Bachelor thesis

Artificial Intelligence

Radboud University

Comparing distance-based point feedback in vocabulary acquisition

Author:
Thomas de Lange
S1016514

Supervisor:
dr. F.T.M. Leoné
Artificial Intelligence
f.leone@donders.ru.nl

Second reader:
dr. A.F.J. Dijkstra
Centre for language studies,
Donders Institute
t.dijkstra@donders.ru

February 18, 2023

Abstract

Feedback is an important part of education, yet proper feedback given by a teacher is time consuming. For this reason other methods such as automatically giving points for answers have been used. However, these points are often in a binary form and do not give much detail. Therefore, this study investigates the effects of using different forms of orthographic similarity-based feedback on vocabulary acquisition as they could possibly be more informative. I compare two forms of feedback based on the word distance to the correct answer. In this study participants learned 15 words in Swahili on a 2D map that visualized the distances between words and were given points for their guess with regards to the correct answer. One condition received points based on the distance regardless of their previous answer. The other condition gave points relative to their previous and the correct answer. After the learning session, as well as three days later, participants were tested on their knowledge of the vocabulary using a translation recognition test. No significant difference was found between the two learning methods. However, the results do indicate a possible preference for exclusively positive feedback to increase motivation.

Introduction

One of the most important aspects for effective education is providing feedback, as receiving proper feedback has been shown to increase how well information is remembered (Cyr & Anderson, 2018; Metcalfe, 2017). At the same time it is important for the motivation of the student (Burgers et al., 2015). It has been found that more detailed feedback fosters a better understanding of the mistakes made (Metcalfe & Finn, 2012). Thus, not only the correct answer should be provided but also clear reasoning on why that is the correct answer. Metcalfe et al. (2009) showed that feedback does not have to be immediate after error generation to be effective. However, it is important the student pays adequate attention to the feedback to ensure it's effectiveness.

Ideally, feedback is given through personal supervision or collaborative effort (Gallien & Oomen-Early, 2008). However, this is not always feasible given the time and costs associated. Therefore, other methods of feedback have been used (Keuning et al., 2016). For example, some systems make use of points based on the users answers to follow the progress of the student. However, these points are often not informative to the user as they only supply a binary 'right' or 'wrong'. In particular, in the learning of vocabulary there are opportunities to expand on the detail of feedback provided to the student in the form of points. Therefore, this research will focus on how orthographic similarity can be used to improve feedback for vocabulary learning.

In order to establish the relation between feedback and information retention, I will explore how information, in particular vocabulary, is stored in the brain, how feedback affects learning, and how these can be combined in order to provide a more detailed similarity-based feedback system.

Vocabulary learning and similarity

The mental lexicon describes the organization of words in the brain (Aitchison, 2012). It contains not only the words meaning, but also the associated form in both orthography and phonology. Orthography concerns the written form (i.e. the combination of letters that form the word) and phonology the spoken form (i.e. the sounds associated with the combination of letters). A lexical representation can be said to be of high quality if it contains the complete orthographic representation of the word as well as a redundant phonological representation, covered by both a representation from spoken language and from orthographic-to-phonological mapping (Perfetti & Hart, 2002). This results in words being complete and reliable on retrieval.

It is known that words similar in form are grouped together and that the activation of one word causes activation of others. Words will elicit co-activation of similar words causing a competition between the word and its neighbors (Collins & Loftus, 1975; De Groot, 1983).

Learning words is not a matter of binary knowledge. Instead words will be partially available in the brain and therefore be learned gradually over time (Laufer & Goldstein, 2004). How this gradual learning is shaped has an effect on both the learning performance as well as the motivation for the learning task.

In learning a distinction can be made between generating errors or not. This leads to two the-

ories on learning. Errorful learning and errorless learning. In errorful learning making mistakes is encouraged and corrective feedback is provided. In errorless learning knowledge is presented in a way such that mistakes cannot be made and the correct representation is therefore kept error free (Scheper et al., 2019). Both errorful (Cyr & Anderson, 2018; Metcalfe, 2017) and errorless (Scheper et al., 2019; Warmington et al., 2013) learning have been shown effective for studying. However, the context is important in deciding which method is best. Potts et al. (2019) suggests that errorful learning is beneficial when studying vocabulary in particular.

