Modeling Age of Acquisition Norms Using Transformer Networks

Antonio Laverghetta Jr. and **John Licato**
Advancing Machine and Human Reasoning (AMHR) Lab
Department of Computer Science and Engineering
University of South Florida
Tampa, FL, USA
{alaverghett,licato}@usf.edu

Abstract

The age at which children acquire words is an important psycholinguistic property for modeling the growth of children's semantic networks. Much work over the years has explored how to effectively exploit statistical models to predict the age at which a word will be acquired, ranging from simple linear regression to LSA and skip-gram. However, thus far no work has explored whether transformers are any better at modeling word acquisition, despite the superior performance they have achieved on a wide variety of natural language processing (NLP) benchmarks. In this paper, we explore using several transformer models to predict the age of acquisition norms for several datasets. We evaluate the quality of our models using various experiments based on prior work and compare the transformers against two baseline models. We obtain promising results overall, as the transformers can outperform the baselines in most cases.

Introduction

The normative age of acquisition (AoA) of a word is the age at which the word is typically learned, and is often defined as the age at which 50% of children are said (by caretakers familiar with them) to be able to understand or produce the word. Within psycholinguistics, AoA is thought to be an important variable in predicting the lexical processing of words, along with concreteness and affectiveness (Paivio, Walsh, and Bons 1994; Zevin and Seidenberg 2002; Kousta et al. 2011). For instance, AoA is thought to affect how fast words are read (Juhasz 2005), and how fast pictures can be named (Brysbaert and Ellis 2016). 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 combining word embeddings with human judgment ratings can allow us to model human perceptions related to health behavior and risks (Richie, Zou, and Bhatia 2019).

Early studies on AoA were small in scale, focusing on a handful of words picked for certain properties they possessed (Gerhand and Barry 1999). While this 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, or rather are special cases (Brysbaert, Keuleers, and Mandera 2014). The difficulties of gathering AoA norms by hand have led to much interest in applying distributional models to extrapolate the AoA of new words. Work in this area has so far focused on older non-contextual models, especially LSA (Deerwester et al. 1990), HAL (Lund and Burgess 1996), and skip-gram (Mikolov et al. 2013). Within the NLP community, however, these models have been eclipsed by deep contextual models. Models based on transformers (Vaswani et al. 2017) have achieved impressive performance on a wide variety of NLP benchmarks, often surpassing their non-contextual counterparts. An interesting question then is whether transformers can do any better than older distributional models at modeling acquisition norms.

In this paper, we investigate using transformers to predict acquisition norms for English words, using techniques that have been shown to work well in prior work. We use BERT (Devlin et al. 2018) and RoBERTa (Liu et al. 2019), two popular transformers. BERT is probably the most well-known transformer architecture, as it introduced pre-training objectives that have become common. RoBERTa 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.

We perform our experiments using two AoA datasets. The first is Kuperman's AoA ratings (Kuperman, Stadthagen-Gonzalez, and Brysbaert 2012), which contains acquisition norms for over 30,000 English words. The original dataset was later expanded to include data from several other studies (Bird, Franklin, and Howard 2001; Stadthagen-Gonzalez and Davis 2006; Cortese and Khanna 2008; Schock et al. 2012), bringing the total size up to over 50,000 words. The second dataset comes from Wordbank (Frank et al. 2017), which is a database of responses to MacArthur-Bates Communicative Development Inventory (CDI) (Fenson 2002) 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. All code to reproduce our results can be found on Github.[1]

Main Contributions: We present an analysis of two popular transformer models on the task of predicting AoA. To our knowledge, we are the first to investigate the use of BERT for predicting AoA and the first to use RoBERTa to predict any psycholinguistic variable. We find that in the majority of cases the transformers achieve superior performance to the baselines. We hope our work will stimulate further interest in the use of transformers for predicting psycholinguistic properties.

