Blockworld: A study of task-oriented language evolution

Theo Higham

4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2021

Abstract

Ever since the first primitive human languages were created, they have steadily evolved and mutated over time, branching out into the thousands of modern languages we see today.

In order to model the creation and evolution of new languages, we have developed the online web game Blockworld for experimental use, based on the original code base by Omar Abarca Arriaga.

This report details development of the web game to facilitate its use for gathering research, followed by several experiments with participants using Blockworld, and an analysis of their results.

The experiments address the issue of how a new language can most effectively communicate a message to a user, in order to aid them in the completion of set tasks. The results will help us to better understand how language adapts to suit the environment for which it is used.

Acknowledgements

I would like to thank my supervisor Chris Lucas for offering continuous support throughout the project, and helping with all of my questions. Thank you to my family and friends for supporting me during the project. Finally thank you to all of the participants who took the time to produce results for the experiment.

Table of Contents

Chapter 1

Introduction

In order to carry out a task, a participant requires information. This information may come in the form of a message in a communicative language from the task giver. To succeed at the task, the user must associate the content of the message with the task giver's intended meaning. This is required in order to understand the context of the task, objects involved, and criteria for its completion. In linguistics, a language task may be defined as consisting of four main parts: A focus on practical meaning, a guiding language resource for the participant to utilise, a 'gap' i.e. a mental leap that the participant must make, and finally a clearly defined outcome [1]. To complete the task successfully, the participant must correctly use the information in the language resource provided, in order to bridge their mental gap in understanding regarding the new and unknown task environment. In the case of the Blockworld experiments presented here, participants must use logic to bridge a reasoning gap in order to deduce the relation of the language's words to the block puzzles that are presented to them. This allows us to study how different ways of presenting the task instructions of the language to the participant can affect their ability to bridge the reasoning gap and successfully correlate linguistic understanding with task performance.

1.1 Motivation

Language learning games have become increasingly popular in recent years. For example, in 2020 the language learning game Duolingo became the most downloaded education app worldwide [8].

With the increased interest in gamifying language learning comes the question: What method of presenting a new language to a user leads to their greatest success in the game?

Furthermore, many languages are adapted to suit the environment in which they are used, and the tasks that come with living in such an environment. One example is the Amazonian Pirahã language, which features no notions of the complex abstract principles commonly used in urban society, such as counting, naming colours, or past and future tense [17]. Rather, the Pirahã language is better adapted for the tasks of hunt-

ing and survival in the Amazonian jungle, since it can be quickly expressed through whistling [9]. Cases such as this of language evolving to adapt to a unique set of tasks raises the question, in what way do new features emerge that support its usefulness in the relevant context? This report aims to answer these questions through the process of experimentation.

1.2 Blockworld

Blockworld is an online single player web game that we have developed for experimental use based on the initial codebase by Omar Abarca Arriaga. In the game, users are presented with a series of tasks, which involve the placing and arrangement of blocks. For each task, an instruction is provided in a new language, which is unknown to the user. The user must then utilise the information given by the instruction to attempt to complete the task. Later, the user is asked to provide their own instructions for specific tasks, using words from the language to explain how to complete it. These instructions may then be passed on to the next participant in order to form the next generation of the language by evolution.

1.3 Objectives

This experiment aims to explore the following questions:

1. How is task performance affected when participants are given language learning tasks in a curriculum order, vs in a random order?
2. How is language comprehension affected when participants are given language learning tasks in a curriculum order, vs in a random order?
3. As new instructions are passed on to evolve the language, how does their semantic meaning change?
4. How does the generational evolution of a task-oriented language affect the length of its instructions?

1.4 Hypotheses

Listed below are the hypotheses for the experiment:

1. Participants from the Curriculum Task Order group will have a higher average task success rate than those from the random order group.
2. Furthermore, the participants from the Curriculum Task Order group will take fewer attempts per task, and will require less time to complete each task, compared to the random order group.
3. The semantic meaning of words added to the language will change from those of the previous generation.

4. As the language evolves over multiple generations the instruction length will decrease.

Chapter 2

Background

2.1 Curriculum Learning

Humans are known to learn concepts far better when the teaching examples presented to them are in a meaningful order rather than a random one [13]. A meaningful order for optimal learning would introduce concepts gradually, slowly progressing from the most simple to the most complex concepts. Furthermore, using a curriculum appropriate to the tasks at hand is key, because an incorrectly suited curriculum could hamper rather than aid a participant's learning, for example by confusing them with incorrect or contradictory information. When learning from a well made curriculum, we would expect a participant to have a faster speed of convergence when it comes to their understanding of the topic at hand, allowing them to more quickly adapt to the demands of the scenario.

There are several types of curriculum for human learning, each with its own specialised goal [13]. A knowledge-based curriculum aims to maximise the quantity of information that the participant will learn. A skill-oriented curriculum will focus on improving the participant's ability to complete a particular task. A creative curriculum will teach the participant how to use concepts from the subject area freely, so that they can then go on to apply their own ideas to the subject and make new creations. Finally, a thematic curriculum will provide structural context to learning by linking each piece of content to a relevant theme.

2.2 Language Evolution

Language is known to evolve over generations of use [5]. When a parent passes down language to their child, often the passed down language includes mutations which make it differ from the original. These mutations can be produced either by error, local dialect, or as a way to improve the language's efficiency or descriptiveness for the context in which it is used. [10].