In addition, the way words are represented during learning has an effect on how well words are remembered. Baxter et al. (2021) have shown that contrasting words of similar orthographic form sharpens the differentiation of these words and thus leads to better retention of the learned words. It has been shown that grouping semantically related words (e.g. *cat* and *dog*) can have a negative affect on word retention (Stahl et al., 1992). Additionally, cross-language similarity can also have effects on word learning (Dijkstra et al., 2010).

Earlier work

Recent work by Ansteeg et al. (2023) tried to establish a connection between generating errors and an improved retention of vocabulary given different forms of similarity-based feedback. They performed research comparing binary versus gradual point rewards in vocabulary learning. The experiment consisted of participants learning thirty Finnish nouns repeating six times in an online setting. During the learning sessions participants received points for their answer. In both the binary and gradual condition 100 points were attributed for a correct response, however in the binary condition an incorrect response earned 0 points. In the gradual condition participants received points based on how far their given answer was from the correct answer based on the normalized Levenshtein distance. The normalized Levenshtein distance describes the amount of differences in characters between two words, using additions, deletions and/or substitutions (Levenshtein, 1966). Participants would always score points regardless of their previous answers.

The researchers found that that participants in the gradual condition group committed errors more often during learning, the post-test and the delayed post-test. Although no significant difference was found between the two groups for the immediate test performance, the gradual condition group did perform better on the one-week delayed test. This suggests that gradual feedback might improve memory retention over time. Moreover, it was shown in both groups that the learning of words was gradual. In addition, they highlight the advantages of gradual feedback in being able to follow progress in an online setting, as well as being more motivating and enjoying for the student.

The role of distance in learning and motivation

In general, feedback plays a role in the motivation of the student. Negative feedback (i.e. being punished for mistakes) has been shown to reduce the motivation of students for the task (Zhuang et al., 2017). So, although negative feedback might improve memory retention during learning it can negatively affect the long-term effectiveness if student motivation drops.

There seem to be advantages to gradual feedback over binary feedback as seen in the work by Ansteeg et al. (2023). Yet, it is unknown how this gradual feedback should be given. There are two theories on learning and motivation that could be applied as a basis to using similarity in learning. These are prediction errors and the learning progress hypothesis.

Firstly, prediction errors can be described as errors that occur when results of a given action are surprising (Ouden et al., 2012). In the case of motivational prediction errors neurons react the the valence (sign) of the error with the regards to the expectation. Meaning that neurons react differently when the outcome was positively versus negatively surprising. In particular, neurons increase their firing rate when the outcome was better than expected, do not change when the outcome was expected and fall silent when an expected reward is left out.

This can be related back to the way similarity can be used to provide feedback in word learning. If errors are rewarded according to the distance they were away from the correct word it could reinforce a gradual style of learning by slowly increasing reward as the target word is approached.

Secondly, the learning progress hypothesis suggests people are intrinsically motivated to pursue activities in which predictions are improving

(Oudeyer et al., 2016). Moreover, humans are nudged towards tasks of intermediate complexity. Tasks that are not too hard, yet not too easy and provide for a learning opportunity. According to the theory, people are motivated to learn by a curiosity to receive information. In other words, people are motivated to learn as long as the experience progress in their knowledge. This means that the more informative feedback might increase motivation as it provides more learning opportunities and thus a greater sense of progress.

When integrating this with similarity measures it is feasible that a form of feedback that is relative to the previously given answer is more informative. It not only gives insight to the distance to the target word, but also gives information on the relation between the target word and the last given answer. Thus it might be more motivating to receive feedback in this way.

Discerning effects of distances

The research by Ansteeg et al. (2023) sets the stage for the work in this thesis. There are clear reasons to believe gradual feedback is advantageous over binary feedback. Yet it is unknown how this gradual feedback should be given. The researchers used what I will call an absolute form of rewarding points. Participants score points based on distance regardless of their previous answers. As highlighted earlier, a form of relative feedback is possibly more informative to the user and therefore more motivating. With a relative form points are based on the previously given answer for that specific word. This means that points could be taken away if the distance increased and less points will be rewarded for improving only a small distance. This type of feedback looks to be more accurate to the actual learning that takes place and thus more motivating.

Therefore this work will focus on comparing an absolute versus relative reward system in vocabulary learning. The aim is to investigate how effective both methods are for vocabulary acquisition as well as how motivating and enjoying for the participants. I hypothesize that the relative scoring condition will be more effective and motivating as it will be more in line with the learning that occurs. This hypothesis is based on the learning progress hypothesis by Oudeyer et al. (2016) that suggests that experiencing a sense of progress is vital for a motivating learning experience.