Related Work

Factors that contribute to word acquisition have been studied extensively over the years. It has been shown that word frequency (Steyvers and Tenenbaum 2005), length (Hills et al. 2009), polysemy (Casas et al. 2018), and part of speech (Hills et al. 2010) 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 (Stella and Brede 2016). Inspired by these insights, there has been much work on modeling AoA using machine learning. Stella (2019) 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%. Russo (2020) 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 (Chang and Deák 2020), there has been an interest in using distributional models as a way to model various linguistic feature norms, including AoA. Mandera, Keuleers, and Brysbaert (2015) extrapolated AoA ratings using LSA, HAL, and skip-gram models. They achieved about 73% correlation with human norms using the skip-gram model. Mohler et al. (2014) combined a distributional model with Wordnet (Miller 1998) to create an algorithm for expanding psycholinguistic datasets in a semi-supervised fashion. Bestgen and Vincze (2012) 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 do not examine AoA. Kolovou, Iosif, and Potamianos (2017) used a network-based distributional model to study how affective word features influence early language development. Alhama, Rowland, and Kidd (2020) trained SVD and skip-gram models on child-directed speech, and evaluate 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 Bhatia and Richie (2020), 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.

Methodology

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[2] 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.[3] 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:

1. **Frequency:** How often the word occurs in language. We use the frequency counts of words in the OpenSubtitles database (Barbaresi 2014), since it has been shown this dataset is more suitable for studying psycholinguistic phenomena than other corpora (Brysbaert and New 2009). For words not present in the data, we set the value to 1.
2. **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.
3. **Whether the word is a noun:** In the Kuperman norms this data is already present. For Wordbank since the

[1] https://github.com/Advancing-Machine-Human-Reasoning-Lab/modeling-acquisition-norms

[2] https://github.com/huggingface/transformers

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

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

Table 1: 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 $p > 0.05$.

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 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.

4. **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. 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.

Results

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

Table 2: Results of the t-test on *bert-large* SVR per validation correlations and baseline decision tree correlations.

Kuperman Table 1 shows results on the Kuperman norms. We experimented with a variety of regression models, all implemented in sci-kit learn (Pedregosa et al. 2011). We use Pearson (Pearson 1895) and Spearman (Spearman 1961) 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 tuneable 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:

- *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 the majority of 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 with 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 (Sheynin 1995) 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 2. 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.

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

Table 3: Final Wordbank dataset statistics.

Wordbank For this dataset, we used an evaluation based on prior work which framed AoA as a classification task (Stella 2019). We first bin the Wordbank AoA norms into a set of 3 discrete labels. Table 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 (Matthews 1975) to measure performance. To address the class imbalance, we weighted the

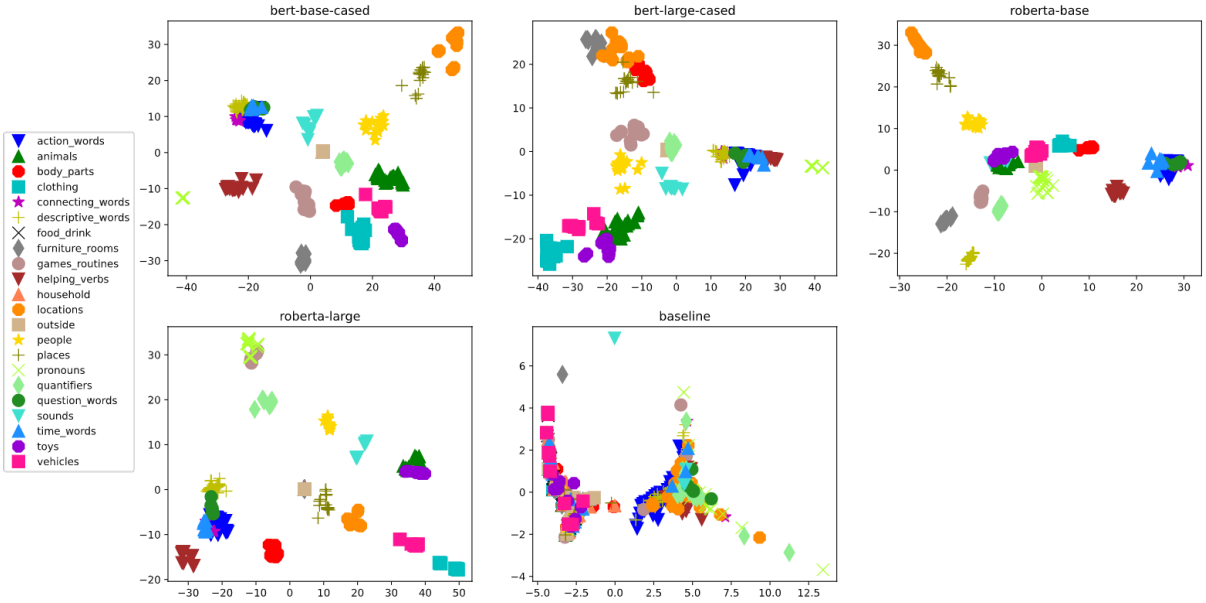

Figure 1: Isomap projections of all featuresets.

