine the previously mentioned features. In both the 28-pair and 40-pair conditions, the logistic regression and random forest model did not achieve the accuracy score of the *context effect* model. The combined models do however perform better than the *reaction time* models.

Table 24. Performance of models in the Zipfian condition for combining the individual predictors

Model	**28 word-object pairs**		**40 word-object pairs**	
	10-fold cross validation average accuracy	Test set accuracy	10-fold cross validation average accuracy	Test set accuracy
ZeroR (accuracy baseline)	-	.56	-	.73
LR frequency only	.67	.66	.77	.76
RF frequency only	.66	.66	.77	.75
LR reaction time	.61	.60	.74	.75
RF reaction time	.62	.59	.76	.75
LR freq. + context effect	.74	.74	.79	.79
RF freq. + context effect	.71	.73	.78	.78
LR combined model	**.71**	**.73**	**.77**	**.77**
RF combined model	**.69**	**.71**	**.76**	**.76**

In the individual feature testing phase, the AUC score of the context effect in the Zipfian condition ranged between .77 and .79 for both the 28-word and 40-word conditions (section 4.1.5.). In the same phase, the AUC score of the reaction time ranged between .65 and .70 for both conditions (section 4.1.4.). Figure 22 shows that the current model with *reaction time*, *context effect,* and *frequency* as mediators scored better than *reaction time* alone, but there is no improvement compared to the model with *context effect* only. It can be concluded that the combined models do not achieve additional predictive power compared to the individual models.

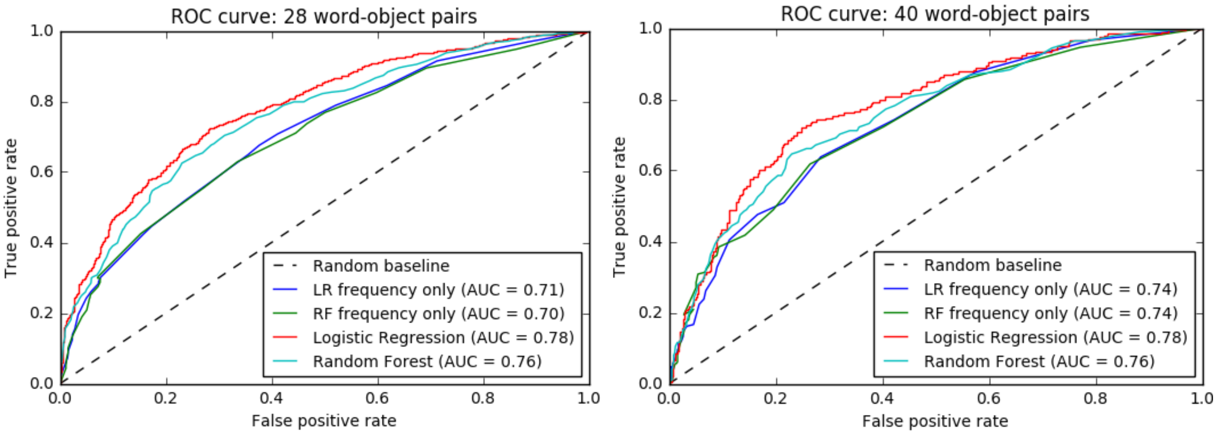


Figure 22. Receiver operating characteristic curve in the Zipfian condition in the experiment with 28 word-object pairs (left) and in the experiment with 40 word-object pairs (right). The figure shows the results of the combined test from the tests that were predictive individually: reaction time and context effects. Frequency is also added in the combined model as a mediating variable. The graph also shows the models with frequency only.

Figure 23 shows the importance per feature for both the conditions. In both conditions, the three features that contain the descriptive statistics regarding reaction times are the most important features to the random forest model. Moreover, frequency seems to be a more important predictor than *number of screens without uncertainty* (CE 2) in the condition with 40 pairs, but this is not the case in the condition with 28 pairs.

