Toki Pona Language Models: Investigating Learning, Communication, and Translation with Limited Vocabulary

Adam McDaniel
Tickle College of Engineering
University of Tennessee, Knoxville
Knoxville, Tennessee
amcdan23@vols.utk.edu

***Abstract*—The aim of this research project was to investigate the potential of highly-regular, low-vocabulary constructed languages as a medium for language models to comprehend, translate, and communicate through text. Toki Pona is a constructed language with a vocabulary of 123 words; this means that sentences written in Toki Pona use a much larger percentage of the overall vocabulary in the dictionary compared to English, where many words are very rarely ever said. This should make Toki Pona more ideal medium of communication for a language model than English. Through an iterative process, multiple language models were trained to translate, comprehend, and produce Toki Pona content. To evaluate the effectiveness of our models, accuracy metrics of next-token predictions were measured each epoch, and translations were qualitatively assessed. The research findings provide insights into the adaptability of language models in learning and utilizing languages with reduced lexicons, shedding light on their potential in future language model research.**

***Index Terms*—toki pona, language model, deep learning**

CONTENTS

I. INTRODUCTION

Language models have recently transformed many aspects of the world. These models are usually trained to speak one or many natural languages which have evolved over millennia in order to be optimal for human communication. A medium of communication which is specially suited for the human brain, however, might not be optimal for a machine learning model. There are thousands upon thousands of useful English words with which models can understand their training tasks, but the distribution of the *frequencies* of these words is terrible: the 614th most common word, "describe", appears 159,521 times across the Corpus of Contemporary American English, while the 45,003th and 100,060th most common words "thawing" and "druggy" only appeared 115 and 89 times respectively[5]. The vast majority of the words that these language models are expected to learn usually appear very infrequently in the training data. Toki Pona is a constructed language with a lexicon of 123 words, created by Sonja Lang in 2001[2]. The grammar is regular, and there are no conjugations or articles. Ideas are expressed through the composition of frequently used root words; an argument is a "talk fight", a city is a "house territory", a musical instrument is an "entertaining sound tool". The name "Toki Pona" itself means "good talk" or "good language".

This research aims to investigate the potential of Toki Pona, or similarly related small constructed languages, as a medium of communication for language models. Are there benefits from training a language model with a significantly smaller lexicon? If Toki Pona is well suited for language models, then a minified 10,000 word variant of English could be possibly more suitable than standard

English. Additionally, this research offers insights into the potential of language models in scenarios where linguistic resources are scarce; there is little Toki Pona content on the internet in comparison to a natural language like English.

Toki Pona content was scraped from several locations on the internet and coalesced into a database suitable for continued future use. Several iterations of models were trained to learn Toki Pona from this data over the course of the experiment. The models were quantitatively analyzed with metrics gathered throughout the training epochs and qualitatively assessed by examining the models' caliber of translations and responses to questions.

II. Previous Work

A. Automatic Translation of Toki Pona and Other Constructed Languages

There have been limited attempts in the past to automatically translate from Toki Pona to English using transformer based models[7;11]. These attempts are mostly based in applying transfer learning to pretrained translation models for natural spoken languages, such as the Spanish-to-English OPUS-MT model[7;17]. Additionally, the Tatoeba.org dataset these models were trained on was much smaller and has since grown: the most recent Toki Pona language model found was only trained on 15,579 Toki Pona sentences[7].

There has also been related automatic translation work for other constructed languages with vastly larger vocabulary sizes than Toki Pona, such as Esperanto, one of the most widely used and studied constructed languages[8;9;14]. Esperanto is much more complex than Toki Pona: its dictionary is at least 56.91 times larger (7000 root words alone in Esperanto compared to 123 total words in Toki Pona nimi pu). Esperanto also has articles, noun phrases, noun cases, conjugations for tenses, and conjugations for different parts of speech, all of which Toki Pona lacks. These high level language features make Esperanto sentence structures more analogous to natural spoken languages; specifically, Esperanto can be almost word-for-word translated between Indo-European languages[8], which would be a significantly more difficult undertaking to implement for Toki Pona. In fact, most of the automatic translation work that exists for Esperanto uses the technique of breaking apart the syntax tree of the source language input and substituting words with their equivalent translation according to its part of speech and a dictionary in the target language[8;9]. Due to the limited vocabulary, Toki Pona requires an understanding of context to translate effectively, making this technique much less feasible. A Toki Pona translator has to comprehend all of what has been said so far in order to translate the next sentence. Furthermore, Esperanto translation to and from Indo-European languages is simplified due to its fundamentally similar grammar structure which Toki Pona does not share: Esperanto has a synthetic grammar structure (common in Indo-European languages) rather than an analytical grammar structure (Toki Pona; common in Asian languages)[10]. The relationships between words in synthetic grammar structures are mostly governed by inflection and word morphology rather than word order and grammatical particles[10]. This gap between Indo-European languages and Toki Pona makes it even more difficult to apply other existing techniques that are used for Esperanto.

