Behavior Research Methods
https://doi.org/10.3758/s13428-022-01797-5

Age of Exposure 2.0: Estimating word complexity using iterative models of word embeddings

Robert-Mihai Botarleanu[1] · Mihai Dascalu[1,2] · Micah Watanabe[3] · Scott Andrew Crossley[4] · Danielle S. McNamara[3]

Accepted: 12 January 2022

Abstract
Age of acquisition (AoA) is a measure of word complexity which refers to the age at which a word is typically learned. AoA measures have shown strong correlations with reading comprehension, lexical decision times, and writing quality. AoA scores based on both adult and child data have limitations that allow for error in measurement, and increase the cost and effort to produce. In this paper, we introduce Age of Exposure (AoE) version 2, a proxy for human exposure to new vocabulary terms that expands AoA word lists through training regressors to predict AoA scores. Word2vec word embeddings are trained on cumulatively increasing corpora of texts, word exposure trajectories are generated by aligning the word2vec vector spaces, and features of words are derived for modeling AoA scores. Our prediction models achieve low errors (from 13% with a corresponding R^2 of .35 up to 7% with an R^2 of .74), can be uniformly applied to different AoA word lists, and generalize to the entire vocabulary of a language. Our method benefits from using existing readability indices to define the order of texts in the corpora, while the performed analyses confirm that the generated AoA scores accurately predicted the difficulty of texts (R^2 of .84, surpassing related previous work). Further, we provide evidence of the internal reliability of our word trajectory features, demonstrate the effectiveness of the word trajectory features when contrasted with simple lexical features, and show that the exclusion of features that rely on external resources does not significantly impact performance.

Keywords Age of acquisition · Age of exposure · Word embeddings · Word exposure

Introduction

Age of acquisition (AoA) is a measure of word complexity that attempts to account for the age at which a child learns a word. Previous studies have demonstrated that AoA scores predict readers' text processing and comprehension (Crossley et al., 2017), lexical decision times (Kuperman et al., 2012), and measures of writing quality (Crossley & McNamara, 2011) above and beyond other measures of word complexity such as word frequency. AoA norms have been collected from adult populations (see Alonso et al., 2015; Cortese & Khanna, 2008; Kuperman et al., 2012; Montefinese et al., 2019; Moors et al., 2013; and Stadthagen-Gonzales & Davis, 2006) and child populations (see Álvarez & Cuetos, 2007; Brysbaert & Biemiller, 2017; Chalard et al., 2003; Frank et al., 2017; Grigoriev & Oshhepkov, 2013; Morrison et al., 1997). AoA norms based on adult data allow for errors in the estimation because they are based on adults' perceptions and memory of word learning. On the other hand, AoA norms from child data may result in lists with specific types of words (Morrison et al., 1997), introduce error from parent reporting (Frank et al., 2017), or rely on older datasets that require updating (Brysbaert & Biemiller, 2017). Finally, experiments to compile AoA scores are time-consuming, expensive, and require periodic updates to account for language change.

An automated method for estimating AoA scores has the potential to alleviate the problems of current AoA collection methods. First, an automated method can simulate the word learning process, potentially removing the limitation of relying on adults' memory. Second, automating the process drastically increases the number of words for which an AoA score can be estimated. Finally, automating the process can significantly reduce the time and cost of

✉ Mihai Dascalu
mihai.dascalu@cs.pub.ro

1 University Politehnica of Bucharest, Bucharest, Romania

2 Academy of Romanian Scientists, Bucharest, Romania

3 Arizona State University, Tempe, AZ, USA

4 Georgia State University, Atlanta, GA, USA

Published online: 15 February 2022

collecting human word ratings. Thus, this paper presents an automated method capable of simulating human AoA ratings through a combination of features extracted from word embeddings trained on corpora of increasing sizes together with lexical features (such as hyponym and hypernym tree characteristics, synonym set cardinalities, and statistical attributes of words).

Our work expands on the Word Maturity (WM) approach introduced by Landauer et al. (2011) and on the Age of Exposure (AoE) model described by Dascalu et al. (2015). Our objective is to examine the accuracy of a new version of the AoE algorithm (AoE 2.0) that not only exposes semantic models to an increasing number of texts, but also considers the ordering of text by readability. We assess the strength of our approach by comparing the results of AoE 2.0 to AoE 1.0, WM and to various AoA measures, such as those reported by Kuperman et al. (2012), Bird (Bird et al., 2001), Bristol (Stadthagen-Gonzalez & Davis, 2006), Cortese (Cortese & Khanna, 2008), Shock (Shock et al., 2012), and Morrison (Morrison et al., 1997). Besides features generated from AoE trajectories, we also consider non-AoE lexical features to more effectively simulate human ratings. Our approach is inspired by Crossley et al. (2013), wherein traditional word features such as word length, frequency, hypernymy, and polysemy were used in conjunction with features extracted from WordNet and LSA dimensions in order to predict human word ratings. The addition of these features to the AoE model increases the predictive power of our method, enabling us to better model human AoA ratings.

Age of acquisition

AoA is an estimate of the average age at which a word's meaning is acquired by an average speaker of the language. AoA scores are derived from human ratings and provide a linear scale for comparing words, with terms acquired in later age groups having a higher complexity compared to those acquired earlier. AoA scores correlate with other measures of word complexity such as familiarity and concreteness (Gilhooly & Logie, 1980), word frequency, word length, and ease of production (Kuperman et al., 2012) with higher AoA scores indicating the words are more complex. AoA scores have been demonstrated to significantly predict readers' text processing and comprehension (Crossley et al., 2017), lexical processing speed (Johnston & Barry, 2006), and have been incorporated within models of writing quality (Crossley & McNamara, 2011).

While AoA scores are valuable tools, there are limitations. The first challenge of AoA is the significant human effort required to generate AoA scores, since the lists are usually crowdsourced. For example, Kuperman et al. (2012) generated an English AoA list of 30,000 words through crowdsourcing experiments, with 1960 participants tasked with estimating the age at which they learned a set of words. Other AoA word lists include: Bird (Bird et al., 2001) with 2522 terms collected from 45 British English speakers with a diverse age distribution ($M = 60.7$, $SD = 15.5$); Bristol (Stadthagen-Gonzalez & Davis, 2006) with 3353 terms that constitute norms collected from 20 undergraduate subjects that were combined with previous G&L norms (Gilhooly & Logie, 1980) which provided 1944 AoA scores estimated by 36 student volunteers; Cortese (Cortese & Khanna, 2008) with 2999 terms estimated by 32 undergraduate participants enrolled in a psychology course; and Shock (Shock et al., 2012) with 3000 terms collected from a study with 32 participants enrolled in undergraduate psychology courses.

The crowdsourced process illuminates a second limitation: AoA lists are derived from adults' estimates of when a word was learned. Previous studies have demonstrated adults' ratings correlate with children's ratings and estimates of when words are known (see Ghyselinck et al., 2004). However, these correlations are run on small subsets of words, and often only on nouns. For instance, Morrison et al. (1997) and Chalard et al. (2001) presented approximately 300 object pictures to 14 groups of children and an adult group. The researchers used the children's object naming accuracy to calculate an objective AoA norm for each word. The objective AoA norm, which was derived from the data from children, was highly correlated with traditional AoA scores (Chalard et al., 2003; Morrison et al., 1997). Furthermore, Chalard et al. (2003) reported evidence that objective AoA norms derived from children accounted for more variance in a lexical decision response-time task than adult AoA ratings. These studies provide evidence that AoA norms derived from ratings by children who are learning the language may be more accurate than AoA norms from adults who are mature language users and recalling their exposure to words. However, collecting ratings from children is infeasible at large scales, and previous attempts to collect AoA ratings from children have used picture and object naming tasks to estimate when children learn a word, which limits the targeted AoA scores to nouns (see Álvarez & Cuetos, 2007; Chalard et al., 2003; Grigoriev & Oshhepkov, 2013; Morrison et al., 1997).

There have been previous attempts to estimate AoA norms for large quantities of words from child-derived data. Frank et al. (2017) utilized parent-reports of children's word acquisition, which allows researchers to estimate AoA scores (Braginsky et al., 2016). Brysbaert & Biemiller, 2017 created a test-based AoA estimate from previous studies on children's word acquisition using a regression to convert from grades to AoA scores. They compared the AoA estimates to the Kuperman adult AoA ratings, reporting a correlation of .757. While both

databases are valuable resources, there is the possibility of error in estimation either due to parent-reporting (Frank et al., 2017), or the use of measures not specifically designed to estimate AoA (Brysbaert & Biemiller, 2017).

These limitations demonstrate the need for AoA scores that are easier to collect and based on proxies for the word learning processes. An automated method for estimating AoA scores can potentially alleviate problems inherent in the representation of AoA, as well as the time, cost, and subjectivity of collecting human word ratings. The concept of expanding AoA word lists using Machine Learning models has been explored in the past. Mandera et al. (2015) investigated the usage of various algorithms for constructing semantic spaces, namely LSA, LDA, HAL (Lund & Burgess, 1996), and a skip-gram based method (Mikolov, Le, & Sutskever, 2013c), in combination with either *k*-nearest neighbors or random forest models for extrapolating various subjective ratings. Authors report correlations of up to .737 with Kuperman AoA ratings using two random splits of the data, using a corpus compiled through downloading 204,408 documents containing English film and television subtitles from the Open Subtitles database (http://opensubtitles.org). However, the authors also reported that some of the extrapolation methods introduced artifacts to the data, potentially producing different conclusions from human ratings in the context of two lexical decision tasks. While the method presented by Mandera et al. (2015) has many similarities with the approach adopted in the current study, they relied solely on word embeddings generated through different algorithms using a single training corpus, whereas the current study models AoA by leveraging exposure trajectories which aim to simulate the way language learners acquire words. In the following section, we describe word learning processes, specifically in relation to the impact of increasing exposure to words. We then describe two prior machine learning approaches to simulate increasing word exposure, the Word Maturity model (Biemiller et al., 2014; Landauer et al., 2011) and the AoE 1.0 model (Dascalu et al., 2015).

Word learning and exposure

Children's ability to learn a particular word depends largely on exposure to that word in their language environment (Hills et al., 2010; Hoff & Naigles, 2002; Roy et al., 2015). As the primary source of word input, caregivers essentially manifest the language environment during the child's infancy (Weisleder & Fernald, 2013). Caregiver names and concrete objects are typically among the first learned by children because of the constant exposure infants have to those words in their early language environment (Roy et al., 2015). Furthermore, a greater number of words in infants' language environments predicts larger and more expressive vocabularies several months later, indicating that mere exposure to a greater number of words affords infants opportunities to learn more words (Hoff, 2003; Pan et al., 2005). These studies demonstrate that word learning is dependent on word input from outside sources, which for children is primarily adults (Hills et al., 2010; Hoff & Naigles, 2002).

As children grow, the relationship between word exposure and word learning can be observed both in speech and in written language. Children whose peers are more linguistically skilled are more likely to learn new words compared to children whose peers are less linguistically skilled (Justice et al., 2011; Webb, 1991). In addition to speech, literate children will begin working with written language, and reading will provide children the opportunity to learn new words from the context of written passages (Nagy et al., 1987; Teng, 2019). Both studies demonstrate that exposure to new sources of words, either through peers or written language, affords students the opportunity for greater word learning. Even for adults, the ability to learn words in a second language depends on exposure to new words (Eckerth & Tavakoli, 2012).

Overall, the literature on word learning indicates that increased exposure to words over time is a foundational aspect of word learning; and yet estimating the rate of word exposure throughout childhood is fraught with barriers. Thus, computational simulations of increasing word exposure over time have strong potential to contribute to research and practice in areas related to literacy. Previous efforts to estimate increasing word exposure include the Word Maturity model and the AoE 1.0 model, each described in the following sections.

Word Maturity

The AoE 2.0 model builds on the Word Maturity model (Biemiller et al., 2014; Landauer et al., 2011), an automated model constructed to reflect the word exposure process by deriving word-occurrence patterns from incremental subcorpora of texts. The incremental subcorpora approximate the growing language environment experienced by children. As the subcorpora increase in size, the word-occurrence patterns change based on the repeated exposures to words, and associations of trained words to novel words.

Word Maturity uses Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) to transform words into vector representations within a semantic vector space. LSA uses a term-document occurrence matrix whose dimensionality is reduced using Singular Value Decomposition. The newly formed vector representations hold latent information on the relationships between words such that words which co-occur in similar texts have similar meanings. LSA does not

directly represent the probability of a word appearing in similar documents, but rather it deconstructs the meaning of a paragraph as a sum of the meanings of its component words. Thus, each word is represented using a high-dimensional vector with numerical values, such that the decomposition optimizes the least squares criterion.

Word Maturity quantifies the evolution of any word's complexity throughout a speaker's process of language acquisition and approximates the manner in which language readers are exposed to new texts during their learning process. The method described by Landauer et al. (2011) consists of splitting a text corpus into cumulative subcorpora. For each of these incremental subcorpora, LSA semantic spaces are trained, and then the generated intermediary vector representations are aligned to the mature semantic space via Procrustes rotation. Next, the Word Maturity model computes the averaged vector for all paragraphs containing a word for each intermediate semantic space; these averaged vectors are then used to measure the cosine distance to the mature vector representation. These cosine values for the incremental subcorpora model each word's trajectory such that subcorpora gather the word's vector representations across intermediary models, up until the mature space (for which we inherently have a perfect overlap, thus a cosine of 1). These trajectories can be either compared between words or viewed for singular words.

These trajectories show equivalent sensitivity to word frequency as do conventional psychometric tests, as indicated by the strong correlations that were reported. This was measured through evaluating the relationship between human vocabulary knowledge and the various word metrics on a number of vocabulary tests, resulting in rank order correlations that ranged from 0.73 for the Kaufman Brief Intelligence Test-II (Kaufman & Kaufman, 1990) and 0.76 for the Peabody Picture Vocabulary Test-III (Maddux, 1999), to 0.81 for the Kaufman Assessment Battery for Children-Verbal Knowledge (Kaufman & Kaufman, 1983) and 0.83 for the Kaufman Assessment Battery for Children–Expressive Vocabulary (Kaufman & Kaufman, 1983).

The results indicate that the relation between the trajectories taken by individual words in the Word Maturity model can be viewed as an indicator of the word's complexity—i.e., a consequence of the amount of *exposure* needed before the latent representation of a model matures. Words with low complexity, acquired early in a learner's vocabulary, have latent representations in earlier intermediate models which are highly similar to those found in the mature model. Thus, the model requires fewer texts to adequately represent the word. In contrast, the meaning of complex words is acquired only after a certain degree of exposure to a language, which is not necessarily measured by grade level or age.