Method

Design

The experiment concerned a within-subject, 2x2 factorial design which was counterbalanced to remove ordering effects. Participants were randomly distributed over different combinations of the two lists and two conditions. Participants were free to perform the experiment at home. They were instructed to perform the experiment on a suitable PC or laptop and to complete it in one session.

		Word lists	
		List 1	**List 2**
Conditions	**Absolute**	Group A	Group B
	Relative	Group C	Group D

Table 1: 2x2 design of the conditions combined with the two lists of words. Groups indicate with which pair of conditions the participant started.

Participants

The group of participants consisted of 17 participants aged between 18-35 years old(8 female, mean age: 22.2). They were all Dutch native speakers and had no disorders that could influence their vocabulary learning capabilities. Additionally, none had prior experience with Swahili. All participants participated voluntarily and received no compensation.

Materials

Word lists

Two word lists were made consisting of fifteen translations Dutch to Swahili. Swahili was chosen as it shares the Latin alphabet with Dutch making it easier for participants to learn the language without prior knowledge. Moreover, it is practical to find participants that do not yet have knowledge of Swahili. The number of fifteen words was chosen as it balances the relative short time it takes to learn the words as well as a large enough amount in order to make inferences on performance. The words were chosen from a large dataset on Swahili words by Goertz et al. (2023), which included the values for the selection criteria.

Words were selected based on the following criteria: concreteness, frequency and word length. Concreteness describes how tangible a word is. Words were selected on concreteness, as more tangible words are learned easier than abstract words (Fliessbach et al., 2006). Frequency describes how often words are used in daily conversation. Word frequency is an important predictor for learning effectiveness (Brysbaert et al., 2011). Finally, the word length is an important factor as it can effect how well words are remembered and later recalled (Baddeley et al., 1975). Thus lists selected were of similar average values for these criteria. The two lists were balanced to have similar values on these criteria as well.

List 1		List 2	
Swahili	**Dutch**	**Swahili**	**Dutch**
mizizi	wortels	fasihi	boek
mchuma	geweer	farasi	paard
mkono	hand	funguo	sleutel
kupanda	fabriek	mchange	zand
duara	wiel	mlima	berg
degaga	bril	miguu	voeten
chupa	fles	ramani	kaart
chura	kikker	kinywa	mond
bahasha	envelop	bamba	bord
bahari	zee	kamba	touw
langa	deur	dhahabu	goud
shimo	gat	chanjo	schaar
kanisa	kerk	chikiru	vogel
nyasi	gras	dhoruba	storm
nyoka	slang	barua	brief

Table 2: The two word lists used to test the experiment.

Survey

To guide the participants through the experiment a Qualtrics survey was used. The survey gathered consent as well as basic information such as age, gender (if specified), learning disorders and knowledge of both Dutch and Swahili. Additionally, it contained the post-test. Finally, the survey contained a set of qualitative questions asking participants on their experience and motivation during the two conditions. They had to give a score on a Likert scale ranging from one to five for both conditions and could also describe their reasoning behind it.

Post-tests

The vocabulary post-test contained three sets of pairs of Dutch and Swahili words. These sets were chosen as follows. One set contained all the correct translations Dutch to Swahili. Another set contained pairs where the Swahili words were similar to the actual translation. For example, the Dutch word for 'slang' was coupled with 'nyasi', whereas the correct word for 'slang' is 'nyoka'. The last set contained pairs where the Swahili words were dissimilar to the actual translation. These three sets were combined and randomized so participants saw 45 pairs of which 15 contained the correct translation. The task was to identify whether a pair contained the correct translation or not.

MindSort

For learning the vocabulary participants made use of MindSort. MindSort is an in-development brain-inspired model-based approach for language learning (Leoné, 2023). This online tool allows for users to study translations of words using a location-based system. Users can upload their own lists to study and study using a variety of modes. The software contains multiple methods for generating the maps that users can interact with.

MindSort is based on research on the cognitive organization of information as well as the use of visuospatial bootstrapping (Darling et al., 2020). As explained in the introduction, information in the mental lexicon is stored based on similarity. MindSort makes use of this by grouping words together based on their similarity. This contrasting elicits co-activation of neighboring words and strengthens their representations (Collins & Loftus, 1975). Visuospatial bootstrapping describes the effect that the embedding of spatial information has on remembering information (Darling et al., 2020). The maps used in MindSort aim to make use of this effect by again grouping words together to create an overview that aids in remembering words.