B. Toki Pona NLP Analysis

Little work has been done on performing NLP analysis on Toki Pona text, although interesting metrics have been observed. Toki Pona, like natural spoken languages, displays the Zipfian distribution in accordance with Zipf's Law[15]. This is unexpected due to the supposition that the vocabulary size would skew the distribution of word occurrences. This appears to be the most extensive and recent example of analysis on Toki Pona text; this subject is certainly ripe for future exploration.

C. Datasets

Several related models have previously been trained on the Tatoeba.org dataset[7;16], but there doesn't appear to be extensive research using a larger corpus of text, or texts larger than just sentence translations. Additionally, a large metacorpus has been collated together with links to smaller corpuses[6]. These corpuses consists of various types of unsorted blog posts and their contents: it appears that much of this content has been collected by Toka Pona enthusiasts to read and share. Any previously existing models trained on this vast amount of data were not found in the research for this paper; the models in this exploration appear to be the first.

III. Technical Approach

A. Pretrained Base Models

Two small to medium sized language models were chosen for the experiment: DistilGPT2 and GPT2. GPT2 is a language model produced by OpenAI with 117 million parameters[12]. A diagram of GPT2's architecture is depicted in Figure 1. DistilGPT2 is a model released by HuggingFace which is trained on GPT2 prompts and responses[13]; it is a slightly less capable copy of the original. DistilGPT2 is much lighter than GPT2 with only 82 million parameters, however; this allows for quick training[13]. This was advantageous, as it allowed for faster development and improvement of the prompting instructions before training the larger and more time consuming GPT2 models. Both of these base models were chosen primarily for their suitability to the hardware constraints set by the training environments available; future research should be done with these techniques applied to larger language models such as ChatGLM[18]. Multilingual language models trained in languages with analytical grammars might be best suited to generating Toki Pona content.

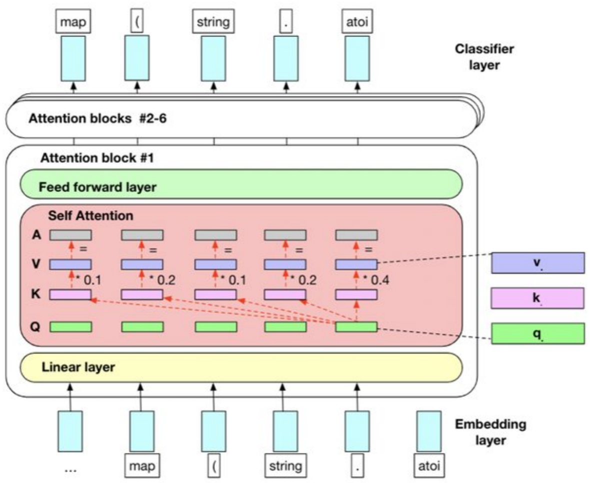

Fig. 1. Network architecture of the GPT2 language model.

B. Prompt Engineering

Multiple prompt engineering techniques were applied according to the type of model being finetuned. For the iterations of the experiment using the base DistilGPT2 model, a simple prompting format was used created to cue the language models for different tasks.

1) Initial Prompting Technique: Because the base models are not instruct models and are not designed to be prompted in any particular manner, tokens were chosen to indicate desired outcomes to the models. The list of tokens ”>>>>>” was chosen to denote translating source text into another target language, and ”=====” was chosen to signal to the model that the previous text is an instruction to be followed. Instructions are supplied in both Toki Pona and English for different prompts, and translation can be either to or from Toki Pona. Prompts were generated from the dataset with this technique and were used as the training data for the base DistilGPT2 models. Some example prompts are given in Table I.

2) Alpaca-Based Prompting Technique: Another, separate prompting technique was applied to the GPT2 models. It was hypothesized that using an instruction based model would be better suited to learning semantic information about Toki Pona; the model would already be pretrained to interpret the instruction portion of the prompt, now in Toki Pona, as a command, and the rest of the output as the desired result. A pretrained GPT2 model finetuned on GPT4 outputs from the Alpaca dataset was found and further trained with the Toki Pona datasets formatted using the Alpaca prompt template syntax. Examples of these prompts can be seen in Table III.

C. Training the Models

The models are trained in a series of successive steps. First, the prompts are created from the overarching Toki Pona dataset. These prompts are splintered into training and test datasets, fed into the tokenizer of the respective base model being trained, divided into sequences, and split into batches. The models were trained on the batches of tokenized sequences, and evaluated after each epoch against the validation data.

