March 2021

Exploring the Use of Neural Transformers for Psycholinguistics

Antonio Laverghetta Jr.
University of South Florida

Exploring the Use of Neural Transformers for Psycholinguistics

by

Antonio Laverghetta Jr.

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: John Licato, Ph.D.
Sriram Chellappan, Ph.D.
Tempestt Neal, Ph.D.

Date of Approval:
March 15, 2021

Keywords: Language Models, Psychometrics,
Natural Language Processing, Age of Acquisition, Deep Learning

Table of Contents

List of Tables

List of Figures

Abstract

Deep learning has the potential to help solve numerous problems in cognitive science and education, by providing us a way to model the cognitive profiles of individual people. If this were possible, it would allow us to design targeted tests and suggest specific remediation based on each individual's needs. On the flip side, employing techniques from psychology can give us insight into the underlying skillsets neural networks have acquired during training, addressing the interpretability concern. This thesis explores these ideas in the context of transformer language models, which have achieved state-of-the-art results on virtually every natural language processing (NLP) task. First, we study the ability of transformers to model Age of Acquisition, an important variable in predicting word processing in humans. We then examine the broader challenge of using transformers to model the human responses to a test of linguistic competencies, this time employing measures from psychometrics as part of our evaluation. Compared to simpler models, we find that transformers can achieve superior results in both cases, suggesting they are more suitable for modeling psychological properties. The use of psychometric measures also allows us to study what linguistic skills transformers *cannot* learn, suggesting where future architectures can make improvements.

Chapter 1: Introduction

Psycholinguistics studies how humans process language and represent it in the mind and brain [97]. Work within this field has led to many theories on how language is acquired and comprehended. While this work is clearly of interest to psychologists, artificial intelligence (AI) has benefitted as well. Work dating back decades has investigated applying ideas from psycholinguistics to create NLP systems [98]. In the context of neural networks specifically, this area of research has sometimes been called "connectionist psycholinguistics" [99].

Recently, NLP has seen the rise of a new class of neural language models, based on the transformer [12]. Transformers rely on an attention mechanism to perform a sort of feature extraction on the input, thus encoding rich structural and semantic information. Attention in neural networks is designed to mimic attention from cognitive science, by allowing the network to focus on import parts of the input and mask out the rest. It is a general design mechanism in deep learning that has been used in fields outside of NLP, such as computer vision [106]. Within NLP, while various forms of attention have been used extensively in prior work [100], transformers take this to an extreme by using *only* attention in its encoder and decoder layers, throwing out recurrent [101] and convolutional [102] layers entirely. A transformer encoder thus consists of only a self-attention layer and a fully connected layer, separated by residual connections [107] and layer normalization [108]. The entire architecture consists of multiple encoders stacked on top of each other, with a decoder before the output layer that could be as simple as a linear layer. This greatly reduces the complexity of the model and the computational

costs of training. Moreover, architectures based on the transformer have achieved state-of-the-art results on a vast number of NLP tasks.

Given this great success, a pertinent question is whether transformers are any better than previous methods for modeling psycholinguistic properties. An answer to this question would be of interest to both psychology and AI. On the one hand, transformers could improve prior connectionist approaches for modeling psycholinguistic effects, enriching our understanding of language processing in humans. The techniques psychologists have used for decades to model latent cognitive processes, in particular, psychometrics,[1] could also be quite valuable to AI for addressing interpretability concerns. However, perhaps because transformers are still a relatively new class of architecture, very little work has addressed their psychological plausibility.

This thesis explores using transformers for modeling psycholinguistic properties of language, using both traditional techniques from machine learning and NLP, as well diagnostics originally developed in psychometrics for human evaluations. We begin in Chapter 2 by reviewing related work, which covers the use of machine learning to model psycholinguistic features, previous attempts to merge psychometrics with AI and work on the interpretability of transformers. In chapter 4 we examine how well transformers can model Age of Acquisition (AoA), which is the age at which a word is typically learned by humans. Within psycholinguistics, AoA is thought to be an important variable in predicting the lexical processing of words, along with concreteness and affectiveness [1-2]. For instance, AoA is thought to affect how fast words are read [3], and how fast pictures can be named [4]. AoA and other psycholinguistic norms provide a powerful data source for modeling various aspects of human behavior using techniques from NLP. For example, research within psychology has shown that

[1] https://www.psychometricsociety.org/what-psychometrics

combining word embeddings with human psycholinguistic judgment ratings can allow us to model human perceptions related to health behavior and risks [54]. We use BERT [10] and RoBERTa [11] for these experiments, two popular transformer models. BERT is probably the most well-known transformer architecture, as it introduced pre-training objectives that have become common in related architectures. RoBERTa is described as a "robustly optimized BERT pre-training approach". It uses the same architecture as BERT but makes various careful optimizations to the pre-training strategy that have led to improved performance on various benchmarks. We compare the transformers against two baselines, one which simply makes random predictions, and the other which uses a set of handcrafted features known to correlate highly with AoA.

In chapter 4, we expand the scope of our experiments by testing the reasoning capabilities of transformers on a wide number of linguistic skills. However, unlike previous similar work, we make use of psychometric measures to study the performance of transformers compared to a human baseline. Unlike the single-valued performance measures common in machine learning (accuracy, F1, etc.), psychometrics allows us to model performance on a test as being affected by multiple latent variables, which in this case are underlying linguistic competencies. We can thus study how these variables relate to each other, giving us a way to check whether certain linguistic skills in transformers depend on having acquired others first. We can also study how closely this sequence of skill acquisition mirrors human data. We evaluate our transformers on GLUE [64], a well-known NLP benchmark. We specifically use the benchmark's diagnostic, which is one of the most comprehensive tests of linguistic reasoning devised by the NLP community thus far.

Chapter 2: Related Work

2.1 Applying Distributional Models to Age of Acquisition

Factors that contribute to word acquisition have been studied extensively over the years. It has been shown that word frequency [16], length [18], polysemy [20], and part of speech [15] are highly correlated with AoA. Other work has used techniques from network science to generate lexical graphs of words and found that associations within these networks could predict AoA quite well [48]. Early studies on AoA within psycholinguistics were generally small in scale, focusing on a handful of words picked for certain properties they possessed [5]. While this type of factorial design allows for specific variables to be studied very precisely, it is unclear whether the words being examined have properties typical of all the words in the vocabulary in question, or rather are special cases [6]. To address these difficulties, much work has used machine learning as both a way to expand existing psycholinguistic datasets automatically and to analyze the predictive power of those datasets. [17] used a handcrafted set of psycholinguistic features to train machine learning models to predict AoA. They find that a logistic regression model achieves up to 72% accuracy on this task, with the random baseline being 50%. [22] similarly uses handcrafted features to train a linear regression model to predict the AoA of Italian words.

Because children are thought to utilize co-occurrence information during lexical processing [23], distributional models have been especially popular for modeling acquisition norms. Work in this area has so far focused on older non-contextual models of semantics, especially LSA [7], HAL [8], and skip-gram [9]. [24] extrapolated AoA ratings using LSA,

HAL, and skip-gram models. They achieved about 73% correlation with human norms using the skip-gram model. [25] combined a distributional model with Wordnet [26] to create an algorithm for expanding psycholinguistic datasets in a semi-supervised fashion. [27] used LSA to estimate several psycholinguistic variables, by predicting a word's rating as the average rating of the word's k-nearest neighbors in the LSA space. They achieved a strong correlation for several of the variables tested, though they did not examine AoA. [41] used a network-based distributional model to study how affective word features influence early language development. [42] trained SVD and skip-gram models on child-directed speech and evaluated the model's ability to predict AoA norms. They achieved a modest and significant correlation on two evaluation tasks.

