Fall 2024

Understanding Human Statistical Learning in Language Acquisition: Insights from Neural Networks

Anna Kopec

Understanding Human Statistical Learning in Language Acquisition: Insights from Neural Networks

Anna Kopec, University of Connecticut, anna.kopec@uconn.edu

Dr. James Magnuson, Department of Psychological Sciences, james.magnuson@uconn.edu

Abstract

This project explores how the cognitive mechanisms associated with human statistical learning in language acquisition align with computational processes in three kinds of neural networks: feedforward networks (FFN), simple recurrent networks (SRN), and long short-term memory (LSTM) recurrent networks. Prior research in infants has provided evidence of statistical learning in discovering word boundaries within continuous spoken speech. Replicating statistical learning tasks using neural networks could allow for a better understanding of the fundamentals of these parallel processes in our brains and neural networks alike. This project tested the ability of FFNs, SRNs, and LSTMs to make syllable-by-syllable predictions from sequential data in order to determine if the network could accurately attune to word-like structures. Preference for words over part-words and non-words was measured to see if the network could understand transitional probabilities in the same way that human infants can. The results showed that all three networks could perform as well or better than infants on the same word segmentation tasks, where the LSTM was able to achieve the highest proportion better values for both non-word and part-word tasks, followed by the SRN, and finally the FFN. These results suggest that neural networks, specifically the LSTM, develop internal structures that could behave analogously to the cognitive mechanisms behind human statistical learning in human language acquisition. Additionally, they may provide a foundation for continuing work

where I will investigate the limits of each network using more complex learning paradigms.

Introduction

The way that human infants begin to acquire language, one of the most vital parts of human culture, is to this day yet to be fully understood. It is still being explored by scientists, with one proposed theory of infant language acquisition being that of 'human statistical learning.' 'Human statistical learning' refers to the cognitive ability in humans to attune to implicit statistical structures from one's environment without prior instruction or even awareness (Sherman et al., 2020). This theory largely originated with a study by Saffran et al. (1996), a foundational work in the field of human statistical learning and its relation to language acquisition. Scientists designed an experiment based on the familiarization-preference procedure (Jusczyk & Aslin, 1995), in which preverbal infants are first familiarized with auditory stimuli and then tested with both the stimuli and specifically novel altered versions of those stimuli. In the Saffran et al. experiment, 8-month old preverbal infants were exposed to a continuous spoken stream of made-up syllables, such as 'bidakupadotigolabubidaku' that, despite sounding random, actually forms four made-up words. Separating this sequence produces the words 'bidaku', 'padoti', 'golabu', and 'bidaku'. Each of these words occurred together in a pseudo-random order during the familiarization sequence, with the only constraints being that each word appeared forty-five times, and no word could occur twice in a row.

The indication that words exist in this sequence lies in their transitional probabilities, a measure of probability describing the likelihood that a syllable Y will proceed given syllable X: $(Y|X) = \frac{frequency\ of\ XY}{frequency\ of\ X}$. All words had the same pattern of internal syllable-syllable transitional probabilities (syllable transitional probabilities were 1.0 within words, and 0.333 at word

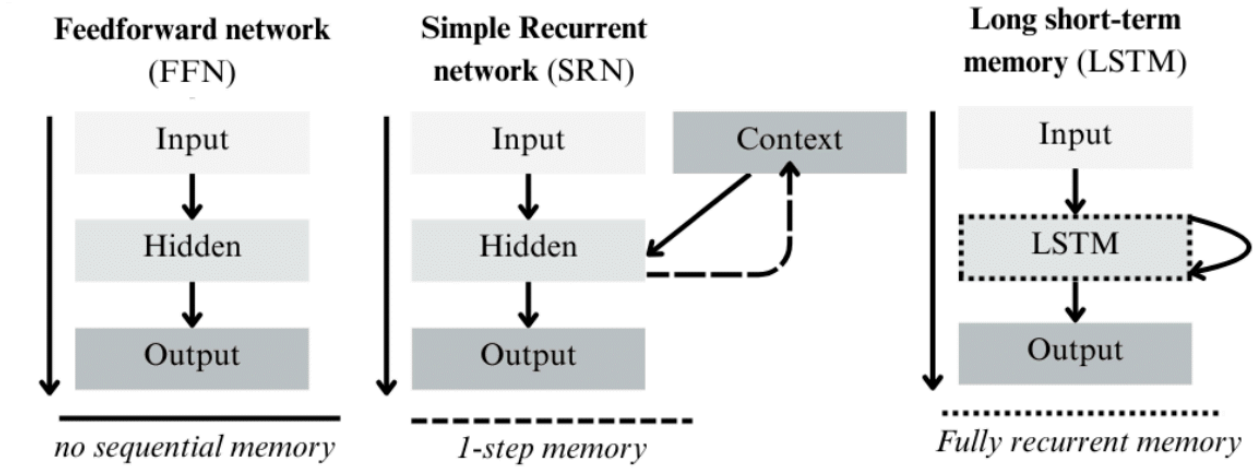
boundaries). During testing infants were presented with both words and two novel types of stimuli: non-words, which consisted of three syllables that never occurred together during the exposure, and part-words, which contained the last syllable of one word and the next two syllables of another word. Infants displayed significantly longer listening times for the novel stimuli as opposed to the familiar stimuli, showing that despite the short time-frame of only two minutes of exposure, the infants were still able to attune to the word-like patterns embedded in the sequence of syllables solely based on statistical relationships between syllables (Saffran et al. 1996).