Age of Exposure version 1.0

AoE 1.0 (AoE; Dascalu et al., 2015) was developed to provide an alternative measurement of a word's complexity. One limitation of the Word Maturity model is that it is proprietary and thus neither the code nor the word complexity estimates were available to the public. The overarching objectives driving AoE 1.0 were to provide an open-source model and at the same time improve on the Word Maturity model. AoE 1.0 did so by utilizing latent topic probability distributions (Blei et al., 2003) in place of LSA. In contrast to LSA, Latent Dirichlet Allocation (LDA) establishes latent topics in which words have corresponding probabilities; higher probabilities of a word in multiple different topics are indicative of potential different word senses (i.e., LDA accounts to some extent for word polysemy). AoE provides measures of a word's relative complexity at various points during the iterative training by matching topics across intermediate models, with more difficult words having topic distributions further away from those of the mature semantic space.

AoE 1.0 was computed by generating sequentially increasing corpora of documents that simulate a human's exposure to language. For each of the generated subcorpora, LDA models were trained. Each LDA model learned topics for both documents and words, which were then aligned to the mature model. The topic distributions were quantified by applying a flow algorithm over a bipartite graph consisting of the intermediate topics and the mature topics, with edges having weights computed as the Jensen-Shannon divergence between word probabilities corresponding to the two topics. After this matching, various features were extracted from the aligned intermediate and mature topic spaces by measuring the cosine similarities between the word's representation in the intermediate model and its topic distribution in the fully-trained model. Some of these extracted features included (a) the inverse average similarity between intermediate and mature models; (b) the inverse linear regression slope of the similarities; (c) the index of the intermediate model that first exceeds a similarity threshold of .3 to the mature model (i.e., the "index above threshold"); (d) the index of the first intermediate model, for which a grade 3 polynomial fit gives a score above a .4 threshold; and (e) the inflection point of the polynomial (see Dascalu et al., 2015, for further explanations). All features extracted in the AoE model were intended to capture the way in which a word's representations evolve as more documents are used to train incremental LDA models. Dascalu et al. (2015) reported high correlations between AoE indices and various word features (e.g., Kuperman AoA, .716 to .893; word frequencies, −.599 to −.774; word entropies, −.615 to −.780; word naming latencies, .611 to .779; lexical decision

latencies, .616 to .766) as evidence for the method's adequacy in building word features that capture the way a word's representation evolves as a greater portion of the dataset is used.

Age of Exposure Version 2.0: The current study

Our objective in the current study is to enhance the accuracy of AoE estimates of human ratings by incorporating alternative computational approaches and assessing the advantages of two alternative methodological approaches. Computationally, first, AoE 2.0 incorporates the use of regression models to predict word features, enabling the generalization of AoA scores to words that are not included in the original AoA word lists. Second, AoE 2.0 leverages word2vec, which affords estimates of syntactic and semantic relations between words using vector geometry. Methodologically, we examine the impact of randomly introducing the texts to the model compared to introducing text sequentially using an automated readability score, Flesch Reading Ease (Flesch, 1948).

AoE 2.0 leverages the generalization capabilities of regression models to expand human-collected AoA word lists by first training on the existing words in the list and subsequently generating predictions for other words from the vocabulary that are not available within an AoA list. This can be especially beneficial for AoA word lists that have a limited numbers of words (see e.g., Łuniewska et al., 2016). To simulate incremental word exposure, AoE 2.0 predicts words' AoA scores using features generated by word2vec models that are incrementally exposed to increasing corpora of text.

Word2vec is an algorithm for training neural networks to estimate word vector representations through self-supervised learning. Context windows of words with fixed sizes are considered across the training corpus and word embeddings are learned. The principal difference between word2vec and LDA (used in AoE 1.0) is that LDA is a topic modeling technique, whereas word2vec generates word embeddings. The word vectors that word2vec generates reflect syntactic and semantic relations between words through vector geometry, which have been shown to be superior to LSA (Baroni et al., 2014; Lenci et al., 2021; Levy et al., 2015; Mikolov, Chen, et al., 2013a; Mikolov, Sutskever, et al., 2013b). Our method requires the alignment of vector spaces generated by models trained at different corpus exposure levels as well as the measurement of word similarities. Word2vec is well suited for this task because of the inherent arithmetic properties of the word representations. Additionally, LDA is computationally expensive on very large corpora, whereas word2vec learns embeddings in a distributed fashion through stochastic gradient descent.

In addition to incorporating computational advantages, we examine the impact that the *order* in which texts are used during training has for the performance of the models. In essence, we compare an arbitrarily random order to an ascending order in which more readable texts are introduced first to the model. This sorting of paragraphs was performed using their Flesch Reading Ease (Flesch, 1948) readability score which provides an indication of the readability of a given text as a score in the range of 0–100.

The accuracy of the models is measured using a 10-fold cross-validation with a random forest regressor (Breiman, 2001), a support vector regression, linear regression, and lasso regression. We selected these regressors because we found that other models, such as multilayered perceptron, produce weaker results on the datasets analyzed in our work. Moreover, random forest and linear regression models are more interpretable.

To assess the generalizability of the models, we compare AoE 2.0, AoE 1.0, and Word Maturity estimates of AoA using the word list built by Kuperman et al. (2012) in terms of mean absolute error (MAE), correlations, and R^2. Second, we assess the accuracy of AoE 2.0 predictions of AoA scores using various other word lists, namely Bird (Bird et al., 2001), Bristol (Stadthagen-Gonzalez & Davis, 2006), Cortese (Cortese & Khanna, 2008), Shock (Shock et al., 2012), and the AoA scores derived from children by Morrison et al. (1997). Third, we analyze the effectiveness of the generated scores in the context of estimating textual difficulty and compare the results to Word Maturity and AoE 1.0. Finally, we explore the impact of the different types of features used for training the regressors by performing an ablation study in order to measure the potential benefits of adding WordNet and word trajectory features to the lexical feature set.

Notably, the AoE 2.0 model is fully reproducible and has been released as an open-source project available at: https://github.com/readerbench/Age-of-Exposure. The generated AoE scores for the entire English vocabulary using our most predictive model are also available within the previously mentioned repository.

Method

Text corpus

A large text corpus was used to simulate the manner in which people are exposed to new words during reading and listening by splitting the full corpus into incrementally increasing subcorpora. Our collection of texts consists of a combination of *TASA* (Touchstone Applied Science Associates, Inc.), *COCA* (Corpus of Contemporary American

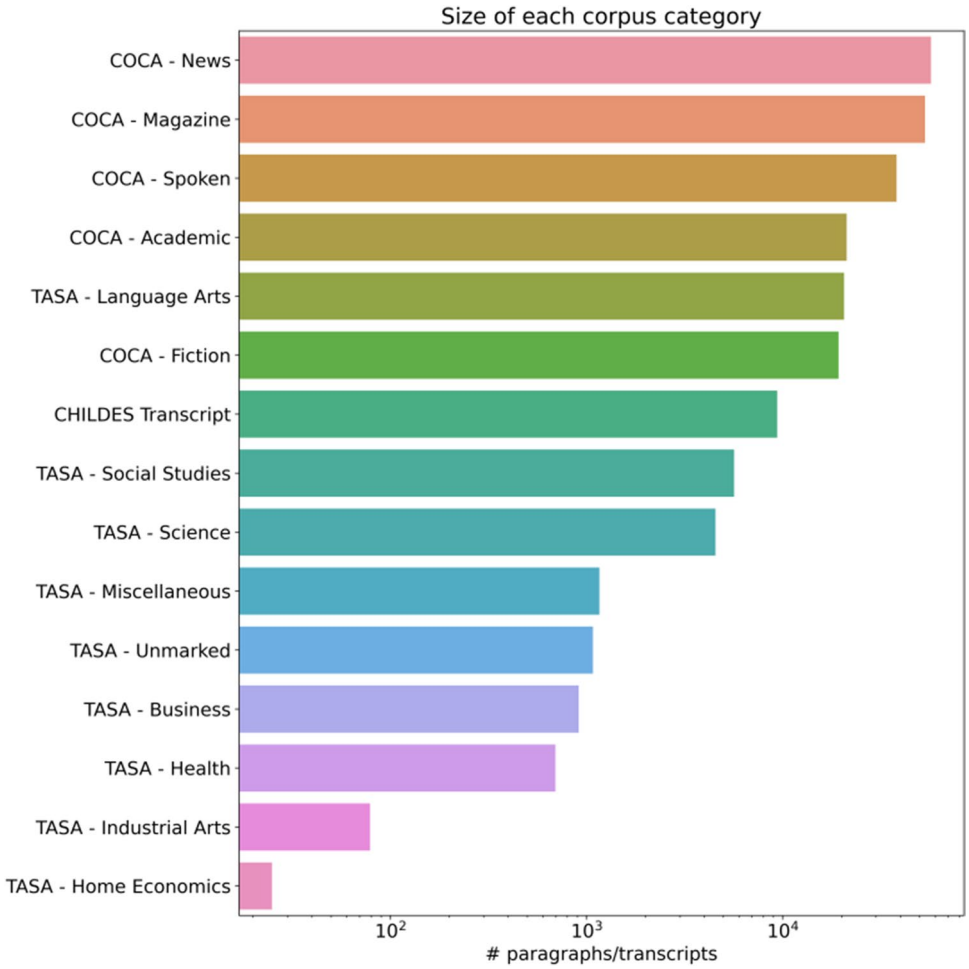

Fig. 1 Histogram showing the number of paragraphs in the TASA, COCA, and CDS corpora as a function of text type on a logarithmic scale

English), and the *Child Directed Speech* (CDS) corpora selected from the Child Language Data Exchange System (CHILDES) dataset (MacWhinney, 1996). *TASA* contains short fragments from text documents that aim to provide a representative sample of educational English for various topics such as health, industrial arts, home economics, social studies, language arts, and others (Ivens & Koslin, 1991). *COCA* contains short- and medium-length documents on a variety of topics—academic journals, fiction (i.e., short stories and plays), magazine articles taken from a variety of popular magazines, newspapers articles, as well as transcripts from TV and radio programs (Davies, 2008-). The CDS corpora consist of transcripts of conversations between young children and mature language speakers. The CDS datasets comprise dialogues between young children, typically before the beginning of formal education, and various interlocutors such as their parents. We include the CDS corpus because it includes words with low AoA scores (i.e., words acquired at an early age). A large-scale database for CDS is the CHILDES (MacWhinney, 1996) system which offers transcripts of various CDS experiments in different languages. In total, we selected the texts from 56 datasets found under the "North America" and "United Kingdom" English sections (please see the index found at https://childes.talkbank.org/access/ for the individual corpora and their citations).

The texts from TASA and COCA were split into their constituent paragraphs, whereas the CDS datasets were split at the transcript level, to generate individual samples roughly equal in length. This was required since certain COCA texts were considerably larger by an order of magnitude in comparison to typical TASA texts. For CDS, we elected to use the entire transcript instead of individual utterances because the child utterances tend to be only a few words long. This approach rendered the transcripts, as a collection of child utterances, comparable in length to the TASA and COCA paragraphs. In total, 9391 CDS transcripts were added together with the combined TASA/

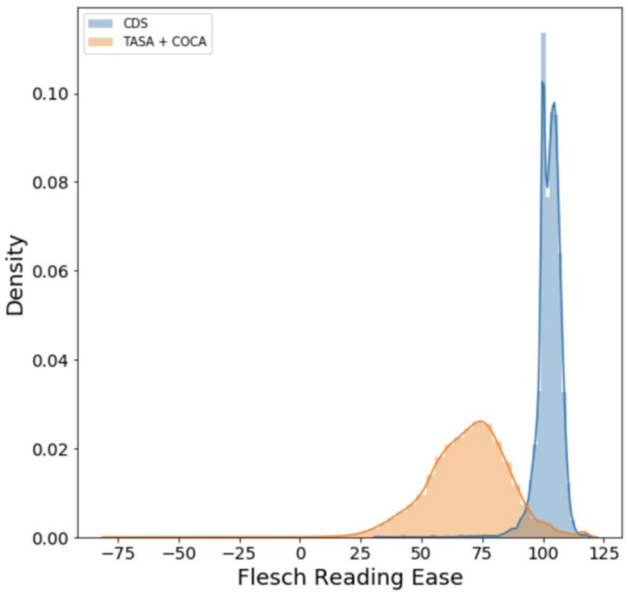

Fig. 2 Readability of CDS transcripts. Scores outside of the 0–100 range correspond to errors in the way sentences are separated during parsing caused by formatting errors

COCA dataset to form our final training corpus of 233,060 documents (see Fig. 1 with the number of paragraphs for each category displayed in a logarithmic scale). We opted to use the logarithmic scale so that the largest categories would not dominate the plot, obscuring the least frequent collections on a linear scale.

Figure 2 depicts the readability scores using Flesch Reading Ease for the selected transcripts, as computed on transcripts as a whole. The density plots from Fig. 2 show that the CDS corpora have texts of lower complexity than the ones present in the COCA and TASA datasets. Notably, however, because they are transcripts of dialogue, there are significant differences between the structure of these documents and the paragraphs extracted from TASA and COCA.

Our aggregated dataset comprising of three corpora (TASA, COCA, and CDS) contains documents written at various grade levels and on various topics. The purpose of combining the three corpora is to provide an adequate representation of common texts in English that are representative of language that might be encountered by children in their home and school environments. These texts are representative of potential exposure to language from various sources such as school, television, internet, and interactions with caregivers, peers, and teachers. Overall, we aimed to simulate a variety of situations in which the usual language learner is exposed to new words. Nonetheless, the AoE 2.0 computational methodology can be applied to virtually any corpus if the aim is to build corpus-specific models.

Building iterative word embedding models

Multiple word2vec models were trained on an incrementally increasing datasets of documents from the aforementioned corpus, with a total of 465,682,595 tokens of which only lemmatized verbs, nouns, adjectives, and adverbs were retained. Lemmatization and the selection of only content words was performed to reduce noise from stop words, reduce all word forms into their dictionary form, and ensure alignment with words from the AoA word lists. The spaCy (https://spacy.io) framework was used for tokenization, part-of-speech tagging, and lemmatization with the largest available English model being selected. Of note is that we did not check the accuracy of this model's tagging and lemmatization. The use of part-of-speech (POS) taggers for CHILDES or other CDS datasets, or the confirmation of spaCy's accuracy on child transcripts may be of interest for future work that uses CDS data as a larger portion of the total corpus. Additionally, we did not check the accuracy of spaCy's taggers on the COCA and TASA documents either.

To perform the iterative word embedding modeling, two components are required: a function describing the dataset *growth*, and a relative *ordering* function of the documents. The growth function simulates the manner in which people are exposed to an increasing number of texts during their lifetime. Because exposure to the content of texts and the number of texts themselves accumulate and have some potential to relate to previous knowledge, our model is also exposed to all texts from the previous stages. Our assumption is that a model trained at step t will also contain most of the semantic and syntactic information on the vocabulary words from models at previous steps ($t' < t$).