Toki Pona Text	Generated Prompt for Training
mi mute li ken ala ken lon ni?	mi mute li ken ala ken lon ni? >>>>> Can we live here?
	Can we live here? >>>>> mi mute li ken ala ken lon ni?
tenpo ni la sina ken ala ken kute e mi?	tenpo ni la sina ken ala ken kute e mi? >>>>> Can you hear me now?
	Can you hear me now? >>>>> tenpo ni la sina ken ala ken kute e mi?
jan mute li nasa lon tomo pali mi. ona li nasa tan ni : jan lawa pi mi mute li ike. taso mi ali li jan pali pi kulupu lawa pi ma Kanse. tan ni la jan lawa pi mi mute li ken ala weka e jan pali. ni li musi mute tawa mi. mi pilin pona tan ni : mi jan lawa ala. a a a.	o pali e lipu musi ===== jan mute li nasa lon tomo pali mi. ona li nasa tan ni : jan lawa pi mi mute li ike. taso mi ali li jan pali pi kulupu lawa pi ma Kanse. tan ni la jan lawa pi mi mute li ken ala weka e jan pali. ni li musi mute tawa mi. mi pilin pona tan ni : mi jan lawa ala. a a a.
	Write a story ===== jan mute li nasa lon tomo pali mi. ona li nasa tan ni : jan lawa pi mi mute li ike. taso mi ali li jan pali pi kulupu lawa pi ma Kanse. tan ni la jan lawa pi mi mute li ken ala weka e jan pali. ni li musi mute tawa mi. mi pilin pona tan ni : mi jan lawa ala. a a a.

TABLE I
EXAMPLE PROMPTS ENGINEERED TO TRAIN THE BASE DISTILGPT2 MODELS.

D. Iterations of Models for Experiment

An overview regarding each of the models trained for the experiment are shown in Table II; the iterations of models are described further in detail in the following sections of the paper. The datasets for each model and prompting techniques were chosen according to the issues displayed by the models at previous iterations in order to produce a better model. The first model displayed little comprehension of the text, so the next iteration was trained for longer and on a wider variety of prompts to help the model better associate words with their semantic meanings. The second model displayed sparks of comprehension of the texts, but the model would sometimes ignore prompts and generate random chains of text. To remedy this, a larger language model pretrained on the Alpaca dataset was found; if the model was pretrained to take the semantic meaning of the instructions and apply it in the response, maybe this would guide the model to better learn the semantic meaning of Toki Pona text used as instructions as well as in the response. This resulted in an unexpected, dramatic effect on the performance of the last two iterations of the experiment.

#	Base Model	Train — Test split		Description
1	DistilGPT2	256303	28479	This model was finetuned with instruction + response prompts, translation prompts, and to complete Toki Pona content.
2	DistilGPT2	264662	29407	This model uses techniques applied to the previous DistilGPT2 finetune, but also adds more complexity in training data by adding variations in phrases for each type of prompt.
3	GPT2	166650	18517	This model was trained on GPT4 output from the Alpaca prompt dataset, and then all the Toki Pona prompts were formatted with the same template. This model's training data is smaller because it does not include encyclopedia data.
4	GPT2	259420	28825	This model is based on the prior model, but it includes the encyclopedia data.

TABLE II
ITERATIONS OF MODELS TRAINED FOR THE EXPERIMENT, THEIR TRAIN — TEST SPLITS, AND VARIATIONS IN PROMPT DATASETS.

IV. DATASET AND IMPLEMENTATION

A. Dataset Collection, Sources, and Content Quality

The sona pona ("good knowledge") metacorpus contains many smaller corpuses which contain almost entirely unsorted data [6]. This corpus contains all kinds of content: translations of texts, stories, articles, poems, books, songs, and rants [6]. Much of the data is stored in file formats which are less convenient to extract text from (such as PDFs), and many of the works have graphical data which cannot be used for training and remove context from the text. The raw text content was collated from this list of corpuses and separated out according to genre by hand. Another large source of text content is lipu tenpo ("time document" in English), a regularly produced magazine for Toki Pona speakers [1]. These documents are also stored in PDF format and are also difficult to extract text from, but the content quality is very high and 19 magazines have