However, nearly 30 years later, how and to what extent statistical learning influences language acquisition is still yet to be understood (Romberg & Saffran, 2010). Taking an interdisciplinary approach by bridging together computer science and cognitive science perspectives to this problem yields a unique pathway for investigating this mystery. According to Elman's 1990 paper 'Finding Structure in Time,' various neural networks can display human-like statistical learning capabilities when presented with linguistic tasks such as predicting the next word in a sequence (in this paper, Elman introduced Simple Recurrent Networks [now frequently called 'Elman networks'], and the method of next-word prediction training, which is still the core principle driving modern-day Large Language Models like ChatGPT). These models provide a framework for developing and testing theoretical proposals regarding the mechanisms behind behaviors in humans that can be observed but not directly understood (Mirman et al., 2010). The convoluted nature of neural and behavioral data has led to the creation of complex theories of processes like language acquisition. Neural networks are a key framework for modeling learning and subsequent processing, through which they can provide predictions about theories that are too complicated to derive analytically. This project leverages three types of neural

networks (FFN, SRN, LSTM; see Figure 1) to better understand the logical framework underlying human statistical learning in language acquisition, as well as to better determine the relative limits and strengths of different neural networks in human statistical learning tasks related to language acquisition. Figure 1 shows the differences in architecture of each neural network, with a focus on recurrence.

Figure 1

Internal Structures and Differentiation by Memory Capacity of FFN, SRN, and LSTM RNN



The FFN is a very simple form of neural network. The term ‘feedforward’ refers to its fully connected, linear progression through input, hidden, and output nodes respectively (Staudenmeyer & Morris, 2019). Due to its linear nature, it is typically limited to being able to work with adjacent sequences (Jurafsky & Martin, 2024). As this project deals with complex sequential analysis tasks, the FFN will make for an interesting benchmark.

The SRN is identical to the FFN in structure except for the existence of additional unique context units (Figure 1) which operate using limited recurrence. These cyclic context nodes preserve the previous internal state of the network, and in addition to the input nodes, are fully

connected to the hidden nodes. This allows the SRN to factor in complex variables like context in language, and complete more difficult tasks such as non-adjacent sequential predictions (Elman, 1991).

The LSTM RNN, the most sophisticated network of this group, has special architecture that allows it to learn from much longer sequences than an SRN can, and which prevents common problems with training fully recurrent networks (standard RNNs can suffer from 'vanishing' or 'exploding' gradients when trained on very long sequences; Staudmeyer & Morris, 2019). Special features of LSTM nodes (Figure 1), specifically the innovation of adding a memory cell and internal gating parameters that allow cells to dynamically shift how much weight they give to previous states, new inputs, and connections from other LSTM nodes, allow its recurrence to be significantly more powerful. The network is able to not only consider its previous state one timestep ago, but all of its previous timesteps. It has many moving parts, such as ‘constant error carousels’ to help keep the backpropagation process stable, and gates which control information flow to the LSTM’s cells. All of these allow it to complete very difficult cognitive learning tasks such as speech and handwriting recognition (Staudemeyer & Morris, 2019).