The 2D map describes the dissimilarity between words. Meaning words that are similar are closer together than those that are dissimilar. For this experiment the distance between words is calculated using the normalized Levenshtein distance. As explained before the Levenshtein distance counts the amount of additions, deletions and/or substitutions are needed in order to convert one word into another word (Levenshtein, 1966). This results in a value where the distance is a fraction of the maximum length of the two words.

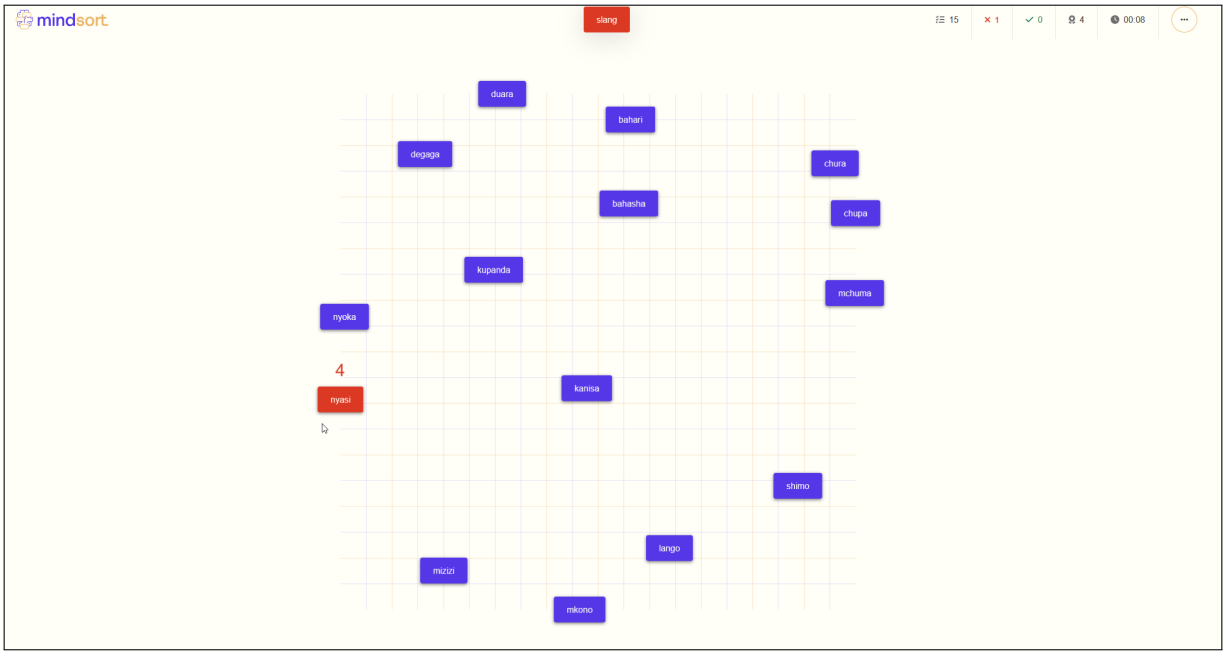

Figure 1: Example of game in MindSort. At the top a target word is shown and feedback is provided once an answer has been selected. The top-right displays the amount of incorrect and correct guesses as well as the points earned so far and time spend. In this example the user selected the wrong answer and got 4 points based on the distance to the correct word.

The map is constructed using multi-dimensional scaling (MDS) on all the distances between words. MDS is a method to visualize pairwise relations, in this case distances, between different data points (Douglas Carroll & Arabie, 1998). MindSort makes use of metric MDS which aims to minimize the so-called stress. This stress measures the squared differences between the ideal distances of the new data and the actual distances of the original data (Kruskal, 1964).

MindSort contains different modes for learning words. In all modes the 2D map of words is used, but the methods of interaction differ. In the MapToFlip mode users get to test themselves on words. They can click on a word to reveal the correct translation and decide for themselves if it was recalled well enough. In the ItemToMap mode users saw a target word at the top of their screen in their L1 language, they are then tasked to select the correct translation in the L2 language. This was the mode used for this experiment. Finally, there is the MapToType mode in which users get a target word and they have to type the correct translation.

Software changes

During the experiment participants saw a map of words as can be seen in figure 1. The map contained all the words of the set they needed to learn in the L2 language (in this case Swahili). At the top-middle of the screen participant saw a target word in the L1 language (in this case Dutch). In the bar on the top-right of the screen they could see the amount of incorrect and correct attempts as well their score and the time they had spend on the learning task.