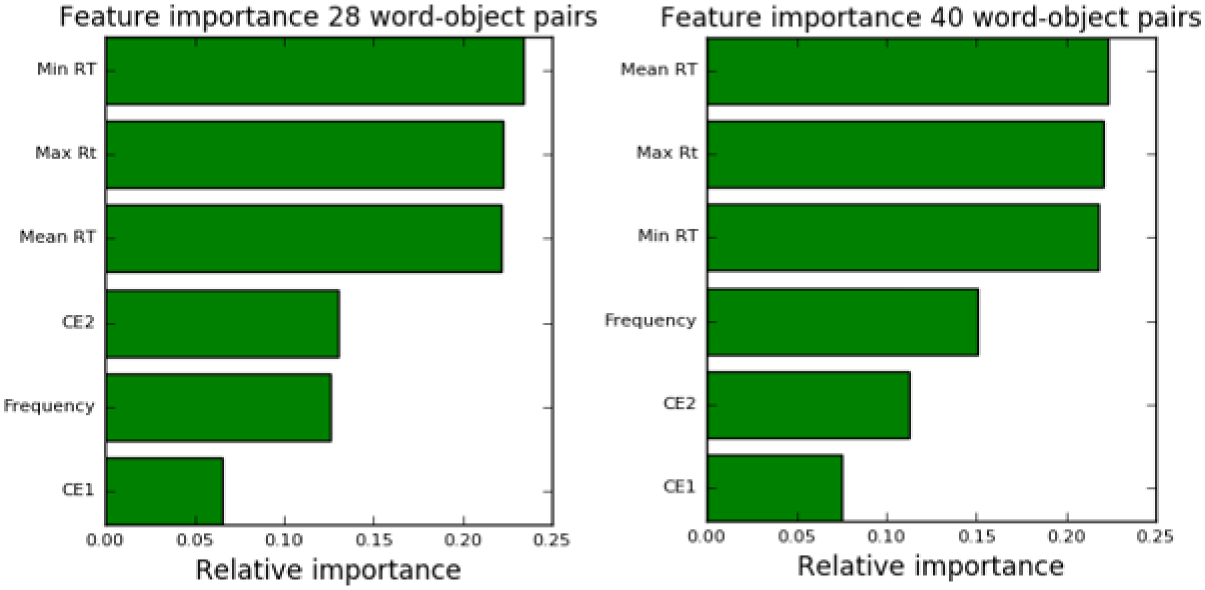


Figure 23. Feature importance in the Zipfian condition for the 28-pair condition (left) and 40-pair condition (right). This figure includes the statistics of reaction time: minimum reaction time (Min RT), maximum reaction time (Max RT,) and average reaction time (Mean RT). It also contains context effect 1 (CE 1) and context effect 2 (CE 2), which represent the two variables used in the individual context effect test: last correct pair and number of screens without uncertainty.

4.3. Part III – Subject clusters

Until now, all analysis and predictive models were realized per word-object pair. The results in this section were achieved by building predictive models per subject for each feature that was tested in the feature testing part. As explained in the methods section (section 3), a model for each subject per feature provides an opportunity to investigate whether some subject groups benefit from a specific feature. This section shows how histograms were built to demonstrate the predictive power of the models per subject.

4.3.1. Gaussian mixture model

The Gaussian mixture model was run multiple times for different numbers of clusters. The highest silhouette scores were obtained with each component that had its own general covariance matrix. Moreover, the K-means initialization method was used to initialize the means, weights and precisions. Figure 24 below shows the silhouette plots for four models with different numbers of clusters.