Collectively, the success of this work indicates that distributional models are a promising way to model feature norms. However, very little work so far has used deep contextual models for this purpose, despite the great success they have achieved on NLP tasks. An important exception is the work by [28], which fine-tuned BERT on feature norms (not including AoA) and demonstrated that the fine-tuned model could predict novel concepts and features quite well. Most interestingly, they investigated the psychological plausibility of BERT by testing it on a wide variety of classic psychological experiments. In fourteen out of a total of sixteen tests, BERT was able to produce human-like responses to the stimuli in a statistically reliable fashion. While these experiments alone are not sufficient to state that BERT is a psychologically plausible model of human cognition, they do indicate that BERT may be superior to older distributional models for psycholinguistic applications.

2.2 Applications of Psychometrics to Machine Learning

Psychometrics is a field study dedicated to developing quantitative measures of psychological properties. Such properties include knowledge, attitudes, and personality traits,

among others. Research within this field has led to the creation of sophisticated models for measuring performance on tests, including diagnostic classification models (DCM) [50], and item response theory models [49]. These techniques can give us a rich understanding of the relationships between underlying cognitive skills because they provide a way to measure how those skills *relate to each other*. In other words, we test whether certain skills depend on having first acquired other skills before they can be learned. This property has made psychometrics models popular in designing assessments in education [51], since it gives us a more nuanced understanding of a student's performance than any single metric can.

Given that we can use psychometrics to build cognitive profiles of humans, can we also use it to build profiles of neural models? Interest in unifying AI with psychology traces as far back as [52], and many others have pursued this unification since then [59-61]. However, despite these efforts, the amount of work joining AI with psychometrics is quite limited. [53] augmented the DINA [62] and DINO [63] cognitive diagnostic models with a feedforward neural network using a semi-supervised learning objective. This architecture can achieve superior results to multiple baselines, on both synthetic and real-world assessments. [54] created a deep learning architecture for extracting psychometric dimensions from the text, which achieved superior performance to prior techniques. [55] investigated how to automatically create a corpus of psychometric data from natural language text, sidestepping the need to explicitly gather survey responses. [57] used psychometric measures to study the impact of question difficulty on the performance of deep neural networks. [56] used deep neural networks to generate data for item response theory models. The generated data achieves moderate to high correlation with actual human data. [58] used item response theory models to efficiently assess chatbots, reducing the

amount of data needing to be annotated by humans. This work clearly demonstrates the potential benefits of psychometrics to NLP, but further work is needed to bring it to fruition.

2.3 BERTology and Interpretability of Transformers

Understanding the inner workings of neural networks has been a topic of study for years. The great success of transformers has also led to a tremendous amount of research on how they operate internally, which collectively is sometimes called BERTology [88]. [90] investigated what kinds of information are being encoded within BERT. They found that BERT implicitly recreates the classical NLP pipeline within its encoder layers. [91] studied the numerical reasoning capabilities of various state-of-the-art language models, including BERT. They found that BERT struggles with forming good representations of floating-point numbers and fails to generalize to numbers not seen in the training data. [92] used BERT's masked language modeling pre-training task to study the model's knowledge about the world. They determined that BERT is competitive with traditional knowledge bases for extracting certain types of information. [93] studied how RoBERTa's predictions on various tasks changed when part of the input was permuted to be nonsensical. Surprisingly, the transformers still produced high confidence guesses even with meaningless inputs, although it was shown the models could be trained to be more robust to this type of permutation.

In summary, prior work has given us a great deal of insight into how transformers process data internally. However, these findings must also be taken with a grain of salt. For example, studying the self-attention mechanism in transformers is a popular approach for interpreting the model's predictions, but it is unclear how interpretable attention is [95-96]. It has also been shown that results on a probing task can change depending on how the task is structured, meaning that a single test is insufficient for drawing strong conclusions [96]. We add to this

literature by applying psychometrics to the interpretability problem, which we hope will lead to new insights into the kinds of tasks transformers are best suited for.

Chapter 3: Modeling Age of Acquisition Norms

In this chapter, we study how well transformers can model AoA norms, compared to several baseline models. We perform our experiments using two common AoA datasets. The first is Kuperman's AoA ratings [14], which contain acquisition norms for over 30,000 English words. Kuperman was able to gather this large-scale norm dataset by employing workers on Amazon Mechanical Turk and demonstrated through several experiments that the norms gathered in this fashion are just as reliable as norms collected in laboratory settings. The original dataset was later expanded to include data from several other studies [30-33], bringing the total size up to over 50,000 words. The second dataset comes from Wordbank [13] which is a database of responses to MacArthur-Bates Communicative Development Inventory [29] (CDI) questionnaires, taken by the caregivers of children around the world. This is a self-reported form of language proficiency of the child as observed by the caregiver and allows us to study the AoA of developing children.

3.1 Methodology and Results

We first perform some preprocessing on our datasets. For Kuperman, we use only the lemmatized version of each word and drop any duplicate words or words which have no AoA rating. For Wordbank, we use data for only English-speaking children and computed the normative AoA of each word. This is the age at which at least 50% of the respondents could produce the word. In total, we have 600 words in Wordbank and about 30,000 words in Kuperman after preprocessing.

We use the Transformers [77] implementation of each of our BERT and RoBERTa models. We use the *bert-base*, *bert-large*, *roberta-base* and *roberta-large* community models from Huggingface.[2] These are all the pre-trained models described in their respective papers. We take the average of the activations for the second to last transformer hidden layer of each token in the input sequence as the word embedding, giving us a 768-dimensional vector for the *base* models and 1024 for the *large* ones. Taking the average ensures that the word embeddings are always fixed to these lengths, which is important because some words consist of several words (for instance "give me five" in Wordbank). Of course, how to best obtain word vectors from contextual embeddings is an open question, and future work is planned to examine how different embedding strategies impact downstream performance.

We compare the transformers against a handcrafted set of psycholinguistic features known to correlate with AoA:

3.1.1 Psycholinguistic Features

- Frequency: How often the word occurs in language. We use the frequency counts of words in the OpenSubtitles database [35], since it has been shown this dataset is more suitable for studying psycholinguistic phenomena than other corpora [34]. For words not present in the data, we set the value to 1.
- Polysemy: The number of senses a word has. We obtain this by counting the number of synsets of the word in Wordnet. For words not present in Wordnet, we set the value to 1.
- Whether the word is a noun: In the Kuperman norms this data is already present. For Wordbank since the dataset is small we manually annotate the word's part of speech based on the category it is assigned to (food, toys, helping verb, etc). In most cases, it is

[2] https://huggingface.co/models

trivial to determine whether the word is a noun. In any case where the part of speech is ambiguous, we set it based on the part of speech of the majority of the word's synsets in Wordnet.

- Length: the number of characters in the word.

We additionally compare all models against a random baseline, where the predicted label is simply assigned randomly in the range of possible labels for the dataset. If the transformers have captured any useful properties for this task, they should be able to consistently do better than the random baseline. In the following sections, when we say "baseline features" or "baseline" we are referring to the psycholinguistic features, and "random baseline" is this random classifier.