Adjustments to the original software have been made in order to test the hypothesis. When selecting an answer users receive points based on the distance between the correct word and the selected word based on the condition. These points are added to the total score. Depending on the conditions the points are treated differently. In both conditions points are calculated as a fraction of the distance between the selected answer and the correct answer ranging from 0 to 10. The formula for this is as follows: $(1 - distance) * 10 = score$. In the absolute score conditions points are always added regardless of how many points were earned in the previous attempts for that word. In the relative score condition points are related to the previously given answer. If the current answer has a greater distance than the previous answer (so it is further away) than the difference in points will be deducted. If the distance is closer than the previous answer the difference in points will be added instead.

For example, in the previous round 5 points

were scored for the word *house*, this round only 3 points were scored as the distance was further away. Therefore the score will be adjusted with -2 points. In the absolute condition this would have added 3 points on top of the already gained 5 points.

Procedure

During the learning task participants looked at a screen with a map of words in Swahili which are part of the learning list as can be seen in figure 1. At the top of the screen a target word was shown in Dutch. Participants were instructed to select the correct translation of the Dutch word to Swahili from the map. Once an answer had been selected they would receive feedback as the selection being either correct or incorrect, and points based on the distance from the target word and the condition. Participants got to see each word seven times, regardless of whether they had it correct or not. Once all words had been shown seven times, users would see a screen with all the words and how often they correctly identified the translation. After this they were done with the learning task and would move on to the post-test.

After each learning session participants were tested on their knowledge of the vocabulary. During the test participants were randomly shown a pair of Dutch and Swahili word from the list they learned. Participants were tasked to identify if the pair contained the correct translation or not. This was done until all words from the test set were shown. No feedback was given during this part of the experiment. Once participants had completed the vocabulary test they would repeat the experiment again, but with the other scoring system and list.

Three days after participants completed the first part of the experiment, they received a retention test in order to test delayed recognition. This test was the same as the initial post-test except the order in which the lists were tested was randomized as well as the order of word pairs within the lists. This retention test was done in order to measure how well the different learning methods perform at retaining information over a longer period of time. Three days was chosen for practical reasons to keep the data collection period as short as possible. However, this three-day period can also be reasoned for based on the Ebbinghaus' forgetting curve (Murre & Dros, 2015). The Ebbinghaus' forgetting curve describes how information is forgotten over time after the initial learning phase. The curve suggests that after three days about 50% of the information learned has been retained.

Analysis

In order to test the hypothesis, performance was measured by the amount of correct and incorrect translations in both the immediate post-test, as well as the delayed recognition test after three days. An ANOVA analysis was performed to examine the effects of the conditions on the average scores of participants (Miller Jr, 1997). Before the the ANOVA analysis was performed the data was examined to see if it conforms to a normal distribution and any possible outliers were removed.

In addition to test scores, the participants distances to the target throughout the learning tasks were collected. This was used to examine how the different scoring systems influence learning progress. Finally, the survey measured the enjoyment of the two learning conditions on a five-point Likert scale, and a free form for participants to provide their opinion on the feedback methods.

Results

Descriptive statistics

Participants completed the task in 32 minutes on average (SD = 9.4). Three participants spend over an hour on the experiment, which seems to indicate they left the experiment and later returned. They were not considered for the average time statistic, but were included in the performance analysis as their scores did not deviate from the rest of the group. Participants correctly identified $\sim 83\%$ (SD = 10%) of the word pairs during the tests on average, regardless of which list or condition was associated. The lowest score by an individual part was 62% and highest score 100% on the immediate post-test and 58% and 100% on the three-day retention test.

As highlighted in figure 3 participants improved their guesses with each attempt in both conditions. An important factor in this is that it also includes those who found the correct answer and thus had a distance of zero. This means that perhaps the progress did not improve as gradual as is suggested by the graph, but rather that a larger proportion of participants selected the correct answer.