Toki Pona Text	English Text	Generated Prompt for Training
mi wile sona e nimi pi ijo ni.	I want to know what this is called.	Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: Translate the following English text into Toki Pona: I want to know what this is called. ### Response: mi wile sona e nimi pi ijo ni.
lipu kule li tan wile pi jan mute. lipu kule li tan pali pi jan mute. jan mute li lon nasin mute. taso jan pali ali li wile e ni: lipu ni o nasin pona. mi ken sitelen mute e nasin pona. taso mi sitelen e ni kepeken toki lili. o kama sona e nasin pona pi lipu kule. mi la, 'pona' li seme? - mi wile e ni: **lipu ni li pona tawa jan mute pona.** jan pona li lon nasin wan ala. (kin la, o kepeken nasin ike ala.) …	Books come from the desires of many people. Books come from the work of many people. Many people are on many paths. But everyone wants this: may this book be good. I can write many good things. But I write them using few words. Please learn the good ways of this book. To me, what does 'good' mean? - I want this book to be good for good people. Good people are not on one path. (Also, don't use bad ways.) …	Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: Translate the following English text into Toki Pona: Books come from the desires of many people. Books come from the work of many people. Many people are on many paths. But everyone wants this: may this book be good. I can write many good things. But I write them using few words. Please learn the good ways of this book. To me, what does 'good' mean? - I want this book to be good for good people. Good people are not on one path. (Also, don't use bad ways.) … ### Response: lipu kule li tan wile pi jan mute. lipu kule li tan pali pi jan mute. jan mute li lon nasin mute. taso jan pali ali li wile e ni: lipu ni o nasin pona. mi ken sitelen mute e nasin pona. taso mi sitelen e ni kepeken toki lili. o kama sona e nasin pona pi lipu kule. mi la, 'pona' li seme? - mi wile e ni: **lipu ni li pona tawa jan mute pona.** jan pona li lon nasin wan ala. (kin la, o kepeken nasin ike ala.) …

TABLE III
EXAMPLE PROMPTS ENGINEERED TO FURTHER TRAIN THE GPT2 ALPACA FINETUNED MODELS.

been produced thus far. `pdftotext` was used to parse the PDF files into text documents, but the parsing errors had to be repaired by hand due to the structure of the magazine content on the pages. Another Toki Pona magazine exists called lipu kule ("color document"), but it's content is stored in Markdown files which are much more convenient to access. Wikipesija is an online encyclopedia in the vein as Wikipedia, except that all of its content is in Toki Pona[4]. Wikipesija has an export feature, which allows users to download the page contents of a list of supplied page names. This was used to collect several tens of thousands of high quality sentences. There also exists a Toki Pona Bible translation project which has translated several books. Translations of the Gospel of John, Genesis, and multiple translations of the sermon on the mount (from the Gospel of Matthew) were collected to train the models. These texts are very high quality, and they have many equivalent alternative English translations. The Tatoeba.org dataset was also used in this project, and the Toki Pona sentence pairs have also increased since the last paper: 48,740 sentence pairs were collected for this dataset. Additionally, Chinese translations for all Toki Pona sentence pairs were scraped as well; since Toki Pona is an analytical language like Chinese, training with Toki Pona and Chinese translations might help a multilingual model learn the language easier.

B. *Dataset Organization*

The dataset is collected in a manner so that it can be used for future Toki Pona language model exploration. First, the items of the dataset are assigned labels which denote their kind of content. Information about each of the different content labels is shown in Table IV. The data is also grouped into four main formatted mini-datasets that are depicted in Table V. The separate groupings of the data according to content format and type help streamline the prompt engineering process: the detailed labels make it much simpler to design generic prompt formats to plug the datapoints into.

C. *Training the Models*

Each of the models were trained with a low learning rate of 0.00002 to only make small and incremental changes to the existing weights; the model should learn ideally Toki Pona by understanding it in terms of its existing understanding of English. Batch sizes of 1000 were used, and the first 500 steps of training were used as warmup steps to make the training smoother over the entire dataset; the model shouldn't change its weights too much towards a given minima at the outset. Google Colab was used to train the models with their standard T4 and premium V100 GPUs. The standard T4 was used for the DistilGPT2 models, which trained in 2 hours and 5 hours on the standard GPU respectively. With the premium V100 GPU, the GPT2 models took about 8 hours each to train.

Content Type Label	Description
article	Individual articles from magazines or blogs.
encyclopedia article	Content from 964 handpicked articles from Wikipesija using the "Export Pages" tool.
blog post	Content scraped from Toki Pona Livejournal blogs.
magazine	Entire magazines from lipu tenpo and lipu kule.
biblical text	Chapters of Toki Pona translations of the Bible, sermons, and the Gospel of John.
story	Miscellaneous stories told in Toki Pona, translations of fables, translation of The Little Prince.
poem	Poems and songs in Toki Pona.
screenplay	Monty Python and the Holy Grail translated into Toki Pona.
chat	Chat logs in Toki Pona.
other	Miscellaneous content, sentences from Tatoeba.org.

