university of groningen / faculty of science and engineering / artificial intelligence

Cross-situational word learning in conditions of limited time

Bachelor's Project Thesis

Mircea Negru, s2945126, m.negru@gmail.com,
Supervisors: Dr. Stephen M. Jones

Abstract: The cross-situational word learning (CSWL) process involves aggregating information across multiple ambiguous situations to acquire a new word. There are two leading theories on the specific mechanisms of CSWL. The propose-but-verify mechanism is thought of as a type of conscious learning by which only one hypothesis about the meaning of a word is stored for later testing. On the other hand, the gradual association mechanism is a type of statistical and subconscious learning through which not just one, but many possible word-to-referent pairs are learned simultaneously. We discuss how the time constraint influences the learning strategy. We investigate the interaction between the two competing strategies through an experiment in which participants try to find the meaning of novel words in ambiguous situations under conditions of limited time. A stricter deadline facilitates subconscious learning. Therefore, we expect to see more evidence for gradual association as less time is given. However, more time to make a choice would show more evidence for the propose-but-verify account. The time given to make the choice is the condition. We use 3 different conditions for comparison pseudo-randomized within participants: 2s, 3.5s and 5s. Our results show stronger learning as more time is given. Furthermore, our findings are in line with the propose-but-verify mechanism in each condition. When modeling the interaction between condition and accuracy, our results do not reach statistical significance.

1 Introduction

When learning new words, children usually have to rely on referents from their immediate environment. They learn an association between a word and a referent, whether that is an object, an action, or an abstraction. It is easy to imagine how a child can hear the word "ball", while holding a ball, and make the association between the word and its immediately accessible referent. However, what if there are many possible referents? Usually, there is uncertainty about which referent is the appropriate one to be linked with a new word, given the complex environment in which the word was heard (Quine, 1964). When a child hears the word "fork", they could be seeing a fork, a spoon, a plate, and a cup in the same place. Even pointing to an object may be ambiguous because the word might refer to the object, its parts, an action, a state, or an attribute. Children often have to disambiguate between multiple word-to-referent mappings, yet they eventually succeed in learning the word.

1.1 Cross-situational word learning

Children succeeding in learning new words despite referent ambiguity is explained by the fact that they do not learn a new word through one learning instance. Instead, the same word will be learned only after it was encountered across multiple learning instances. This learning paradigm has been coined as cross-situational word learning (CSWL). For example, every situation in which the word "fork" was encountered might be ambiguous. Despite this, a fork will be present in every such situation. By recalling which object was present in every situation in which "fork" was uttered, one will be able to deduce the appropriate word-to-referent mapping.

Thus, in CSWL a new association will be learned across multiple ambiguous situations, in which the word-to-referent pair is present (Pinker, 1989;

Siskind, 1996). The correct referent (target) will co-occur with the word more often, while irrelevant items (distractors) will not always be present. The learner will extrapolate from the many exposures until they narrow down to the right association.

In a typical CSWL experiment participants are exposed to a number of words repeatedly in an ambiguous setting. Specifically, there are many possible referents for one word, and there is no feedback or clue as to which is the correct word-to-referent mapping. During a trial the participant will see or hear a word and have a number of images to select from as possible referents. The participants have to narrow down the right word-to-referent mappings by recalling previous co-occurrences. The context of previous encounters influenced the participants choice during later exposures to the same word, showing evidence that they keep in mind some information from that context (Roembke et al., 2023; Yu & Smith, 2007; Dautriche & Chemla, 2014). At each trial, the participant is typically required to click on an image out of a number of images shown on a screen, even if they have to choose at random. The number of times they clicked on the target divided by the total number of trials is the accuracy.

1.2 Gradual association

The initial assumption was that CSWL is a type of statistical associative learning. Namely, during an ambiguous learning situation, the learner will keep in mind many possible word-to-meaning mappings (Yu & Smith, 2007). At later learning instances for the same word, some associations will weaken, others will get reinforced by repeat referent-word co-occurrence. The correct association can eventually be narrowed down by disproving all the other association through repeated exposure. It has been shown that adults, as well as children, will retain multiple associations from a single learning instance (Yu & Smith, 2007; K. Smith et al., 2011; Dautriche & Chemla, 2014). This learning mechanism is known as *gradual association*.

1.3 Propose-but-verify

However, an opposing account was proposed that requires less memory load, namely the *propose-but-verify* (hypothesis-testing) mechanism (Medina et al., 2011; Trueswell et al., 2013). According to this account only one word-to-referent pair (a hypothesis) will be selected by the learner during an exposure. During later learning instances, this so called hypothesis may be confirmed and reinforced, or falsified and replaced by a new one. According to propose-but-verify one and only one association will be remembered during an encounter. It is therefore less memory intensive, but also less effective. Namely, in the case where the hypothesis was rejected because the referent is no longer present together with the word, a new hypothesis will be selected at random. Since no other word-to-referent pairings were retained, the new choice will not be influenced by the previous context or distractors. This contrasts with the gradual association account, in which multiple possible referents are retained. Thus, when the learner finds out that one of the referents is wrong, they will still have a higher than chance accuracy. Here, the chance level depends on the amount of images present during a trial (1 target plus the number of distractors).