Table 3.2 shows results on the Kuperman norms. We experimented with a variety of regression models, all implemented in sci-kit learn [46]. We use Pearson [36] and Spearman [37] correlations to measure performance. To ensure statistical significance we shuffled the dataset and ran 10-fold cross-validation on all models. The reported correlations are the mean correlations of these trials for each model. For the baseline experiments, we first standardized the features by removing the mean and scaling to unit variance. For any model which had tunable hyperparameters, we first ran a grid search, using a separate validation set held out from the training set, and used the following settings found to be optimal:

3.1.2 Kuperman Hyperparameters

- SGD: elasticnet penalty, squared loss, adaptive learning rate, eta0 = 0.001, alpha = 0.01
- Decision Tree: at least 4 samples per leaf, min impurity decrease of 0, max depth of 5
- k-NN: number of neighbors equal to 25
- SVR: C = 3.26, epsilon = 0.81

All other hyperparameters are left at their defaults. In most cases, the transformers either outperform or perform just as well as the baseline features. In most cases, *bert-large* performs somewhat better than *bert-base*, which is to be expected given the larger size of this model. The same trend holds for the RoBERTa models, however, both variants perform noticeably worse overall than the BERT models. The transformers perform much better than the random baseline, which only gets very weak correlation using both measures. For most folds on the random baseline, the correlation is also not statistically significant.

While the best model is the decision tree using the baseline features, the difference is small, as *bert-large* using SVR comes within 10% of the Pearson correlation and 5% of the Spearman correlation. To determine whether this difference in correlation was statistically significant, we performed a t-test [40] on the per-fold reported correlations for the *bert-large* SVR model and the baseline decision tree model. We performed this test on the Spearman and Pearson correlations separately, results are in Table 3.1. The difference for Pearson correlation is clearly statistically significant, but results are less certain for Spearman. While the p-value is less than 0.05, it comes close to this significance cutoff, as the exact value is 0.0496. Overall, it appears that the baseline features are achieving a modestly stronger correlation than the best transformer model, though the difference is quite small.

For Wordbank, we used an evaluation based on prior work which framed AoA as a classification task [17]. We first bin the Wordbank AoA norms into a set of 3 discrete labels. Table 3.3 shows the class assignments and the number of examples per class for the resulting dataset. Since the large majority of words are acquired at around 20 to 25 months old, we could not use uniform ranges for the bins without having classes with an extremely small number of examples. We therefore manually tuned the ranges to balance out the number of examples per

class as much as possible, although one class still has less than half the number of examples as the other two.

We trained various classification algorithms, again using both the transformers and the baseline features. We used Matthews correlation [43] to measure performance. To address the class imbalance, we weighted the input samples to be inversely proportional to the class frequencies. We use the following hyperparameter settings (all others are left at their defaults):

3.1.3 Wordbank Baseline Hyperparameters

- Logistic Regression: C=0.3, L2 penalty, newton-cg solver
- Decision Tree: Gini impurity, max depth of 200, log2 max features, use the best split
- SVC: C=0.2, gamma set to auto, rbf kernel
- KNN: chebyshev distance metric, 15 nearest neighbors

3.1.4 Wordbank Transformer Hyperparameters

- Logistic Regression: C=1.0, L2 penalty, sag solver
- Decision Tree: entropy impurity, max depth of 15, log2 max features, use the best split
- SVC: C=5.0, gamma set to scale, rbf kernel
- KNN: manhattan distance metric, 15 nearest neighbors

Table 3.1 Results of the t-test on *bert-large* SVR and baseline decision tree correlations

Correlation	t-statistic	p-value
ρ	2.17	< 0.05
r	5.3	< 0.01

Table 3.2 Spearman ρ and Pearson r correlation on Kuperman norms. SGD is linear regression with stochastic gradient descent. k-NN is k-nearest neighbors regression. All correlations in the table are significant, with $p < 0.001$. For the random baseline, we obtain 0.01 correlation on average using both measures, and 95% of the trials have $p > 0.05$.

Model	bert-base ρ	bert-large ρ	roberta-base ρ	roberta-large ρ	baseline ρ	bert-base r	bert-large r	roberta-base r	roberta-large r	baseline r
Linear	0.53	0.54	0.37	0.41	0.40	0.54	0.55	0.38	0.42	0.44
Ridge	0.53	0.54	0.37	0.45	0.39	0.54	0.55	0.38	0.42	0.44
SGD	0.53	0.45	0.28	0.32	0.40	0.54	0.45	0.28	0.33	0.44
k-NN	0.50	0.48	0.3	0.31	0.53	0.51	0.48	0.3	0.32	0.62
Decision Tree	0.36	0.31	0.18	0.21	0.59	0.37	0.33	0.19	0.21	0.64
SVR	0.53	0.56	0.39	0.42	0.46	0.54	0.58	0.4	0.43	0.53