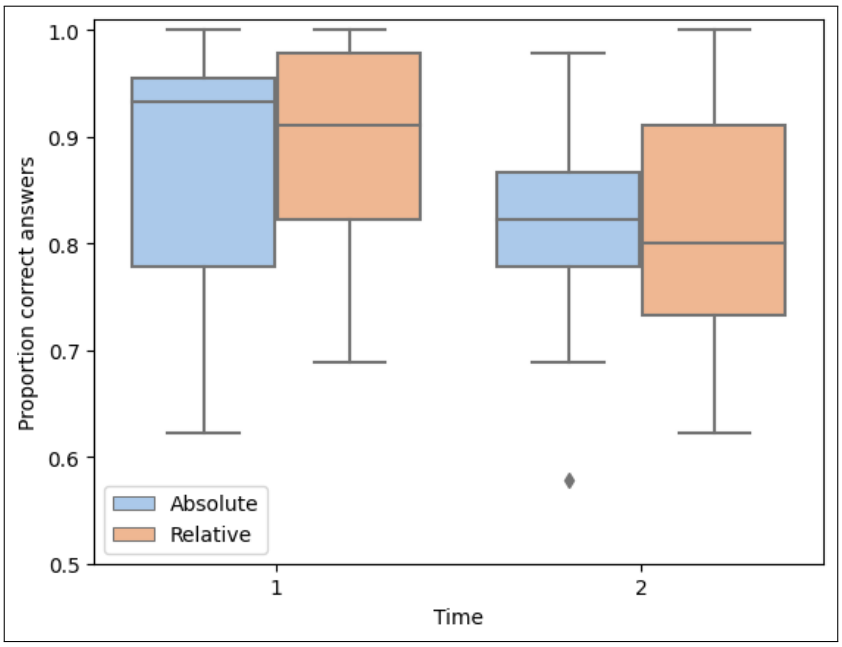

Figure 2: Comparison of results. The left side shows the scores of the absolute condition with the two lists. The right side displays the scores of the relative condition

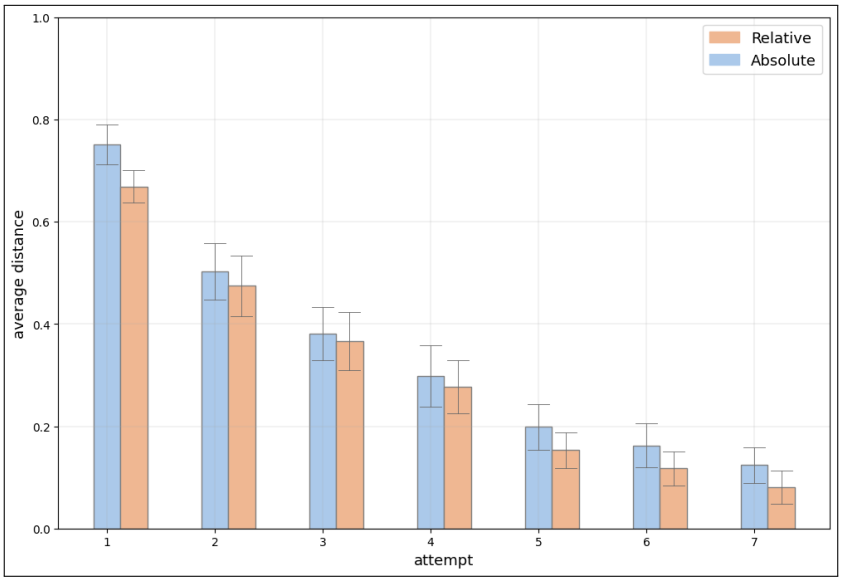

Figure 3: Comparison of average normalized Levenshtein distance over all words per attempt

Test performance

The hypothesis was that relative feedback increases test performance. Based on figure 2 the relative feedback resulted in a slightly better performance. An ANOVA anlysis was performed to test results using the following model as it provided the best prediction based on the Akaike information criterion (AICc):

$$correct \sim condition * time + list$$

The lists had no influence on test results ($F(1) = 0.0001$, $p = 0.990$). Moreover, a small interaction effect was found between the condition and time of the post-test ($F(2) = 3.209$, $p = 0.047$). Finally, no significant difference was found between the conditions ($F(1)=0.0237$, $p = 0.628$) and scores.

Feedback enjoyment

Participants graded both forms of feedback on a Likert scale from 1 to 5. Participants reported a

mean enjoyment of 3.235 (SD=0.876) for the absolute condition and a mean enjoyment of 2.764 (SD=0.806) for the relative condition. Using an independent t-test a significance of $p = 0.123$ was found.

Based on a free text feedback asking participants about the different forms of feedback we can see that participants enjoyed the absolute condition more as they were not punished for their mistakes unlike in the relative condition. Many also noted that they felt there was little effect of the points on the learning experience.