Trueswell et al., 2013 found experimental evidence for the Propose-but-Verify mechanism, using a typical CSWL experiment paradigm. During a trial, the participants were shown 5 images along with a novel word. As shown in Figure 1.1 the distractors change with each trial, yet the target is always present with the word it is referring to. Namely, a picture of a bear is always present with the word "zud". If the participant has clicked on the bear in the first trial, then they have guessed correctly. In the subsequent "zud" trials they are likely to reinforce their hypothesis and select the bear again. It has been shown that in the cases in which the participant has made a correct choice in a previous trial, by clicking on the target ("previously correct"), they have a higher than chance accuracy in subsequent trials with the same novel word (Medina et al., 2011; Trueswell et al., 2013; Dautriche & Chemla, 2014). Note that the likelihood of being correct by chance in a trial, as shown in Figure 1.1 is 20%, considering that there are 5 images present on the screen during a trial.

However, the "previously correct" accuracy measure is consistent with both gradual association and Propose-but-Verify mechanisms. So, in order to disambiguate between the two mechanisms, Trueswell et al., 2013 looked at cases in which participants clicked on a distractor in the previous trial ("previously wrong"). For example, if the participant

Figure 1.1: Example of a typical CSWL experiment layout (Figure 1 from Trueswell et al. (2013)). This figure shows a sequence of two learning trials in which the word "zud" meant "bear". Note that the distractors are different, while the target stays the same, even though the target image may vary. In between these 2 trials there was other trials for other novel words.

clicked on the dog in the first trial in Figure 1.1, in the next "zud" trial the dog was no longer there. According to the gradual associative account, the participant will have memorized other images in addition to the one they clicked on. Importantly, they can recall the presence of a bear and notice that it is present again. Thus, they will click on the bear again. According to the gradual associative account, the participant will have a higher than chance accuracy in the "previously wrong" cases, in addition to the "previously correct" cases.

In contrast, the Propose-but-Verify account predicts that one and only one hypothesis will be retained. Thus, after having clicked on the dog, the participant will not be able to remember other images and will make a random choice in the subsequent trial. According to the Propose-but-Verify account the accuracy in the "previously wrong" cases will be at chance. With the described setup (see Figure 1.1), Trueswell et al., 2013 found the accuracy to be at chance level (20%), confirming the Propose-but-Verify account.

Dautriche & Chemla, 2014 used a similar paradigm to Trueswell et al., 2013, except that during each trial there were 3 distractors instead of 4, resulting in a total of 4 images instead of 5. Dautriche & Chemla, 2014 also used a new measure of accuracy. Instead of looking at all "previously wrong" cases, Dautriche & Chemla, 2014 focused on the cases in which the referent selected in the previous trial for the same word is absent in the subsequent trial. Specifically, they did not use accuracy in relation to the target, but in relation to the fact that the previously clicked image was present or not. In this measure, the cases where the previously clicked on referent is present again ("present") show a positive learning curve. This is congruent with both the gradual associative and the Propose-but-Verify accounts. Conversely, in the cases where the previously clicked on referent is no longer present ("not present"), a higher than chance rate of selecting a previously present referent would show that the participant is able to recall referents other than the one they clicked on. That is precisely what Dautriche & Chemla, 2014 found, supporting the gradual associative account.

This new measure is important, as in both the designs of the experiments of Trueswell et al., 2013 and of Dautriche & Chemla, 2014, due to the limited set of images, distractors sometimes repeat in subsequent trials. In these cases, due to the lack of feedback, the participant is prone to learning the distractor as if it were a target. The new measure proposed by Dautriche & Chemla, 2014 counts these cases as successful learning instances.

1.4 The mixed account

On the one hand, the memory load required for a perfect gradual association (where most of the relevant context is memorized) seems implausible. On the other hand, to just consider the propose-but-verify view is to leave out a lot of useful contextual information that could greatly facilitate the learning process. In light of the presence of evidence supporting the two opposing views (gradual association and propose-but-verify), it was suggested that the two accounts are both present as learning mechanisms working in parallel during CSWL (the mixed account) (Roembke et al., 2023; Kachergis et al., 2014; Roembke & McMurray, 2016; Wang, 2020). The degree of prevalence of one mechanism over the other may vary with different factors, such as the number of distractors, the familiarity with the novel words and/or referents, task instructions, time pressure to respond and people's beliefs and confidence during the learning process.

1.5 Research question

In this study, our aim is to clarify the relationship between the two learning mechanisms, gradual association and Propose-but-Verify, by bringing new evidence into the discussion. In particular, our research question is : *Does the time pressure to respond influence the learning mechanism the learner might use*, whether subconsciously or deliberately.

To predict the effect of time pressure on accuracy in a CSWL task, one might assume that with more time given to choose a word-to-referent pair, more context will be memorized. After all, there is more time to analyze and be attentive to the context, which favors gradual association. Meanwhile, little time makes for a fast and random choice, leaving no room to pay attention to the context, engaging in a propose-but-verify type of learning. However, the reverse could also be argued for. With more time, the learner will focus more on the one choice they make, forcibly ignoring the context. Therefore, with a more lenient deadline, the propose-but-verify mechanism is more likely to be used. Respectively, with less time they might focus less on any one hypothesis, yet some context might be remembered subconsciously, supporting the gradual association account.