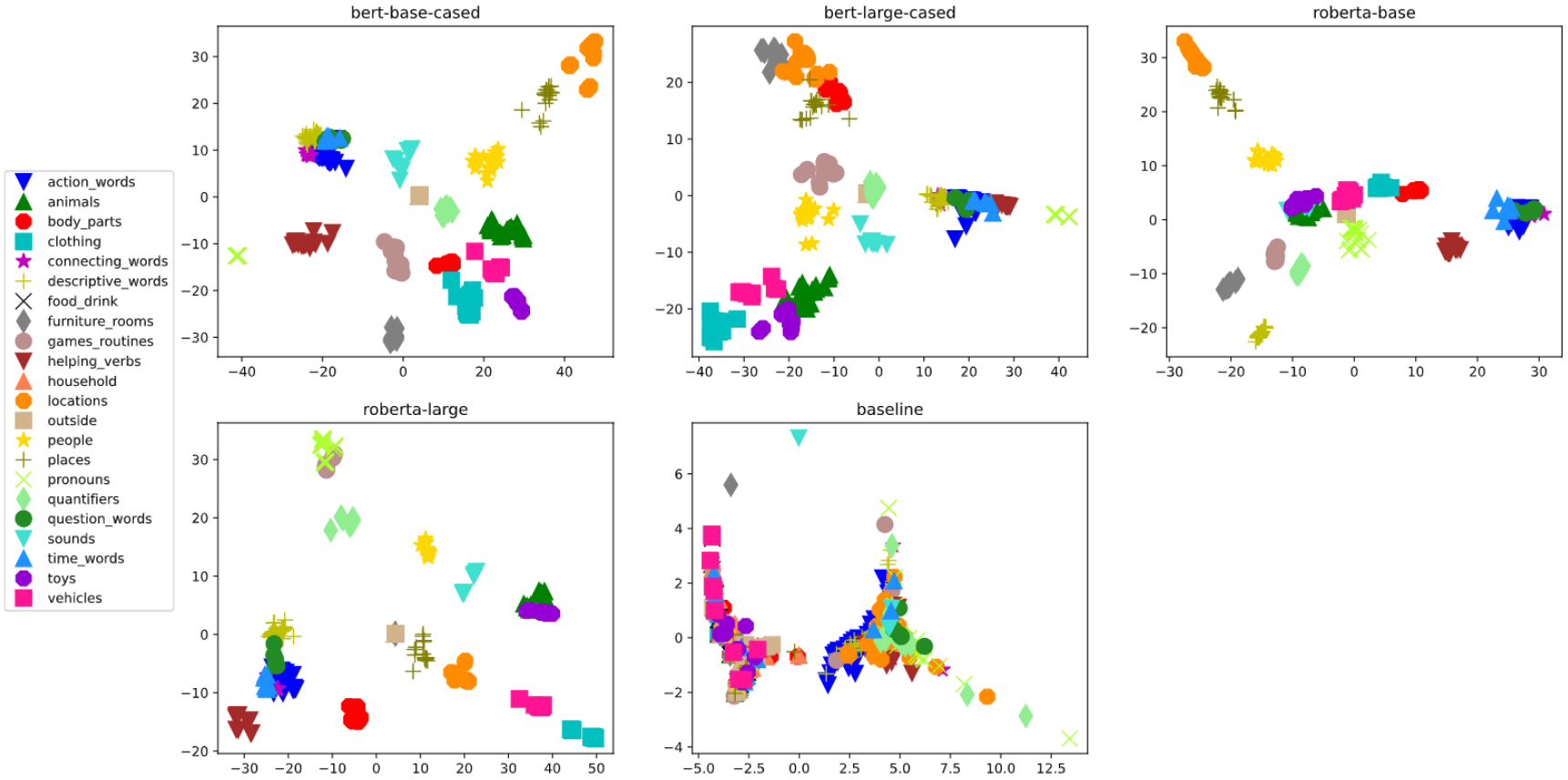

Figure 3.1 Isomap projections of all featuresets

Table 3.3 Final Wordbank dataset statistics.

AoA Range	Label	Count
(0, 20]	0	83
(20, 25]	1	254
(25, 52]	2	263

Table 3.4 shows the results of our experiments. We ran 10-fold cross-validation on all models using the optimal hyperparameters, correlations shown are of the average across all folds. Getting a strong correlation on this dataset is much more challenging since there are only a few hundred examples and the class distribution is imbalanced. We obtained only weak correlation regardless of the configuration, however, this time *bert-large* achieves superior performance to both baselines, getting as high as 0.14 correlation. We again find that the random baseline achieves very weak correlation, which all transformers can surpass using at least one of the classification models.

Unlike in the Kuperman dataset, Wordbank also groups words into semantically related categories. For example, there is an “animals” category that has the names of various animals. We performed an additional qualitative analysis on this dataset by projecting both the baseline features and the transformer embeddings into a 2-dimensional space using isometric mapping [39]. Figure 3.1 shows the resulting clusters for all feature sets, color-coded by the word’s assigned category in Wordbank. We experimented with other manifold dimensionality reduction algorithms but found that isomap gave the most meaningful clusters overall. Even without any task-specific fine-tuning, *bert-base* is clearly segmenting the words along semantically meaningful dimensions, as words belonging to the same category are consistently grouped

together. It also appears that the space is roughly organized by imageability, which is defined as how easily "words arouse a sensory experience", or in this case how easily the word can be visualized [44]. Abstract concepts (actions, descriptive words, connecting words, etc.) are skewed negative along the x-axis, while concrete concepts (toys, animals, vehicles, etc.) are skewed positive. Previous work has found that imageability and AoA are at least moderately correlated with each other [38], so if BERT has learned to distinguish words by this feature that may partially explain the observed performance. A similar trend is seen in the other transformers, though the clusters are not always grouped in similar locations. We also see this trend using the baseline features, however, the clusters are less compact and closer to each other, suggesting that BERT has learned to distinguish this semantic feature more effectively.

Table 3.4 Matthews correlation on the Wordbank norms. The random baseline gets -0.05 correlation.

Model	baseline	bert-base	bert-large	roberta-base	roberta-large
Logistic Regression	-0.01	-0.01	0.08	0.00	0.05
Decision Tree	0.02	-0.03	0.07	0.01	0.06
SVC	0.07	0.01	0.14	0.03	0.03
KNN	0.01	-0.05	0.08	0.01	0.02

3.2 Discussion

Age of acquisition is an important psycholinguistic property known to influence lexical processing. While much work over the years has studied how distributional models can be used to model AoA, the most recent advances in NLP are seldom used. In this chapter, we have

addressed this deficit by exploring the use of state-of-the-art transformers to model AoA. Our results overall are promising, but not sufficient to definitively state that transformers are superior to the baseline psycholinguistic features. On the Kuperman norms, we were able to achieve better correlation using the transformers for many of the models we tested, but the best performing model used the baseline features. Our t-test confirmed that the higher correlation obtained using the baseline features was statistically significant.

Results on Wordbank are also unclear, while the transformers achieve the highest correlation on this dataset, the best correlation was still quite low. Not surprisingly, the transformers achieve consistently better performance than the random baseline on both datasets, which suggests they must encode at least some features predictive of AoA. We generally observed that the larger versions of the transformers outperformed their smaller counterparts. This was expected, since adding more encoder layers and self-attention heads usually improves a transformer's predictive capabilities. However, while RoBERTa is theoretically a superior architecture to BERT, we found that the RoBERTa models performed consistently worse than BERT. This is in line with prior work in interpretability which has found RoBERTa does not always perform better than BERT on diagnostic tasks [45]. It is reasonable to think that not all transformers are equally good at modeling psycholinguistic properties, and these results suggest that BERT may be a better model for predicting such properties of language. We cannot be certain, however, since other properties (concreteness, affectiveness, etc.) were not examined.

Probably our most interesting results were the visualizations of the word embedding spaces. The transformers clearly showed more meaningful organization of the words than the baseline features, which makes it more surprising the transformers could not consistently achieve the highest correlation.

Chapter 4: Using Transformers to Predict Psychometric Properties

The results from the previous chapter indicate that we can use transformers to model psycholinguistic properties of language with reasonable predictive power. However, some of our results were mixed, making it difficult to determine whether transformers are actually better than previous models. What then is the extent of a transformer's ability to model psycholinguistic phenomena? Are there certain properties of language transformers are better at capturing than others? In this chapter, we investigate this question using experiments conducted on the GLUE diagnostic.

GLUE and its extension SuperGLUE [65] are suites of NLP tasks designed to test the general capabilities of language models across a wide range of different domains. The main tasks within GLUE test understanding of sentiment [66], semantic similarity [67], and natural language inference [68]. However, most relevant to this work is the benchmark's diagnostic, which is a small dataset created by NLP experts meant to evaluate the fundamental linguistic reasoning capabilities of models. The diagnostic consists of questions covering four main categories of linguistic competencies: *lexical semantics, predicate-argument structure, logic*, and *knowledge and common sense*. These categories are further divided into multiple sub-categories, each of which covers a specific and interesting phenomenon in language. For instance, within *logic*, there is a category called *propositional structure*, which tests a model's ability to reason over propositional logic occurring in natural language.

We gathered results both from transformers and human participants on this diagnostic. We compared the neural models against the human results using various psychometric measures.

Since the diagnostic tests many different linguistic skills, our hope was that using tools from psychometrics would allow us to study how well transformers correlate with human responses on each specific skill, giving us a better understanding of performance.

4.1 Methodology

4.1.1 Language Model Experiments

To evaluate our models, we selected a subset of the diagnostic questions that were a member of only one sub-category. This ensured that the questions were testing a single specific cognitive skill. In most cases, there were enough questions in a single sub-category that we could just drop all questions that belonged to multiple categories. However, there were three cases where we needed to merge members of one category into another category to prevent overlap:

4.1.1.1 Merged GLUE Categories

- *negation* and *double negation* questions were merged into *morphological negation*.
- *symmetry/collectivity* was merged into *core arguments*.
- Questions in both *world knowledge* and *named entities* were merged into *named entities*.