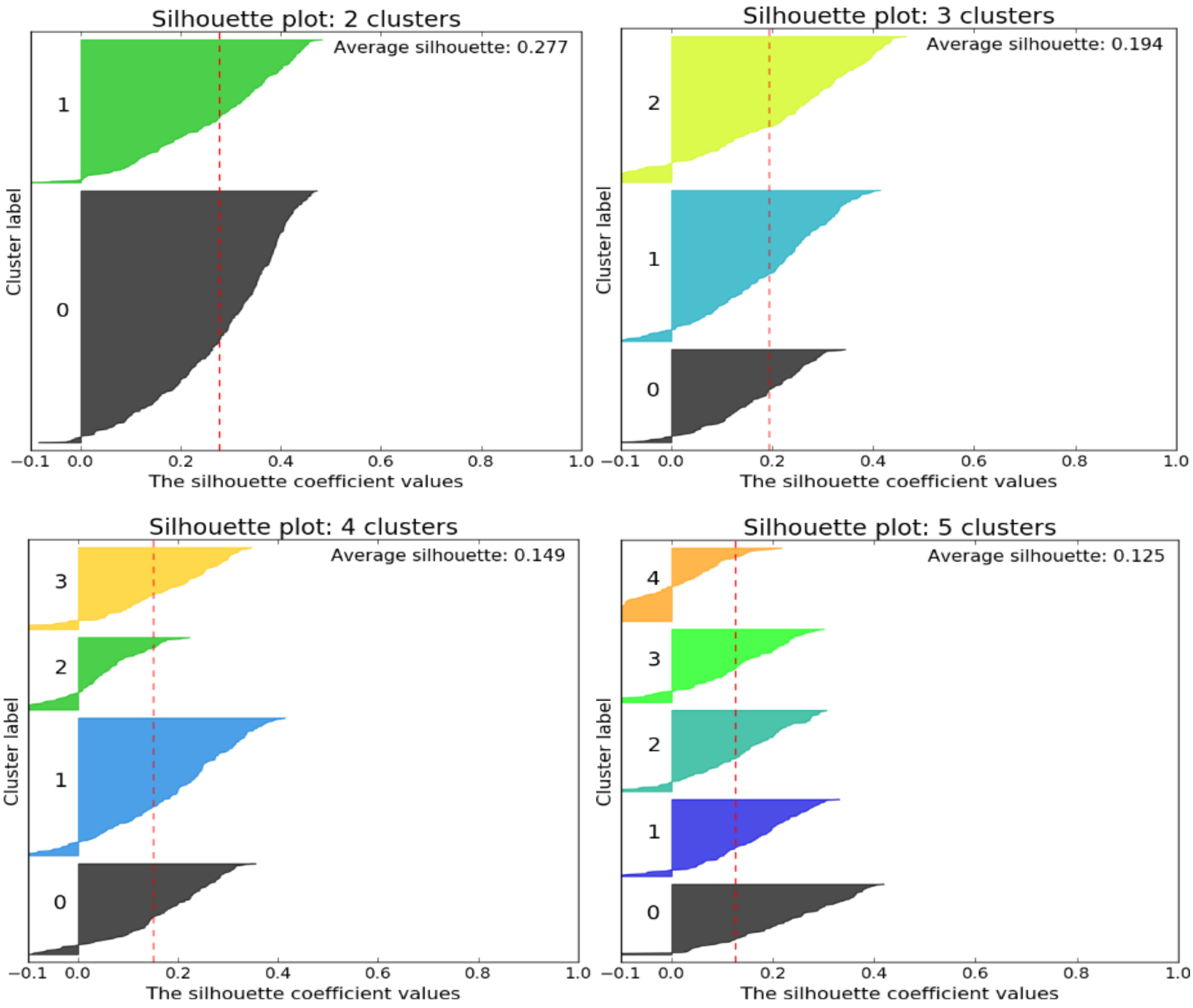


Figure 24. Silhouette coefficient values in the Gaussian mixture model for different numbers of clusters. The striped red line is the average silhouette score across all individual clusters, and each spike represents a sample in the data. The thickness of a cluster in the graph illustrates the size of a cluster.

The graph shows that the model that separated two clusters performs better than the models with more than two clusters. The average silhouette score of the model with two clusters is .277, which indicates a weak and possible artificial structure. None of the individual clusters found across the plots indicate a strong separated cluster. Moreover, the negative silhouette scores indicate that some samples may be assigned to the wrong cluster. Due to the poor performance of the Gaussian mixture model, it is not valuable to interpret and examine the distinct clusters.

4.3.2. Hierarchical clustering

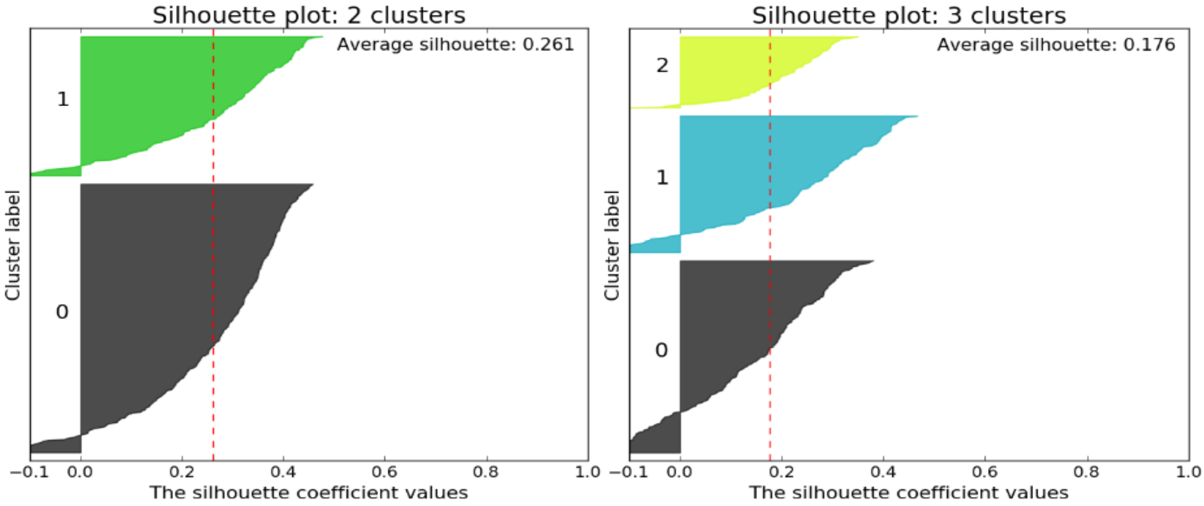
The second clustering approach was conducted with a hierarchical clustering model. Next to a fixed number of clusters, this clustering algorithm can be optimized with a various number of parameters. The settings that achieved the highest silhouette scores were explored before the silhouette plots were applied. Table 25 illustrates the combination of the parameters and the corresponding silhouette score.