Because languages are transmitted in this way, they are required to be well suited for the process of learning and real-world usage, and such necessities are shown in the lan-

guage's structure [14]. The specifics of these necessities may may vary across different geographical, cultural, and demographic environments, and thus these non-linguistic environmental differences can systematically define the language's linguistic features. Furthermore, by taking a probabilistic analysis of the factors that influence language structure, insights can be made into the causes and limits of linguistic diversity [15].

2.3 Linguistic Analysis of Language

A language consists of a set of signs. Every sign has at least both an exponent: the object it refers to, and a meaning: the semantic notion undertaken by the object [12]. In a written language, a sign could represent any amount of text, although it most commonly refers to a single sentence. A sign can be further broken up into four parts: its phonology, morphology, syntax and semantics [12]. A sign's phonology refers to the patterns of sound that constitute its message, including the phonemes from which it is constituted. A phoneme is the smallest base unit of sound which provides enough information to differentiate one word from another, for example between the words 'tap' and 'tab', the letters 'p' and 'b' are phonemes. By combining phonemes together, more complex signs can be created, first by constructing words, and then by arranging these words into a sentence according to syntax. Syntax defines the arrangement of words in a language according to the language's rules, and two sentences with the same words but arranged differently can have completely different meanings [12]. The meaning of a sentence is known as its semantics and this refers to the message that the sentence creator is trying to convey to it's consumer.

When comparing a past version of the language with it's latest evolution, it is important to know out of all the possible changes, which changes are simply aesthetic or 'dialectic', and which changes are relevant to its semantic meaning. One way to compare versions of the language is to analyse a new word with a change of only one phoneme from its previous version, yet that has a different meaning, and this is known as a minimal pair [12]. By analysing minimal pairs in a language, we can see which changes had a real effect on its use and meaning.

2.4 Inference and Clustering

Humans are very good at finding regularities and patterns within data. A way in which they practice this skill is by putting similar objects together into groups. This technique is known as clustering [2]. Once they have formed clusters, a person can then make predictions about new stimuli based on similarity to a previously defined cluster. Upon encountering a new stimulus, such as a sentence in an unknown language, a human participant is most likely to first compare it to any similar examples that they understand, in order to simplify the learning process and narrow the scope of their search for understanding [2]. Therefore, a new sentence which is similar to an already learned sentence albeit with minor alterations will likely be much easier for a participant to understand, since they can 'bootstrap' its meaning based on their current understand-

ing. For this reason, it is rare for new language features to evolve suddenly or with large increments, since it takes time for the speaker to accumulate these changes into their preconceived clusters of understanding [7]. Given an event x with probability P, the effectiveness of a clustering model designed to predict the event's outcome can be measured by calculating its information content: $log(1/P(x))$ [2].

Furthermore, language clustering is very useful as an approximation tool for lossy compression, in order to quickly and concisely convey a message [2]. For example, there exist many large plants with roots, trunk and branches of wood, adorned with leaves, fruit or acorns, but if a messenger wanted to quickly refer to such a plant, they could simply call it a 'tree'. This way, the message receiver could use their clustering knowledge of trees to immediately gain an approximate understanding of the object in question, without needing to hear all of it's particular details.

2.5 Related Work

2.5.1 Curriculum Learning

Since the work of Elman [4], the concept of using a curriculum for training learning machines has been of interest to the machine learning community. Elman's experiments involved teaching a simple grammar to a recurrent network. It was found that in order for the grammatical structure to be successfully learned, at first a limited framework with restricted complexity must be taught, before gradually moving on to more complex resources [13].

While the use of a curriculum for machine learning is widely studied, there exists less research into the effects of curriculum learning on humans, in particular when it comes to learning skills for a specific task. Curriculums for human learning are widely utilised in schools and educational institutions, and are a vital part of the teaching process, in particular for teaching new and complex skills to children, such as reading and writing. This experiment aims to help better understand the consequences of utilising a curriculum when it comes to task-oriented language learning.

2.5.2 Language Evolution Experiments

An example of a similar language evolution experiment is the Language Evolution Simulator by fatiherikli, found at: https://fatiherikli.github.io/language-evolution-simulation/ . This project simulates three islands which constantly derives new words using small chances of mutation. When agents from different islands intersect, they can combine their language styles to create new derivations. The Blockworld experiments of this paper differ in that human participants are recruited in order to complete tasks and evolve the language.

Chapter 3

Development of Blockworld for Experimental Use

The original blockworld codebase developed by Omar Abarca Arriaga is an innovative proof of concept providing the framework for a web interface for conducting task-oriented language experiments. The codebase uses HTML, JavaScript and CSS to create an online web game that can be played in the browser.

In order to prepare the Blockworld webgame for experimental use, key updates were made to facilitate the input of real users, and additional functionality was added to suit the requirements of the project, in particular the capacity to evolve the language across several generations. Furthermore the project was updated to utilise the Laravel PHP framework, so that it could be hosted on a web server and accessed via a domain, and to have user results saved in a MySQL database for later analysis.

3.1 Prototyping

In order to accurately identify the areas which most needed improvement, a prototype run was first carried out with participants.