Each of these was a case where the sub-categories tested closely related skills, and thus overlapped highly. This process gave us a set of 811 questions from the diagnostic, which we used to evaluate the linguistic capabilities of our models. We gathered performance metrics on the diagnostic for a wide array of transformer models, including BERT [10], RoBERTa [11], T5 [69], ALBERT [70], XLNet [71], ELECTRA [72], Longformer [73], SpanBERT [74], DeBERTa [75], and ConvBERT [76]. Each of these models differed from the others along one or more factors, including underlying architecture, pre-training objective and data, or the general category

the model belongs to (autoregressive, autoencoding, or sequence-to-sequence). This allows us to treat each model as effectively being a different individual, which might have a radically different cognitive profile from its counterparts. We use the Transformers implementation of all models listed above. We experimented with different publicly available versions of each of these models, with differing numbers of parameters, transformer layers, and self-attention heads. As a baseline, we use an LSTM [78] architecture implemented in PyTorch, which was specifically designed for SNLI.[3] We use 50- dimensional Glove [79] word embeddings for the LSTM. We ran a non-exhaustive grid search to generate a population LSTM baselines, changing the number of recurrent layers, size of the hidden layers, learning rate, and dropout [109] probability. To evaluate the neural models, we experimented with four different training regimes:

4.1.1.2 Model Training Regimes

- Zero shot: The model is initialized with random weights in the hidden layers and is evaluated on the diagnostic without any training. This is meant to test whether there is any property of the architecture itself which is useful for solving the diagnostic.
- Pre-train, no finetune: The model is pre-trained but not finetuned for NLI.
- No pre-train, finetune: The model weights are initialized randomly, but we finetune the model on the NLI datasets before evaluating it.
- Pre-train and finetune: The standard way to evaluate models on the diagnostic, where it is pre-trained and then finetuned on the NLI datasets.

We use the SNLI [80], MNLI [81], and ANLI [82] training and dev sets to finetune our models, using Matthews correlation as the evaluation metric. When finetuning, we performed one trial that included all three datasets, and then a trial using just SNLI and MNLI. Since ANLI

[3] https://github.com/pytorch/examples/tree/master/snli

is a challenging task, we wished to study its effect separately from the other two tasks. We finetuned our models on these datasets for between 5 to 15 epochs, stopping whenever we found further training did not improve correlation on the dev set. We used a learning rate of $1*10^{-5}$ and a max sequence length of 175. We found these settings allowed our models to get consistently strong results on the combined dev sets of our NLI datasets, always achieving a Matthews correlation of at least 0.5, and often 0.7 or higher, indicating that they had learned to solve the NLI task well. It is important to note that our goal in finetuning was not necessarily to optimize the model's performance on these NLI datasets. Rather, since the diagnostic is formatted as an NLI task, we hoped that finetuning would help the models to learn what the output labels should be.[4]

For BERT, we experimented with both the pre-trained models from [10], and a BERT model we trained from scratch. Our BERT model had an identical architecture to *bert-base*, and was pre-trained on Google's One Billion Words corpus [89]. We trained the model for 52 epochs, using a learning rate $4*10^{-5}$, a max sequence length of 128, a warmup ratio of 0.01 and a weight decay of 0.01. We used Transformers to pretrain this model, and saved every end of epoch checkpoint. We then used every 10^{th} checkpoint as a separate individual to gather diagnostic data on, using the previously mentioned training regimes. All training was done on three Tesla V100 GPUs with 32GB of memory each. Wherever possible, we used Apex[5] to speed up training.

In summary, our approach allowed us to vary the underlying architecture, number of parameters, and amount of data used in each trial. Because these changes might lead to radically

[4] Because T5 is a sequence-to-sequence model, finetuning is necessary, otherwise its outputs would be completely random.
[5] https://github.com/NVIDIA/apex

different performance on the diagnostic, our hope was that this would lead to a large variation in the profiles of the various transformers tested.

4.1.2 Human Studies

In addition to evaluating the underlying linguistic skills of language models, we also wished to gather the same data for human participants. To do this, we recruited workers on Amazon Mechanical Turk[6] to complete a subset of the diagnostic questions. While these platforms make conducting large scale human studies convenient, there are also well-documented problems with participants not completing surveys in good faith, and instead adversarially answering questions to complete them as quickly as possible [83-85]. This was especially problematic for our experiments because we could not filter out participants just because they performed poorly on a given category. Therefore, we carefully designed our human studies based on recommendations from prior work, so that we could detect and filter out bad faith participants.

We first gathered "attention check" questions, sometimes called "instructional manipulation checks" [86], which were very easy questions used to assess whether participants were paying attention to the survey. We used questions from the ChaosNLI dataset [87], which gathered over 450,000 human annotations on questions from SNLI and MNLI. Each question in ChaosNLI was annotated by 100 different workers. The large number of responses per question gave us more confidence that, if the inter-annotator agreement for a given question was high, that question was likely extremely easy to solve. These questions were also in the exact same format as the diagnostic questions, which made it less likely that workers would realize they were being

[6] https://www.mturk.com

given an attention check. We gathered 36 questions from ChaosNLI where the agreement for the correct label was at least 90%. The labels for this subset were perfectly balanced. These were enough questions to ensure that each phase of our trials used a unique set of attention check questions.

The human studies were split up into 5 phases, and workers who did sufficiently well in one phase were given a qualification to continue to the next phase:

4.1.2.1 Human Study Phases

- On-boarding: A qualifying HIT open to any worker located in the United States, who had completed at least 50 HITs with an approval rating of at least 90%. The HIT consisted of 5 attention check questions, given to each worker in the same order. We gathered responses from up to 200 workers and paid each participant $0.50.
- Phase 1: Included questions from *morphological negation,* and 3 attention checks. We gathered up to 45 responses and paid workers $3.60.
- Phase 2: Included questions from *lexical entailment* and *prepositional phrases,* as well as 6 attention checks. We gathered up to 36 responses and paid workers $7.20.
- Phase 3: Included questions from *quantifiers* and *propositional structure,* as well as 6 attention checks. We gathered up to 27 responses and paid workers $7.20.
- Phase 4: Included questions from *richer logical structure* and *world knowledge,* as well as 6 attention checks. We gathered responses from all accepted workers from Phase 3, and paid workers $7.20.

We selected these sub-categories based on how much the average performance of the language models improved after pre-training and finetuning since a substantial performance improvement indicated this category was actually solvable by the models. Some categories, like

restrictivity and *core arguments*, showed only a very small performance improvement even after being fully pre-trained and finetuned, and thus were excluded.

In each phase, questions were randomly ordered, except for attention checks which were spread evenly at specific locations throughout the survey. We used Qualtrics[7] to create the surveys for each HIT and collect the responses. Participants were first presented with instructions for the task and some examples, which were based on the instructions originally given to annotators for the MNLI dataset.[8] We used the label *can't tell from information provided* instead of *maybe correct* for the *neutral* label. The diagnostic questions from each category were a randomly chosen subset of the questions tested on the language models for that category. For each question, workers also had to provide a short justification for why they believed their answer was correct, which was used to help filter out bad faith participants. To validate the responses to our surveys, we developed the following authentication procedure based on insights from prior work:

4.1.2.2 Human Studies Codebook

- Stage 1: Look for duplicate IPs or worker IDs, indicating that the worker took the HIT more than once. If there are any, reject the second and future HITs, but keep the first submission.
- Stage 2: If the worker's overall score was less than 40%, reject the HIT. If their overall score was greater than 60%, accept the HIT. For workers who scored between 40% and 60%, we still rejected the HIT if they got less than 75% of the attention checks correct. Otherwise, proceed to stage 3.