However, for more precise predictions we take a look at the existing literature. Generally, the Propose-but-Verify mechanism is associated with explicit memory (Roembke et al., 2023), while gradual association is a type of statistical learning associated with implicit memory (Frost et al., 2019). In fact, it was shown that awareness is not required for subjects to show learning by gradual association in a cross-situational ambiguous learning instance (Yu & Smith, 2007).

With less time, less deliberate decision making will be made, and more of the subconscious visual memory will play a role in future learning instances. This study aims to show which strategy may be described as more often subconscious than the other.

We are not trying to separate the two, given that explicit and implicit memory can be used simultaneously (Paller et al., 2007). So can both propose-but-verify and gradual association strategies be used in parallel, in the mixed account (Roembke & McMurray, 2016). Instead, we are looking at how time pressure influences both strategies simultaneously.

Voss et al. (2008) found that introducing a 2 second deadline lowers accuracy in an explicit memory task and increases it in an implicit memory one. In general a stricter deadline should boost implicit memory detection, while inhibiting explicit memory mechanisms. Thus, we hypothesize that, even though overall accuracy will increase with a more lenient deadline, the two strategies will have separate correlations. Specifically, *with stricter time pressure, gradual-association will be more prevalent, while propose-but-verify will be used more as time given is increased.*

We use an experiment paradigm closely resembling the experiments performed by Trueswell et al., 2013 and Dautriche & Chemla, 2014. Furthermore, we introduce time pressure as an independent variable. We use the measure of accuracy proposed by Dautriche & Chemla, 2014. Specifically, we look at the accuracy in the "present" and "not present" cases. If accuracy is higher than chance in the "present" cases, it shows that participants retain memory of what they have clicked on. A higher than chance accuracy in "not present" cases would show that they retain information about the other stimuli as well, which is in line with the gradual associative account. Conversely, the "not present" accuracy being at chance level would be in favor of the Propose-but-Verify account.

Taking these measures, while manipulating the time given from stimuli onset (t), we predict the following. As t increases, the "present" accuracy increases as well, showing a positive learning curve. Furthermore, the "not present" accuracy decreases as t increases, showing that learning by gradual association is favored as a stricter deadline is applied. If the opposite happens, then at a low t we would see low accuracy in the "not present" cases, whereas at a higher t, accuracy in the "not present" cases would be higher as well. This would mean that propose-but-verify is favored with a stricter deadline and gradual association is favored with more time given. It also possible that we find no effect of t on the accuracy in the "not present" cases. This means that either accuracy is above chance or at chance, but always the same across all values of t. The former would be evidence in support of gradual association, while the latter is evidence in support of the propose-but-verify account.

2 Methods

Our experiment design closely resembles the experiments performed by Dautriche & Chemla (2014) and Trueswell et al. (2013). The goal is to find out what learning mechanism is used during a CSWL task under varying time pressure. Specifically, we want to look at how often and how many objects does a participant retain other than the one they clicked on, and compare data between different amounts of time given from stimulus onset. The two mechanisms, gradual association and propose-but-verify, differ in the cases in which the participant was wrong in the previous learning instance for the same word. Thus, we will look at accuracy in as a function of accuracy in the previous block.

2.1 Participants

We recruited 10 participants (4 male, 6 female). The participants were mostly university students, aged between 21 and 34, with 1 exception, aged 58. We performed the experiment on adults, replicating existing studies, in which reliable performance in typical CSWL tasks was found in adults (Dautriche & Chemla, 2014; Trueswell et al., 2013), allowing for robust experiment comparison. Additionally, there is the advantage of easier access to a larger number of adult participants. The participants are all English language speakers. Six participants are fluent in other languages as well, two of which are Dutch speakers.

The participants were recruited through social media and advertisements. The experiment was done anonymously. The participants had the option of receiving a small reimbursement of 5 euros.

2.2 Stimuli

We randomly selected 36 words from the Novel Object and Unusual Name (NOUN) Database provided by Horst & Hout (2016) (*vash, blicket, moop, adet, gaddle, smick, beag, slique, dite, tupa, krad, dax, vab, jate, koba, fote, wost, kinch, sarn, deld, cheem, smope, doud, quan, tulver, isot, geap, sprock, glark, aned, yok, sount, judpe, tust, plail, goke*). The NOUN Database provides many advantages. That is, the words were checked for novelty and similarity, resulting in a set of phonetically legal English words, that are nevertheless completely new and nonsensical. Since most participants are fluent in Dutch, as well as English, we checked all the words for possible usage in both languages. This way, we can control for possible association with other words in english or dutch that could facilitate learning. Another advantage of the NOUN Database is that many researchers use the NOUN database, including in the design of CSWL experiments, which facilitates experiment reproducibility. The words used by Dautriche & Chemla (2014) and Trueswell et al. (2013) are also part of the NOUN Database, allowing for a robust experiment comparison.