3.1.1 Findings

1. Participants would often over-think the tasks and spend too long trying to find the answer, using techniques such as trial and error. For this experiment, we wanted to focus on the aspect of how participants use instructions to complete tasks, rather than trying to 'game' the system.
2. Some participants would write down the translations of words to help them remember, such as 'wewo = orange'. This goes against the experiment objectives, since the game is supposed to test how well the user can learn and remember the language, without aids such as writing.

3.1.2 Changes To Address Findings

To address Finding 1, a time limit was added to the game in order to motivate players not to overthink their decisions, and to inspire natural answers, more akin to how language would be used in the real world.

To address Finding 2, a notice was added to the instructions politely asking participants not to write down translations of the language. Furthermore, this finding implied that users were struggling to learn the meanings of words, and therefore the game was not sufficiently teaching them. To address this, the tasks curriculum was edited to provide more of a focus on reinforcement learning, particularly in the earlier tasks, and this is issue is discussed further in section 4.3.1.

Chapter 4

Experiment Design

The Blockworld web game and experiment consists of two parts, the details of which are described below.

4.1 Part 1

In part 1, the player must complete 24 tasks. The task interface consists of the following features. On the left is a work-space containing blocks for the player to use. On the right is the game stage where the blocks can be arranged. At the top is an instruction, as shown in figure 4.1, explaining how the player should arrange the blocks to complete the task.

Blocks can be dragged from the work-space to any position in the game stage, where they can also be stacked on top of another block. Blocks can also be removed from the game stage and dragged into the work-space. The task is successfully completed if the arrangement of blocks on the game stage is equal to one of the predefined success

criteria. On the bottom is the number of attempts remaining. If the player has 5 unsuccessful attempts at the task, that task is marked as failed and they will move on to the next task. Figure 4.2 shows the status at the start of the second task in Part 1.

Figure 4.1: Instructions for Part 1

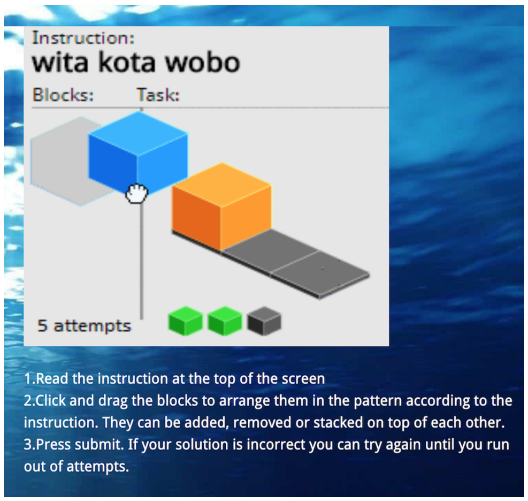
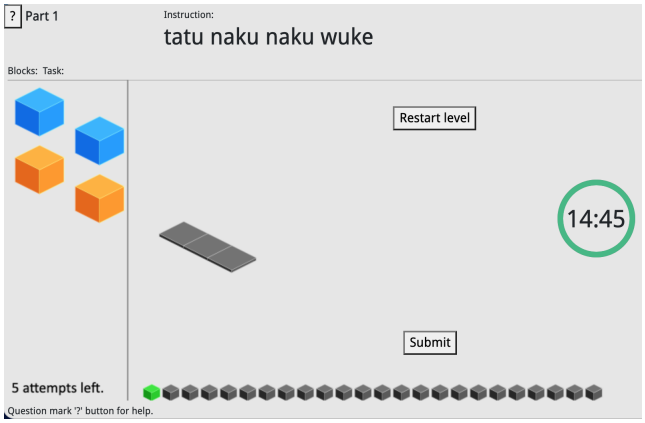


Figure 4.2: Example of a Part 1 Task

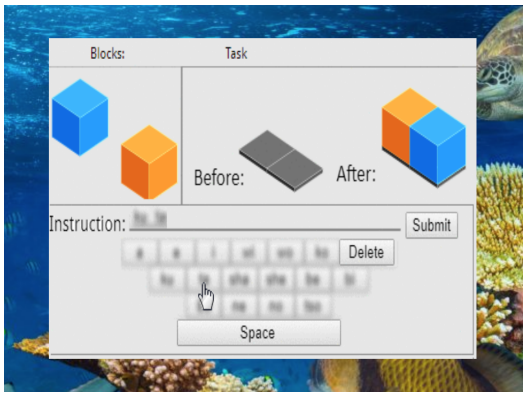


4.2 Part 2

In part 2 of the experiment, the user will now demonstrate their understanding of the language from part 1, and contribute towards the language's evolution.

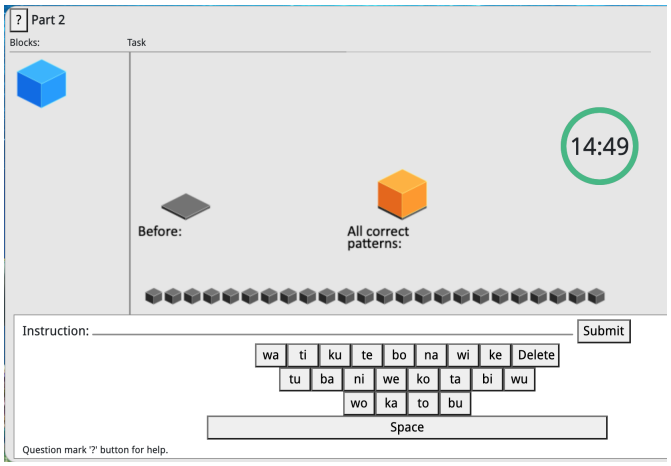
For each task, the user is shown the before and after states for a specific task, to illustrate how this task is successfully completed. See Figures 4.3 and 4.4.