Table 25. Combination of parameter settings with corresponding silhouette score

Affinity	**Linkage**		
	Ward	Complete	average
Euclidean	.261	.246	.207
Cosine	-	.239	.259
Manhattan	-	.237	.217
L1	-	.237	.216
L2	-	.237	.207

Note. Table of parameter combinations which can be used in the hierarchical clustering model. The ward linkage can only be used in combination with the Euclidean distance. The highest silhouette score was selected per parameter combination after comparing a different number of clusters (2 to 5 clusters).

The highest silhouette score was obtained with a combination of the Euclidean distance and ward linkage. Since the ward parameter minimizes the variance of merged clusters, those settings are used in combination with a different number of clusters. The results of the silhouette plots for a different number of clusters are shown in Figure 25.



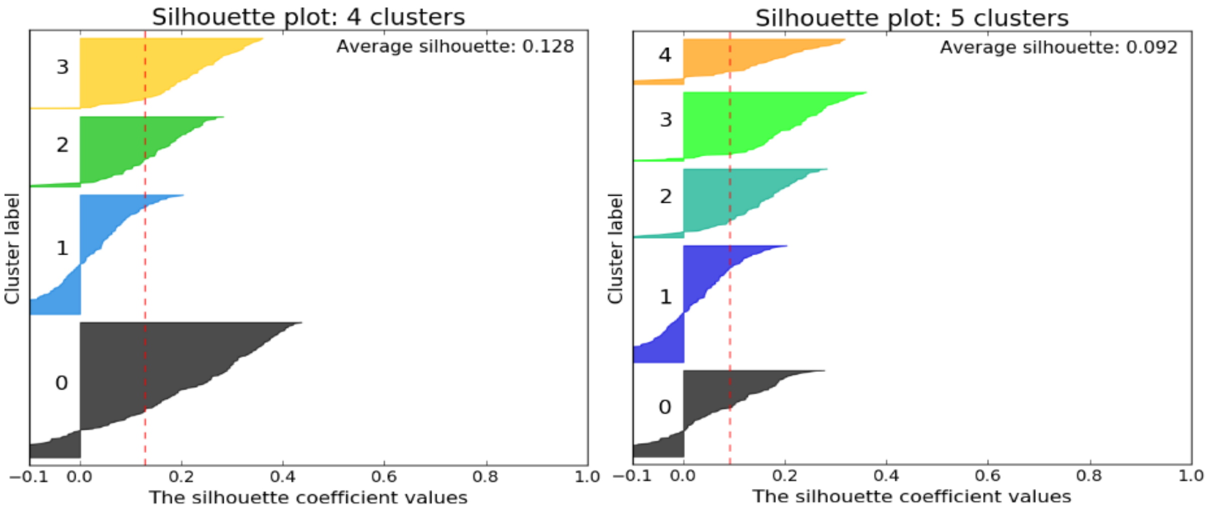


Figure 25. Silhouette coefficient values in the hierarchical clustering model for different numbers of clusters. The striped red line is the average silhouette score across all individual clusters. Each spike represents a sample in the data, and the thickness of a cluster in the graph illustrates the size of a cluster.

The silhouette plots illustrate that the highest silhouette coefficient is from the model with two clusters, and the average silhouette score across the two clusters is .261. The plot also shows that in the model with two clusters, one of the clusters is twice the size of the smaller one. None of the individual clusters found across the plots indicate a strong separated cluster. Furthermore, all clusters contain negative silhouette scores that may indicate incorrectly assigned clusters. A score of .261 indicates a weak and possible artificial structure in the data. In other words, the clusters are separated poorly, and the outcome of the clustering analysis cannot be interpreted or examined. Thus, theoretical conclusions cannot been draw in the hierarchical clustering analysis.

5. Discussion

The first goal of this thesis project was to identify features that could be used to predict whether a word-object pair was learned correctly. Based on the literature review, five interesting potential predictive features were found: word distribution, first presentation, final presentation, reaction time, and one context effect. The second goal was to investigate whether different types of learners could be found, and the five potential predictive features were used to make a clustering analysis. In this section, the findings from the results section (see Section 4) are discussed.

This study could not demonstrate that the presentation distribution of a word-object pair can be used to predict correctly guessed words. In all conditions, the models with word distribution as a predictor achieved poor results below chance. In the literature, most studies have agreed that an interleaving learning method resulted in fewer errors compared to block training. In this research project, no proof was found that the tightness of a word-object's presentations predicts whether that pair will be guessed correctly in the testing phase. A possible explanation for these results may be the lack of adequate differences between interleaving and blocking in these experiments. The moment of a word-object appearance was randomly set up in the available data for this thesis, and thus, the data was not distributed to fit exactly an interleaving or blocking learning setting. Nevertheless, some predictive power was expected as the tightness and distribution between different words fluctuated. Different thresholds were used to determine which number of the same word-object presentations repeated close to each other represents a tight cluster, but none of the thresholds made better predictions than the ZeroR predictor.