In order to simulate CSWL, the referents to be associated with the words must be real objects that are physically possible, yet completely new, meaning they do not have a word already associated with them or they cannot be named by the vast majority of the population. Although Trueswell et al. (2013) and Dautriche & Chemla (2014) used familiar objects as referents, we decided to change the design in favor of the typical CSWL experiment design. The main reason is that, in order to amplify the subtle difference between the two mechanisms (gradual association and propose-but-verify), we should facilitate the use of subconscious learning (so that participants are able to remember referents, besides the one they clicked one, predicted by the gradual association account). In this regard, novel objects, that can not be easily described, were found to increase the likelihood for participants to use implicit learning (Wang, 2020).

To this end, we selected images from the NOUN Database, as well as 3 images of dog toys, resulting in 36 different objects. Since these images are widely used in research, the same reproducibility advantages apply (Horst & Hout, 2016). The NOUN Database images, as well as the ones we selected, depict multipart, multicolored, real three-dimensional (3-D) objects. All images are novel and not already associated with a specific name.

In order to facilitate CSWL learning of objects, instead of memorizing the images, or specific parts or attributes (color, brightness, rotation, size), we created variations for each object. Specifically, for each of the 36 objects, we manually edited 5 variations, by changing color, rotation, reflection and size to random values. All 180 images are available for free use under the CC BY-NC 4.0 public copyright license

(https://drive.google.com/drive/folders/1VnYjcWvjrBBTF8Q8767PeLIcq27E-lg3?usp=sharing).

2.3 Design

Our aim is to compare the effects of different levels of time pressure on performance. Thus, the condition (independent variable) for each trial is the time given from stimuli onset (t). t is a continuous variable, however for experimental simplicity we reduced the condition to just 3 different t values to be compared. Trueswell et al. (2013), using eye-tracking, found that accuracy was already reliable starting at t=1000ms. However, additional time is necessary to move the cursor and click on an image. Wang (2020) found a 55% accuracy rate in a CSWL experiment format with a deadline of t=1500ms. Additionally, after piloting the experiment at t=1500ms we found that in many trials the time was not enough for participants to react. Thus, we settled on a lower bound of t=2000ms. In a memory retrieval experiment, Voss et al. (2008), found that given the instructions to take as much time as they need, participants averaged at about 7500ms of response time. For the upper bound, 7500ms will make the experiment unnecessarily lengthy. Ultimately, we settled on 3 conditions: *t=2000ms, t=3500ms* and *t=5000ms.*

Other studies have reported variability between individuals, as determined by attention span (Yu & Smith, 2011, 2012; L. B. Smith & Yu, 2013) or language skills (Vlach & DeBrock, 2017). Furthermore, there have been found differences in performance, when looking at groups, such as monolinguals and bilinguals (Escudero et al., 2016; Poepsel & Weiss, 2016), children with developmental language disorder (McGregor et al., 2022) or children with autism (Hartley et al., 2020). To control for variability between individuals and groups, we ran all 3 conditions for each subject, which also has the advantage of providing more data points per participant, thus requiring fewer participants for a robust statistical analysis.

However, individuals can get better at CSWL tasks after repeated exposures (Roembke et al., 2023), meaning that by the third condition participants could perform better than during the first condition. Other factors, like attention exhaustion and task assimilation could influence performance across conditions. To control for these differences

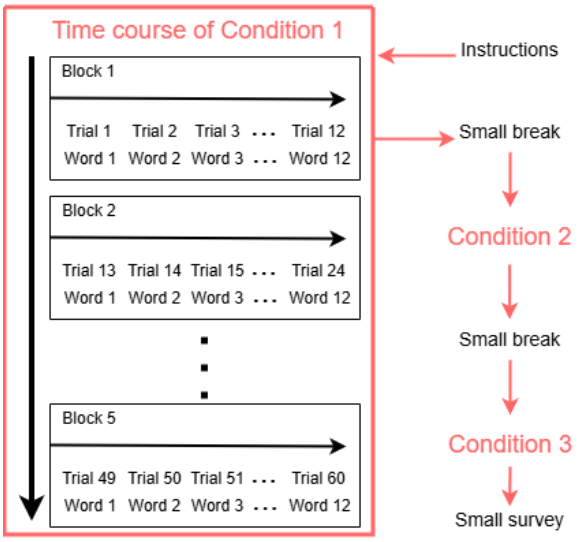

Figure 2.1: Flow chart of the experiment time course. Note that same time course as shown for condition 1 applies to conditions 2 and 3, but with different sets of words and images. Conditions 1, 2 and 3 are different for each participant (in a counterbalanced order).

across conditions we sequenced the conditions in a counterbalanced order across participants. Specifically, each participant gets a different condition order, while keeping the number of participants per condition order similar (maximum difference in participant number is 1).

We developed a computer program in C-sharp that pseudo randomizes the stimuli for this specific experiment design. The algorithm checks for multiple parameters in parallel and outputs appropriate datasets of stimuli for each participant. The sets of 36 words and 36 objects are each shuffled and split into 3 sets (1 per condition). This way, we control for any possible differences in performance across words or images, due to possible familiarity. Similar to Trueswell et al. (2013) and Dautriche & Chemla (2014) experiments, each condition consists of 5 blocks, with 12 trials each (one trial per word). Thus, there are 60 trials during a condition, such that trials for the same word always appear 11 trials apart, as shown in Figure 2.1.