Figure 4.3: Part 2 Instructional graphic



Then, the user must enter an instruction which can be used to guide another player to complete the task. It is expected that the user will try to replicate the instruction(s) that they saw in part 1, with some amount of variation based on how well they were able to learn and remember the language. It is also possible that users implement their own changes on purpose, as a way to update the instruction's grammar for easier understanding. Either way, these changes will be taken into account and used to evolve the language over time.

Figure 4.4: Example of a Part 2 Task.



4.3 Language Generation

The first generation of the language is created from 20 tokens, including a blank space, ' '. These are shown in Figure 4.5.

Figure 4.5: Tokens used to form syllables of the language.

```
,wa,ti,ku,te,bo,na,wi,ke,tu,ba,ni,we,ko,ta,bi,wu,wo,ka,to
```

To implement the initial language generation, words in new the language are associated with instructions in English, as shown in Figure 4.6. For this experiment, we need

words to describe all of the actions relevant to our experiment's functionality. This includes words for adding a block to the game stage, removing a block from the game stage, and stacking a block on top of another one. We also need words to differentiate between orange and blue blocks, then finally a utility word to represent all blocks and a word to represent the space between other words. Once

all of these words are generated, we will have a sufficient language to create instructions for the user.

Figure 4.6: Assignment of English words to 'token' numbers.


```
[
{"word":" ","token":"0"},
{"word":"add","token":"2,10"},
{"word":"remove","token":"15"},
{"word":"on","token":"19"},
{"word":"all","token":"12,1"},
{"word":"blue","token":"5,18"},
{"word":"orange","token":"9,10"}
]
```


Now, for each of these words, one to two unique 'token' number(s) from 0 to 20 are assigned.This number represents an index from the list of tokens shown earlier in Fig. 4.5. Excluding the space (' ') token which remains in position 0, the list of tokens shown in Fig. 4.5 is then randomised, and for each word a translation is produced, consisting of tokens at the specified indexes from the randomised token list.

For example, if the word 'add' from Fig. 4.6 were generated using the token ordering from Fig. 4.5, then its translation in the new language would be 'tiba'.

4.3.1 Curriculum Design

The curriculum was tested and designed in order to best introduce new concepts to the participant one at a time. Thus we utilised a skill-based curriculum, focusing on how to best teach the skill of interpreting concepts such as adding, removing and stacking blocks.

The first group of tasks consists of adding a single blue block. During early prototyping, it was found that many participants struggled to remember the difference between the names for the blue and the orange blocks. This also resulted in participants writing down translations, which does not keep in line with the experiment's goals. Therefore, the curriculum was edited to focus on reinforcement learning.

So to help participants gain confidence and learn the most basic words for the colour of the blue and orange blocks, we used the following curriculum, which has many simple

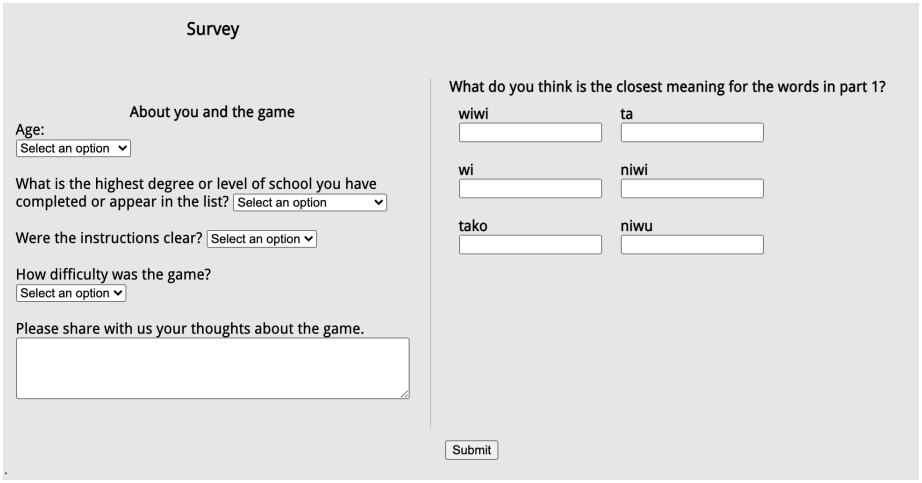
repeated tasks in the beginning of the experiment.

```
"add blue"
"add blue"
"add blue"
"add blue blue"
"add blue blue"


"add orange"
"add orange"
"add orange"
"add orange orange"
"add orange orange"


"add orange blue"
"add orange orange blue"


"add all"
"remove all"
"remove all"
"remove all blue"
"remove all orange add blue"


"add all blue"
"add all blue orange"
"add all orange add blue"


"add blue on orange"
"add blue add blue on blue"


"remove all blue remove orange add blue on orange"
"remove blue add all orange add all blue on orange"
```

After reinforcing the words for 'add', 'blue', and 'orange' to the user we move onto more sophisticated tasks which introduce new words.