In the experiments that follow, we utilize a simple linear growth scheme, where each iteration of training adds an equal-sized portion of the total corpus. In Appendix A, *Impact of the Growth Scheme*, we describe in greater detail how different growth schemes can potentially impact the performance of the model. The order in which documents are introduced to the learner also plays an important role. The first approach involves random sampling of documents, whereas the second method orders documents by their Flesch Reading Ease score. The equation for this scoring method is given below:

$$\textit{Reading Ease Score} = 206.835 - 1.015\left(\frac{\textit{tota words}}{\textit{total sentences}}\right) - 84.6\left(\frac{\textit{total syllables}}{\textit{total words}}\right) \quad (1)$$

The ordering makes simpler documents appear in earlier stages, while more complex texts appear in later stages. The purpose is to simulate the notion that people usually start with more readable texts or simpler language, and then progress to more difficult texts and language as their language proficiency increases. Moreover, less difficult words should

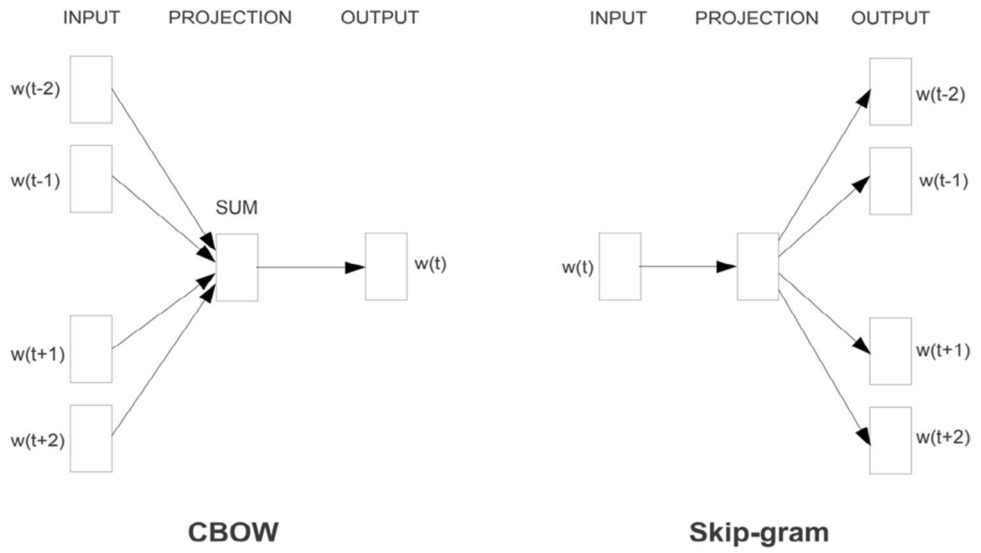

Fig. 3 The two variants of the word2vec algorithm (Mikolov, Chen, et al., 2013a)

be seen early in model training by placing texts with higher readability at the beginning.

The Flesch Reading Ease readability score provides a score from 0 (very difficult to read) to 100 (very easy to read[1]) that is a function of the length of the sentences and the number of syllables per word, which is often used as a proxy for the readability level of a text. The Flesch Reading Ease score it does not consider multiple aspects that may impact text readability (e.g., see McNamara et al., 2014), but it is easily computed for any given text and is commonly used as a measure of a text's readability.

Training self-supervised word embedding models

The word2vec (Mikolov, Chen, et al., 2013a; Mikolov, Sutskever, et al., 2013b) model uses a shallow neural network with a single hidden layer that provides the embeddings and an output layer consisting of a softmax activation over the vocabulary. There are two main variants of training a word2vec model: continuous-bag-of-words (CBOW) and Skip-gram. These are illustrated in Fig. 3. In both cases, the context window of a word is composed of the neighboring terms from the text. For CBOW, the model is trained to predict the target word given an input consisting of the window words, for which the order is disregarded. However, for the Skip-gram model, the target word is given as input and used to predict which words are most probable to appear in the same context as the target word. Thus, the word2vec neural network learns embeddings that capture the local co-occurrence of words.

Both the CBOW and Skip-gram algorithms generate models that reflect a probability distribution over the entire vocabulary by considering an output layer having a number of neurons equal to the size of the vocabulary, combined with a softmax activation, described in Eq. 3:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}}, for\ i = 1..N\ and\ with\ z \in \mathbb{R}^N \tag{2}$$

The softmax activation transforms an array of numbers into a probability distribution, such that the resulting values have a sum of 1 and higher initial values correspond to higher values in the new distribution.

In general, CBOW tends to have better performance on more common words, while Skip-gram is better at representing rare words (Mikolov, 2013). The word2vec model has numerous computational advantages because it does not require the computation of the global occurrence matrix; however, its main benefit consists in the vector representations themselves. As presented in the initial studies conducted by Mikolov, Chen, et al. (2013a), word embeddings capture both syntactic and semantic properties of words. Of particular interest are the highlighted mathematical properties, such as the ability to measure the cosine similarity between two words to determine if they have similar meanings, or the ability to perform algebraic operations on the vectors in order to determine semantic relationships between words (e.g., by adding the vector for "king" and subtracting the representation for "man" to the vector of "woman," the nearest representation in the vector space can be found to be "queen").

[1] In certain cases, the Flesch Reading Ease score may exceed the 0–100 range for ill-formed texts (i.e., missing sentence breaks or whitespaces).

In contrast to the topic distributions from LDA, word2vec representations are additive and can be more easily employed to measure semantic similarity. While LDA captures word polysemy in terms of word associations to multiple topics, it has a problem identifying the optimal number of corresponding topics. This, coupled with a bipartite matching of topics that is more rigid than transformations on word embeddings as found in word2vec, supports the use of word2vec over LDA in developing AoE 2.0. Word2vec models have also been shown to provide embeddings that are qualitatively superior to alternatives such as LSA (Mikolov, Chen, et al., 2013a).

There are numerous options for generating word embeddings besides using semantic models such as LSA, LDA, and word2vec. Ruas et al. (2019), for example, proposed an algorithm for generating word embeddings that are able to capture multiple senses through context-aware disambiguation. Similarly, Li and Jurafsky (2015) proposed that multi-sense embeddings can help improve natural language understanding and introduced an algorithm for training such embeddings. Furthermore, state-of-the-art natural language processing models that are based on transformers, such as BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019), opt to learn embeddings at the level of sub-words which can greatly help with rare words by tokenizing them into statistically more common components. Our choice of using word2vec instead of other embedding methods, such as the multi-sense algorithms and the sub-word embedding schemes outlined previously, was motivated by the fact that word2vec is efficient for large datasets, which constitutes a major advantage in the context of training multiple word embeddings for a single corpus. Additionally, AoA scores are not typically separated by word senses, which potentially renders word2vec a better fit over multi-sense word embeddings. With this in mind, we elected to use word2vec, the most common neural embedding scheme, as a method for modelling holistic exposure trajectories for words.

For each of the 10 incremental sets, a word2vec model was trained. Given the dataset split, each model encapsulates all of the words encountered by the previously trained models. The final word2vec model (i.e., trained on the "mature" dataset) holds the final word vectors, which are used as reference when analyzing all other previous models. We opted to rely on CBOW representations instead of Skip-gram because the overall performance of our model was better served by having higher quality word embeddings for common words. Each word2vec model is trained using the CBOW algorithm for five epochs, with a window size of 5 and a vector size of 300 (see Appendix B, *Impact of Word Embedding Size*, for a discussion on the impact the word embedding size has on the final performance). This means that a model uses 5-grams (i.e., five consecutive words) and is tasked to predict the central word, given the surrounding contextual words. For each word, a 300-dimensional embedding vector is learned over five passes over the texts in a subcorpus. A subcorpus consists of paragraphs assigned to a certain training stage, with the first training stage consisting of approximately 10% of the total corpus. Each consecutive training considers a larger corpus that includes all texts used in previous stages (i.e., the second stage will have around ~20% of the texts, including the 10% used in stage 1).

Generating word features

The cosine function between intermediate and mature representations of terms is used to model the evolution of a word across stages. The vocabulary and the number of word vectors in the word2vec vector space increase as the corpus increases and, as such, result in different vector space alignments at each intermediate step; in return, this leads to the need for a method of aligning the intermediate vector spaces to the mature vector space (i.e., the word2vec model trained on the entire corpus). This is achieved by using the Procrustes alignment (Gower, 1975; Krzanowski, 2000) performed using all words common between all intermediate models as pivots.

The Procrustes alignment translates, scales, and rotates the word vectors of all words present in intermediate models to their corresponding mature values. The components of the Procrustes algorithm are an orthogonal rotation, a reflection, and a scaling transformation which are applied to the intermediate vector space, aligning it to the mature model. Aligning the two vector spaces is necessary primarily due to the stochastic nature of the word embeddings generated by the word2vec neural network. Because the intermediate models may not have seen certain words, 0-vectors are used to bring each intermediate vector space to the same dimensionality as the mature one. Another option might be to use the average vector of known word embeddings as a substitute for unseen words; however, this was not explored in this work. After the Procrustes transformation is applied to the intermediate vector space, the approximate directions and magnitudes of the vectors representing the same word in the two vector spaces should match. Inherent discrepancies between representations provide a measure of the dissimilarity between how a word is embedded in a certain intermediary stage, and how it is embedded when the entire corpus is considered.

A demonstration of the process is displayed in Fig. 4 which illustrates the following steps: (a) the two vector spaces are translated so that both their origins are 0; (b) the spaces are scaled by their respective Frobenius norms; and then (c) the first vector space is transformed using an orthogonal matrix that minimizes the distance between the reference words using orthogonal Procrustes (Schönemann, 1966).

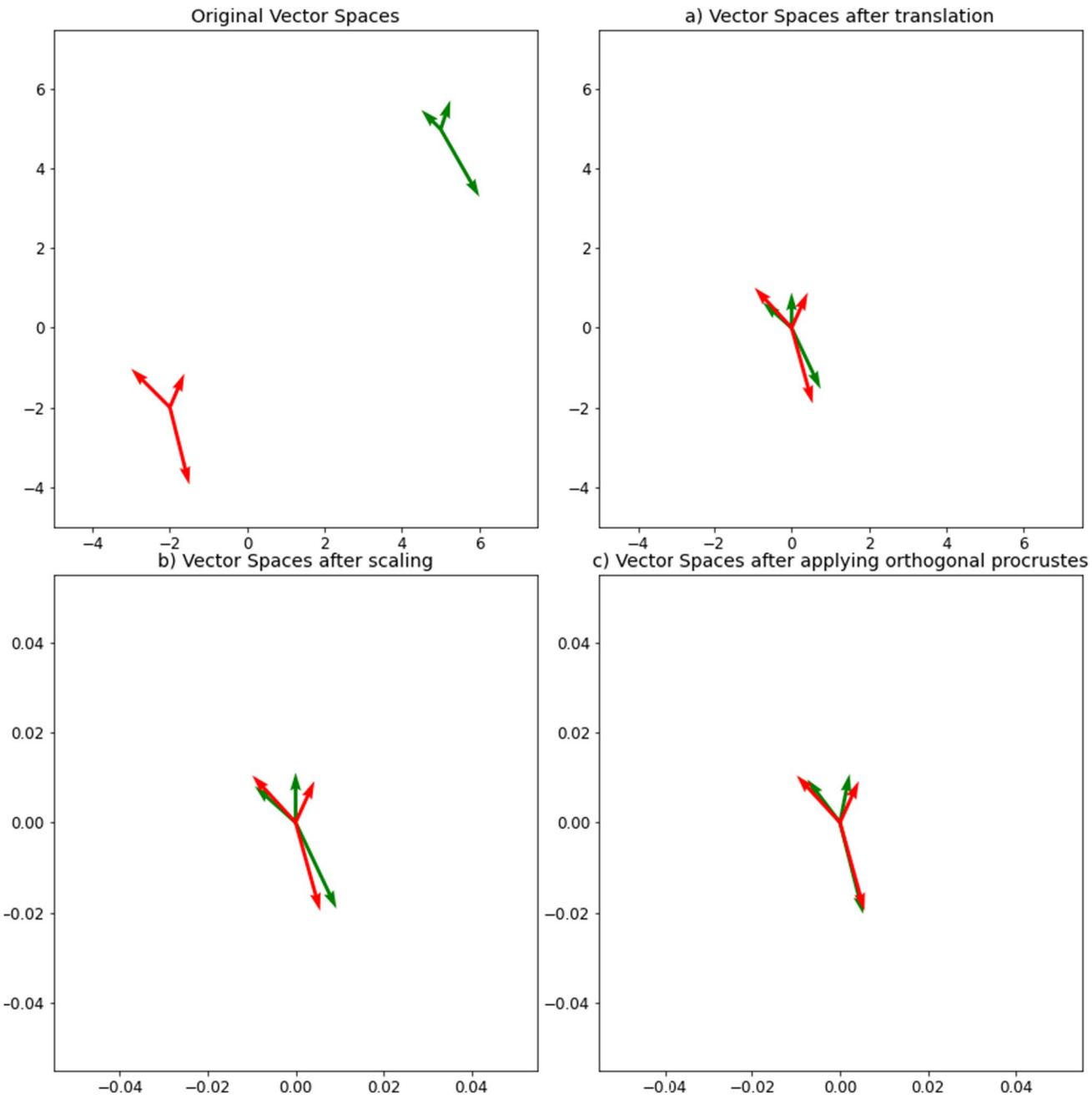

Fig. 4 Procrustes rotation demonstration

The formal definition of the Procrustes alignment is the following:

0. Given two matrices A and B, our aim is to align B to A.
1. Normalize the matrices corresponding to the entire vector spaces and translate the data to the origin.

$$A = \frac{A - \overline{A}}{\|A\|}, B = \frac{B - \overline{B}}{\|B\|} \quad (3)$$

2. Find the orthogonal matrix R that most closely maps A to B using SVD.

$$BA^T = U\Sigma V^T \ (\textit{Singular Value Decomposition}) \quad (4)$$

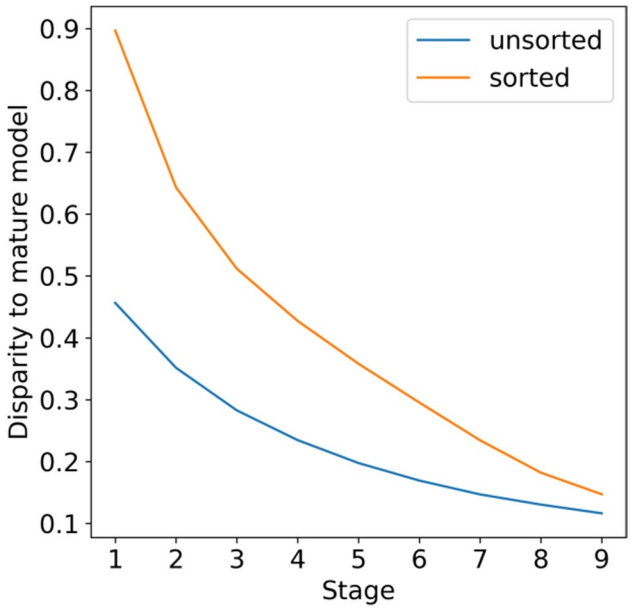

Fig. 5 Disparities of intermediate models for different corpora subsets

$$R = UV^T \tag{5}$$

3. Apply R to B to align the two matrices.

$$B' = B * R^T \tag{6}$$

4. Compute disparity.

$$d = \|A - B\prime\|_F^2 \tag{7}$$

The disparity between each intermediate vector space and the mature vector space can be computed when performing the Procrustes rotation. Disparity measures the sum of the square post-alignment differences between the two vector spaces, which is also the metric minimized by the rotation itself. Our hypothesis is that earlier vector spaces have higher disparities than the later ones, which gradually tend to move closer to the mature model as more of the texts in the corpus are used for training. Figure 5 displays the disparities for the two scenarios introduced using the ordering functions (i.e., unsorted vs. sorted by readability).