In each trial there are 4 images (1 target and 3 distractors). Our algorithm picks distractors in a pseudo-random way to minimize the co-occurrence across subsequent trials of the same word. All ob-

jects in the same trial are different. In total, participants see 180 different images during a condition (4 images * 12 trials * 5 blocks). For repeated object occurrences across trials, all object variations are used in a shuffled and spaced out order. The order of appearance of the four images on the screen during a trial is randomized.

Finally, the stimuli dataset is applied to a jsPsych experiment. jsPsych is an open source project tailored for creating behavioral experiments that run in a web browser (de Leeuw et al., 2023). The experiment starts with a short survey, by which we collect relevant data about participants: age, sex and spoken languages. Next, the following instructions are shown: *In this experiment you will learn new words. Click on the object that you think the word might be referring to. You will have limited time to do so. Please use whichever device you feel more comfortable with, a mouse or track-pad. You will be asked which one you're using. Please use a laptop or a PC, NOT a tablet or phone. You will start with 2 examples.*

After the instructions there are two example trials, one with recognizable animals, and the other with recognizable fruits. These examples are there to get the participant used to the flow of the experiment and to the time pressure. These examples appear at the beginning of each condition.

Immediately after the examples, the trials begin. A trial starts with the participant being asked to click an a cross in the middle of the screen, as shown in Figure 2.2. This is done so their cursor is located in the middle of the screen at stimuli onset. Upon clicking on the cross, the participant will see a short sentence containing the word to be learned and 4 images of unfamiliar objects located equidistantly from the cross in the middle of the screen (see Figure 2.2). In this way we can control for the time it takes to move the cursor to an image. The images are equidistant from the cross in the middle of the screen, so that there is no bias towards selecting an image, and it takes equal time to move the cursor to any of the images.

After the images appear on screen a timer starts, according to the current condition (2, 3.5 or 5 seconds). During this timer, if the participant has clicked on an image, then the next trial starts. Otherwise, if the participant didn't click on any image before the time runs out, then there is a red screen with the text *"Too late!"* for 2 seconds before the next trial.

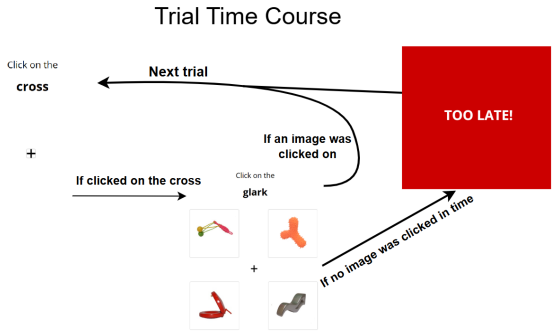

Figure 2.2: Example of the time course of a trial. This loop continues for every trial until the end of block 5. Note that the "Too late!" red screen appears only in case no image was clicked on.

2.4 Procedure

The experiment was hosted on a JATOS (Just Another Tool for Online Studies) server. JATOS is a tool that helps set up and run online studies directly on a personal or public server (Lange et al., 2015). JATOS provides complete control over the research data and helps ensure compliance with ethical guidelines.

The participants received a link, through which they could access the experiment online, on our JATOS server. The experiment was done on a computer or laptop, using a mouse or track-pad. The experiment took 15 minutes on average. After the participants completed the experiment, the data was sent to the JATOS server, ready to be downloaded.

2.5 Data processing

No participants were excluded from the data analysis. There were no apparent rules violations. All participants used the appropriate devices (laptop or PC). There were no obvious outliers when looking at reaction time. Each participant is a fluent English speaker. No participants were detected to have a bias towards an image position on the screen. We collected a total of 1800 data points, each corresponding to a trial (180 per participant, 10 participants). Trials in which participants did not click

on any image in the given time (timeouts) were excluded from the main data analysis.

2.6 Data analysis

Participants' responses were registered as "Correct", if they clicked on the target, and "Wrong", if they clicked on a distractor. However, all objects for each trial were also recorded, in order to be checked for distractor repeats between subsequent blocks. The object that was clicked on was recorded, to be checked if it was present in the previous block. Other trial data was also recorded, such as image variation, response time, condition, target and image position on the screen.

3 Results

All data analysis was performed in R Studio, using the data files created by the jsPsych script as input data. All plots and graphs were done using the ggplot2 package (Wickham, 2009).

3.1 Timeouts

Since the time given from stimulus onset (t) was very limited, it was to be expected that some timeouts will happen. Any such occasion, would reduce the amount of data points. Furthermore, even if no image was selected, the participants will have seen the images, and might have engaged in learning. Yet, we can't capture that, since no response is recorded.

Thus, to reduce the number of timeouts, we informed participants that there will be limited time at the beginning of the experiment. We also informed participants at the beginning of every condition that the time pressure will change. Furthermore, we provided 2 mock examples at the beginning of each condition (see Figure 2.1), to get the participants used to the time pressure. To check for any data anomalies and to take a look at participant interaction with the conditions, we take a deeper look at the timeouts.