4.4 Survey

Finally, at the end of the experiment the participant fills out a survey, as shown below. The first part of the survey is used to record basic non-identifying demographic information about the participant including their age bracket and education level, to provide further context and supplement analysis of their task performance.



The second part of the survey is important for better understanding participant's comprehension of the language, as it asks them to describe what they believe the word means. For a high performing participant, this information can provide an important distinction between whether their success was due to a comprehensive understanding of the language's intended meaning, or alternatively if they grasped the general meaning intuitively, or even if they had no idea what the words meant and their performance was simply a fluke. Results of participant's language comprehension from the experiments are analysed in section 6.2.

Chapter 5

Method

Two experiments were carried out, the first using a using a single language generation, so all participants received instructions with the same morphological structure, with the only difference being the particular tokens used.

In the second experiment, the language evolved between participants.

5.1 Task Ordering

In both experiments, the participants were split equally into two groups. They were not informed of the split, or which group they were in. The first group, were given tasks to complete in the order of a hand-crafted curriculum. The second group were given tasks in a completely random order.

5.2 Experiment 1

5.2.1 Participants

In this study, there were 20 participants. The participants were recruited voluntarily from friends and family. 70% of participants were university students between the ages of 18-24, and the remaining 30% of participants were between 55-64 years old and had a Masters or Doctorate degree. All participants were native English speakers. The participants all agreed to the consent form before partaking in the study.

5.3 Experiment 2

5.3.1 Participants

For this study, once again there were 20 participants. Participants were recruited using the Amazon Mechanical Turk platform. Each participant was compensated an average of £5. 75% of participants were between ages 25-34, 15% were 44-54, 10% were 18-24, and 5% were 65-74. All participants had strong fluency in English.

5.4 Conditions

For part 1 and part 2 of the experiment, the participant had a time limit of 20 minutes to complete all tasks. This amount of time was decided upon based on testing from the prototyping stage, it was found to provide enough time for the participant to reasonably complete the experiment without rushing, whilst also incentivising the participant not to overthink their decisions so that their results are more natural and intuitive.

Chapter 6

Results

Two experiments were carried out, firstly a Single Language Generation experiment, in which the instructions provided were generated beforehand, and remained consistent across all participants. Secondly, a Multi Generational Language experiment was carried out, in which each participant wrote instructions to describe the tasks. Then these instructions were passed on to the next participant to create the next language generation.

For each experiment, two unique groups of participants were tested on. The first group was given tasks according to a Curriculum Order. The second group was given tasks in a Random Order.

For each task, several data points were collected from the participants input. Below is a table describing all of these data points for disambiguation.

Data Point	Description
Result	Result equals 1 if the task was completed successfully, otherwise it equals 0.
Time	The amount of time that the participant took to complete the given task.
Attempts	The number of attempts that the user made at the task, before either succeeding or reaching the maximum of 5.
Task Attempts	All of the configurations of blocks that the participant submitted during their attempts, either successful or not.
Shown Instruction	The instruction that was shown to the participant when they attempted the task.
New Instruction	The instruction written by the participant to describe the task.

6.1 Single Generation Language Experiment

Here results from the first experiment are analysed, in which participants were given the same instructions created from a single generation.

Below are tables showing the mean, standard deviation and variance for each of the above numeric data points, totalled across all participants from their respective task order group. Time is shown in seconds.

Data Point	Mean	St. Dev	Variance
Result	0.78	0.42	0.17
Attempts	2.10	1.70	2.89
Time	13.82	11.56	133.53

Figure 6.1: Results of Curriculum task order group

Data Point	Mean	St. Dev	Variance
Result	0.60	0.49	0.24
Attempts	2.92	1.86	3.47
Time	17.70	14.97	224.10

Figure 6.2: Results of Random task order group

As we can see the Curriculum task order group had an overall task success rate of 78%, whereas the Random task order group had an overall task success rate of 60%. Figure 6.3 shows a bar chart of the mean task result for both groups, including error bars for standard deviation.

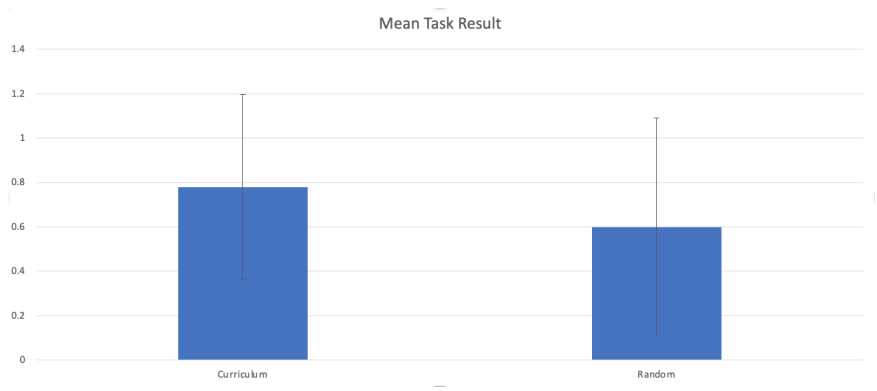


Figure 6.3: Bar chart of mean task result for both groups

6.1.1 T-test Analysis

In order to determine whether the difference in results witnessed between the two groups is due to the ordering of their tasks, or due to random coincidence or fluke,

we carry out a t-test. This test quantifies how unlikely it would be to receive these results, were it not for the underlying condition of task ordering. Thus in calculating the t-test probability we assume that the task ordering had no effect on the participant's results. Since the samples come from separate unique groups, we carry out an independent t-test. We assume a normal distribution amongst the samples. We did not assume that the variance across the groups would be equal, because in the random task order group the particular task ordering that a participant encounters may have an effect on the variance of their results.