Other methods for handling word embeddings to represent change over some dimension of the corpus exist. For example, temporal word embeddings with a compass (TWEC) is designed to aid in the analysis of diachronic shifts in word meaning, is based on word2vec, and uses atemporal reference embeddings, called "compasses," when training temporal-aware embeddings (Di Carlo et al., 2019). The hypothesis is that the majority of terms in a vocabulary do not change their meaning over time, while the ones that do present diachronic shifts will appear in the context windows of those stable ones. In our work, the trained word embeddings would not shift due to temporal changes, but because of their exposure to increasing quantities of texts. At each step, the word embeddings are trained on a section of the dataset that includes all documents used in previous steps. Because the embeddings at each level of exposure become increasingly more accurate (i.e., closer to the mature model as seen in Fig. 5), we elected to utilize the Procrustes method for aligning the vector spaces with the assumption that they represent human exposure trajectories, wherein an individual's mastery of a language increases cumulatively with exposure to that language.

One way to interpret these disparities is that steeper slopes indicate a greater degree of change in terms of vector spaces, from the initial model until the mature one, which may potentially result in more useful word trajectory features. This would follow from the assumption that sorted models would ideally have a more natural progression, with less complex words appearing earlier than more complex ones. Unsorted models, on the other hand, may see most words in the initial training stages, which would result in the vector space being populated early, which may result in a reduction in the amount of change observed from one stage of training to another.

After the intermediate vector spaces are rotated, the feature vector for a word is formed by measuring the cosine similarity between the vector representation of the word for each intermediate model in relation to the mature model. This can be defined as:

$$f_w = \left\{1 - cosine\left(\boldsymbol{w_t}, \boldsymbol{w_T}\right) \middle| \, t \in 1..T-1 \right\} \tag{8}$$

where T is the number of stages considered and $\boldsymbol{w_t}$ is the word2vec representation for a word at a given training step t.

Predicting age of acquisition

The last step consists of training a regressor model to predict the Kuperman AoA scores. If the cosine similarities between the intermediate word representations and the mature word representation are indicative of a word's evolution as language proficiency increases and learners are exposed to more texts, then they can be used as indicators of a word's AoA. Besides the cosine values for the nine intermediate models, Table 1 provides a list of six other informative features that were also reported in the initial AoE model by Dascalu et al. (2015).

In addition to the AoE features derived from word trajectories presented in Table 1, we also considered several word-level lexical features provided in Table 2. These features are frequently used to reflect word complexity and may impact the age at which a word is acquired (Nelson et al., 2012). Simple statistical measures, such as the number of syllables and the number of characters, are indicators of how difficult a word may be perceived, with longer

Table 1 AoE 2.0 features extracted from word trajectories

Feature name	Formula	Description
Inverse average	$1 - \overline{CosSim_t}$	The complement of the mean cosine similarity, showing the average dissimilarity of intermediate embeddings to the mature word embedding.
Highest cosine similarity	$\max_t CosSim_t$	The highest cosine similarity, denoting how well intermediary stages best match the mature model
Inverse slope	$\frac{1}{a}$, for the linear regression fitting the cosine similarity values: $ax_t + b = CosSim_t, \forall t \in \{1..T\}$	The inverse of the slope as measured while performing a linear interpolation between the cosine values. This feature approximates how quickly the intermediate word embedding models learn a representation of the word while matching the mature one
Continuous index at threshold	t, *where* $CosSim_t \geq thresh$	The continuous index, or the index of the intermediate model that contains a representation of a given word with a cosine similarity above a certain threshold. Moreover, the successor model also has to exhibit the same property, for consistency. This is computed across multiple threshold values, ranging from 0.3 to 0.7, with increments of 0.05. This feature identifies the stage at which a word is well conceptualized
Word-vocabulary integration	$\lvert\{t, CosSim_t \geq 0.3\}\rvert, \forall t \in \{1..T\}$	The number of words with which the target word has a cosine similarity of at least 0.3 in the mature vector space, as well as the average of these cosine similarities
Top 3 cosine similarities	$top_3\{CosSim_t\}, \forall t \in \{1..T\}\}$	The cosine similarities of the three closest words in the mature vector space, together with their mean

Table 2 Word-level features independent of the word2vec models

Feature name	Formula	Description
Average hyponym eccentricities	$\overline{eccentricity\left(HyonymTree_{word}\right)}$	The average eccentricity of the hyponym trees of the word. The eccentricity is the largest distance from the word to any of its hyponyms.
Average hypernym eccentricities	$\overline{eccentricity\left(HypernymTree_{word}\right)}$	The average eccentricity of the hypernym trees of the word. The eccentricity is the largest distance from the word to any of its hypernyms.
# synset	|synsets_word|	The number of synonym sets of the word (i.e., word senses).
# syllables	|word_syllables|	The number of syllables of the word.
# chars	|word|	The length of the word expressed in number of characters.
Term frequency – Stage i	$tf_i(word)$	Term frequency of the word in the training corpus, normalized per 1 million terms at a given intermediate stage.
Term frequency - average	$\overline{tf_i(word), i = 1..T}$	Average normalized term frequency for the intermediate models.
Term frequency – Standard Deviation	$\sigma[tf_i(word)], i = 1..T$	Standard deviation of the normalized term frequency for the intermediate models.

words being acquired later in life. The number of synsets (i.e., word senses that are linked through synonym chains) from WordNet (Miller, 1995) can also impact an individual's ability to infer a word's meaning in a given context. Finally, the eccentricities of the word hyponym and hypernym trees derived from WordNet indicate the genericity of that concept. Words with high hypernym eccentricities are words that are very specific, whereas words with low hypernym eccentricities tend to be very generic terms that may be acquired earlier (e.g., the difference between learning the meaning of "bird" versus "owl"). Conversely, high hyponym eccentricities correspond to words having a wide semantic field, which means that their acquisition may occur later on (e.g., "lincoln" or "steed"). In addition, we also include the term frequency of words in the training corpus, normalized per 1 million terms, as term frequency was previously shown to be a strong determinant of acquisition (Brysbaert & New, 2009).

Some of these features, namely the word and syllable lengths, are also variables used in the Flesch Reading Ease equation. However, this does not introduce a circularity in our method due to a number of reasons. First, the Flesch

Table 3 Split-half reliability Spearman correlation coefficients

Feature	Mean	Standard deviation	Spearman correlation coefficient
Average	0.50	0.29	.94
Inverse average	0.50	0.29	.94
Highest cosine word similarity	0.57	0.12	.77
Slope	0.09	0.037	.62
Inverse slope	12.78	30.54	.59
Continuous index above .30 threshold	2.43	2.43	.86
Continuous index above .35 threshold	2.53	2.43	.85
Continuous index above .40 threshold	2.67	2.45	.85
Continuous index above .45 threshold	2.86	2.49	.85
Continuous index above .50 threshold	3.11	2.56	.86
Continuous index above .55 threshold	3.42	2.65	.86
Continuous index above .60 threshold	3.81	2.76	.87
Continuous index above .65 threshold	4.27	2.85	.88
Continuous index above .70 threshold	4.81	2.89	.89
Term Frequency – Stage 1	37.87	451.06	.80
Term Frequency – Stage 2	203.18	1713.13	.93
Term Frequency – Stage 3	393.38	3108.25	.96
Term Frequency – Stage 4	561.79	4271.31	.97
Term Frequency – Stage 5	722.88	5304.95	.97
Term Frequency – Stage 6	887.03	6244.37	.98
Term Frequency – Stage 7	1065.98	7150.91	.98
Term Frequency – Stage 8	1273.14	8141.60	.98
Term Frequency – Stage 9	1488.82	9170.99	.98
Term Frequency – Stage 10	1638.18	9938.52	.98
Term Frequency – Average	827.22	5487.43	.98
Term Frequency – Standard Deviation	523.32	3118.65	.98
Word-vocabulary integration	709.60	744.61	.79
2nd highest cosine word similarity	0.53	0.11	.77
3rd highest cosine word similarity	0.51	0.11	.77
Top 3 cosine similarities	0.54	0.11	.78
Intermediate cosine similarity 1	0.04	0.23	.69
Intermediate cosine similarity 2	0.24	0.40	.85
Intermediate cosine similarity 3	0.37	0.43	.87
Intermediate cosine similarity 4	0.46	0.43	.89
Intermediate cosine similarity 5	0.54	0.42	.90
Intermediate cosine similarity 6	0.61	0.39	.90
Intermediate cosine similarity 7	0.69	0.34	.90
Intermediate cosine similarity 8	0.76	0.27	.91
Intermediate cosine similarity 9	0.81	0.20	.92

Note: all $p < .001$

Reading Ease uses the length information as aggregates together with sentence information, whereas the features our models use refer only to individual words. Second, even if there is a degree of overlap, these features are only used as word features for the regressors, while the Flesch Reading Ease will influence only the way the word2vec models are trained through the order in which they are shown the documents in the corpora.

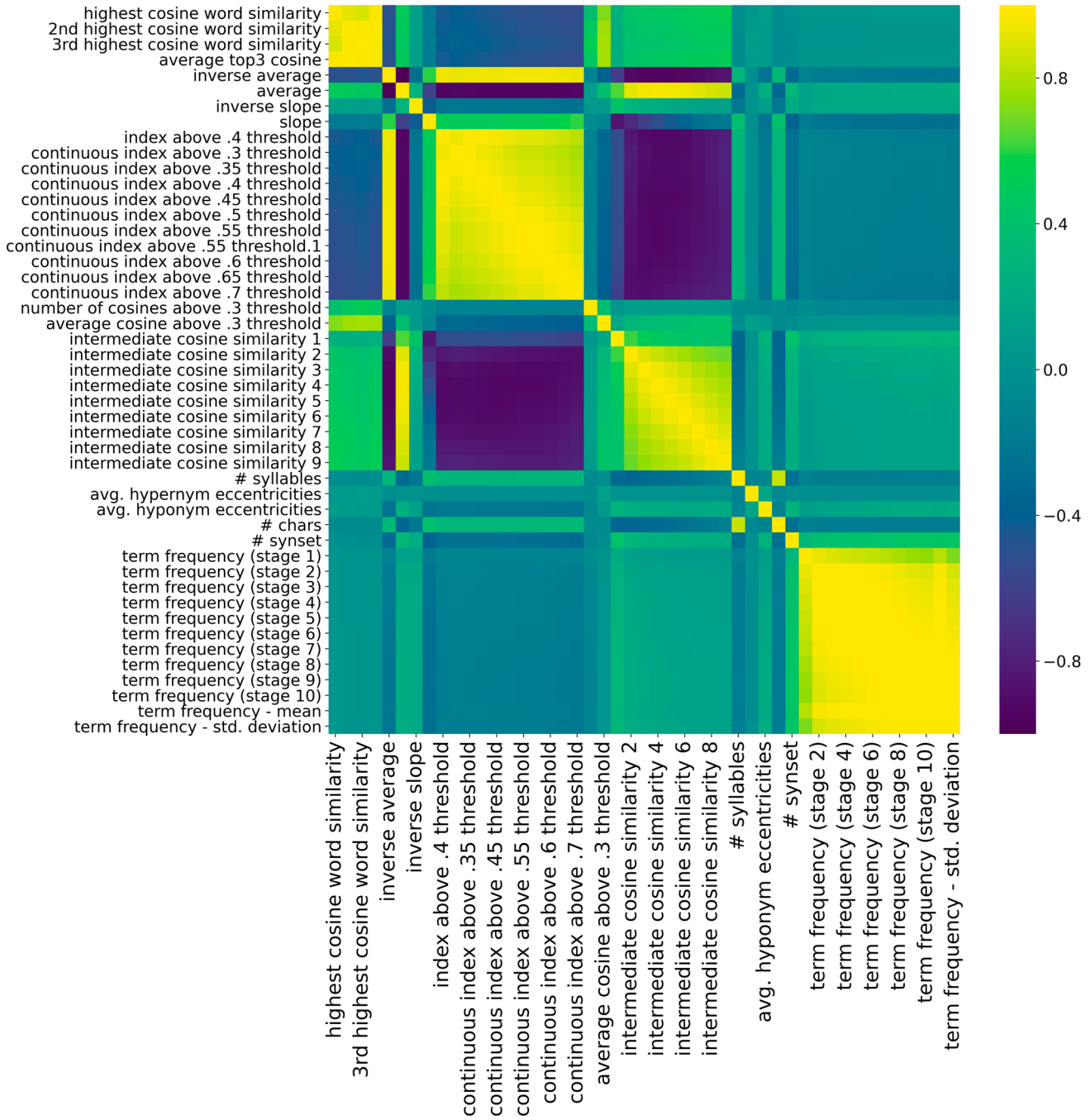

Fig. 6 Pearson correlation coefficient heatmap between word features

Results

Internal reliability

The means, standard deviations, and internal reliability estimates are provided in Table 3 for the indices used within the analysis. A split-half correlation analysis was conducted to evaluate the internal reliability of the indices. We follow a linear growth scheme to generate two parallel datasets from our corpus. Each of these datasets contains half of the paragraphs described in the complete corpus, with paragraphs being assigned to only one of the two halves. This means that the two halves of the dataset are independent. All previous indices are recomputed for each half, thus resulting in two sets of indices per word. The split-half reliability for our feature building method uses the Spearman rank-order correlation between these two index observation vectors for