We checked for the number of timeouts, by taking the percentage of timeouts out of all trials in a block (12 trials). One participant was an outlier, with an average of 70% timeouts, but only in the fastest condition (t=2000ms). Other than that, the timeout rate never exceeded 40%, and stayed below 17% in 95% of the total number of blocks.

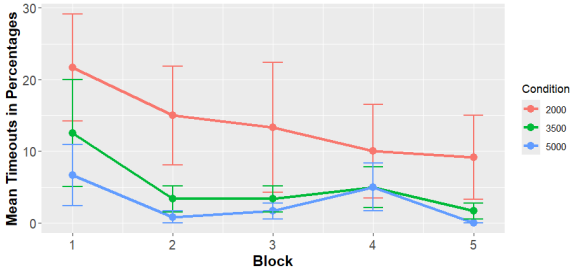

Figure 3.1: Mean timeouts per block in percentages. The percentage of timeouts is the number of timeouts in a block divided by the total number of trials in the block (12). Error bars show the standard error to the mean.

When comparing timeouts between conditions we see that with a lower t, there tends to be a higher number of timeouts as shown in Figure 3.1. Furthermore, participants showed a tendency to adapt to the deadline. As seen in Figure 3.1 there is a downward trend of the number of timeouts from block 1 to 5.

3.2 Accuracy

In each trial, the response was registered as "Correct" when the participant has clicked on the target (the intended referent, the image that appears with the same word in every block within a condition) or "Wrong" if the participant clicked on a distractor. Timeouts were registered as "NULL" responses and omitted for the purposes of the accuracy analysis. For each block, we measure accuracy as the ratio of "Correct" responses over the total responses for the respective block. Using this measure of accuracy, we calculate the mean accuracy across all blocks to be 28.22%, with a standard error of 1.25. Importantly, accuracy is found to be higher than chance. Chance is at 25%, because the chance to click on the target out of four images shown on the screen is 1/4.

In order to test our hypothesis, we look at the accuracy as a function of the accuracy in the previous block. Specifically, for each word we compare the accuracy in block N between cases in which the response was "Correct" ("previously correct") in block N-1 and cases in which the response was "Wrong" ("previously wrong") in block N-1 for the

same word. In this measure we also exclude all trials in which there was a timeout in the trial for the corresponding word in block N-1, resulting in a total number of trials of 1274. The resulting mean of the accuracy is 30.1% (higher than chance), with a standard error of 1.73.

The mean accuracy in "previously correct" cases is 35.97% with a standard error of 3.06. Notably, accuracy is significantly higher than chance, showing successful learning in "previously correct" cases. The mean accuracy in "previously wrong" cases is 24.38% with a standard error of 1.47, which is at chance level. This shows that after clicking on a distractor participants chose a new referent at random, which is in line with the propose-but-verify mechanism. Note that according to the propose-but-verify account only hypothesis of the word-to-referent associations is retained at a given moment. Thus, when wrong, a new hypothesis has to be chosen at random.

Furthermore, we look at differences among conditions. As seen in Figure 3.2, for every condition in "previously wrong" cases the accuracy mean is at chance level (25%). Thus, the results for every condition are in line with the propose-but-verify account. Note that in order to see evidence for learning by gradual association, we would need to find a higher than chance accuracy in the "previously wrong" cases.

On the other hand, looking at "previously correct" cases across conditions, there seems to be a bigger difference between the "previously correct" and "previously wrong" cases in the slowest condition (t=5000ms), than between the "previously correct" and "previously wrong" cases in the fastest condition (t=2000ms) (see Figure 3.2). This affirms that with more time given, participants were able to retain word-to-referent pairs more often.

Because the dependent variable is binary and we are looking at the interaction between two independent variables (condition and accuracy in the previous block), we use a mixed effects logistic regression to model our data. The effect of the "previously correct" condition on the accuracy does not reach statistical significance ($z = 1.36$, $p = 0.174$). Furthermore, looking at the effect of the condition on the accuracy in the "previously correct" cases we find for t=3500ms, $z = 1.214$, $p = 0.225$, and for t=5000ms $z = 1.381$, $p = 0.167$. Neither interaction reaches significance. We suspect that the statistical

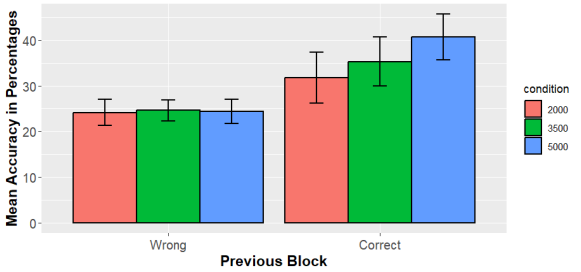

Figure 3.2: Mean accuracy per block, measured in percentage of correct responses (the participant clicked on target) over total responses in the block (timeouts omitted). Accuracy is taken as a function of accuracy in the previous block when looking at the trial with the same word. "Wrong" means that the participant clicked on a distractor, and "Correct" means that the participant clicked on the target in the previous block. Error bars show the standard error to the mean.

analysis is underpowered, given the small number of participants (N = 10) and the relatively high variability between participants, in both accuracy and timeouts.