The critical probability selected is p = 0.05. If the probability shown by the t-test is lower than this then we will reject the null hypothesis, otherwise it will be accepted.

We define the first null hypothesis H_n^1, that the mean task result of the Random order group will be higher than the mean task result of the Curriculum Order group.

A t-test for the Data point Result found a probability of 9.76E-06 that H_n^1 is true.

Therefore the null hypothesis H_n^1 can be rejected.

We define the second null hypothesis H_n^2, that the mean attempts of the Random order group will be lower than those of the Curriculum Order group.

A t-test for the Data point Attempts found a probability of 3.78E-07 that H_n^2 is true.

Therefore the null hypothesis H_n^2 can be rejected.

Finally, we define the third null hypothesis H_n^3, that the mean Time of the Random Order group will be lower than that of the Curriculum Order group.

Carrying out a t-test on the Data point Time found a probability of 0.0008 that H_n^3 is true.

Therefore the null hypothesis H_n^3 can be rejected.

In summary, the rejection of all three null hypotheses shows that there is a statistically significant difference between the results of participants from the Curriculum Order group, compared to those from the Random Order group. It was found that participants from the Curriculum Order group had a higher mean Result score, lower mean number of Attempts per task, and lower mean time to complete each task. And it was shown that these differences occurred as a result of the participant's task ordering group rather than by random chance, with 95% certainty.

6.1.2 Language Comprehension

In the final survey, participants write their interpreted meaning of each word in the language. Now we analyse how similar these interpretations are to the intended meanings of the words.

We aim to answer: How does language comprehension correlate with task score, and how does the ordering affect comprehension?

Below are two spider web diagrams, one for participants from the curriculum order group, and for those from the random order. Shown are the links between the intended meaning of the word 'add', and what participants thought that it meant. Next to each is the score of the participant who suggested that word. The symbol x^*2 shows that two participants chose the word x, in which case their scores were averaged.

Figure 6.4: Curriculum task order group

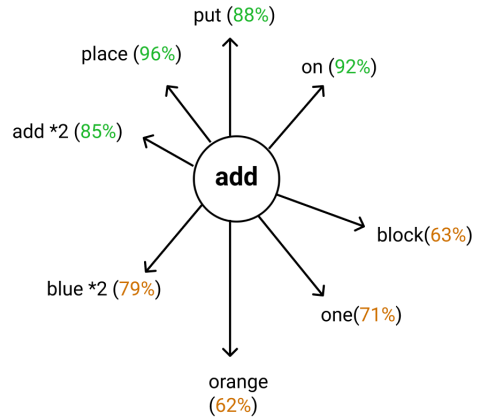
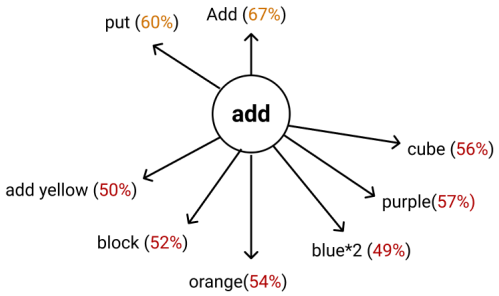


Figure 6.5: Random task order group



As the spider web diagram shows, participants with good language comprehension who understood the approximate meaning of the word add on average performed better than those who did not.

Furthermore, participants from the random order group had overall fewer connections to words with similar meanings to 'add', and their scores were lower.

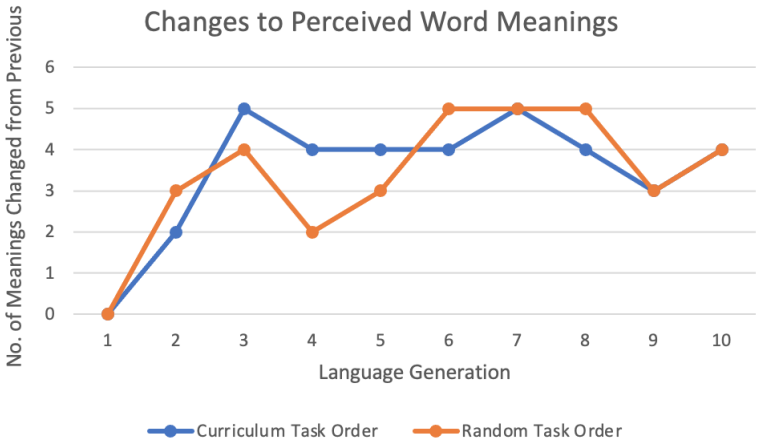
6.2 Multi Generational Language Experiment

Here results from the second experiment are analysed, in which participants receive instructions created by the participant who preceded them.

6.2.1 Changes in perceived word meaning

As the experiment progressed across multiple language generations and the words used to describe the environment changed, the meaning that participants associated with each word also changed. In the survey at the end of the experiment, we showed 5 words from the language to the participant and asked what they thought the meaning of each word was. If the previous participant gave a meaning to a word, then the next participant defined the same word with a different meaning then we count this as 1 change that has occurred. Figure 6.6 shows the number of such changes in perceived word meanings over each generation of the language.