No relationship between the first or last presentation and correctly guessed words was found. These experiments did not detect any evidence that correctly guessed words could be predicted by first or last presentation. This outcome is not in line with the theory of Ebbinghaus (1913) who found that the first few and last few items are recalled best in his serial-position effect theory. There are several possible explanations for this result. For example, most of the presentations of a word-object pair occurred in the middle of the training phase. Moreover, the current study only examined the first and last moment that a word-object pair was presented, and it is possible that a word-object pair was presented in the first few screens but did not return for several screens. Perhaps the referential uncertainty was too high in the beginning of the training phase, resulting in that the serial-position effect not applying because the real word-object pair had not been learned yet. In future research, the serial-position effect could be examined by measuring at which screen in the training phase a word-object pair is learned correctly, and counting that screen as the first presentation.

For the reaction time experiment, the results showed that reaction time in the training phase can be used to predict the final correctly guessed words. Reaction time was a predictor of correctly guessed words in the Zipfian and uniform experiments for both the 28-pair and 40-pair conditions. The effect of reaction time on correctness was similar in all experiments and conditions. In the uniform condition, higher response times were seen in the first four times a word-object pair was shown, indicating a higher chance that the pair was learned correctly. After the first four presentations, higher response times indicated that a word pair was less likely to be learned. These results could suggest that taking time to combine the correct word with an object in the first four presentations results in higher odds of learning the correct pair. In contrast, if more time is needed after the four presentations, this indicates that it is more likely that the word-object pair will be learned incorrectly. As mentioned in the literature review, Rubenstein (2013) showed that a fast response time indicated a higher chance of errors. In this thesis project, fast reaction times only indicated more errors in the first four presentations of a word-object pair. After the first four presentations, fast reaction times indicated higher odds that a specific pair was learned correctly.

Regarding the context effect, the results show that the more frequently a word-object pair occurs without uncertainty, the more likely it is that this pair will be learned correctly in all conditions. These results are consistent with the literature in the related work section (see Section 2). Brent and Siskind (2001) found that learners learn words more effectively when exposed to them in an isolated setting compared to words learned in a referent uncertain setting, which is consistent with the current research. From all the individual features that were tested, this context effect was the strongest predictor of correctly guessed words.

Considering the complete models part, the findings of the model of the individual predictive features were disappointing. The individual tests showed that both reaction time and the examined context effect were predictors of correctly guessed words, but these features combined resulted in slightly more accurate predictions only in the 28-pair condition in the uniform experiment. In all other conditions and experiments, the model did not result in more accurate predictions, but in worse predictions. The reason for this is not clear, but overfitting cannot be the reason since the cross-validation outcome scores were lower than the final scores on the test set. One possibility is that the models become too complicated and there are more benefits from a simpler learning model. Another possibility is that the achieved results are the highest possible results for the dataset that was used. This seems unlikely, however, as many more potential features can be examined, but this can be the case for the features that are found in this current thesis project.

Moving to the second research question, the subject clustering results showed poorly separable clusters. The results of this clustering analysis did not support the theory of different types of learners, though this is not evidence that different types of learners do not exist in cross-situational word learning. One valid reason that the clustering models performed poorly could be the feature selection. In the individual feature testing part, only two of the five tested features were predictors of correctly guessed words. Thus, the non-predictive features may not be meaningful in the clustering analysis. One suggestion for future research is to identify more predictive features and make a new attempt to discover structure between subjects. In the clustering analyses conducted in this thesis project, the intercept and coefficient outcomes of the logistic regression algorithm were used for each feature and for each subject. Other methods to retrieve weights per subject could be implemented in future research.

One limitation of this thesis project is the number of individual features that were tested, as five individual potential features were tested. In future investigations, it might be possible to use different features as predictors, as the available data for this thesis project offers many opportunities to extract potential features. In this research project for example, only one context effect was examined due time limits. This thesis project demonstrates that environmental factors influence word learning.

This chapter begins by discussing the results of the individual feature testing part and shows that only reaction time and the examined context effect were predictors of correctly guessed words. Overall, all the results were similar regardless of the condition (28 vs 40 words) and experiment type (Zipfian vs Uniform). The second part of the result section, the complete models, is then discussed. It was unexpected that adding reaction time and the examined context effect in a complete model did not achieve better results. Moreover, adding the non-predictors of word distribution, first presentation, and final presentation to a complete model did not result in a predictive model. These results are consistent with the results of the individual feature testing part. Finally, the results of the subject cluster analysis showed too much noise in the clusters, which made further analysis impossible.