The specific goal of this project is to analyze the capabilities and limitations of the FFN, SRN, and LSTM in emulating human statistical learning to discern word boundaries. This will be achieved by first training and testing the neural networks to do syllable-by-syllable predictions on a primary sequential dataset. The neural networks will also be tested on their ability to differentiate between three kinds of sequential data that vary in internal syllable-to-syllable transitional probabilities. These three neural networks were chosen due to their due to their similarities as connectionist models commonly used for studying human statistical learning, as

well as their differences in recurrence, with the FFN having no recurrence, the SRN having a single time step back in recurrence, and the LSTM having full recurrence. I predict that the LSTM will be able to minimize loss most effectively, followed by the SRN, and finally the FFN. This is because of the relative degree of recurrence present in each model. I believe that the higher the degree of recursion present, the more information the network will have at its disposal to produce nuanced sequential analyses (this higher-scale working memory is particularly important for longer sequential dependencies). Additionally, a higher degree of recurrence is more conceptually inline with the ability of the human mind to generate complex recurrent patterns (Dedhe et al., 2023).

Methods

Datasets

All data was generated using code I developed using the Python programming language. Outputs were stored in plain text files, with comma separated values.

Primary Dataset: The primary dataset used for training and testing was a sequence of words, built from syllables. This sequence was designed to mimic the training data used by Saffran et al. (1996). Although numerical parameters were identical, syllables were represented by 12 single unique characters and corresponding one-hot encoded vectors rather than 12 strings (tu-pi-ro = A-B-C). One-hot encoded vectors are a series of *n* one-dimensional arrays of *n* elements which are used in machine learning to numerically represent categorical variables. Each array will be filled with zeros except for the presence of a single one, whose placement in each array will vary, allowing the network to distinguish between arrays using its position. In human experiments, human familiarization with language must be taken into account, and extra

measures like the made-up language from Saffran et al. (1996) must be implemented. However, the objective of these neural networks is to focus solely on extracting statistical structures from the data. Therefore, the specific content of the words is arbitrary, as the primary goal is to assess the neural networks' ability to identify patterns without relying on actual linguistic meaning. Four words were formed by grouping together three unique syllables (tupiro, golabu, bidaku, padoti = ABC, DEF, GHI, JKL). These words were combined into a sequence of 180 words, where each word occurred 45 times, and no word appeared twice in a row. During training and testing, each network was provided with the entire sequence split into three shuffled mini-batches that preserved sequential structure but also helped to avoid overfitting.

Saffran et al. 1996 Test Dataset: This dataset directly mimics the word, part-word, and non-word data used in the Saffran et al. (1996) familiarization periods and preference tests. Like the primary dataset, syllable data exists in the form of unique text characters and corresponding one-hot encoded vectors. Although Saffran et al. (1996) does not specify all combinations of non-words and part-words used, this dataset contains all possible combinations of said data types in order to make the neural network word and part/non-word preference tests more comprehensive.The following tables show the limited data provided by the Saffran et al. paper, and explain the logical development and properties of the part-word and non-word datasets for this project.

Table 1: Test Conditions and Corresponding Syllable Data

Condition A	*Condition B*

	2 non-words	2 part-words
Saffran et al.	dapiku, tilado	tudaro, pigola
symbolic representation	HBI, LEK	IJK, LHI
syllable ordering in parent word	223, 322	312, 312

	2 non-words	2 part-words
Saffran et al.	tupiro, golabu	pabiku, tibudo
symbolic representation	LBH, KEG	LGH, IJK
syllable ordering in parent word	322, 221	312, 312

Table 2: Transitional Probabilities

	Syllable 1 → 2	*Syllable 2 → 3*
word	1.0	1.0
non-word	0	0
part-word	0.33	1.0

Table 3: Non-word/Part-word Formulas used in Implementation:

1. *x* and *y* are 2 words
2. ($x \neq y$)
3. s(a, b) means syllable of numerical position *a* in word *b*:

possible non-word structures:	*possible part-word structures:*
· s(2, x) + s(2, y) + s(3, x) · s(3, x) + s(2, y) + s(2, x) · s(2, x) + s(2, y) + s(1, x) · s(3, x) + s(2, y) + s(1, x)	· s(3, x) + s(1, y) + s(2, y)
((4 + 4) * 3) * 2 = 48 possibilities	(4 * 3) = 12 possibilities

Network Implementation

Python and the machine-learning framework PyTorch were used to implement the FFN, SRN,

and LSTM. The hyper-parameters for each of the networks are listed in Table 4.

Table 4: Network Hyper-parameters

	FFN	*SRN*	*LSTM*
optimizer	SGD	Adam	Adam
criterion	MSE	MSE	MSE
Learning rate	0.05	0.005	0.0001
Activation Function	RELU	tanh	tanh
Input size	12	12	12
Hidden size	12	12	12
recurrence	n/a	1 timestep	Full sequence

Each network has a corresponding functionality class which provides it with attributes and methods for data initialization, training, testing, results storage, and data visualization.