Figure 6.6



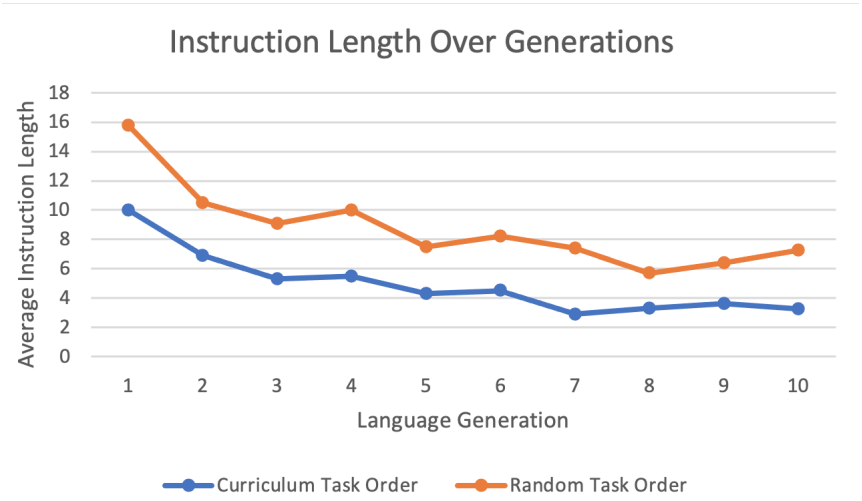
As we can see, the number of changes first started at 0 for the initial generation. Then, for both Task Order groups there was a medium increase to a moderate amount of changes for the second generation. Since these instructions were based on the originals, it makes sense that word meanings would remain fairly similar. After this, the Curriculum Order group saw a steep increase to the maximum value of 5 changes in the third generation, and the Random Task Order group saw a gradual increase to 4 changes for the third generation. At this point the Curriculum Task Order group mostly settled in the region between 4 to 5 changes. The Random Order group saw a dip to 2 changes in generation 4, then a gradual increase back to 5 changes in generation 6, at which point it also mostly settled within the 4 to 5 changes region.

Overall, the number of changes increased as generations passed, settling near the maximum number of possible changes. Since the participants had no moderation of their instructions, it makes sense that their instructions may become inconsistent and difficult to interpret by another reader. This also showed that around 10 generations of usage, the language was still volatile, with its meanings constantly changing, so perhaps with more generations the instructions would become more stable and uniform.

6.2.2 Changes in Instruction Length

Figure 6.8 shows how the Average Instruction Length (in tokens) changes with each passing generation of the language, for both the Random Task Order group and the Curriculum Task Order group.

Figure 6.7



As the language generations progressed, instruction average instruction lengths for both groups consistently decreased, until reaching a plateau around generation 8. The Random Task Order group maintained a higher instruction length, roughly 4-6 tokens longer throughout. One possible explanation for this finding is that participants from the Random Task Order group are not presented with the language concepts in a coherent manner, so their intuitive understanding of the language may be weaker than that of the Curriculum Task Order group. Therefore, they may need to use additional complexity in their instructions in order to fully explain the task. The overall decrease in instruction length over generations may be a sign that the language is becoming more efficient, however this could also be explained by other factors such as participants gradually exerting less effort when writing instructions, as the instructions that they see become less consistent over time.

Furthermore, the number of tokens in use by the language also saw a decrease. In the Curriculum Task Order group, the number of tokens in use decreased from 5 to 4 with generation 6. In the Random Task Order group, the number of tokens in use decreased

from 5 to 4 with generation 4. For both groups we saw a particular token being used less and less, until it became 'extinct' once a participant did not include it in their instructions.

Chapter 7

Summary and Conclusions

In this project, we further developed Blockworld, an online web game made from HTML, CSS and JavaScript. In the game, players are given instructions in a new and unknown language to aid them in the completion of block arrangement tasks. The game was upgraded to be used for task-oriented language learning experiments. We carried out two such experiments, in the first of which participants were given equal instructions created from a single generation. Then, in the second experiment each participant was given instructions created by the participant who preceded them. For both experiments, the participants were split equally into two groups, with one group being given the tasks in a predefined Curriculum Order, and the other group being given tasks in a Random Order. By analysing the results of these experiments, we aim to explore our Experiment Objectives and to prove or disprove our Hypotheses.

To analyse the effects of using a Curriculum Task Order vs a Random Task Order, several t-tests were carried out. The t-test analysis of 6.1.1 rejected all null hypotheses, thus confirming the Hypotheses 1 and 2. Hence it was shown that participants from the Curriculum Task Order group did have a higher average task success rate and took on average less time and fewer attempts to complete each task, when compared with those from the Random Task Order group. Regarding Objective 1, the experiment found that giving language learning tasks to participants in a curriculum order, rather than a random one, increased task performance. Therefore it is shown that the use of a context appropriate skills-based curriculum does provide a benefit to users when it comes to completing language learning tasks. Furthermore regarding Objective 2, the participants who had a Curriculum task Order showed higher rates of language comprehension when it came to understanding the word 'add' as shown in 6.1.12.