3.3 Alternative measure

So far we have looked at accuracy as a function of the participant having clicked on the target or distractors. However, this measure does not capture all possible successful learning instances. Sometimes distractors may also repeat in learning instances for the same word, but no more than twice in a row. In these cases the participant might have learned the distractor as a correct referent. This could be considered successful learning, since the participant does not receive feedback and has no way of knowing which is the intended target.

To capture these cases Dautriche & Chemla (2014) have proposed a new measure that accounts for all distractors that appear sequentially in the learning instances for the same word. This time we count the repeating distractors as possible targets to be learned. In this measure, chance level depends on the number of images that repeat in subsequent learning instances. Thus, chance level will vary for every trial, but can only have values of 25%, 50% or 75%, depending on the number of repeating ob-

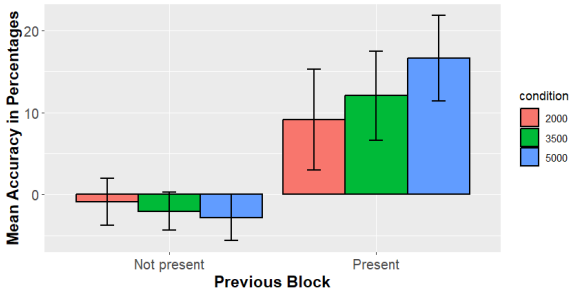

Figure 3.3: Mean accuracy measured in percentage point minus chance level. Chance level varies per trial (can only have values of 25%, 50% or 75%). "Not present" means that the participant clicked on an object that was not present in the previous learning instance, and "Present" means that the participant clicked on an object that was present in the previous learning instance. Chance level is at 0%. Error bars show the standard error to the mean.

jects.

In this measure, we calculate accuracy for each trial as the proportion of responses that belong to the set of repeating stimuli minus chance (which varies per trial). Figure 3.3 shows the accuracy as informed by previously seen referents, where chance is at 0%.

Instead of comparing between "previously wrong" and "previously correct" cases, we now look at whether the referent clicked on in the previous learning instance is present again. Thus, we differentiate between "Present", cases in which the referent the participant clicked on in the previous block is present again, and "Not present", cases in which the chosen referent is no longer present. Figure 3.3 shows accuracy separately for "Not present" and "Present" cases.

This new measure does not change the results. We see higher accuracy in "Present" cases as t increases, just like in the "previously correct" cases, affirming that more time given facilitates learning. Meanwhile, the "Not present" cases show accuracy at chance, in line with the propose-but-verify mechanism.

We use a mixed effects linear regression to model our data. The effect of the "Present" condition on the accuracy does not reach significance ($t = 1.479$, $p = 0.139$). Furthermore, looking at the effect of the time pressure condition on the accuracy in the "previously correct" cases we find for t=3500ms, $t = 1.448$, $p = 0.148$, and for t=5000ms $z = 1.302$, $p = 0.193$. Neither interaction reaches significance, just like in the previous measure.

4 Discussion

Our findings point to a possible effect of the time pressure on the accuracy. Namely, with more time given from stimuli onset learning is improved. However, with no statistical significance found, these conclusions are only preliminary. Furthermore, participants used their previous choice to guide their choice in subsequent learning instances, shown by the difference in accuracy between "previously correct" and "previously wrong". This effect did not reach statistical significance either. The main reason is that we had a limited sample of only 10 participants. A bigger sample would improve the statistical power of the data analysis and would be able to reveal the more subtle effects, such as small differences in "previously wrong" cases across conditions.

Importantly, our results show an at chance (25%) accuracy in the "previously wrong" cases for all conditions (see Figure 3.2). This is in line with Trueswell et al. (2013) findings affirming the propose-but-verify model. However, our experiment closely follows the experiment design by Dautriche & Chemla (2014), and yet contrary to their findings, we do not find evidence of learning by gradual association. The gradual association mechanism would be supported by a higher than chance accuracy in "previously wrong" cases. The only relevant difference that sets Dautriche & Chemla (2014) experiment apart from ours is that, although they used novel words, they used pictures of familiar objects from common categories, such as animals, dishes and clothes. Images of familiar objects would facilitate learning, since they are easier to remember. On the other hand, Wang (2020) found a higher likelihood that participants use implicit learning if the objects shown could not be easily described. Manipulation of object familiarity could show more interaction with time pressure.

The new measure proposed by Dautriche & Chemla (2014) did not reach statistical significance either. Even though this new measure captures

more of the possible successful learning instances, it did not reveal anything different. That is explained by the low number of participants and high variability between individuals.

Our randomization of word and object selection ruled out the possibility of a bias towards choosing a word-to-referent pair by familiarity or semantic closeness. Participants did not report to have encountered familiar words or objects during the experiment.

During the feedback phase (at the end of the experiment), the majority of the participants complained that they did not have enough time to make a choice and think about their answer. However, the tight deadline was important to facilitate the detection of a subconscious learning process. From the timeout analysis, we found that the number of timeouts was acceptable for us to proceed with the data analysis (see Figure 3.1). However, each timeout is a lost data point, for which the learning process could not be recorded.