[7] https://www.qualtrics.com
[8] https://nyu-mll.github.io/GLUE-human-performance/mnli.html

- Stage 3: Finally, we examined the justifications of all workers not previously rejected to decide whether workers should be accepted. Here we were looking for simple, but clear, reasons for why workers chose their answer. We included this step because we found that workers sometimes provided nonsensical justifications for their answers even when they did well on the survey, making it hard to tell whether they were truly paying attention. We checked that the justifications appeared relevant to the question (some workers seemed to paste random text from other websites into the justification), that they did not paste part of the question for their justification, that they did not use the same justification for every question regardless of whether it was relevant, and that they did not use short nonsensical phrases for their justification (some workers simply wrote “good” or “nice” as their justification). Workers who gave responses that were not questionable based on these criteria were accepted.

Using the above procedure, we gathered human responses for all the diagnostic categories of interest. In total, 27 out of the original 200 workers passed all phases of the human study, and we used responses from these workers for our experiments.

4.2 Results

Using the data from the previous steps, we performed several experiments to study the psychometric properties of the response data. Our main interest here is determining if there are any commonalities between the human data and the transformer data. If so, it would suggest transformers may be able to model some cognitive properties of humans, and therefore be useful diagnostic tools in education and other domains.

First, we study how effective the transformers are at modeling simple problem difficulty, which is defined as how many members of each population (transformer, human, etc.) get a

given question on the diagnostic correct. For each question *i* given to the human participants, we calculated the percentage of the humans who got that question right. We then did the same for the transformer and LSTM models. As an additional baseline, we also include a random classifier that simply chooses a label randomly for each question. Table 4.1 shows the results of this experiment. We see that transformers generally correlate much better with the human data than either the LSTMs or the random baseline. The main exceptions are *morphological negation* and *richer logical structure*, where they do not achieve significant correlation. In both of these categories, humans perform considerably better than either transformers or LSTMs on average than they do on any other category (above 90% in both cases versus 60% and 43% respectively). These categories also test skills transformers are known to have difficulty mastering. *Richer logical structure* involves numerical reasoning, and as discussed in Chapter 2 transformers do not generally form good representations of numbers. *Morphological negation* involves reasoning over negation, which transformers are notoriously bad at [105]. Overall, results from these experiments indicate transformers could be useful as a model of problem difficulty in humans.

One important idea which DCM psychometric models build on is that two items (questions) that rely on the same underlying skills should have similar chances of being answered correctly. To determine whether two items have the same skill requirements, we can use inter-item correlation (IIC). A high IIC indicates the items have similar skills, and is calculated by taking the correlation (Spearman or Pearson) between all pairs of items.

In our second experiment, we used IIC as a distance metric for clustering the diagnostic problems. This allowed us to study how closely the transformer's estimates of the required skills

for a given question matched the same estimates obtained using human data. To do this, we converted the IIC *c* into a distance metric by taking 1 – *c*.

Table 4.1 Spearman correlation and p-value for transformer, LSTM, and random estimates of problem difficulty, compared to the human estimates.

Category	Transformer, p	LSTM, p	Random, p
Morphological Negation	0.05, > 0.5	0.03, > 0.5	0.93, > 0.5
Prepositional Phrases	0.78, < 0.001	0.23, < 0.5	-0.33, < 0.5
Lexical Entailment	0.68, < 0.01	0.31, < 0.5	-0.26, < 0.5
Quantifiers	0.50, < 0.1	-0.33, < 0.5	0.11, > 0.5
Propositional Structure	0.89, < 0.001	-0.25, < 0.5	-0.20, < 0.5
Richer Logical Structure	0.09, < 0.1	0.32, < 0.5	-0.57, < 0.05
World Knowledge	0.85, < 0.001	-0.06, > 0.5	-0.15, > 0.5

We applied k-medoids clustering to these data points, and used the silhouette method [103] to find the optimal value of k. We calculated the optimal clusters on each sub-category, using the human, transform, LSTM, and random data separately. To check how well the resulting clusters matched across populations, for each pair of items *i*,*j* we define $C_{i,j} = 1$ if *i* and *j* are in the same cluster, and 0 otherwise. We calculated these scores for every populations, and finally take the Pearson correlation between the resulting vectors from different populations. Results are in Table 4.2. In most cases, we see moderate and significant correlation with the human data using the transformer models. The correlation for LSTM and random models is consistently

insignificant, which further suggests transformers provide a better way to model cognitive profiles. The one exception to this trend is on *Morphological negation*, where we again see weak and insignificant correlation using transformers. This time however, the LSTMs achieve significant correlation on this category, unlike all the others.

Table 4.2 Pearson correlation and p-values for how closely the transformer, LSTM, and random models match the clusters for the human responses.

Category	Transformer, p	LSTM, p	Random, p
Morphological Negation	0.18, < 0.5	0.40, < 0.01	-0.14, < 0.5
Prepositional Phrases	0.31, < 0.01	-0.15, < 0.5	-0.01, > 0.5
Lexical Entailment	0.64, < 0.001	-0.03, > 0.5	-0.16, < 0.5
Quantifiers	0.22, < 0.05	0.001, > 0.5	0.06, > 0.5
Propositional Structure	0.51, < 0.001	0.03, > 0.5	0.04, > 0.5
Richer Logical Structure	0.60, < 0.001	0.00, > 0.5	0.04, > 0.5
World Knowledge	0.37, < 0.001	0.00, > 0.5	-0.09, < 0.5

4.3 Discussion

Our results on these experiments indicate transformers are better at predicting some psychometric properties, compared to non-contextual baselines. The correlations on individual categories reveal interesting patterns, for instance, transformers do not model *Morphological negation* well at all. This, along with prior work, strongly indicates that transformers do not handle negation well, and suggests an avenue for how they can be improved. Transformers

generally perform well on the other categories, consistently achieving modestly strong and significant correlation. Compared to other baselines which seldom achieve above random correlation, they clearly correlate much better with human psychometric data.

It is important to stress however that further study is needed before we can draw strong conclusions. We had data for only 27 human participants and 111 neural models for these studies, which is a very limited sample size. We found that, for transformers, individual variances on problem difficulty did not correlate at all, which indicates that our population of neural models may not be as diverse as we hoped. Part of the problem may be that due to the immense computing resources required to finetune very large transformers (e.g. T5-11B), the largest transformer we trained had only about 700 million parameters. While transformers are a clear improvement over the previous generation of non-contextual models, compared to humans they are still absurdly inefficient learners that require much more data to achieve robust knowledge. Therefore, we may need data for much larger transformers before more interesting patterns start to emerge. Regardless, we believe our findings are quite encouraging, and are optimistic that further study will continue to reveal interesting parallels between the transformer and human responses.

Chapter 5: Conclusion

In this work, we have investigated the use of transformers to model psycholinguistic and psychometric properties of language. We examined experimental results from several related lines of research, the first on using transformers to model Age of Acquisition, and the second on applying psychometrics models to measure how the linguistic capabilities of transformers relates to humans. Unlike prior work in interpretability which tended to focus on a handful of models, we experimented with as many transformers as possible to draw broad conclusions about the capabilities of this class of architecture. Our results are quite encouraging, as transformers consistently either achieved superior results to the baselines examined or did no worse.