6. Conclusion

This section briefly answers the research questions formulated in the introduction section (see Section 1). The first research question was:

Research question 1 (RQ$_1$): *Which features predict cross-situational word learning?*

The first research question aimed to identify features that could predict cross-situational word learning. The outcome of the results section showed that word distribution, first presentation, and final presentation were not predictors of correctly guessed words. The models performed worse than the ZeroR baseline model, and were poor predictors in both the 28 and 40 word-object pair conditions and in the uniform and Zipfian experiments. Moreover, adding all individual non-predictive features to a single model did not lead to a predictive model.

This project determined that the reaction time of the subjects in the training phase can be used to predict whether a word-object pair is guessed correctly, as this was the case in all conditions and experiments. The last individual feature that was tested was a context effect, the number of screens without uncertainty. The models showed that word-object pairs that appear without uncertainty can be used to predict correctly guessed words, and the feature was a good predictor in all conditions and experiments. Combining all individual predictive features into a complete model only added slightly more accurate predictions in the 28-pair, uniform condition. All other conditions did not benefit from the combined model.

The second aim of this study was to investigate whether different types of learners could be identified. One suggestion was that some features may impact learner A, but not learner B. The following research question was defined:

Research question 2 (RQ$_2$): *Can evidence be found to prove the existence of different types of learners?*

The results of the clustering analysis from both the Gaussian mixture model and the hierarchical clustering model provided no evidence of different types of learners. In both clustering models, the distinct clusters were separated poorly and only a weak and artificial structure was found. Nevertheless, no evidence was found that different types of learners do not exist.

References

Azizyan, M., Singh, A., & Wasserman, L. (2015, February). Efficient sparse clustering of high-dimensional non-spherical gaussian mixtures. *Artificial Intelligence and Statistics*, 37-45.

Beckham, C. J. (2015). Classification and regression algorithms for WEKA implemented in Python.

Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, *9*(Sep), 2015-2033.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*, 59-68.

Blythe, R. A., Smith, K., & Smith, A. D. (2010). Learning Times for Large Lexicons Through Cross-Situational Learning. *Cognitive Science*, *34*(4), 620-642.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145-1159.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33-B44.

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*(5), 671-682.

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*, 161-168.

Dunn, R. S., & Dunn, K. J. (1978). *Teaching students through their individual learning styles: A practical approach*. Prentice Hall.

Ebbinghaus, H. (2013). Memory: A contribution to experimental psychology. *Annals of neurosciences*, *20*(4), 155.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, *20*(5), 578-585.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Hendrickson, A. T. & Perfors, A. (2016). Cross-situational learning in a Zipfian environment. Manuscript submitted for publication.

Horst, J., & Hout, M. (2015). The novel object and unusual name (NOUN) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, 48(4), 1393-1409.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90-95.

Kachergis, G., Shiffrin, R., & Yu, C. (2009). Frequency and contextual diversity effects in cross-situational word learning. *Proceedings of the Cognitive Science Society*, 31.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2017). A bootstrapping model of frequency and context effects in word learning. *Cognitive science*, *41*(3), 590-622.

Kirschner, P. A. (2017). Stop propagating the learning styles myth. *Computers & Education*, *106*, 166-171.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323-1329.

McKinney, W. (2015). Pandas: A Python data analysis library.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological science in the public interest*, *9*(3), 105-119.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825-2830.

Quine, W. V. (1960). Word and object. Cambridge, Mass.

Raschka, S., & Mirjalili, V. (2017). *Python machine learning*. Birmingham, England: Packt

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, *24*(3), 355-367.

Rott, S. (2007). The effect of frequency of input-enhancements on word learning and text comprehension. *Language Learning*, *57*(2), 165-199.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Rubenstein, A. (2013). Response time and decision making: An experimental study. *Judgment and Decision Making*, *8*(5), 540.

Schein, A. I., & Ungar, L. H. (2007). Active learning for logistic regression: an evaluation. *Machine Learning*, *68*(3), 235-265.

Schotter, A., & Trevino, I. (2014). *Is response time predictive of choice? An experimental study of threshold strategies* (No. SP II 2014-305). WZB Discussion Paper.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480-498.

Tarpey, T. (2007). Linear transformations and the k-means clustering algorithm: applications to clustering curves. *The American Statistician*, *61*(1), 34-40.

Technavio. (2015, September). *Global Education Apps Market - Market Study 2015-2019.* Retrieved from http://www.technavio.com/

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, *66*(1), 126-156.

Vogt, P. (2012). Exploring the robustness of cross-situational learning under Zipfian distributions. *Cognitive Science*, 36, 726–739.

Walt, S. V. D., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, *13*(2), 22-30.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18, 414–420.