An array of improvements could facilitate timely responses and reduce the number of timeouts. These can include more detailed instructions about the expected time constraints and the nature of the experiment. We kept the instructions to a minimum, in order to facilitate subconscious learning. Kachergis et al. (2014) showed, in a CSWL experiment, that adults were able to acquire new word-to-referent mappings even in the absence of explicit instructions. Another improvement to reduce the number of timeouts would be an increased number of examples and/or mock trials, which may help participants get used to the limited time.

Other studies (Escudero et al., 2016; Poepsel & Weiss, 2016) found performance differences between bilinguals and monolinguals. Since there was great variability in the number of languages our participants possess, controlling for that would be an improvement to the current experiment design, but this would require access to considerably more participants. Furthermore, possession of other languages could interfere with word familiarity. Some words could sound familiar in languages, other than we controlled for (English and Dutch). Thus, participants should be chosen based on the languages they speak, ideally possessing only languages checked for word familiarity, and all speaking the same languages (to decrease variability).

There is evidence of performance differences in typical CSWL tasks between various age groups (Fitneva & Christiansen, 2017). Therefore, to reduce variability across participants, they should all be part of the same age group. Our participants age ranged from 21 to 58 years.

In conclusion, we find that after being incorrect, participants choose at random in the next learning instance. The implication is that participants were memorizing only one hypothesis and no context (distractors). These results are in line with the propose-but-verify account. Furthermore, in the cases in which participants choose correctly they are more likely than chance to choose the correct referent again. This means that they were able to successfully learn the word-to-referent mappings. However, the effects of time pressure on accuracy did not reach statistical significance. Therefore, the results are inconclusive. Further research with an improved experiment design could reveal more about the effects of time pressure on CSWL.

References

Dautriche, I., & Chemla, E. (2014). Cross-situational word learning in the right situations. *Journal of Experimental Psychology*, *40*, 892–903.

de Leeuw, J., Gilbert, R., & Luchterhandt, B. (2023). jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8(85)*, 5351.

Escudero, P., Mulak, K. E., Fu, C. S. L., & Singh, L. (2016). More limitations to monolingualism: bilinguals outperform monolinguals in implicit word learning. *Front. Psychol.*, *7*, 1218.

Fitneva, S. A., & Christiansen, M. H. (2017). Developmental changes in crosssituational word learning: the inverse effect of initial accuracy. *Cogn. Sci.*, *41*, 141–161.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: a critical review and possible new directions. *Psychol. Bull.*, *145*, 1128–1153.

Hartley, C., Bird, L. A., & Monaghan, P. (2020). Comparing cross-situational word learning, retention, and generalisation in children with

autism and typical development. *Cognition*, *200*, 104265.

Horst, J., & Hout, M. (2016). The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behav Res*, *48*, 1393-1409.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2014). Cross-situational word learning is both implicit and strategic. *Front. Psychol.*, *5*, 588.

Lange, K., Kühn, S., & Filevich, E. (2015). 'just another tool for online studies' (jatos): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE*, *10(6)*.

McGregor, K. K., Smolak, E., Jones, M., Oleson, J., Eden, N., Arbisi-Kelm, T., & et al. (2022). What children with developmental language disorder teach us about cross-situational word learning. *Cogn. Sci.*, *46*, e13094.

Medina, T., Snedeker, J., J.C., T., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proc. Natl. Acad. Sci. U.S.A.*, *108(22)*, 9014-9019.

Paller, K., Voss, J., & Boehm, S. (2007). Validating neural correlates of familiarity. *Trends Cogn Sci.*, *11(6)*, 243-50.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.* Cambridge, MA: MIT Press.

Poepsel, T. J., & Weiss, D. J. (2016). The influence of bilingualism on statistical word learning. *Cognition*, *152*, 9-19.

Quine, W. V. O. (1964). *Word and object (vol. 4).* Cambridge, MA: MIT Press.

Roembke, T. C., & McMurray, B. (2016). Observational word learning: beyond propose-but-verify and associative bean counting. *J. Mem. Lang.*, *87*, 105-127.

Roembke, T. C., Simonetti, M. E., Koch, I., & Philipp, A. M. (2023). What have we learned from 15 years of research on cross-situational word learning? a focused review. *Frontiers in Psychology*, *14*, 1-9.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39-91.

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

Smith, L. B., & Yu, C. (2013). Visual attention is not enough: individual differences in statistical word-referent learning in infants. *Lang. Learn. Dev.*, *9*, 25–49.

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

Vlach, H. A., & DeBrock, C. A. (2017). Remember dax? relations between children's cross-situational word learning, memory, and language abilities. *J. Mem. Lang.*, *93*, 217–230.

Voss, J., Baym, C., & Paller, K. (2008). Accurate forced-choice recognition without awareness of memory retrieval. *Learn Mem.*, *15(6)*, 454-9.

Wang, F. H. (2020). Explicit and implicit memory representations in cross-situational word learning. *Cognition*, *205*.

Wickham, H. (2009). ggplot2: Elegant graphics for data analysis. *New York, NY: Springer.*.

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. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Dev. Sci.*, *14*, 165-180.

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word–referent learning: prior questions. *Psychol. Rev.*, *119*, 21-39.