Discussion

The main goal of this study was to determine whether there was a difference between absolute versus relative feedback for performance and motivation. I hypothesized that relative feedback would be more motivating as it supplies a greater sense of progress in accordance with the learning progress hypothesis of Oudeyer et al. (2016), which could lead to a greater performance.

The study did not succeed in finding an effect of feedback forms on vocabulary retention or motivation. There are several possible reasons the experiment failed to find an effect. Most importantly, each list only contained 15 words for participants to learn. Combined with the recognition test used to test the performance there was a possible skill ceiling. This means that the task was too easy for participants and thusly any effects, if present, were not able to be distinguished.

However, some small surrounding effects were found that could prove useful for future research which will be explored in the coming sections.

Effects of point rewards on performance

No effect was found of the effect of points on the performance of vocabulary recognition. The relative feedback form resulted in a marginally better performance, but there is no significant effect. Future research could look to expand on the tests used to also test for recall of the vocabulary as well as having a larger set of words to learn in order to find possibly significant effects.

Effects of point rewards on motivation

Participants were generally positive of the use of points as feedback. They described the relative condition less motivating as you could lose points for making mistakes, although there was no significant difference between the two. An important question regarding motivation is how it changes over time. Therefore, follow-up work could examine the long-term effects of different similarity-based feedback forms for learning motivation.

Gamification of education

The use of point rewards in education has been studied previously with moderately positive results (Dehghanzadeh et al., 2019). The addition of gradual feedback could widen the applicability of this process. Moreover, the addition of different scoring methods allows for personalization to the students characteristics as well as the educational goals.

Limitations and further research

The foremost limitation of this study is the limited sample size of the study. The sample size made it difficult to properly address the research question in a meaningful way. On top of that, the limited time participants had to practice the words provided did also not benefit the results. Participants only partook in two learning sessions of 10 minutes each. As evident from the data analysis, there was a possible skill ceiling as participants correctly identified over 80% of the pairs in both tests. Future research could track these learning methods on a larger scale, including more sessions and a greater learning dataset.

In addition, the experiment could benefit from being in a controlled setting. Participants were free to do the experiment at a moment and place of their own convenience without a possibility to track their attention to task. Although, attention was paid to particular long learning sessions that could indicate distraction from the task, this is only from inference and does not offer direct evidence for how well attention was paid.

Additionally, there was barely a significant difference in the results between the immediate post-test and the retention test. As was discussed in the procedure, according to the Ebbinghaus' forgetting around 50% of the vocabulary should have been retained. I believe this is possibly due to the

test being a recognition test rather than recall. This sort of test is inherently easier, but might be subject to a different rate of forgetting than that of a recall test (Barclay, 2021).

Finally, as noted by the participants themselves the presentation of the feedback could be improved. Participants had too little time to realize that the correct answer was displayed after their attempt. Moreover, the points given were also shown for a relatively short time and the total progress was not displayed clearly enough leading to participants not paying much attention to them. Future research could specifically look at what forms of presentation have the largest effect.

Conclusion

This study examined the effects of an absolute versus relative point reward feedback on vocabulary acquisition. Although no significant effects were found the results do indicate a possible use for different forms of feedback for personalization of education. Absolute feedback was experienced as potentially more enjoyable than relative feedback possibly because of negative feedback in the relative group. Therefore future research is needed to see whether there are effects of distance-based feedback on motivation and performance on the long term.

Ethics statement

The study was reviewed and approved by Ethics Committee Social Sciences (ECSS) of the Radboud University Nijmegen through the light track. Written informed consent to participate in this study was provided by the participants.

Acknowledgements

I would like to thank my supervisor Frank Leoné for continuing to believe in me during the long duration of my thesis, as well as for giving valuable feedback.

References

Aitchison, J. (2012). *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons.

Ansteeg, L., Leoné, F., & Dijkstra, T. (2023). *Encouraging errors through gradual feedback to improve vocabulary learning (unpublished), donders institute, radboud university*.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, *14*(6), 575–589.

Barclay, S. C. (2021). *Examining the learning burden and decay of second language vocabulary knowledge* [Doctoral dissertation, UCL (University College London)].

Baxter, P., Droop, M., van den Hurk, M., Bekkering, H., Dijkstra, T., & Leoné, F. (2021). Contrasting similar words facilitates second language vocabulary learning in children by sharpening lexical representations. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.688160

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental psychology*.