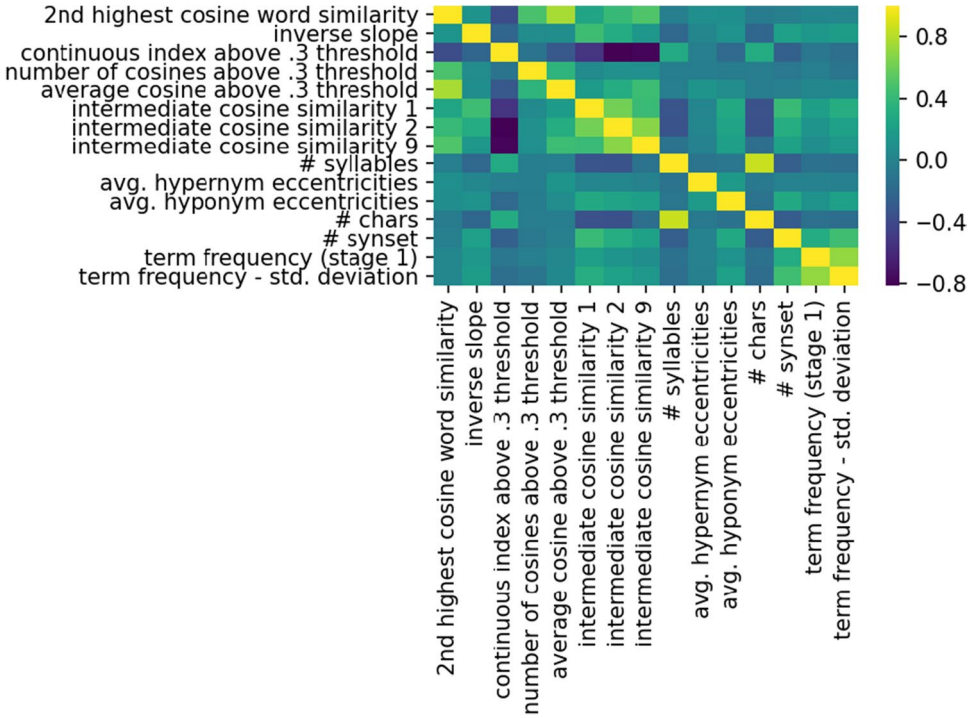

Fig. 7 Pearson correlation coefficient heatmap after variance inflation factor analysis

all words in our vocabulary. All correlation coefficients from Table 3 are statistically significant ($p < .001$) and denote high agreement.

Multicollinearity analysis

We performed a thorough analysis of multicollinearities between the generated features to better understand their similarities and complementarities. This analysis of multicollinearity provides insight into the relationships between different features, highlighting linear dependencies between individual features used in follow-up models. First, Fig. 6 illustrates a heatmap of the linear relations between all pairs of features in terms of Pearson correlations. These are obtained from the features generated by the word2vec models trained on the corpus, sorted by readability.

We can observe that there are blocks of highly correlated features among the "continuous index above a threshold" type features; similarly, "intermediate cosine similarity" type features have lower, but still high, linear correlations between them and the term frequency features. Additionally, the average of the intermediate cosine similarities is correlated with the individual values, whereas the inverse of this mean is correlated with the "continuous index above a threshold" features. The highest, 2nd highest, 3rd highest, and the "average top3 cosine" features create a block of linearly correlated features. There are also correlations between the number of characters of a word and its number of syllables. These results are not surprising, given the definitions of these features. Interestingly, the first stage of training in all three correlation blocks (continuous index, intermediate cosine similarity, and term frequency) appears to have the lowest correlation to the rest of the features in the respective blocks. This may suggest that the first stage of training is the most distinct, a fact that may also be related to the significantly lower reliability for "Intermediate cosine similarity 1" and "Term Frequency – Stage 1" in Table 3, when compared to the counterpart features from later training stages.

An automated method of measuring the degree of collinearity was required to remove features that suffer from multicollinearity. We employed the "variance inflation factor" (VIF; Craney & Surles, 2002), which performs a linear regression on each feature using the other features as predictors. The associated VIF score for a variable is then given by the inverse of the complement of the coefficient of determination r^2 measured for that feature. Craney and Surles (2002) discuss the selection of a cutoff point for a feature's VIF value depending on the desired degree of tolerance. In our case, we opted to use a cutoff point of 5, corresponding to a variability of 80% in a feature being explained by the other features. Fifteen features remain using a VIF cutoff of 5. These relate to the Pearson correlation coefficient heatmap in Fig. 7 as the new correlation heatmap shows low linear relationships between the remaining features.

However, we found that performing a VIF analysis to select features for the nonlinear model leads to a loss in

Table 4 AoE v2 Kuperman AoA prediction results

Texts sorted by readability	Model	MAE	Normalized MAE	Normalized MAEstd. dev.	R^2
Yes	Linear regression	1.75	.08	.00124	.45
Yes	Lasso Lars (AIC)	1.75	.08	.00107	.45
Yes	Lasso Lars (BIC)	1.75	.08	.00108	.45
Yes	**Random forest regressor**	**1.51**	**.07**	.00092	**.57**
Yes	SVR	1.58	.08	.00127	.54
No	Linear regression	2.01	.10	.00127	.29
No	Lasso Lars (AIC)	2.01	.10	**.00086**	.29
No	Lasso Lars (BIC)	2.01	.10	.00105	.29
No	Random forest regressor	1.80	.09	.00121	.41
No	SVR	1.84	.09	.00117	.39

Bold marks the best results

performance when it comes to the ability of the models to perform predictions of the Kuperman AoA scores. Because multicollinearity is a common issue that can impact linear models, we elected to utilize the traditional VIF cutoff of 5 for the linear regression model input features. For the other considered models, namely the support vector regressors and the random forest regressors, we did not perform a multicollinearity filtering on the entirety of the feature set because they are nonlinear models that should not be affected by the presence of multicollinearity in the input feature matrix. However, we did utilize the VIF analysis to minimize the redundancy found when it came to the "continuous index above a threshold" features. By using the same cutoff value of 5, we found that VIF retained only the first and last such indices, namely "continuous index above .3 threshold" and "continuous index above .7 threshold." Similarly, we reduced the term frequency features using VIF to four features: first and last stage term frequency normalized per 1 million words and the mean and average of the term frequencies per word. We elected to utilize this strategy of selectively applying VIF because of its simplicity and the robust results it provided, which slightly boosted the performance of the random forest and support vector regressors without a loss of performance. Random forests, in particular, have been shown to provide superb predictive accuracy, albeit at the cost of not being able to provide insight into the contributions of individual predictors and their interactions (Tomaschek et al., 2018). Given our primary goal of predicting AoA norms with high accuracy, we elected to favor performance over feature interpretability.

Predicting the Kuperman AoA scores

Our primary objective in this study is to predict the Kuperman AoA scores using the predictors described in Tables 1 and 2. The Kuperman AoA scores predicted by our AoE 2.0 model have a distribution (M = 10.98; SD = 3) that has a slight negative skew (−0.2) and a Pearson kurtosis of 2.62, which suggests a platykurtic distribution. The original Kuperman AoA scores form a relatively normal distribution (M = 11; SD = 3.04) with a similar negative skew (−0.2) and a Pearson kurtosis of 2.63.

To predict these scores, a regressor was trained using the features derived from the cosine similarities. Results for the previously described scenarios are displayed in Table 4. The results are measured by averaging over a 10-fold cross-validation with random forest regression (Breiman, 2001), linear regression, and support vector regressor. We found that other models, such as multilayered perceptrons, produce weaker results and decision trees, and support vector regressors tend to outperform neural models for small-medium datasets. Moreover, decision trees and linear models are more interpretable than neural models. Additionally, we also measured the performance of the least-angle regression (LARS) models together with either the Bayes information or the Akaike information criterion for model selection in order to attempt to reduce the complexity of the resulting models (Zou et al., 2007). The mean absolutes errors (MAEs) and their normalized values from Table 4 are reported for the linear, support vector, LARS lasso, and random forest regressor models, as well as whether the texts were sorted by readability before they were split into subsets (i.e., whether simpler texts are seen first). For each experiment, we report the average result for a 10-fold cross validation when training a model using the features presented in Tables 1 and 2 to predict the AoA scores of words.

An ANOVA confirmed that there were significant differences between the models reported in Table 4, $F(2,\ 48{,}540) = 1476$, $p < .01$, growth schemes, $F(1,\ 26{,}976) = 34.5$, $p < .01$, and ordering schemes, $F(1,\ 26{,}976) = 2594$, $p < .01$. Post hoc tests were conducted using a Bonferroni correction, and demonstrated that the

Table 5 Results for predicting Kuperman AoA scores using previous methods

Model	MAE	Normalized MAE	R^2
AoE 1.0 + Random forest regression	1.92	.913	.30
AoE 1.0 + Linear regression	2.02	.096	.23
AoE 1.0 + SVR	**1.84**	**.088**	**.36**

Bold marks the best results

Table 6 Results for predicting Kuperman AoA using the best models available for each method

Model	MAE	Normalized MAE	R^2
AoE 2.0 + RF	**1.51**	**.070**	**.57**
AoE 1.0 + SVR	1.84	.088	.36
Word Maturity indices	N/A	N/A	.46

Bold marks the best results

We also report the results obtained through using the features that were generated using the AoE 1.0 in Table 5. An ANOVA confirmed that there were significant differences between the AoE 1.0 models from Table 5, $F(2, 33{,}869) = 779$, $p < .01$. Post hoc tests using a Bonferroni correction demonstrated that the SVR model, $t(21, 997) = 38.7$, $p < .01$, and the random forest model, $t(21, 197) = 23.1$, $p < .01$) resulted in significantly less error than the linear regression model.

Table 6 introduces a side-by-side comparison of the best AoE 1.0 and AoE 2.0 models, alongside the Word Maturity indices. Because the Word Maturity Index is a single numerical value, we report the squared correlation between the Word Maturity indices and the AoA word list in order to provide a direct comparison to the R^2 of the other models (marked with "*" in the table). An ANOVA confirmed that there were significant differences between the models, $F(2, 34{,}599) = 409$, $p < .01$. Post hoc tests using a Bonferroni correction demonstrated that the AoE v2 model had significantly

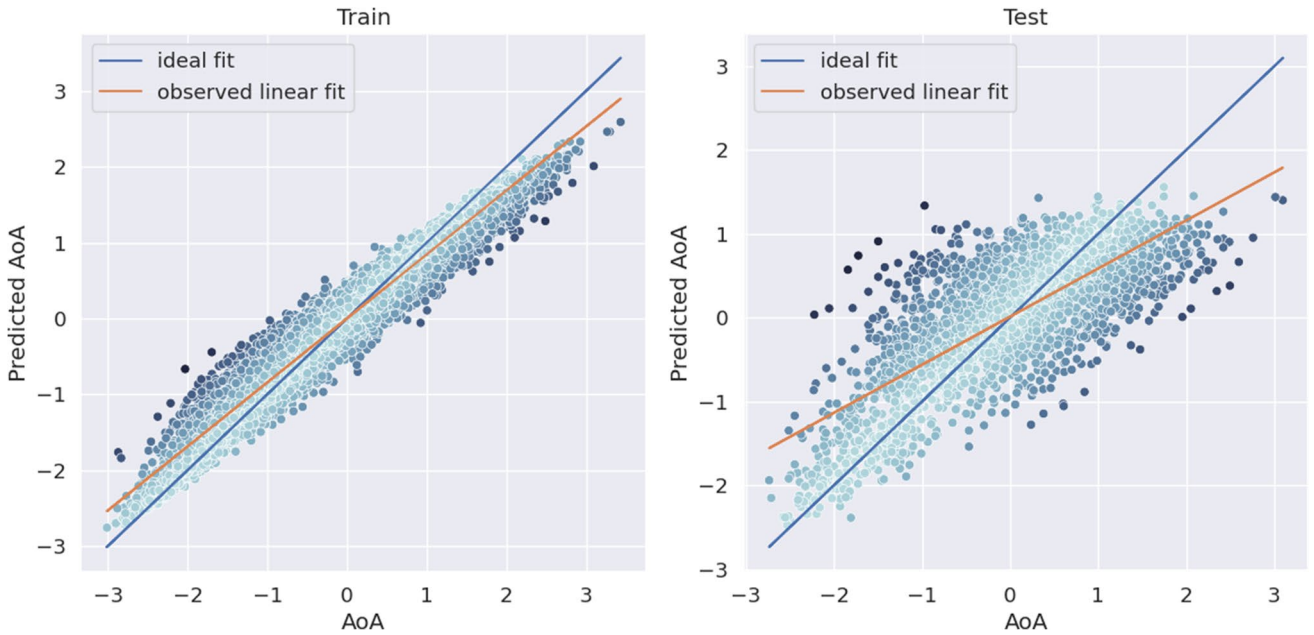

Fig. 8 Scatterplots showing the predicted AoA score and the Kuperman AoA score

random forest model had significantly less error than the linear regression model $t(107{,}907) = 71.2$, $p < .01$, the SVR model $t(107{,}907) = 22.5$, $p < .01$, the AIC model $t(107{,}907) = 62.1$, $p < .01$, and the BIC model $t(107{,}907) = 62.4$, $p < .01$. Finally, the models that presorted text by readability had significantly less error than models that did not presort the texts, $t(269{,}770) = 134.0$, $p < .01$. Thus, sorting the documents by their readability scores appears to improve performance.

less error than the Word Maturity model, $t(17{,}583) = 12.2$, $p < .01$, and the AoE v1 model, $t(17{,}583) = 28.6$, $p < .01$. Our method also marks a significant improvement over AoE 1.0, with R^2 values improving by as much as .21, as well as having a higher correlation with the Kuperman AoA values than the Word Maturity Indices by up to .11.