Training and Testing:

Each network's ability to make syllable-by-syllable predictions correctly after undergoing training was observed. Each network underwent 200 epochs (one epoch is a full-pass through the entire dataset) for the learning process. For each syllable in the primary sequence, the network was given a one-hot encoded vector and asked to predict the next one in sequence. Loss would be determined by comparing the network's predicted output with the data for the actual next syllable in sequence. During training only, the network would undergo backpropagation and optimization to minimize loss. This predictive training and testing is analogous to the 2-minute familiarization

done by infants in Saffran et al. (1996). Testing would use the same syllable prediction criteria, but without any internal parameter adjustments, and using the same set of mini batches of syllable data but re-shuffled to help prevent overfitting.

Preference Test:

After training and testing, each epoch concluded with a preference test to assess how the network would react to familiar words versus part-words and non-words. Mimicking Saffran et al. (1996), the network would undergo three trials, in which during each trial, the network would be exposed to a word vs. non-word preference test and a word vs. part-word preference test. During the non-word test, the network would be exposed to two random words from the primary dataset, and two random non-words. It would be asked to predict the second syllable given the first syllable, and then the third syllable given the second syllable. The part-word test had the same procedure, except instead of two non-words, two random part-words were selected and used for testing. Hidden states were not preserved during these trials, and the networks did not make any internal alterations in parameters in order to ensure that results properly reflected only alterations made during training.

Analysis:

This project uses two measures to characterize network performance: ‘proportion better’ and ‘proportion word choice’, which were calculated based on results from the word preference tests. Part-word vs. word and Non-word vs. word preference test data were analyzed separately for proportion better and proportion word choice over time.

Proportion better: This measure tracks relative-differences, allowing it to be used not

only for model-model comparisons, but also for model-human empirical data comparisons (French et al. 2011). This makes it ideal for this project, as both types of comparisons will be considered. The formula for proportion better is:

$$\frac{b-a}{a+b}$$

where *b* = summed error of part-words or non-words, *a* = summed error of words. Higher proportion better shows that the error for non-words or part-words is higher than the error for words, which is analogous to the infants in Saffran et al. (1996) displaying longer listening times to non-words and part-words than words.

Proportion word choice: This method is a simpler and more intuitive way of displaying word preference, as it is a simulation of making the network 'choose' a word or comparison item (non-word or part-word) based on error (it 'chooses' the item with lower error). Four loss value comparisons are made, where the error of each word is compared to the error of each non-word or part-word from a single trial. If the error of a word is lower than the erorr of a corresponding part-word or non-word, then it is counted as a 'word choice' and one is added to the total number of word choices across all trials for the current epoch. This comparison can be described as,

$$int(a1 < b1) + int(a1 < b2) + int(a2 < b1) + int(a2 < b2)$$

where a1 and a2 are two randomly selected words, and b1 and b2 are two randomly selected part-words or non-words. The formula used in this project for proportion word choice is as follows:

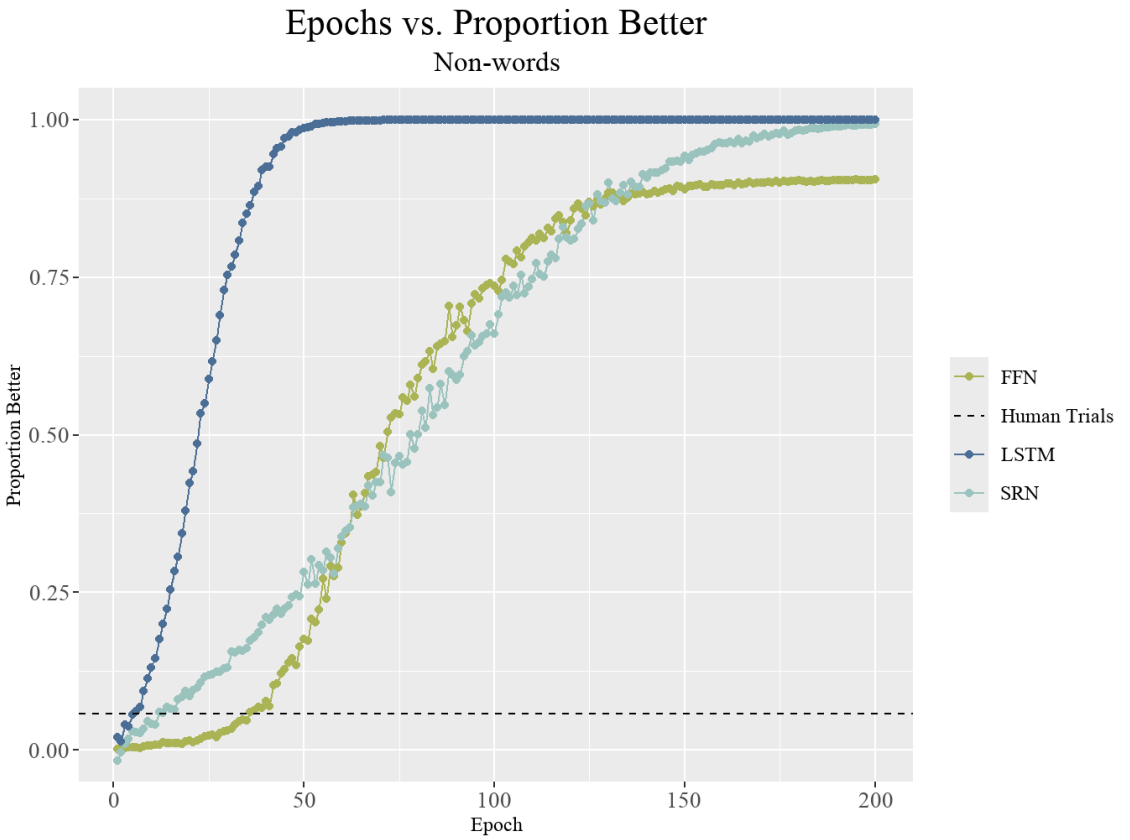
$$proportion\ word\ choice = \frac{word\ choices}{w * c}$$

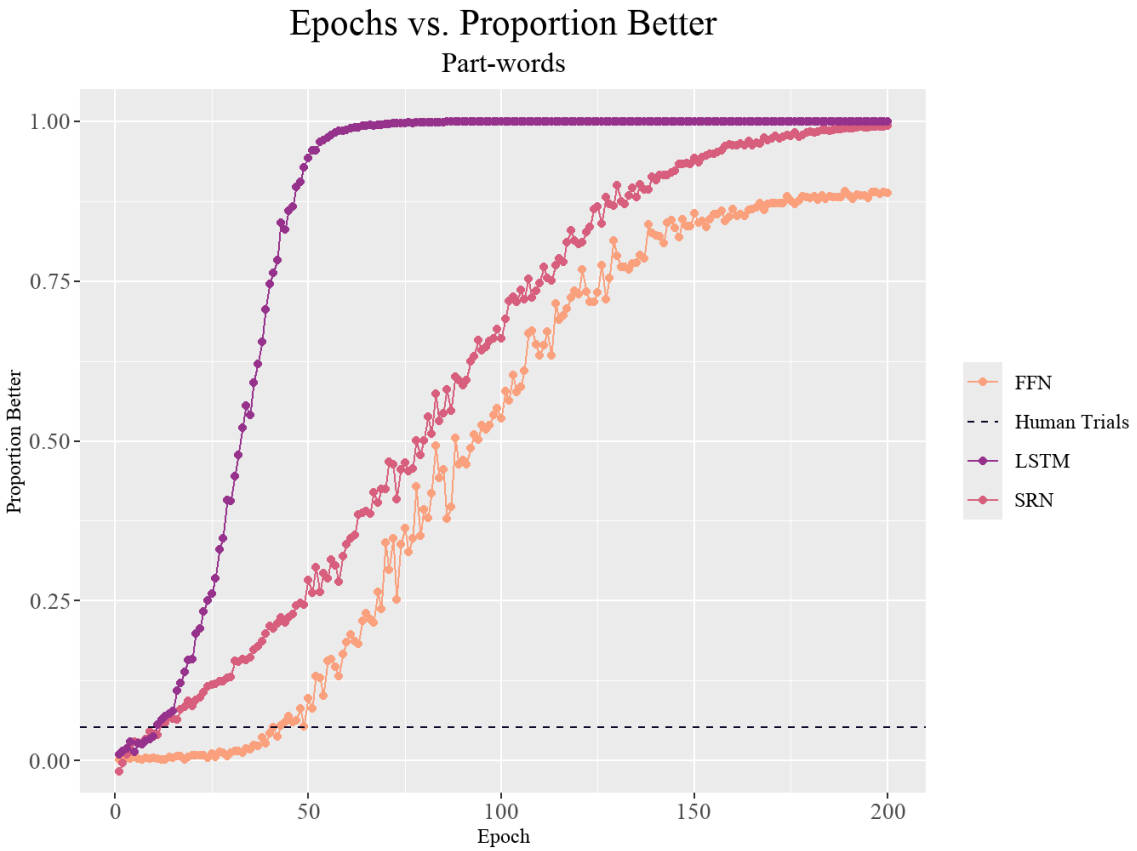
Where *word choices* represents the integer number of words 'chosen' within a trial, *w* represents the number of words and *c* is the number of comparison items. Increased proportion word choice