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word–referent learning: Prior questions. *Psychological review*, *119*(1), 21.

Zipf, G. K. (1949). Human behaviour and the principle of least effort: An introduction to human ecology. Cambridge, MA: Addison-Wesley.

Appendices

Appendix A: Stimulus

Visual stimulus

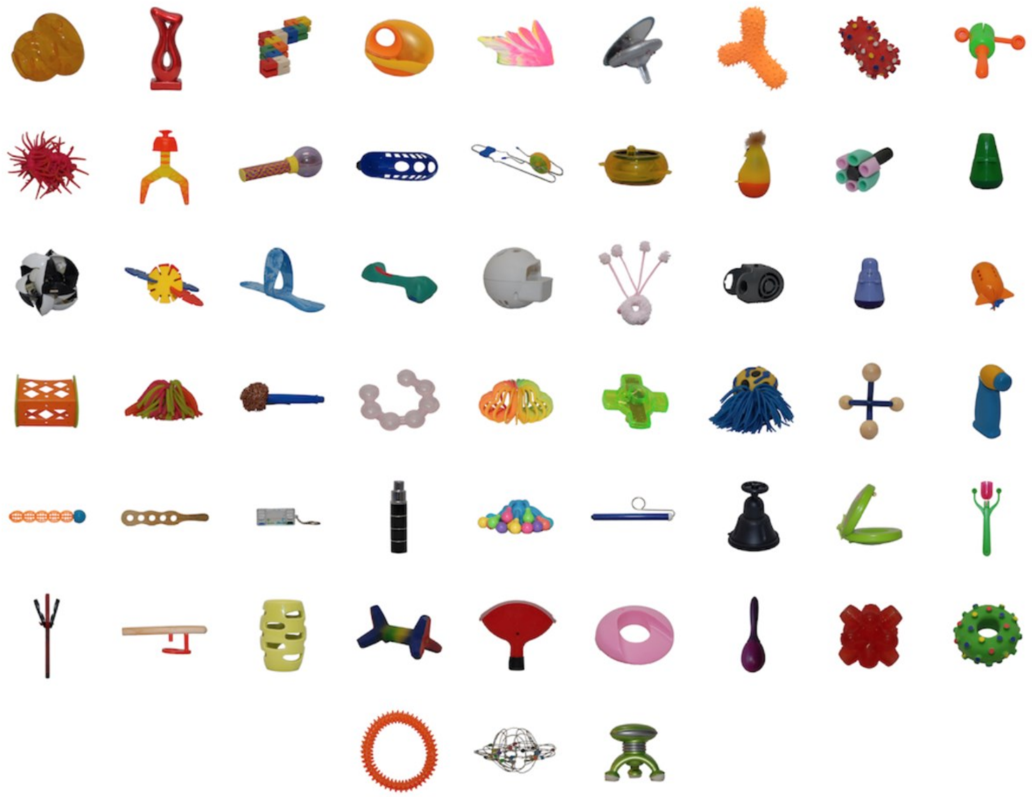


Figure 26. The stimuli which were used for this experiment were obtained from the Novel Object and Unusual Name (NOUN) database (Horst & Hout, 2015). The content of this figure is put together by Dr. A. T. Hendrickson (Personal communication, January 2, 2018).

Audi stimulus

The following pseudo words were used in the experimental data used for this thesis project. This list of pseudo words is put together by Dr. A. T. Hendrickson (Personal communication, January 2, 2018)

"boam", "chave", "glay", "loid", "naft", "heef", "plook", "queed", "rinch", "shug", "thorp", "zye", "ank", "drib", "mav", "troz", "vurl", "wope", "duppy", "enzol", "pazoo", "impet", "mandle", "oidup", "titchiv", "urbick", "vaqua", "wajong", "niftim", "grambay", "zurter", "cheldob", "arturum", "drezelist", "fobivan", "kepata", "silodox", "yunkiter", "epichuf", "prebantik".