TABLE IV
CONTENT TYPES AND INFORMATION ABOUT THE DATA CATEGORIZED UNDER EACH LABEL.

Data Subset	Word Count	Description
documents	1,167,019 words	Entire documents in their completion in Toki Pona. This includes all articles, biblical texts, chats, encyclopedia articles, etc. in their entirety.
chapters	34,077 words	Numbered chapters and subsections of large documents with chapters or divided into scenes.
sentences	402,085 words	Miscellaneous Toki Pona (without translations).
sentence translations	662,611 words	Miscellaneous Toki Pona sentences and their translations in English and Chinese.

TABLE V
EACH MAJOR COLLECTION OF TOKI PONA DATA, THEIR SIZES, AND INFORMATION ABOUT THEIR DATA.

V. EXPERIMENTS AND RESULTS ANALYSIS

A. *Quantitative Results*

Tables VI, VII, VIII, and IX contain metrics taken from the development of each model over the training epochs. The second experiment seems to have the worst performance of all the models *upon first glance*, but this is due to the more complex prompts introduced between the first and second experiments; if the first experiment had been done with the prompts used for the second experiment, the loss would've been greater as well. The third experiment seems to be the most successful quantitatively, the fourth experiment is a close second but it was also trained on more complex data from encyclopedia articles.

Quantitatively, the first two iterations of the experiment (the two DistilGPT2, custom prompt models) did very poorly in comparison to the latter models. These models learned to generate grammatically correct Toki Pona text, but the generated text was not very relevant to the prompts.

The latter models performed much better; the perplexity and validation loss are significantly lower.

Epoch	Training Loss	Validation Loss	Perplexity
1	1.7747	1.6708	5.3164
2	1.6538	1.5588	4.7531
3	1.6185	1.5251	4.5956

TABLE VI
EXPERIMENT #1 (DISTILGPT2): TRAIN THE MODEL WITH THE FIRST DATASET.

Epoch	Training Loss	Validation Loss	Perplexity
1	1.9908	1.8937	6.6439
2	1.8501	1.7470	6.6439
3	1.7636	1.6663	5.29254
4	1.7040	1.6184	5.0450
5	1.6656	1.5890	4.8988
6	1.6331	1.5782	4.8462

TABLE VII
EXPERIMENT #2 (DISTILGPT2): TRAIN THE MODEL WITH MORE VARIATION IN THE PROMPTING DATA, AND WITH A LARGER NUMBER OF EPOCHS.

Epoch	Training Loss	Validation Loss	Perplexity
1	0.8469	0.7308	2.0767
2	0.7526	0.6435	1.9031
3	0.686	0.5936	1.8105
4	0.6331	0.5625	1.7551
5	0.6076	0.5435	1.7220
6	0.5832	0.5318	1.7019
7	0.5825	0.5278	1.6951

TABLE VIII
EXPERIMENT #3 (ALPACA PRETRAINED GPT2): TRAIN THE MODEL WITH ALPACA FORMATTED TOKI PONA PROMPTS.

Epoch	Training Loss	Validation Loss	Perplexity
1	1.0386	0.9873	2.6839
2	0.9368	0.8761	2.4015
3	0.8786	0.8070	2.2411
4	0.8313	0.7617	2.1419
5	0.7885	0.7323	2.0798
6	0.7734	0.7150	2.0441
7	0.7586	0.7089	2.0317

TABLE IX
EXPERIMENT #4 (ALPACA PRETRAINED GPT2): TRAIN THE MODEL WITH MORE COMPLEX TOKI PONA PROMPTS.

B. *Qualitative Results*

Upon interacting with the model from the first experiment, it became clear that the model was undertrained; this is readily apparent when viewing the changes in loss in the training history. The model should've been trained for more epochs, but interestingly the model makes little to no grammar mistakes and mixes in common noun phrases into different sentences. The first model doesn't seem to have a grasp of the actual semantic meaning of the prompts; when the model is asked to translate a small amount of text, it doesn't seem to understand the task and outputs semi-related Toki Pona content. Example outputs are depicted in Table X.

The other models, however, have much more interesting properties compared to the first model. The latter models did not suffer from undertraining, and the second model seems to have a better grasp of the meanings of phrases and their translations compared to the first. The model still seems to constantly diverge from the topic at hand, however; the answers to the prompt contain the relevant information at the beginning of the generated content, but then the model continues to produce random chains of grammatically correct phrases. The model still seems to have a very incomplete understanding of the semantic relationships between the prompt and the desired response, but overall it is an improvement over the first model trained. There are sparks of language comprehension with the second model. Example outputs are shown in Table VII.