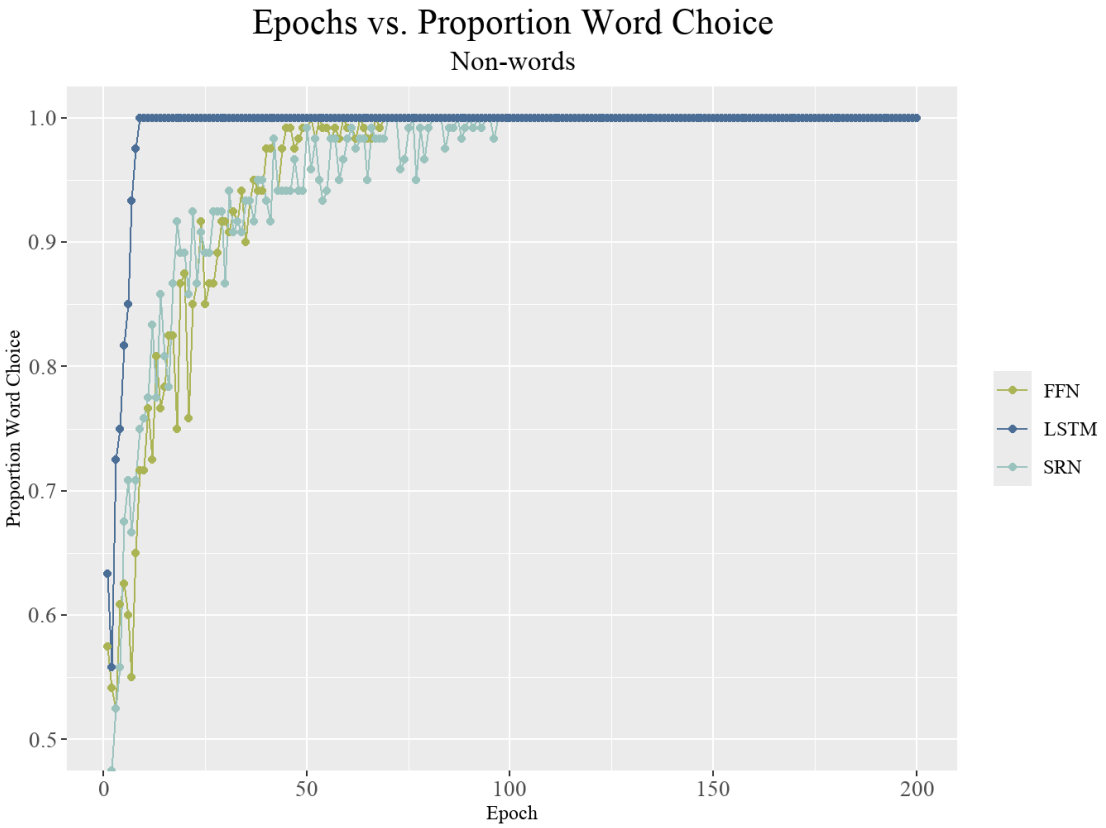
signifies that the network is consistently displaying lower error for words than for non-words or part-words.