In order to better understand how the regressor behaves for different AoA magnitudes, Fig. 8 introduces the scatterplots for a pair of training and test sets randomly sampled from the

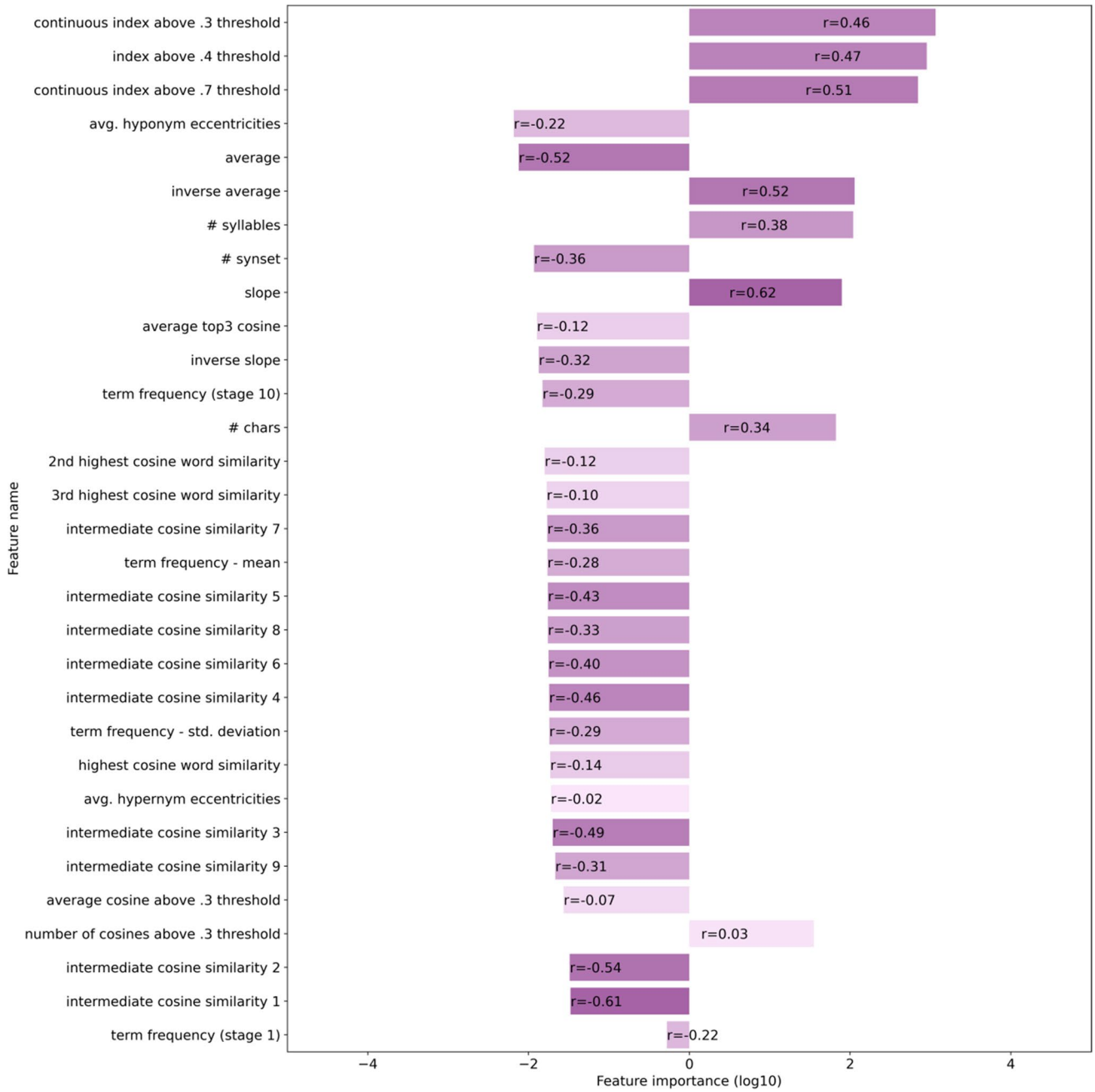

Fig. 9 Importance of selected features within the best-performing random forest model

complete dataset. The blue line represents the ideal fit (i.e., the case where the predicted and the actual AoA scores are equal), while the orange line shows the observed linear fit between the predicted and the real AoA scores. Each data point has a hue that corresponds to the absolute error between the predicted score and the real score. These results suggest that the model produces reasonable approximations of the Kuperman AoA scores, with no perceivable bias towards over or under estimation. The R^2 values indicate that our models are capable of explaining over half of the variance of the Kuperman AoA scores. Of note is that we predicted the mean AoA score for each word in the Kuperman word list, with each word in the study also having an associated standard deviation (SD = 3.04). This is introduced by the fact that the Kuperman AoA list, as well as other AoA lists, is generated by participants estimating the age at which they acquired a word and then aggregating these values. The variance in AoA scores introduces a level of noise in the distribution that limits the performance of models trained to predict such scores, which may explain to a certain degree the errors that the models produce.

Table 7 Results for various AoA scores

AoA metric	Number of words in word list	Normalized MAE of the random forest regressor	R^2 of the random forest regressor
Bird	1973	.09	.61
Bristol	3274	.08	.66
Cortese	2816	.06	.74
Shock	2894	.06	.69
Morrison - Objective AoA (75%) (months)	294	.13	.35

Figure 9 depicts feature importance within the random forest model in descending order, with higher importance reflecting features that are determined to be more relevant for splits in the constituent decision trees. The size of each horizontal bar is determined by the value of the feature importance. These are computed through the impurity decrease of each feature in the constituent trees of the random forest, for each decision node. For a given feature, the Pearson correlation between the feature and the Kuperman AoA is displayed, with the color of the bars corresponding to the absolute value of this correlation and their direction by the sign of the Pearson correlation. The features considered most important by the model are the continuous index values above various thresholds, which indicate that the model uses the inflection point at which an intermediate model generates a vector similar to the mature model. Other features with average importance values are the average and inverse average cosine similarities, various word-level statistical features such as the number of syllables, term frequency features, the number of hyponym eccentricities, the number of synsets to which it belongs, and the number of characters. Features that are deemed least useful by the model are the number of cosines above a certain threshold and specific cosine similarities at various intermediate stages (denoted as "intermediate cosine similarity X," where X is the selected stage), as well as the term frequency measured during the first training stage.

The feature importance values are generated from the random forest regressor and represent the amount of error that each feature reduces across the decision trees that form the forest. Some of these values are intuitive, an example being that the number of synsets for a word is inversely correlated with its AoA score, meaning that words with more synonyms tend to have lower AoA scores and that words with higher term frequencies have lower AoA scores. However, the slope of the cosine similarities between the intermediate vector representations and the mature representation has a positive correlation with the AoA score, suggesting that words that have an initial lower cosine similarity tend to have higher acquisition ages. Continuous indices above certain thresholds are positively correlated with the AoA scores since they are a direct indicator of how many corpus subsets need to be used during the training of the word2vec model before the word representation becomes sufficiently aligned with the mature vector model. Negative correlations, which indicate lower AoA scores, include large hypernym/hyponym average eccentricities which indicate large semantic field tree hierarchies, large values of cosine similarities for the intermediate models, high inverse cosine similarity averages, inverse slopes, and normalized term frequencies.

Generalizing AoE 2.0 to predict other AoA word lists

Our method can easily be applied to alternative AoA word lists, for example: Bird (Bird et al., 2001), Bristol (Stadthagen-Gonzalez & Davis, 2006), Cortese (Cortese & Khanna, 2008) or Shock (Shock et al., 2012). In addition, we also attempt to model the objective AoA scores derived from children (Morrison et al., 1997). Training a random forest regressor and measuring the mean absolute error (MAE) across 10 folds, as before, we obtained the results from Table 7. The purpose of this experiment is to verify that the method is independent of the AoA word list used. Because the previous AoA metrics have differing scales, we also show the scatterplots to better illustrate the performance of the regressor. For each AoA word list, a different random forest model is trained using the same features that were used for predicting the Kuperman AoA scores.

We can see that in the Bird case (see Fig. 10), the reduced number of terms leads to a scattered distribution of values in the test data. On average, the predictions do not appear to have a tendency of solely overestimating or underestimating the Bird scores. However, the model has a tendency to overestimate for low AoA scores, while underestimating high AoA scores.

In the case of the Bristol scores (see Fig. 11), the errors appear to be less scattered (corresponding to the higher R^2 coefficient). Additionally, the errors do not appear to indicate

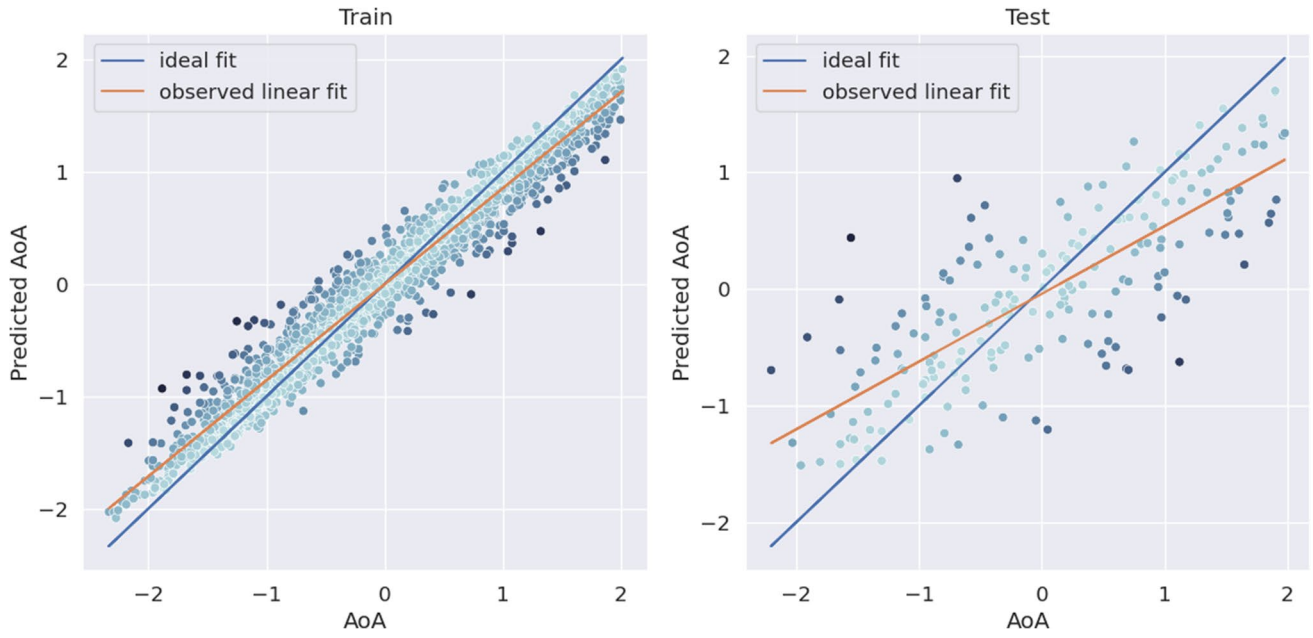

Fig. 10 Results for Bird score prediction

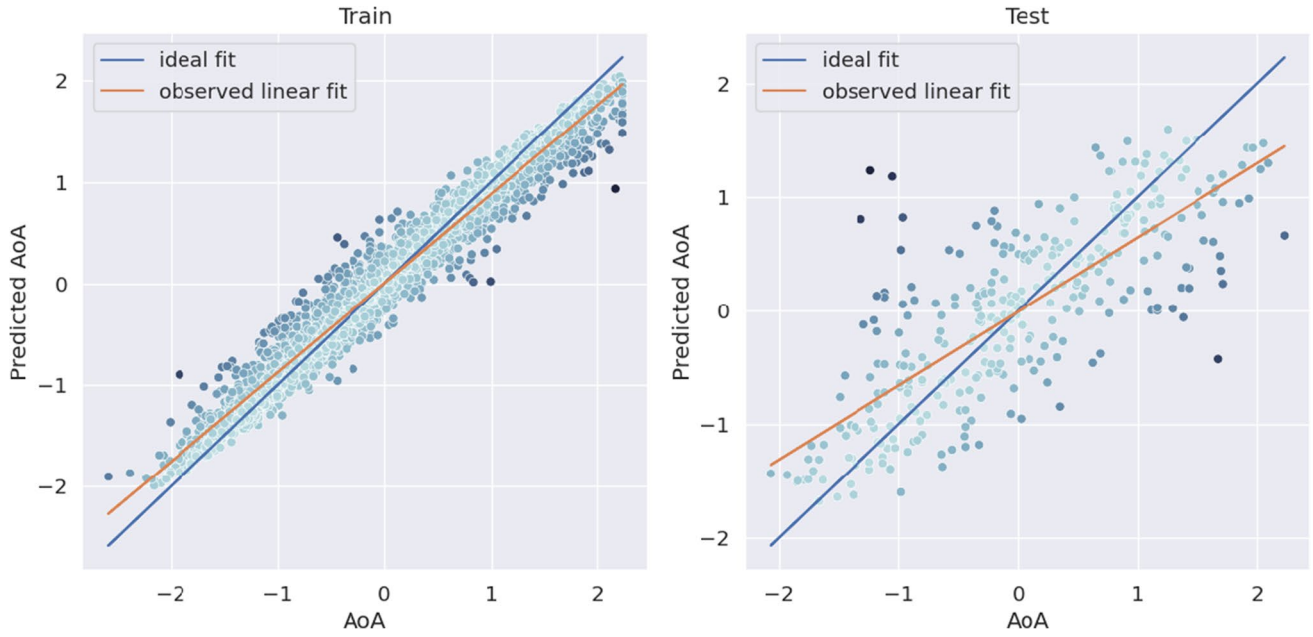

Fig. 11 Results for Bristol score prediction

the same degree of overestimation for low AoA scores and underestimation for high AoA scores observed in the case of the Bird word list.