Regarding Objective 3, participant's perceived meanings of words in the language changed noticeably throughout all language generations, for both participant groups. A reason for this may be that since participants were allowed to completely define the instructions for the next language generation with no moderation, instructions did vary largely between generations. In particular, the instructions passed on to the next participant may have been inconsistent, requiring said next participant to make strategic

choices; for example between a) ensuring that their score was unlikely to be very low by rotating through plausible interpretations or b) aiming for a 'lucky' higher score by sticking with one interpretation. Hence, although the results could be construed as a sign that the meaning of the language evolved over time, without further interviewing participants as to the reasoning behind their instruction choices and intended meanings, Hypothesis 3 could not be conclusively tested, as it is not completely certain what the intended meanings of the newly introduced instructions are.

Regarding Objective 4, the experiment showed that as a language evolved over multiple generations, the length of its instructions decreased, thus confirming Hypothesis 4. For both the Curriculum Task Order and Random Task Order groups, it was shown that average instruction lengths decreased progressively over generations, until stagnating around generation 8. The fact that instruction length decreased may indicate that the language is becoming more efficient throughout the experiment, thus requiring less bits of information to convey its desired meaning. However, the question of whether the language really became more efficient is a complex and multi-faceted problem and cannot be conclusively determined in this report. For example, there may have been a feedback loop effect in that as generations progressed and participants were given shorter instructions, they then felt the need to write shorter, less complex instructions in return, either to match those that they witnessed, or to reduce the amount of effort expenditure required. Since the average instruction length of the language stagnated around the 8th generation, this could possibly be construed as a sign that the language was reaching its limit in terms of simplicity and efficient communication. It is also possible that participants felt less strongly compelled to write complete sets of instructions if they had been exposed to inconsistent information. This is another issue that would be best addressed through more comprehensive follow up questioning.

7.1 Future work

One interesting question to consider in future work is: "How do different types of curriculum affect performance in task-oriented language learning?" This experiment utilised a skills-based curriculum, however other types of curriculum, such as a thematic curriculum, are available, and may have benefits when it comes to performance in tasks of relevant context. For example, the study of the effects of using a thematic curriculum in language learning could be applied to the learning of language in the context of historical events, with new words and terminology introduced to describe each event. Furthermore, the reasons why a curriculum is effective for higher performance in task-oriented language learning could be further studied.

In this experiment, we analysed how the meaning of words changed over generations of task-oriented language use. Further work could analyse how these new meanings affect task performance, as well as which mutations prove useful in describing the tasks at hand.

References

[1] Rod Ellis. Task-based Language Learning and Teaching. *Oxford University Press.*2003.

[2] David J.C. MacKay. Information Theory, Inference, and Learning Algorithms. *Cambridge University Press*.2003.

[3] Mark Atkinson,a Kenny Smith,b Simon Kirbyb. Adult Learning and Language Simplification. *Cognitive Science 42 (2018) 2818–2854.*

[4] Jeffrey L.Elman. Learning and development in neural networks: the importance of starting small. *Cognition Volume 48, Issue 1*.1993

[5] Martin A. Nowak and David C. Krakauer. The evolution of language. *PNAS 96 (14) 8028-8033*.1999.

[6] Joshua B. Plotkin and Martin A. Nowak. Language Evolution and Information Theory. *J. theor. Biol. 205, 147-159*.2000.

[7] Luke Strongman. (2017). Language Evolution, Acquisition, Adaptation and Change. *Interdisciplinary Perspectives,IntechOpen, DOI: 10.5772/67767*.

[8] Cindy Blanco. 2020 Duolingo Language Report: Global Overview `https://blog.duolingo.com/global-language-report-2020`..

[9] G O'Neill.(2014). Humming, whistling, singing, and yelling in Pirahã context and channels of communication in FDG. *Pragmatics. Quarterly Publication of the International Pragmatics Association.*

[10] Amy Perfors, Daniel J. Navarro. (2014) Language Evolution Can Be Shaped by the Structure of the World. *Cognitive Science 38, 775–793*.

[11] Olga Fehér,Nikolaus Ritt,Kenny Smith. (2019) Asymmetric accommodation during interaction leads to the regularisation of linguistic variants. *Journal of Memory and Language 109, 104036*.

[12] Marcus Kracht. (2008). Introduction to Linguistics. *Department of Linguistics, UCLA*

[13] Yoshua Bengio, Jérôme Lourador, Ronan Collobert, Jason Weston. (2009) . Curriculum Learning. *Proceedings of the 26 th International Conference on Machine Learning, Montreal, Canada, 2009.*

[14] Beckner et al. (2009) .Language Is a Complex Adaptive System: Position Paper. *Language Learning 59(s1):1-26*.

[15] Dale & Lupyan. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems 15(03n04)*.

[16] G. K. Zipf (1949). Human Behavior and the Principle of Least Effort. *Wesley Press, Oxford, UK*.

[17] Daniel L. Everett (2005). Cultural Constraints on Grammar and Cognition in Pirahã: Another Look at the Design Features of Human Language. *Current Anthropology*.

[18] Hahn, M., Degen, J., & Futrell, R. (2021, April 1). Modeling Word and Morpheme Order in Natural Language as an Efficient Trade-Off of Memory and Surprisal. *Psychological Review. Advance online publication*.