After finding these results for the first and second models, the Alpaca based prompt engineering approach was utilized. Alpaca based prompting vastly improved the models' responses both quantitatively and qualitatively. In the examples shown in Table XII, this is the first instance where the question "What did Jesus say to Peter?" is directly answered in the response by the model. Additionally, the translations from the model are much more accurate, and the model shows an understanding of the task at hand. When the model hallucinates or confuses itself, it generates text that still makes sense in the context: it substitutes "pimeja ala (not black)" for the color "laso (blue)". The model is clearly comprehending the instructions and translating the text.

The final model shows the best results qualitatively, despite the worse measure of perplexity. This worse measure of perplexity is likely due to the fact that this model was trained on encyclopedia data; the more complex encyclopedia articles were harder to reproduce in the validation set, so the next-word prediction was worse. When asked the question "What did Jesus say to Peter?", not only does it answer the question like the third model, it also directly addresses the question by responding "Jesus says this to Peter: . . . ". Its translation skills are comparable to the third model, but its comprehension of the prompt text is an improvement.

VI. CONCLUSION

In conclusion, the results revealed that the initial models, trained with custom prompts, performed very poorly; the generated content was grammatically correct, but the language model refused to stay on topic. It was difficult to get the model to treat the tokens for "cues" with the proper behavior, even with more training. The subsequent iterations of the experiment performed astoundingly better quantitatively and qualitatively; sometimes the models would diverge from the main topic, but it was often after already clearly addressing the main request. Additionally, when the 3rd or 4th models "run off" from the prompt, it is typically in a manner that is very easy to tell and less confusing: the model doesn't generate random content, it

Prompt	Response	English Translation (With Translation Notes)
What did Jesus say to Peter? =====	jan sewi Jawe li toki e ni: ona li kama sona e toki Epelanto. jan sewi li sona ala e ona. ==kulupu jan == jan li ken ala ken toki kepeken toki Inli? ==lipu lawa== lipu pi lipu ni li lon lipu pi toki pona.	God says this: he is learning Esperanto. God doesn't know it. ==people community== Can people talk using English? ==rules== Documents of this book are in Toki Pona.
Reading books helps you learn! >>>>>	jan li kama sona e toki Epelanto lon toki pona lon tenpo pini. taso ona li pana e sona ni tawa mi: mi sona ala e ni: jan li ken ala ken toki kepeken toki-pona? mi ken kama jo e lipu sona pi toki Inli lon lipu . . .	People have learned Esperanto in Toki Pona in the past. But it gives me this knowledge: I don't know this: can people speak with Toki Pona? I can get an encyclopedia article in English at (document; unfinished expression indicating position in a document). . .
If the fish is red, i love it. If the fish is blue, I hate it. >>>>>	jan li ken ala ken toki kepeken toki Inli?Translate this to Toki Pona: jan Ton li kama sona e toki Kanse. — Tom learned French.o toki e ni tawa jan Ton: ona li lukin e sitelen pi lipu ni.Translate . . .	Can people speak using English? Translate this to Toki Pona: Tom is learning French — Tom learned French. Say this to Tom: he sees images of this document. Translate . . .

TABLE X
MODEL ITERATION #1 RESPONSES TO TEST PROMPTS NOT PRESENT IN THE DATASET

Prompt	Response	English Translation (With Translation Notes)
What did Jesus say to Peter? =====	jan sewi Jawe li toki e ni: 'jan lawa o, sina sona e mi anu seme?' jan Jesu li sona ala e ona. ona li pana e sona tawa jan pi lukin ala. jan mute mute li lon. jan sewi li luken. kipisi lipu nanpa luka luka	God (Yahweh) says this: "oh kings, do you know me?" Jesus does not know them. They teach blind people. Many many people are true. God sees (misspelled). Chapter #10
Reading books helps you learn! >>>>>	lipu mute li pana e sona tawa sina.sina wile ala wile moku e telo nasa mute?mi kama sona e toki Wensa. 1245 Joshua Birns-Sprague tenpo suno ni la mi lukin e sitelen tawa pi toki pona lon lipu ni. taso tenpo ni la . . .	Many books give knowledge to you. Do you want to drink a lot of alcohol? I'm learning Swedish. 1245 Joshua Birns-Sprague Today I saw a Toki Pona video on this site. But now, . . .
If the fish is red, i love it. If the fish is blue, I hate it. >>>>>	soweli li ike tawa mi tan seme? mi sona ala e ni: ona li pilin ike. taso soweli li pona ala tawa ona.sina ken ala ken kepeken e toki Kanse? 845 Sergey Tihonov toki . . .	Why do I dislike animals? I don't know this: it feels bad. But animals aren't good for them. Can you speak French? 845 Sergey Tihonov Hello. . .