Burgers, C., Eden, A., van Engelenburg, M. D., & Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, *48*, 94–103.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Cyr, A. A., & Anderson, N. D. (2018). Learning from your mistakes does it matter if youre out in left foot, i mean field? *Memory*, *26*, 1281–1290. https://doi.org/10.1080/09658211.2018.1464189

Darling, S., Havelka, J., Allen, R. J., Bunyan, E., & Flornes, L. (2020). Visuospatial bootstrapping spatialized displays enhance digit and nonword sequence learning. *Annals of the New York Academy of Sciences*, *1477*, 100–112. https://doi.org/10.1111/nyas.14429

De Groot, A. M. (1983). The range of automatic spreading activation in word priming. *Journal of verbal learning and verbal behavior*, *22*(4), 417–436.

Dehghanzadeh, H., Fardanesh, H., Hatami, J., Talaee, E., & Noroozi, O. (2019). Using gamification to support learning english

as a second language a systematic review. *Computer Assisted Language Learning*, *34*, 934–957. https : / / doi . org / 10 . 1080 / 09588221.2019.1648298

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. https://doi.org/10.1016/j.jml.2009.12.003

Douglas Carroll, J., & Arabie, P. (1998). Chapter 3 - multidimensional scaling. In M. H. Birnbaum (Ed.), *Measurement, judgment and decision making* (pp. 179–250). Academic Press. https://doi.org/https://doi.org/10.1016/B978-012099975-0.50005-1

Fliessbach, K., Weis, S., Klaver, P., Elger, C. E., & Weber, B. (2006). The effect of word concreteness on recognition memory. *NeuroImage*, *32*, 1413–1421. https://doi.org/10.1016/J.NEUROIMAGE.2006.06.007

Gallien, T., & Oomen-Early, J. (2008). Personalized versus collective instructor feedback in the online courseroom: Does type of feedback affect student satisfaction, academic performance and perceived connectedness with the instructor? *International Journal on E-learning*, *7*(3), 463–476.

Goertz, R., Leoné, F., de Groot, R., & Bekkering, H. (2023). The effect of a similarity structure on autonomous second language word learning (unpublished), donders institute, radboud university.

Keuning, H., Jeuring, J., & Heeren, B. (2016). Towards a systematic review of automated feedback generation for programming exercises. *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*, 41–46.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399–436. https://doi.org/10.1111/J.0023-8333.2004.00260.X

Leoné, F. T. M. (2023). Computational modeling as a prescriptive bridge between neuroscience and education (unpublished) donders institute, radboud university.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, *10*(8), 707–710.

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Metcalfe, J., & Finn, B. (2012). Hypercorrection of high confidence errors in children. *Learning and Instruction*, *22*(4), 253–261.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in childrens and adults vocabulary learning. *Memory & cognition*, *37*(8), 1077–1087.

Miller Jr, R. G. (1997). *Beyond anova: Basics of applied statistics*. CRC press.

Murre, J. M., & Dros, J. (2015). Replication and analysis of ebbinghaus forgetting curve. *PloS one*, *10*(7), e0120644.

Ouden, H. E. D., Kok, P., & de Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*, 548. https://doi.org/10.3389/FPSYG.2012.00548

Oudeyer, P.-Y., Gottlieb, J., & Lopes, M. (2016). Intrinsic motivation, curiosity, and learning. *Progress in Brain Research*, *229*, 257–284. https://doi.org/10.1016/bs.pbr.2016.05.005

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. *Precursors of functional literacy*, *11*, 67–86.

Potts, R., Davies, G., & Shanks, D. R. (2019). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 1023–1041. https://doi.org/10.1037/xlm0000637

Scheper, I., de Bruijn, E. R., Bertens, D., Kessels, R. P., & Brazil, I. A. (2019). The impact of error frequency on errorless and errorful learning of object locations using a novel paradigm. *Memory*, *27*, 1371–1380. https://doi.org/10.1080/09658211.2019.1661493

Stahl, S. A., Burdge, J. L., Machuga, M. B., & Stecyk, S. (1992). The effects of semantic grouping on learning word meanings. *Reading Psychology: An International Quarterly*, *13*(1), 19–35.

Warmington, M., Hitch, G. J., & Gathercole, S. E. (2013). Improving word learning in children using an errorless technique. *Journal of Experimental Child Psychology*, *114*,

456–465. https://doi.org/10.1016/J.JECP.2012.10.007

Zhuang, Y., Feng, W., & Liao, Y. (2017). Want more? learn less: Motivation affects adolescents learning from negative feedback. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/FPSYG.2017.00076/ABSTRACT