For the Cortese AoA and Shock datasets (see Figs. 12 and 13, respectively), the observed linear fit and the ideal fit have high overlap, confirming the high R^2 coefficients of .74 and .69. We observe little bias towards overestimation or underestimation for these two datasets, albeit with some significant outliers. By contrast, the model resulted in substantially reduced performance for the objective AoA scores by Morrison et al., 1997 (see Fig. 14; $r^2 = .35$). The limited number of terms (i.e., almost an order of magnitude lower than the other word lists) in the latter dataset resulted in the model having a significantly reduced performance in

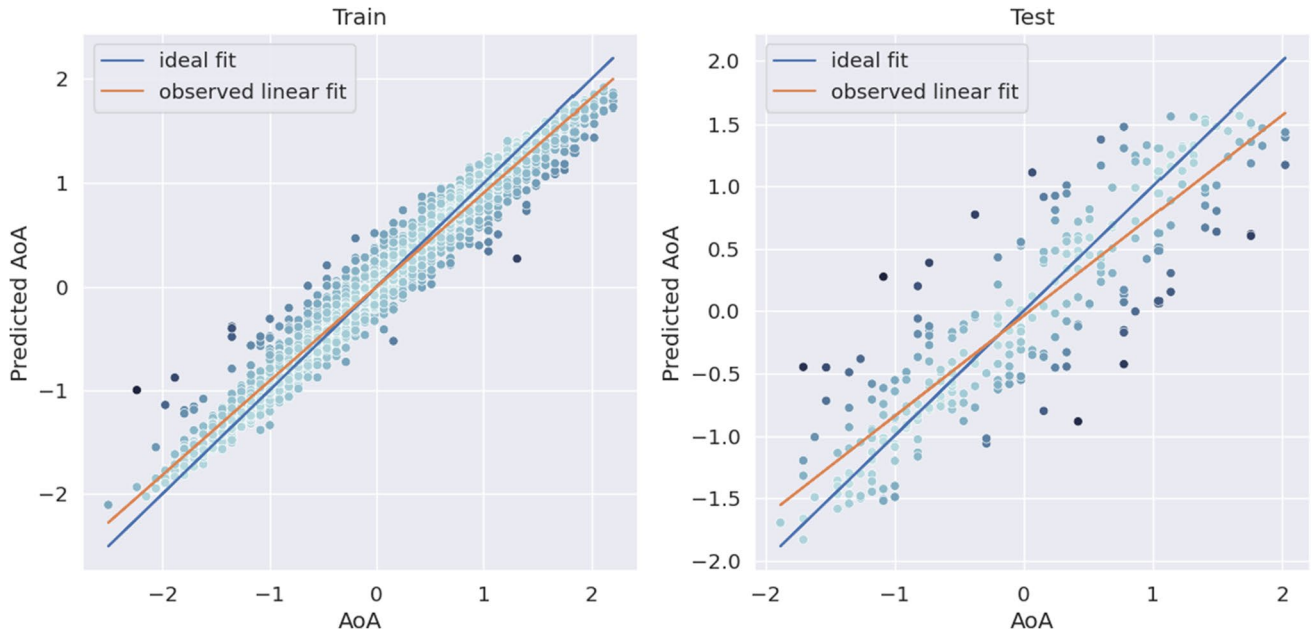

Fig. 12 Results for Cortese score prediction

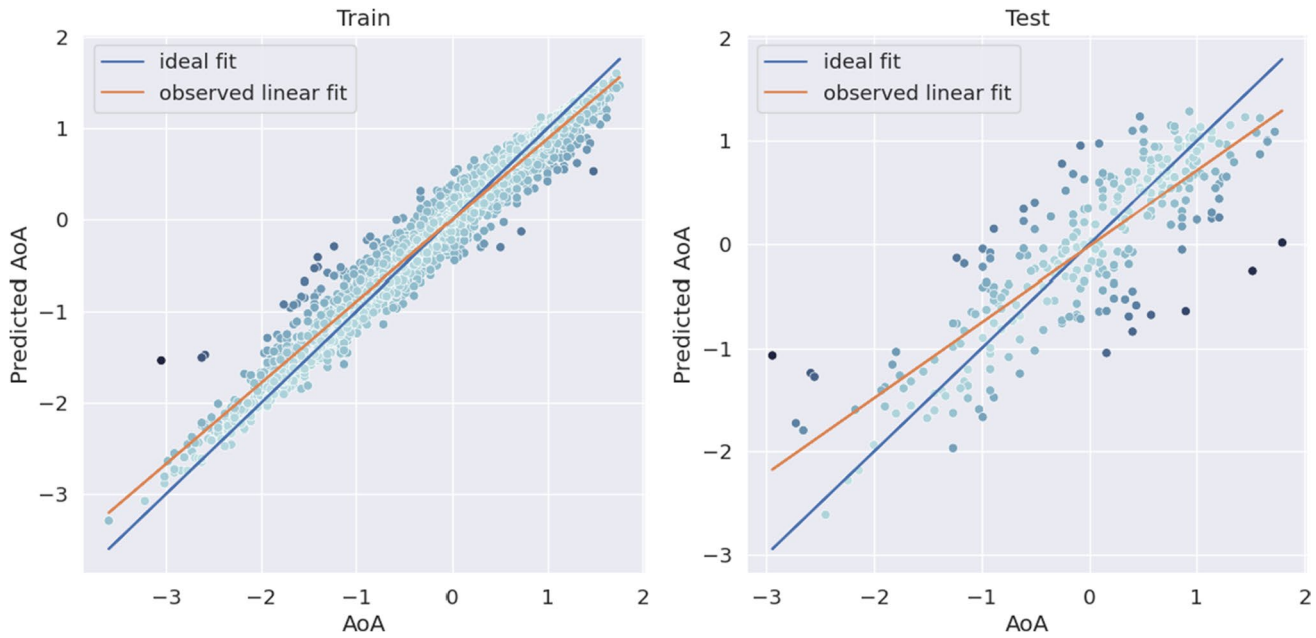

Fig. 13 Results for Shock score prediction

comparison to the other experiments. Nonetheless, collectively, the results indicate that the AoE v2 model generalizes well to a wide variety of AoA scores. However, this method may only be applicable to AoA lists that comprise at least a few thousand words.

Observing the evolution of AoE cosine similarities

Figure 15 illustrates the alignment discrepancy between the intermediate models at each stage and the mature model, as described by the cosine similarity between the intermediate

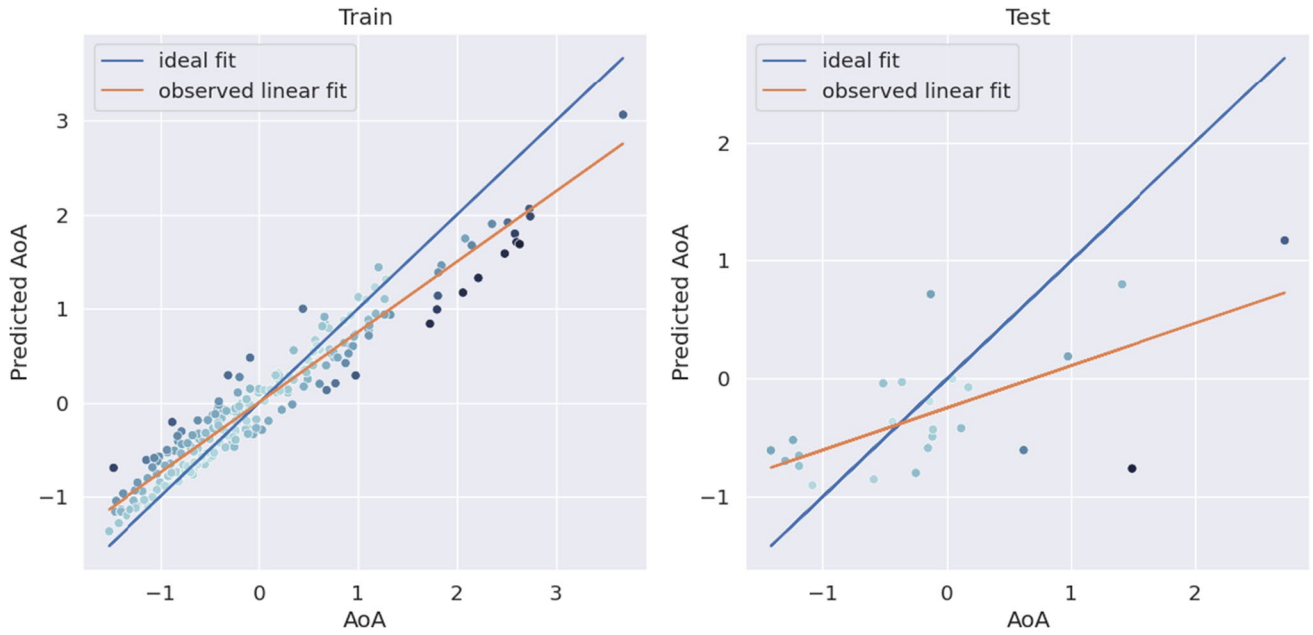

Fig. 14 Results for Morrison Objective AoA score prediction

vector representation of a word and its mature vector representation. The same words from Dascalu et al. (2015) were selected. A cool-to-warm color gradient is used for each word based on its frequency (i.e., the rarest word is "singularity" and the most frequent word is "class," with "chocolate" having the average frequency between the seven selected words). Plots representing the words have an upward tendency, with the cosine similarity rapidly approaching 1, a value which signifies perfect overlap. These cosines are computed after the intermediate vector space was rotated to align it with the mature vector space. Of the seven words shown, there is a clear differentiation of more specialized words, such as "singularity," "clustering," and "virus" and the more common words "chocolate," "happy," "tech," and "class." Slopes tend to be smaller when the intermediate word representation is more similar to the one produced by the mature model. Overall, this figure illustrates that frequent words tend to have higher cosine similarities in early stages in comparison to rarer terms.

Correlation with text difficulty

In order to evaluate the potential utility of the predicted scores, we performed an experiment to measure the correlation between the generated AoE scores and text difficulty. To this end, we analyzed the StairStepper corpus (Balyan et al., 2020; Perret et al., 2017), which contains 162 expository texts rated in difficulty ranging from grade 1 to grade 12. The predicted AoE scores generated by the best-performing models from Table 5 for both AoE 1.0 and AoE 2.0, as well as the original Kuperman AoA scores and Word Maturity, were evaluated. In order to provide a single measurement of text difficulty, we aggregated the scores for the words of that text using a simple average, and then measured the Spearman rank correlation between the predicted scores of the 162 texts and their marked grades.

The results reported in Table 8 indicate that the scores predicted by our method outperform the Word Maturity and AoE 1.0 indices, and even the Kuperman scores themselves. The results also show that the initial version of AoE had a substantially lower correlation with text difficulty. While it is perhaps unexpected for the predicted scores to achieve higher performance than the reference scores, the latter result may be due to the larger vocabulary being covered by AoE 2.0.

Impact of lexical, WordNet, and word trajectory features

In addition to the features derived from the exposure trajectories, we also added several features that are generated using external resources, such as WordNet (Miller, 1995). Because the number of syllables, the hyponym/hypernym trees, and the number of synsets depend on language-specific implementations, they may not be available in all cases. With this in mind, we performed an ablation study on the features used to predict the Kuperman AoA word list by incrementally adding features starting from the simple lexical features and then adding WordNet features and, finally, the word trajectory features. We performed this study on

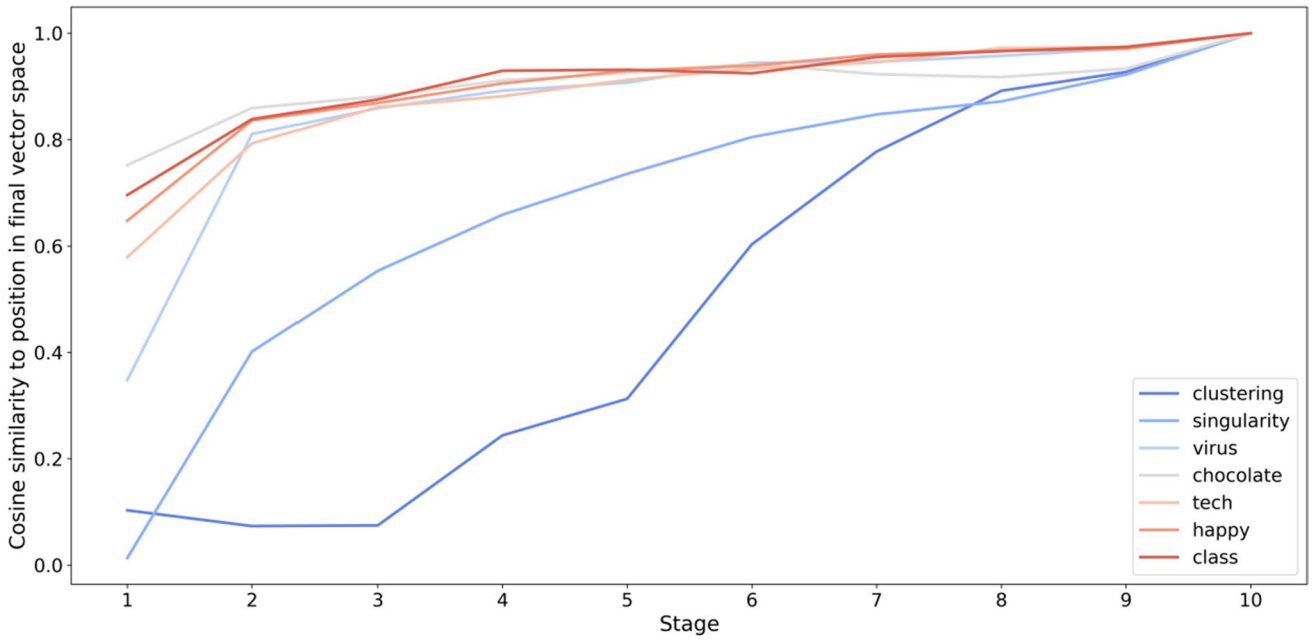

Fig. 15 Word vector evolution over the 10 word2vec models

Table 8 Correlation measurements with StairStepper text difficulty

Word list	Spearman rank correlations
AoE 1.0	.21
AoE 2.0	**.84**
Kuperman	.76
Word Maturity	.75

Bold marks the best results

Table 9 Ablation study on AoE v.2 prediction features

Feature set	MAE	Normalized MAE	R^2
Lexical features	1.612	.077	.513
WordNet features	2.225	.106	.126
Word trajectory features	1.667	.079	.495
Lexical + word trajectory features	1.558	.074	.545
All features	1.513	.072	.572

our best-performing experiment, namely the one in which the texts were sorted in increasing order of readability, and reported the results for the random forest regressor.

The results presented in Table 9 show that the lexical features and the word trajectory features alone capture a significant amount of the Kuperman AoA scores variance, with the combination between lexical and word trajectory features resulting in a significant increase in performance. Additionally, the absence of the WordNet features has a negligible impact on the overall performance. As such, our method can be applied even in contexts wherein certain external word feature resources, such as those reported by WordNet, are not available—with minimal loss of accuracy.

Discussion

The objective of this study was to generate AoA scores using a novel AoE 2.0 model in order to reduce the cost and effort to produce human AoA scores, increase the total number of words available in AoA databases, and reduce the error of adult-based corpora by simulating the word learning process. AoE 2.0 was able to model AoA word lists with R^2 coefficients ranging from .35 for the objective, child derived, AoA norms (Morrison et al., 1997), .57 for the Kuperman AoA list (Kuperman et al., 2012), and up to .74 for the Cortese AoA norm (Cortese & Khanna, 2008). Methodologically, we explored the impact of ordering the documents used to train the word vector models in AoE by using the Flesch Reading Ease score. Our findings suggest that sorting the texts in ascending order of their readability has a significant positive impact on the performance of the model. Despite the influence of ordering documents in the corpus, the results with a simple linear growth and an unsorted corpus are still relatively accurate, which means that good estimations of AoA can be obtained without the need of calculating readability scores.