TABLE XI
MODEL ITERATION #2 RESPONSES TO TEST PROMPTS NOT PRESENT IN THE DATASET

just hallucinates an additional prompt and generates text for that as well. It is very clear from interacting with the latter models that they are comprehending the text and trying to address questions and translate prompts. Even with the small amount of training data available from an estimated 3500 speakers, Toki Pona is very well suited as a viable medium of communication for future language models. Future research should be focused on collecting more data, and possibly finetuning larger language models with a better grasp on analytical languages to speak Toki Pona. Additionally, because Toki Pona is successful with such few tokens to train on, exploring minified versions of English as a medium of communication might be interesting as well. Reinforcement learning with human feedback would also be valuable for these models. There is a large Toki Pona Discord server where an experiment such as this could be implemented using bots to communicate with users and receive feedback through the message reactions.

APPENDIX

A. Workload Distribution

I (Adam McDaniel) implemented all of this project and all of the words on this report despite having three group-members in total: Joshua Byers, Sheikh Uddin, and me. I consistently tried to get both of the others involved and

Prompt	Response	English Translation (With Translation Notes)
Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: What did Jesus say to Peter? ### Response:	jan Jesu li toki e seme? jan kute ona li sona e ni: jan ni li jan Jesu. jan Jesutasu li toke e ni tawa jan Petulo: ‘jan lawa o, jan seme li pali ike tawa sina? jan ni, anu mama …	What does Jesus say? His followers know this: this person is Jesus. Jesus (misspelled) says this to Peter: ”oh kings, who does wrong to you? This person or …
Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: Translate the following English text into Toki Pona: Reading books helps you learn! ### Response:	jan li lukin e lipu la ona li ken kama sona e ni: lipu li pana sona tawa sina. ni li pona mute tawa ona. jan li sona ala e ona la ni li ike mute t…	If people read books, they can learn this: books give knowledge to you. This is very good for them (or it). If people don’t know it, this is very bad …
Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: Translate the following English text into Toki Pona: If the fish is red, I love it. If the fish is blue, I hate it. ### Response:	kala li loje, la mi olin e ona. kala li pimeja ala, la ona li ike mute tawa mi.Below is a text that English text is worth reading.	If the fish is red, I love it. If the fish is not black, I dislike it very much. Below is a text that English text is worth reading.

TABLE XII
MODEL ITERATION #3 RESPONSES TO TEST PROMPTS NOT PRESENT IN THE DATASET

Prompt	Response	English Translation (With Translation Notes)
Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: What did Jesus say to Peter? ### Response:	jan Jesu li toki e ni tawa jan Petulo: ‘jan Simon o, mije lili pi jan Jojane o, jan mute ni li olin e mi. sina olin mute ala mute e mi? tan seme la, sina sona e ni: mi jan pona sina …	Jesus says this to Peter: ”oh Simon (this character’s full name is Simon-Peter), the son of John, this crowd loves me. Do you love me? Why; you know this: I’m your good person …
Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: Translate the following English text into Toki Pona: Reading books helps you learn! ### Response:	jan li lukin e lipu la ona li kama sona e ni: lipu li pana e sona.Below is a translation of the English version of the song. tenpo pini la jan li ken ala lukine e …	If people read books, they can learn this: books give knowledge. Below is a translation of the English version of the song. In the past, people couldn’t see (misspelled)…
Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: Translate the following English text into Toki Pona: If the fish is red, I love it. If the fish is blue, I hate it. ### Response:	kala li loje, la mi olin e ona. kala li pimeja, la ona li ike.Below is a tentative version of the song.	If the fish is red, I love it. If the fish is black, I dislike it. Below is a tentative version of the song.

TABLE XIII
MODEL ITERATION #4 RESPONSES TO TEST PROMPTS NOT PRESENT IN THE DATASET

kept them up to date with all of my progress. I tried to give them assignments for things to do, but they have done very little or not at all. Joshua Byers has responded much better than Sheikh: Sheikh has yet to contact me back since April 18th through Google Chat or by text, but Joshua would occasionally meet with me once every couple of weeks or so after class for about 10 minutes. I assigned Joshua to setup the Google Cloud Compute machine for our project, but nothing ever came of it; every time I’ve asked him for updates there’s been no progress. I told him that I simply wanted him to run the evaluation program that I had already written and to supply it with some interesting English prompts so we could start collecting results for our paper; he told me he would do it, but did not follow through. When I asked him about his final project for COSC 527 the day before the projects were presented, he told me that he had not done it; once I learned this I continued with the project as if it would depend on me alone because I did not believe he would ever follow through. He contributed exactly one thing to the project in total: three weeks ago he made a 17 line Dockerfile which we did not use because he never finished setting up the cloud VM, and I used Google Colab Pro instead. I'm 99% certain both of the other group-mates are entirely unaware of how the code works or what this paper is about in detail; I sent them the shared link to edit this paper on May 8th but neither of them ever opened it. I did all of the lab assignments and the final project by myself this semester in addition to two other 500 level classes and a 20 hour teaching assistant-ship; it was definitely a challenge but I'm happy I survived! I just don't want my grade to suffer for solo work that should've been a team effort.