There are several avenues worth exploring in future work. Regarding the AoA studies, it is possible that applying dimensionality reduction to the transformer features before using them for training may improve the performance of our models, especially since our isomap projections revealed that transformers were clearly segmenting the Wordbank norms in a semantically meaningful fashion. We also have not established how transformers compare against other common distributional models. Finally, we have not determined whether fine-tuning the transformers on AoA data can boost downstream performance. We plan to investigate these possibilities in the follow-up experiments.

Regarding the psychometrics experiments, we plan to increase the size of our populations, by gathering data from more workers and more neural models. For the human studies, we also plan to control age, income, and education level to gather a representative sample, whereas in the studied presented here that data was not collected. We believe that using

mixture models, in particular Google's switch transformer [104], could allow us to scale up to billions of parameters while still being computationally tractable. Like what was done with *bert-base*, we plan to pre-train a switch transformer from scratch on the Colossal Clean Crawled Corpus [69], which contains 800gb of English text. We will scale up the switch model to billions of parameters, as many as possible without running out of memory, and save each end of epoch checkpoint. We hope that a much larger model than the others tested will increase the diversity of the transformers to the point that we can find more interesting patterns in the psychometric data. We will gather this data for the switch transformer, along with other transformers which we did not examine. We will also expand the diversity of our finetuning configurations by finetuning using more partitions of our 3 training sets, for instance using just SNLI or MNLI. As the size of our population increases, we expect that we will be able to use more sophisticated psychometric constructs from item response theory and diagnostic classification models, both of which require much larger sample sizes to get reliable results.

References

[1] Kousta, S.T., Vigliocco, G., Vinson, D., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14.

[2] Paivio, A., Walsh, M., & Bons, T. (1994). Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1196.

[3] Juhasz, B. (2005). Age-of-acquisition effects in word and picture identification. *Psychological bulletin*, *131*(5), 684.

[4] Brysbaert, M., & Ellis, A. (2016). Aphasia and age of acquisition: are early-learned words more resilient? *Aphasiology*, *30*(11), 1240–1263.

[5] Gerhand, S., & Barry, C. (1999). Age of acquisition, word frequency, and the role of phonology in the lexical decision task *Memory & cognition*, *27*(4), 592–602.

[6] Brysbaert, M., Keuleers, E., & Mandera, P. (2014). A plea for more interactions between psycholinguistics and natural language processing research *Computational Linguistics in the Netherlands Journal*, *4*, 209–222.

[7] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis *Journal of the American society for information science*, *41*(6), 391–407.

[8] Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence *Behavior research methods, instruments, & computers*, *28*(2), 203–208.

[9] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111-3119.

[10] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding *arXiv preprint arXiv:1810.04805*.

[11] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach *arXiv preprint arXiv:1907.11692*.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, ., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010).

[13] Frank, M., Braginsky, M., Yurovsky, D., & Marchman, V. (2017). Wordbank: An open repository for developmental vocabulary data *Journal of child language, 44*(3), 677.

[14] Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words *Behavior research methods, 44*(4), 978–990.

[15] Hills, T., Maouene, J., Riordan, B., & Smith, L. (2010). The associative structure of language: Contextual diversity in early word learning *Journal of memory and language, 63*(3), 259–273.

[16] Steyvers, M., & Tenenbaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth *Cognitive science, 29*(1), 41–78.

[17] Stella, M. (2019). Modelling early word acquisition through multiplex lexical networks and machine learning *Big Data and Cognitive Computing, 3*(1), 10.

[18] Hills, T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological science, 20*(6), 729–739.

[19] Braginsky, M., Yurovsky, D., Marchman, V., & Frank, M. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages.. In *CogSci*.

[20] Casas, B., Catala, N., Ferrer-i-Cancho, R., Hernández-Fernández, A., & Baixeries, J. (2018). The polysemy of the words that children learn over time *Interaction Studies, 19*(3), 389–426.

[21] Stella, M., Beckage, N., Brede, M., & De Domenico, M. (2018). Multiplex model of mental lexicon reveals explosive learning in humans *Scientific reports, 8*(1), 1–11.

[22] Russo, I. (2020). Guessing the Age of Acquisition of Italian Lemmas through Linear Regression. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 43–48).

[23] Chang, L., & Deák, G. (2020). Adjacent and Non-Adjacent Word Contexts Both Predict Age of Acquisition of English Words: A Distributional Corpus Analysis of Child-Directed Speech *Cognitive Science, 44*(11).

[24] Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *Quarterly Journal of Experimental Psychology, 68*(8), 1623–1642.

[25] Mohler, M., Tomlinson, M., Bracewell, D., & Rink, B. (2014). Semi-supervised methods for expanding psycholinguistics norms by integrating distributional similarity with the structure of WordNet. In *LREC* (pp. 3020–3026).

[26] Miller, G. (1998). *WordNet: An electronic lexical database*. MIT press.

[27] Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes *Behavior research methods, 44*(4), 998–1006.

[28] Bhatia, S., & Richie, R. (2020). Transformer Networks of Human Concept Knowledge.

[29] Fenson, L. (2002). *MacArthur Communicative Development Inventories: User's guide and technical manual*. Paul H. Brookes.

[30] Schock, J., Cortese, M., Khanna, M., & Toppi, S. (2012). Age of acquisition estimates for 3,000 disyllabic words *Behavior Research Methods, 44*(4), 971–977.

[31] Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words *Behavior Research Methods, Instruments, & Computers, 33*(1), 73–79.

[32] Stadthagen-Gonzalez, H., & Davis, C. (2006). The Bristol norms for age of acquisition, imageability, and familiarity *Behavior research methods, 38*(4), 598–605.

[33] Cortese, M., & Khanna, M. (2008). Age of acquisition ratings for 3,000 monosyllabic words *Behavior Research Methods, 40*(3), 791–794.

[34] Brysbaert, M., & New, B. (2009). Moving beyond Kuvcera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English *Behavior research methods, 41*(4), 977–990.

[35] Barbaresi, A. (2014). *Language-classified Open Subtitles (LACLOS): download, extraction, and quality assessment*. (Doctoral dissertation, BBAW).

[36] Pearson, K.. (1895). Notes on Regression and Inheritance in the Case of Two Parents Proceedings of the Royal Society of London, 58, 240-242.

[37] Spearman, C. (1961). The proof and measurement of association between two things.

[38] Cortese, M., & Khanna, M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words *Quarterly Journal of Experimental Psychology, 60*(8), 1072–1082.

[39] Tenenbaum, J., De Silva, V., & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction *science, 290*(5500), 2319–2323.

[40] Sheynin, O. (1995). Helmert's work in the theory of errors *Archive for history of exact sciences, 49*(1), 73–104.

[41] Kolovou, A., Iosif, E., & Potamianos, A. (2017). Lexical and affective models in early acquisition of semantics.. In *WOCCI* (pp. 40–45).

[42] Alhama, R., Rowland, C., & Kidd, E. (2020). Evaluating word embeddings for language acquisition. In *(Online) Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2020)* (pp. 38–42).

[43] Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme *Biochimica et Biophysica Acta (BBA)-Protein Structure, 405*(2), 442–451.

[44] Dellantonio, S., Job, R., & Mulatti, C. (2014). Imageability: now you see it again (albeit in a different form) *Frontiers in psychology, 5*, 279.

[45] Talmor, A., Elazar, Y., Goldberg, Y., & Berant, J. (2020). oLMpics-on what language model pre-training captures *Transactions of the Association for Computational Linguistics, 8*, 743–758.

[46] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python *Journal of Machine Learning Research, 12*, 2825–2830.