Appendix B: Datasets

Table 26. Description per dataset

Name dataset	Conditions	Screens training	Words to learn	Words per screen
Alfa	• Zipfian • Uniform	70 screens, guess while training	28 word-object pairs	4 words
Bravo	• Zipfian • Uniform	70 screens, guess while training	28 word-object pairs	4 words
Charlie	• Zipfian • Uniform	70 screens, no guessing	28 word-object pairs	4 words
Delta	• Zipfian • Uniform	70 screens, no guessing	28 word-object pairs	4 words
Echo	• Zipfian • Uniform	70 screens, guess while training	40 word-object pairs	4 words
Foxtrot	• Zipfian • Uniform	70 screens, guess while training	40 word-object pairs	4 words
Golf	• Zipfian • Uniform	70 screens, guess while training	40 word-object pairs	4 words
Hotel	• Zipfian • Uniform	70 screens, guess while training	40 word-object pairs	4 words
India	• Zipfian • Uniform	240 screens, one object per screen	28 word-object pairs	1 word
Juliet	• Zipfian • Uniform	70 screens, guess while training	28 word-object pairs	4 words
Kilo	• Zipfian • Uniform	Various experiments with different conditions	28 word-object pairs	1 word or 4 words
Lima	• Zipfian • Uniform	Various experiments with different conditions	32 word-object pairs (4 check items)	1 word or 4 words
Mike	• Zipfian • Uniform	Various experiments with different conditions	32 word-object pairs (4 check items)	1 word or 4 words
November	• Zipfian (one syllable words)	12 screens, 9 training blocks	12 word-object pairs	3 words
Oscar	• Zipfian (one syllable words)	12 screens, 9 training blocks	12 word-object pairs	3 words

Appendix C: Features

Table 27. Description of all features in the datasets

Feature name	Description
Language	The self-reported language of the participant
Age	The self-reported age of the participant
Gender	The self-reported gender of the participant
Country	The self-reported country of the participant
Completioncode	The randomly generated code needed to complete the experiment
Experiment	The ID of the experiment code (sometimes is confusing)
Distribution	The condition of the experiment (many possible values)
Subjectid	The randomly generated ID number for each participant
Breakfrequency	The number of training screens between each break
Typeofdata	The header of this section. Indicates the type of data being recorded
Experiment	The ID of the experiment code (sometimes is confusing)
Trialwithinscreen	The number (from 0 to N) of previous words displayed during the current screen
Screen	The current screen number (0 to N). Each screen can have multiple entries in the datafile, if multiple words were presented on the screen words
Phase	The current phase of the experiment. Should be training or testing
Correctword	The word as a string currently presented
Correctobject	The ID (from 1 to N) of the current correct word
Responsetime	The number of milliseconds until the response was made
Locationselected	The ID (1-N) of the selected object's location object selected
Wordofobjectselected	The word as a string of the current selected object
Accuracy	The accuracy of the current selection
Databasekey	The ID of this entry in the database. A duplicate (even across experiments) indicates a likely data duplication error
Initialtrialinfo words	The list of all words in the experiment. The nth element of this list should convert from word ids to words
Trials	A list of lists. The outer list consists of a list of screens for the experiment. Each inner list contains a list of the ids of the words and objects for the screen. It is possible that neither list is in the correct order.
Typeofdata	The header of this section. Indicates the type of data being recorded
Experiment	The ID of the experiment code (sometimes is confusing)
SubjectID	The randomly generated ID number for each participant
Completioncode	The randomly generated code needed to complete the experiment
Experiment	The ID of the experiment code (sometimes is confusing)
Windowwidth	The width of the experiment screen in pixels
Windowheight	The height of the experiment screen in pixels
Distribution	The condition of the experiment (many possible values)

Note. Table received from Dr. A. T. Hendrickson (personal communication, September 9, 2017)

Appendix D: Parameters predictive models

This current appendix contains the optimal parameters that were used for each test in the result section. Table 28 contains the parameters of the individual tests. Next, Table 29 shows the parameters for the second part of the result section, the complete models. Finally, the parameters of the clustering algorithms are given in Table 30.

Table 28. Optimal parameters for the models in the individual feature testing part

Tests	Random Forest	Logistic Regression

Word distribution	cv = 10
	n_estimators = 20
		cv = 10
		cs = 100
		class_weight = 'balanced'
		penalty = 'l2'

First presentation &
final presentation
	cv = 10
	n_estimators = 20
		cv = 10
		cs = 100
		class_weight = 'balanced'
		Penalty = 'l2

Reaction time &
context effect
	cv = 10
	n_estimators = 50
		cv = 10
		cs = 10
		class_weight = 'balanced'

Table 29. Optimal parameters for the models of non-predictive and predictive features

Tests	Random Forest	Logistic Regression

Complete model of individual non-predictive features	cv = 10
	n_estimators = 50
		cv = 10
		cs = 100
		class_weight = 'balanced'
		penalty = 'l2'

Complete model of individual predictive features	cv = 10
	n_estimators = 40
		cv = 10
		cs = 100
		class_weight = 'balanced'
		penalty = 'l2'

Table 30. Optimal parameters for the models in the clustering analysis

Tests	Logistic Regression*	Gaussian Mixture model	Hierarchical clustering

Subject clusters	cv = 3
	n_estimators = 100
	penalty = 'l2'
		covariance_type = 'full'
		n_components = 2
		init_params = 'kmeans'
			n_clusters = 2
			linkage = 'ward'
			affinity = 'euclidean'

*Note. The logistic Regression algorithm was used to obtain the intercepts and coefficients of all individual tests which are used as input to the clustering algorithms.