input samples to be inversely proportional to the class frequencies. We use the following hyperparameter settings (all others are left at their defaults):

Baseline Features:

- *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

Transformers:

- *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 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.

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 (Tenenbaum, De Silva, and Langford 2000). Figure 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 (Dellantonio, Job, and Mulatti 2014). 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 (Cortese and Khanna 2007), 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.

Conclusion

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 paper, 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

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

Table 4: Matthews correlation on the Wordbank norms. The random baseline gets -0.05 correlation.

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 (Talmor et al. 2020). It's 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 can't 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. There are several avenues worth exploring in future work. First, it's possible that applying dimensionality reduction to the transformer features before using them for training may improve the performance of our models. We also haven't established how transformers compare against other common distributional models, especially LSA and skip-gram. Finally, we haven't determined whether fine-tuning the transformers on AoA data can boost downstream performance. We plan to investigate these possibilities in the follow-up experiments.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award numbers FA9550-17-1-0191 and FA9550-18-1-0052. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

References

Alhama, R. G.; Rowland, C. F.; and Kidd, E. 2020. Evaluating word embeddings for language acquisition. In *(Online) Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2020)*, 38–42. Association for Computational Linguistics (ACL).

Barbaresi, A. 2014. *Language-classified Open Subtitles (LACLOS): download, extraction, and quality assessment*. Ph.D. Dissertation, BBAW.

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

Bhatia, S., and Richie, R. 2020. Transformer networks of human concept knowledge.

Bird, H.; Franklin, S.; and 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.

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

Brysbaert, M., and New, B. 2009. Moving beyond kučera 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.

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

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

Chang, L. M., and Deák, G. O. 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):e12899.

Cortese, M. J., and Khanna, M. 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.

Cortese, M. J., and Khanna, M. M. 2008. Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods* 40(3):791–794.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6):391–407.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Frank, M. C.; Braginsky, M.; Yurovsky, D.; and Marchman, V. A. 2017. Wordbank: An open repository for developmental vocabulary data. *Journal of child language* 44(3):677.

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

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

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

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

Kolovou, A.; Iosif, E.; and Potamianos, A. 2017. Lexical and affective models in early acquisition of semantics. In *WOCCI*, 40–45.

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

Kuperman, V.; Stadthagen-Gonzalez, H.; and Brysbaert, M. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods* 44(4):978–990.

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

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

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

Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405(2):442–451.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Miller, G. A. 1998. *WordNet: An electronic lexical database*. MIT press.

Mohler, M.; Tomlinson, M. T.; Bracewell, D. B.; and Rink, B. 2014. Semi-supervised methods for expanding psycholinguistics norms by integrating distributional similarity with the structure of wordnet. In *LREC*, 3020–3026. Citeseer.

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

Pearson, K. 1895. Notes on regression and inheritance in the case of two parents proceedings of the royal society of london, 58, 240-242.

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

Richie, R.; Zou, W.; and Bhatia, S. 2019. Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology* 5(1).

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*, 43–48.

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

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

Spearman, C. 1961. The proof and measurement of association between two things.

Stadthagen-Gonzalez, H., and Davis, C. J. 2006. The bristol norms for age of acquisition, imageability, and familiarity. *Behavior research methods* 38(4):598–605.

Stella, M., and Brede, M. 2016. Mental lexicon growth modelling reveals the multiplexity of the english language. In *Complex Networks VII*. Springer. 267–279.

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

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

Talmor, A.; Elazar, Y.; Goldberg, Y.; and Berant, J. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics* 8:743–758.

Tenenbaum, J. B.; De Silva, V.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290(5500):2319–2323.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Zevin, J. D., and Seidenberg, M. S. 2002. Age of acquisition effects in word reading and other tasks. *Journal of Memory and language* 47(1):1–29.