B. Code Design

There were three main coding portions of the project: preprocessing the collected data, training the language models on the data, and evaluating the models. A program was written to parse Wikipesija content to remove extraneous XML code that inhibited the models' performance when including the article data, such as "gallery" XML elements or long image hyperlinks. A Jupyter notebook was also written to take all of the raw content collected in the sorted folders and assign labels to all of the data and collect it into a unified format. The notebook goes through the data chunk by chunk: first by adding the individual sentence translation pairs to the dataset (as its own tab-separated-value file). Then, all of the documents are joined and labeled into one TSV file as well. Then, a select few of these documents are parsed for dividers like "Chapter" or "kipisi" in Toki Pona, and these numbered chapters are stored and labeled in the dataset separately as well. For each of these individual sections of the dataset, the notebook reloads the stored copy and verifies that it is exactly identical to the working copy in memory that generated the file.

The implementations for all of the models were trained using Google Colab Pro; it took 200 units of compute to train all of the models due to the VM instances timing out before the models would be fully trained, and due to the fluctuating pricing of the premium GPUs on the platform over the day as the notebooks kept needing to restart. The training notebooks work by downloading the overarching Toki Pona dataset stored on git, and then creating the dataset of prompts engineered for the respective model the notebook is for (with the `collate` function in each notebook for the models). This dataset, which is loaded using pandas, is then converted into a HuggingFace dataset object which can be plugged into the HuggingFace DistilGPT2 model and the other GPT2 pretained model with found on the platform as well. The learning rate, weight decay, batch size, and sequence length for models were used from their "Train your language model" documentation page, which uses common GPT2 tuning parameters[3].

To evaluate the models qualitatively, a program was written to conveniently start an instance of any of the trained models. The program allows the user to start the model with various configurations such as the maximum output length, but by default the program uses a fallback configuration. It's just a REPL loop for the user to prompt each of the iterations of the Toki Pona models.

References

[1] lipu tenpo ¿ lipu tenpo — liputenpo.org. https://liputenpo.org/. [Accessed 10-May-2023].

[2] Toki Pona (official site) — tokipona.org. https://tokipona.org/. [Accessed 14-May-2023].

[3] Transformers Documentation — huggingface.co. https://huggingface.co/docs/transformers/notebooks.

[4] Wikipesija — wikipesija.org. https://wikipesija.org/. [Accessed 10-May-2023].

[5] Word frequency: based on one billion word COCA corpus — wordfrequency.info. https://www.wordfrequency.info/. [Accessed 13-May-2023].

[6] Corpora — sona.pona.la. https://sona.pona.la/wiki/Corpora, 2023. [Accessed 10-May-2023].

[7] José Pablo Baggetto Chamero. *Study and automatic translation of a language with limited resources*. PhD thesis, Universitat Politècnica de València, 2022.

[8] Eckhard Bick. Wikitrans: the english wikipedia in esperanto. In *Constraint Grammar Applications, Workshop Proceedings at Nodalida*, volume 14, pages 8–16, 2011.

[9] Mark Fettes. The esperanto community: A quasi-ethnic linguistic minority? *Language Problems and Language Planning*, 20(1):53–59, 1996.

[10] Martin Haspelmath and Susanne Maria Michaelis. Analytic and synthetic. In *Language Variation-European Perspectives VI: Selected Papers from the Eighth International Conference on Language Variation in Europe (ICLaVE 8)*, pages 3–22, 2017.

[11] Evangelos C Papakitsos and Ioannis Giachos. The study of machine translation aspects through constructed languages. *International Journal of Electronic Engineering and Computer Science*, 1(1):28–34, 2016.

[12] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS* $EMC^2 Workshop$, 2019.

[14] Shigeru Sato and Hideki Kasuya. Automatic translation/speech synthesis of japanese from written esperanto incorporating a linguistic knowledge base editor. In *ECST*, pages 2414–2417, 1987.

[15] Dariusz Jan Skotarek. Zipf's law in toki pona. *ExLing 2020*, page 189, 2020.

[16] Jörg Tiedemann. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*, 2020.

[17] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*,

pages 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[18] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.