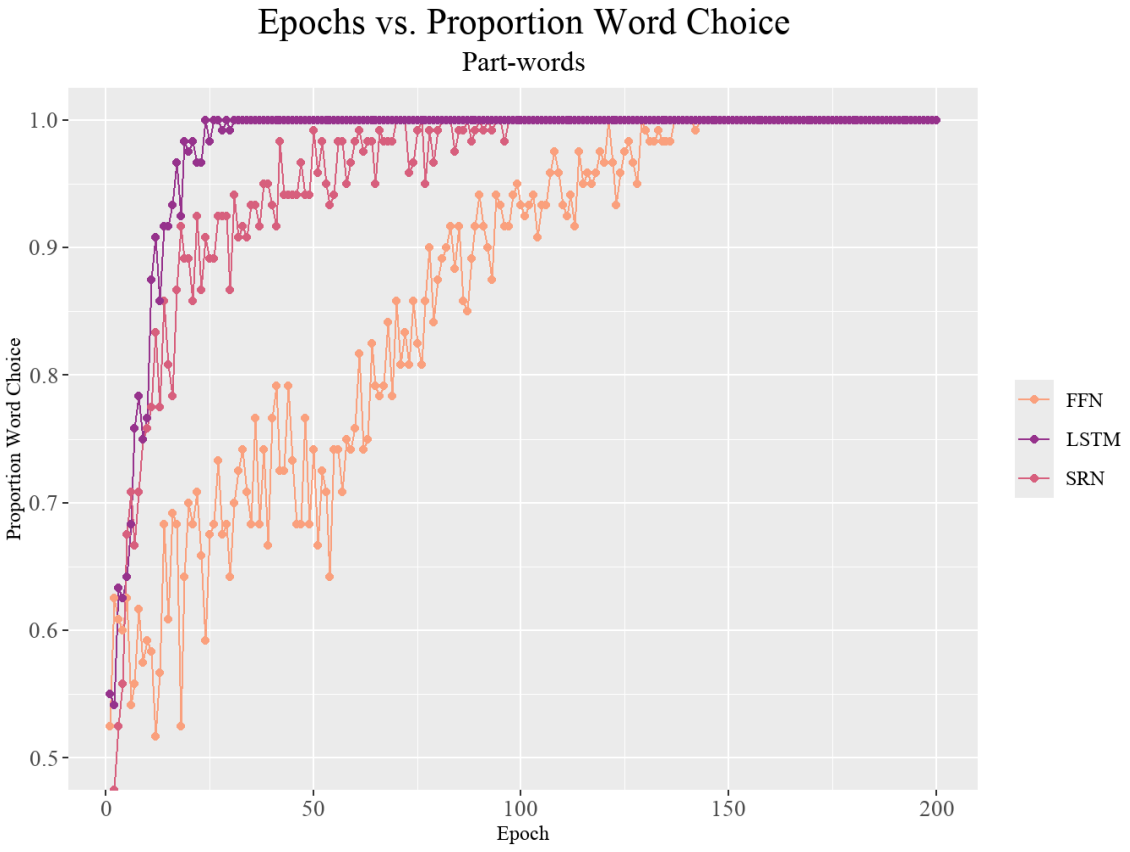
Results

The following graphs show data collected and averaged from the FFN, SRN, and LSTM over the course of ten complete runs through 200 epochs. 'Human trials' line refers to empirical infant performance data from Saffran et al. (1996). Each human trial proportion better value was calculated using the ratios of seconds spent listening to word stimuli vs non-word or part-word stimuli. Note that the level of human performance was statistically greater than chance, but quite low, since humans received the equivalent of just one epoch in their 2 minutes of exposure.


Epochs vs. Proportion Better
Non-words
Proportion Better
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
Epoch
FFN
Human Trials
LSTM
SRN


Epochs vs. Proportion Better
Part-words
Proportion Better
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
Epoch
FFN
Human Trials
LSTM
SRN


Epochs vs. Proportion Word Choice
Non-words
Proportion Word Choice
1.0
0.9
0.8
0.7
0.6
0.5
0
50
100
150
200
Epoch
FFN
LSTM
SRN


Epochs vs. Proportion Word Choice
Part-words
Proportion Word Choice
1.0
0.9
0.8
0.7
0.6
0.5
0
50
100
150
200
Epoch
FFN
LSTM
SRN

Discussion and Conclusions

Results above show each network's increases in proportion better and proportion word choice during the preference tests run at the end of each epoch. In the proportion better graphs, there is a dotted line labeled 'human trials' which refers to infant performance on the non-word and part-word preference tests in Saffran et al. (1996). The results show that the FFN, SRN, and LSTM can all perform as well as or better than humans. The LSTM was able to learn the most effectively in the shortest amount of epochs. The SRN learned more quickly and effectively than the FFN on the part-word tasks. On the non-word tasks, however, the FFN and SRN alternated between having the higher proportion better over three sets of intervals, where the SRN performed better near the beginning and end of the entire learning process, and the FFN performed better in the middle of the learning process. This was the case even though the FFN had the largest learning rate and the LSTM network had the lowest (see Table 4), a point I will discuss in detail below.