[47] Russell Richie, Wanling Zou, & Sudeep Bhatia (2019). Predicting High-Level Human Judgment Across Diverse Behavioral Domains *Collabra: Psychology, 5*(1).

[48] Stella, M., & Brede, M. (2016). Mental lexicon growth modelling reveals the multiplexity of the English language. In *Complex Networks VII* (pp. 267-279). Springer, Cham.

[49] Mark D. Reckase (2009). Multidimensional Item Response Theory Models *Multidimensional Item Response Theory.*

[50] John Sessoms, & Robert A. Henson (2018). Applications of Diagnostic Classification Models: A Literature Review and Critical Commentary *Measurement: Interdisciplinary Research and Perspectives, 16.*

[51] Yi-Hsin Chen (2012). Cognitive diagnosis of mathematics performance between rural and urban students in Taiwan *Assessment in Education: Principles, Policy & Practice, 19.*

[52] Turing, A. (1950). Computing Machinery and Intelligence *Mind, 59*, 433–460.

[53] Xue, K. (2019). Computational diagnostic classification model using deep feedforward network based semi-supervised learning. In *25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) Workshop on Deep Learning for Education*.

[54] Ahmad, F., Abbasi, A., Li, J., Dobolyi, D., Netemeyer, R., Clifford, G., & Chen, H. (2020). A deep learning architecture for psychometric natural language processing *ACM Transactions on Information Systems (TOIS), 38*(1), 1–29.

[55] Abbasi, A., Dobolyi, D., & Netemeyer, R. (2020). Constructing a Testbed for Psychometric Natural Language Processing *arXiv preprint arXiv:2007.12969*.

[56] Lalor, J., Wu, H., & Yu, H. (2019). Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (pp. 4240).

[57] Lalor, J., Wu, H., Munkhdalai, T., & Yu, H. (2018). Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (pp. 4711).

[58] Sedoc, J., & Ungar, L. (2020). Item Response Theory for Efficient Human Evaluation of Chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 21–33).

[59] Bringsjord, S. (2011). Psychometric artificial intelligence *Journal of Experimental & Theoretical Artificial Intelligence, 23*(3), 271–277.

[60] Bringsjord, S., & Schimanski, B. (2003). What is artificial intelligence? Psychometric AI as an answer. In *IJCAI* (pp. 887–893).

[61] Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey *Connection science, 15*(4), 151–190.

[62] De La Torre, J. (2009). DINA model and parameter estimation: A didactic *Journal of educational and behavioral statistics, 34*(1), 115–130.

[63] Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods, 11*(3), 287.

[64] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353–355).

[65] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems*Advances in Neural Information Processing Systems, 32*.

[66] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP* (pp. 1631–1642).

[67] Agirre, Eneko and M`arquez, Llu'is and Wicentowski, Richard (2007). *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics.

[68] Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., & Szpektor, I. (2006). The second pascal recognizing textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*.

[69] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, & Peter J. Liu (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer *Journal of Machine Learning Research, 21*(140), 1–67.

[70] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, & Radu Soricut (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *ICLR 2020 : Eighth International Conference on Learning Representations*.

[71] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, & Quoc V. Le (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems* (pp. 5753–5763).

[72] Kevin Clark, Minh-Thang Luong, Quoc V. Le, & Christopher D. Manning (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR 2020 : Eighth International Conference on Learning Representations*.

[73] Iz Beltagy, Matthew E. Peters, & Arman Cohan (2020). Longformer: The Long-Document Transformer *arXiv preprint arXiv:2004.05150*.

[74] Joshi, M., Chen, D., Liu, Y., Weld, D., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans *Transactions of the Association for Computational Linguistics, 8*, 64–77.

[75] Pengcheng He, Xiaodong Liu, Jianfeng Gao, & Weizhu Chen (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention *arXiv preprint arXiv:2006.03654*.

[76] Jiang, Z.H., Yu, W., Zhou, D., Chen, Y., Feng, J., & Yan, S. (2020). ConvBERT: Improving BERT with Span-based Dynamic Convolution *Advances in Neural Information Processing Systems, 33*.

[77] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, & Alexander M. Rush (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics.

[78] Sepp Hochreiter, & Jürgen Schmidhuber (1997). Long short-term memory *Neural Computation, 9*(8), 1735–1780.

[79] Jeffrey Pennington, Richard Socher, & Christopher Manning (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543)

[80] Bowman, S., Angeli, G., Potts, & Manning, C. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

[81] Williams, A., Nangia, N., & Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1112–1122). Association for Computational Linguistics.

[82] Nie, D. (2020). Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

[83] Adam J. Berinsky, Michele F. Margolis, & Michael W. Sances (2014). Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys *American Journal of Political Science, 58*(3), 739–753.

[84] Adam J. Berinsky, Michele F. Margolis, & Michael W. Sances (2016). Can we turn shirkers into workers *Journal of Experimental Social Psychology, 66*, 20–28.

[85] Melissa G. Keith, Louis Tay, & Peter D. Harms (2017). Systems Perspective of Amazon Mechanical Turk for Organizational Research: Review and Recommendations. *Frontiers in Psychology, 8*, 1359.

[86] David J. Hauser, & Norbert Schwarz (2015). It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks *SAGE Open, 5*(2).

[87] Yixin Nie, Xiang Zhou, & Mohit Bansal (2020). What Can We Learn from Collective Human Opinions on Natural Language Inference Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9131–9143).

[88] Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in bertology: What we know about how bert works *Transactions of the Association for Computational Linguistics, 8*, 842–866.

[89] Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[90] Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601).

[91] Wallace, E., Wang, Y., Li, S., Singh, S., & Gardner, M. (2019). Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5310–5318).

[92] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019, November). Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2463-2473).

[93] Gupta, A., Kvernadze, G., & Srikumar, V. (2021). BERT & Family Eat Word Salad: Experiments with Text Understanding *arXiv preprint arXiv:2101.03453*.

[94] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, & Roger Wattenhofer (2020). On Identifiability in Transformers. In *International Conference on Learning Representations*.

[95] Wiegreffe, S., & Pinter, Y. (2019). Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 11–20).

[96] Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., & others (2019). Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[97] Steinberg, D., & Sciarini, N. (2013). *An introduction to psycholinguistics*. Routledge.

[98] Powers, D. (1983). Neurolinguistics and psycholinguistics as a basis for computer acquisition of natural language *ACM SIGART Bulletin*(84), 29–34.

[99] Chater, M., & others (2001). *Connectionist psycholinguistics*. Greenwood Publishing Group.

[100] Karmakar, P., Teng, S., & Lu, G. (2021). Thank you for Attention: A survey on Attention-based Artificial Neural Networks for Automatic Speech Recognition *arXiv preprint arXiv:2102.07259*.

[101] Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors *nature*, *323*(6088), 533–536.

[102] LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Backpropagation applied to handwritten zip code recognition *Neural computation*, *1*(4), 541–551.

[103] Peter J. Rousseeuw (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis *Journal of Computational and Applied Mathematics, 20*, 53-65.

[104] Fedus, W., Zoph, B., & Shazeer, N. (2021). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity *arXiv preprint arXiv:2101.03961*.

[105] Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models *Transactions of the Association for Computational Linguistics, 8*, 34–48.

[106] Sun, Y., & Fisher, R. (2003). Object-based visual attention for computer vision. *Artificial intelligence*, *146*(1), 77-123.

[107] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[108] Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. *stat*, *1050*, 21.

[109] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.