It is important to note that the LSTM achieved maximum proportion better (1.0) and maximum proportion word choice (1.0) for both tasks. The SRN was close, reaching maximums of about 0.99 for proportion better and 1.0 for proportion word choice. Despite achieving maximum proportion word choice for all tasks and overtaking the SRN at some points during the non-word tests, the FFN could not reach proportion better values above about 0.91 for part-words and 0.93 for non-words. Based on the maximum proportion better values achieved by the networks, the results suggest that the LSTM is most adept at discovering word boundaries based on transitional probabilities, followed by the SRN, and finally by the FFN. This trend shows a positive correlation with the level of recurrence in a network and its proportion better. Interestingly, the FFN and SRN performed incredibly similarly on the non-word preference tests,

while the SRN performed distinctly better than the FFN on the part-word preference tests. Why exactly this is the case is difficult to determine, particularly because the SRN's context layer allows it to keep track of information from both the current time-step *and* one time-step ago, theoretically allowing it to recognize if certain syllables never co-occurred together.

The results suggest that all three neural networks are able to use transitional probabilities to attune to word-like structures within the given paradigm, displaying behavior analogous to human statistical learning. The LSTM's exceptional ability to learn both of these tasks more quickly and more consistently than the other networks suggests that the cognitive mechanisms behind the human statistical learning involved in word segmentation may use a high degree of neural recursion, similar to the structure of the LSTM. However, it is difficult to determine the validity of this statement based on only this specific word segmentation paradigm alone. Other statistical learning paradigms, such as the French et al. (2011) 'box language' paradigm and the Perruchet and Poulin-Charronnat (2012) 'phantom words' paradigm, are much more complex and might provide stronger tests of the three networks. To follow up on this project, I will be conducting future research exploring the ability of the same three neural networks to display human-like preferences within these paradigms.

One largely impactful variable in this project was the learning rate of each neural network. Each network had a different learning rate (FFN=0.05, SRN=0.005, LSTM=0.0001) which greatly impacted how each network performed. If the FFN and SRN's learning rates were too low, they were unable to surpass the human trials benchmark. Meanwhile, if the LSTM's learning rate was too high, it almost instantaneously reached maximum proportion better and proportion word choice. The issue of learning rate is central to neural networks and makes it difficult to determine if a network's performance is enough to suggest that its structure is

analogous to the cognitive mechanisms behind that task in humans.

Additionally, more comprehensive experiments on humans in using transitional probabilities to extract word boundaries would provide better reference material for making model-human comparisons. Currently, no comprehensive experiments on humans exist in which the subjects have been given enough time to reach significantly high accuracy like the neural networks in this project did. Despite epochs possessing a sequential structure, it is still difficult to translate them into time and vice versa. A more comprehensive experiment showing change in human learning over time during word segmentation tasks like the Saffran et al. (1996) paradigm could prove to be a much better reference point for projects like this one.

Acknowledgements

I would like to express my sincere gratitude to Mr. and Mrs. Holster for their generous support of this research through the 2024 Holster Scholars Program. Their funding made this project possible, and their commitment to advancing undergraduate scientific research is deeply appreciated.

Additionally, I extend my thanks to my project mentor, Dr. Magnuson, whose guidance and support throughout the Holster Scholar application process and my project itself have been invaluable. This project would not have been possible without his assistance.

References

Dedhe, A. M., Piantadosi, S. T., & Cantlon, J. F. (2023). Cognitive Mechanisms Underlying Recursive Pattern Processing in Human Adults. *Cognitive Science*, *47*(4), e13273. https://doi.org/10.1111/cogs.13273

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2), 195–225. https://doi.org/10.1007/BF00114844

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614–636. https://doi.org/10.1037/a0025255

Jurafsky, M. (2024). *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/

Jusczyk, P. W., & Aslin, R. N. (1995). Infants′ Detection of the Sound Patterns of Words in Fluent Speech. *Cognitive Psychology*, *29*(1), 1–23. https://doi.org/10.1006/cogp.1995.1010

Magnuson, J. S. & Sahil Luthra. (n.d.). *Simple Recurrent Networks are interactive*.

Mirman, D., Graf Estes, K., & Magnuson, J. S. (2010). Computational Modeling of Statistical Learning: Effects of Transitional Probability Versus Frequency and Links to Word Learning. *Infancy.*, *15*(5).

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability

computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*(4), 807–818. https://doi.org/10.1016/j.jml.2012.02.010

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews. Cognitive Science*, *1*(6), 906–914. https://doi.org/10.1002/wcs.78

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.)*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Sherman, B. E., Graves, K. N., & Turk-Browne, N. B. (2020). The prevalence and importance of statistical learning in human cognition and behavior. *Current Opinion in Behavioral Sciences*, *32*, 15–20.

Staudemeyer, R. C., & Morris, E. R. (2019). *Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks*. https://arxiv.org/abs/1909.09586