AoE 2.0 directly links the word trajectories described in Word Maturity and AoE 1.0 to various AoA scores by generating features from the word trajectories that are then used to train a regressor capable of predicting AoA values for words with low error. This leads to a more straightforward use case than previous work: our method produces scores that have direct interpretation, as given by AoA, and may be used to generalize to words for which studies were not previously performed. The addition of non-AoE lexical features was shown to increase the ability of our method to model human AoA ratings, with our analysis suggesting that our method can perform well even when omitting features that rely on external resources, such as WordNet. While the inclusion of these features can be seen as a step back from the original purpose of AoE models, we believe that the ability of our method to simulate human AoA ratings with low error can be of great use in extending existing AoA word lists with ratings for new words that align well with human ratings traditionally used in AoA.

Moreover, we have investigated the influence that the ordering of documents has on the performance of AoE. We have found that simulating human word exposure by using increasingly more complex texts, as approximated using the Flesch Reading Ease, results in a significant increase in performance. One explanation relates to how humans are exposed to words. When adults speak to children, they typically use child-directed speech, which is simpler and more repetitive than adult-directed speech (Hills, 2013). Once children have been exposed to simpler words and sentences in their early childhood, they can construct a vocabulary (Hills et al., 2010) which affords them the ability to understand more complex words and sentences in adult-directed speech (Hills, 2013). By sorting the texts by readability, the model mimics a potential trajectory of exposure to simple words and sentences, followed by exposure to difficult words and sentences.

AoE 2.0 may also be used for downstream tasks, such as measuring the complexity of a text. In our experiments, we have shown AoE 2.0 scores to be a more accurate predictor of textual complexity than Word Maturity and AoE 1.0. In fact, the scores that our method generates were found to have better correlations to the human textual complexity scores in the StairStepper corpus than to the Kuperman scores that they were initially trained to model. The increased accuracy, in comparison with previous methods, has implications for the creation and evaluation of educational materials. Students may benefit from materials that target their zone of proximal development (Shabani et al., 2010). In the context of reading and text complexity, more accurate AoA scores can potentially be used to match students to appropriate texts by providing a more accurate assessment of when students know, or are able to learn, the words in the text.

There are several limitations to our method that should be noted. First, we did not check the accuracy of spaCy's POS taggers on the documents in our corpora, which may present issues especially for the CHILDES data. Second, the task of training predictors to infer AoA scores falls under the category of semantic norm extrapolation (Snefjella & Blank, 2020) and could be interpreted as a missing data problem. Under this hypothesis, the usage of the inferred AoE scores could be subject to biases that would make them not interchangeable with the empirically obtained AoA scores. Nevertheless, our method shows clear improvements over the existing AoE method and the conducted text difficulty experiment indicates that the generated word scores are at least comparable in practice with their training AoA word list. Another limitation of our method is that it relies on an existing AoA word list to infer scores for words in the vocabulary that the list does not cover. When extending our method to other languages, the availability of AoA word lists, as well as their sizes and distributions, may cause issues. Finally, our method requires the training of word embedding models on increasingly larger subsets of a corpora. While word2vec is an efficient algorithm for such cases, the computational cost is still significant, and the availability of corpora that have sufficient diversity in text difficulty can also be an issue for other languages. However, word2vec can handle large datasets better than the LDA and LSA used in AoE 1.0 and WM because of its small memory footprint and iterative training procedure. Additionally, we assume that the use of alternative word embedding algorithms may improve the quality of the exposure trajectories and should be considered in future research.

Conclusions and future research directions

AoE 2.0 offers an automated way of generating AoA scores that significantly reduce the cost and effort in collecting human AoA scores. AoE 2.0 may also reduce error in AoA scores by simulating the word learning process through increasing word exposure instead of relying on adults' recollection of word learning. We show the AoA norms generated by the AoE 2.0 model can be used to predict a wide array of readability ratings. Our experiments strongly suggest that our method can be applied to different word lists with consistently good performance.

Our AoE 2.0 takes advantage of many of the aspects outlined by Word Maturity and Age of Exposure 1.0. The word trajectories described by Landauer et al. (2011) are modeled through word2vec instead of LSA which, in addition to being more computationally efficient, also has the advantage of generating word embeddings that better capture the relations between words. In contrast to the initial Age of Exposure model, the new version avoids the need for determining

the optimal number of LDA topics, with word2vec being notably more robust to changes in vector dimensionalities. In addition, graph-based topic alignment via bipartite matching using Jensen-Shannon dissimilarity and max-flow algorithms is replaced with the Procrustes rotation used in Word Maturity, which is more straightforward and more robust because it does not rely on the distribution similarities measured using Jensen-Shannon dissimilarity.

Because AoE 2.0 is flexible, accurate, and scalable, multilingual comparisons become possible and the AoE 2.0 approach could be used to generate scores and word trajectories to compare words across languages. The scores generated by our model and their corresponding features may be compared with equivalents in other languages in order to determine which terms are acquired later on in some languages. For example, an initial study of using AoE 2.0 for modeling multiple non-English languages (Botarleanu et al., 2021) shows that the word exposure trajectories are able to reflect some aspects of human word acquisition among various languages. Thus, specialized curricula suitable for particular language speakers may be constructed using AoE 2.0.

Additionally, this method may be applied to other languages to investigate the way words evolve in a similar or dissimilar manner across languages. This may be achieved by training unsorted word embedding models on iterative corpora for different languages, and then analyzing the way equivalent words evolve across translations. Through this, the generation of AoA scores for new languages may be possible, thus offering a multilingual metric that does not require human crowdsourcing efforts. In addition, more specialized documents may be used to model the evolutions of domain-specific words by analyzing the word embedding trajectories generated by the iterative training process specific to the AoE and WM models, thus allowing for the evaluation of how different registers and genres as reflected in various corpora may impact a language user's ability to acquire various terms. Given the reliance of AoE 2.0 on data gathered from human estimations and child-based ratings, there is likely a series of optimal points at which the switch can be made from child-based ratings to adult estimates and then to automatically generated AoA scores when constructing a new AoA word list for a language. This constitutes a possible avenue for future research, in order to find the number of words needed that both minimize the data gathering costs, as well as provide data of sufficient quantity and quality for AoE 2.0 to take over and generate AoA scores for the entire vocabulary.

A notable benefit of AoE 2.0 is that its models are available for public use (cf., the Word Maturity model was not publicly available at the time of our experiments). Our AoE 2.0 model is released as an open-source project available at: https://github.com/readerbench/Age-of-Exposure. The AoE lexical scores based on the random forest regressor model are also available within the repository.

In conclusion, measuring AoA is important to research on language and discourse, including but not limited to lexical decisions, text comprehension, and writing quality. Our objective is to provide a means to estimate AoA scores with a publicly available and flexible model. We believe that this method of modeling word evolutions can have a variety of applications, such as helping to design better learning materials in both English and multilingual settings, improving the performance of downstream tasks such as measuring textual complexity and expanding existing AoA lists of limited size through model inference.

Appendix

Appendix 1. Impact of the Growth Scheme

In order to analyze the way in which the growth scheme of the corpus impacts the performance of the models we propose two growth functions: a linear and an exponential growth scheme (see Appendix Fig. 16). The former describes a constant rate of increase in the size of the corpus to which the model is exposed, while the latter describes an exponential increase.

Both linear vocabulary growth and exponential vocabulary growth were considered because children learn words at variable rates. Previous research suggests a "vocabulary spurt," a period of exponential word growth beginning when children are 18–24 months old (Goldfield & Reznick, 1990; Robinson & Mervis, 1998). In contrast, other research has reported that children demonstrate variable word learning rates (Bates et al., 1995) and re-examination of the vocabulary spurt has demonstrated that not all children experience the same exponential growth (Ganger & Brent, 2004). It is now widely accepted that the word learning rate is variable

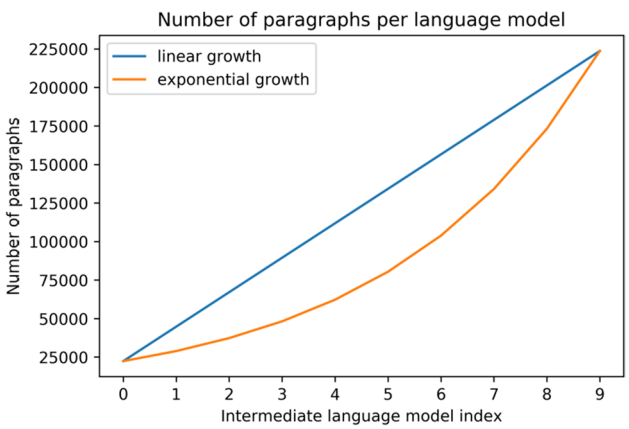

Fig. 16 Word complexity dataset growth schemes

and depends on the individual differences of the learner (Fernald & Marchman, 2012; Rowe et al., 2012).

We split the dataset into 10 stages (9 intermediary models and the last, complete and mature model). The corpus sizes are cumulative and can be described mathematically with the formulas described below. The linear growth scheme can be described as:

$$D_t = \left\{ C_{t\frac{|C|}{T}} \dots C_{(t+1)\frac{|C|}{T}} \mid t \in \{1..T\} \right\} \quad (9)$$

where D_t is the subset at step t, C is the corpus and T represents the number of desired sets within the corpus. In our experiments, we use 10 stages: 9 intermediate models and 1 mature model, which is trained on the entire dataset.

Similarly, for the exponential growth function, the subset at a timestep D_t can be described by the following equation:

$$D_t = \left\{ C_{\frac{|C|}{T}*10^{t/(T-1)}} \dots C_{\frac{|C|}{T}*10^{(t+1)/(T-1)}} \mid t \in \{1..T\} \right\} \quad (10)$$

Training the best-performing models with the different growth schemes indicates that there is little difference in performance between the two methods (see Appendix Table 10). In addition, post hoc tests showed that exponential growth had significantly less error than linear growth,* $t(269{,}770) = 8.96$, $p < 0.01$.

Table 10 AoE v2 Kuperman AoA prediction results – growth schemes

Growth scheme	Model	MAE	Normalized MAE	R^2
Linear	Random forest regressor	1.513	.072	.571
Exponential	Random forest regressor	1.510	.072	.575

Appendix 2. Impact of Word Embedding Size

In order to better understand how the word2vec model configuration affects the performance of the AoA regressors, we performed an experiment on one of the more important hyperparameters of word2vec: the size of the word embeddings. Results from Appendix Table 11 indicate that the vector size has a minimal impact on the final performance for Kuperman AoA prediction. The default value, 300, appears to be an effective choice. All results are reported using the best-performing model: random forest regressor with sorted data. Given the increase in computational cost from a vector size of 300 to a vector size of 1000, the improvement in performance is marginal. As such, we elected to use the default vector size of 300.

Table 11 AoE v2 Kuperman AoA prediction results – word2vec vector size

Vector size	MAE	Normalized MAE	R^2
100	1.514	.072	.570
200	1.515	.072	.572
300 (default)	1.513	.072	.571
400	1.515	.072	.573
500	1.509	.072	.573
1000	1.512	.072	.571

The data and materials for all experiments are available at https://github.com/readerbench/Age-of-Exposure and the experiment was not preregistered.

Acknowledgements This research was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number TE 70 *PN-III-P1-1.1-TE-2019-2209,* ATES – "Automated Text Evaluation and Simplification," the Institute of Education Sciences (R305A180144 and R305A180261), and the Office of Naval Research (N00014-17-1-2300; N00014-20-1-2623). The opinions expressed are those of the authors and do not represent views of the IES or ONR. We would also like to thank Prof. Peter Foltz for providing the Word Maturity indices that were used as a baseline in this paper.

References

Alonso, M. A., Fernandez, A., & Díez, E. (2015). Subjective age-of-acquisition norms for 7,039 Spanish words. *Behavior Research Methods*, *47*(1), 268-274.

Álvarez, B., & Cuetos, F. (2007). Objective age of acquisition norms for a set of 328 words in Spanish. *Behavior Research Methods, 39*(3), 377-383.

Balyan, R., McCarthy, K. S., & McNamara, D. S. (2020). Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. *International Journal of Artificial Intelligence in Education*, *30*(3), 337-370.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* (pp. 238-247).

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. *The Handbook of Child Language, 30*, 96-151.

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T. K., & Foltz, P. W. (2014). Models of vocabulary acquisition: Direct tests and text-derived simulations of vocabulary growth. *Scientific Studies of Reading*, *18*(2), 130-154.

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.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

Botarleanu, R.-M., Dascalu, M., Watanabe, M., McNamara, D. S., & Crossley, S. A. (2021). Multilingual age of exposure. In *22nd International Conference on Artificial Intelligence in Education (AIED 2021)*. Utrecht, Netherlands (Online).

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Philadelphia.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods, 49*(4), 1520-1523.

Brysbaert, M., & 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.

Chalard, M., Bonin, P., Méot, A., Boyer, B., & Fayol, M. (2003). Objective age-of-acquisition (AoA) norms for a set of 230 object names in French: Relationships with psycholinguistic variables, the English data from Morrison et al. (1997), and naming latencies. *European Journal of Cognitive Psychology*, *15*(2), 209-245.

Cortese, Michael J, & Khanna, Maya M. (2008) Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods 40*(3), 791-794.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering, 14*(3), 391-403.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning, 21*(2-3), 170-191.

Crossley, S., Feng, S., Cai, Z., & McNamara, D. S. (2013). Computer simulations of MRC Psycholinguistic Database word properties: Concreteness, familiarity, imageability. In S. Jarvis, & M. Daller (Eds.), *Vocabulary Knowledge: Human Ratings and Automated Measures*. (pp. 135-156). John Benjamins Publishing Company

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes, 54*(5-6), 340-359.

Dascalu, M., McNamara, D. S., Crossley, S. A., & Trausan-Matu, S. (2015). Age of Exposure: A Model of Word Learning. In *30th AAAI Conference on Artificial Intelligence* (pp. 2928–2934). AAAI Press.

Davies, M. (2008). The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/. Accessed 10 Jan 2022.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). ACL.

Di Carlo, V., Bianchi, F., & Palmonari, M. (2019). Training temporal word embeddings with a compass. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 6326-6334). AAAI Press.

Eckerth, J., & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. *Language Teaching Research*, *16*(2), 227-252.

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development*, *83*(1), 203-222.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language, 44*(3), 677-694.

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, *40*(4), 621.

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, *115*(1), 43-67.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12*(4), 395-427.

Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of Child Language*, *17*(1), 171-183.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika, 40*(1), 33-51.

Grigoriev, A., & Oshhepkov, I. (2013). Objective age of acquisition norms for a set of 286 words in Russian: Relationships with other psycholinguistic variables. *Behavior Research Methods*, *45*(4), 1208-1217.

Hills, T. (2013). The company that words keep: comparing the statistical structure of child-versus adult-directed language. *Journal of Child Language*, *40*(3), 586-604.

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

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*, 1368–1378.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development, 73*(2), 418-433.

Ivens, S. H., & Koslin, B. L. (1991). *Demands for reading literacy require new accountability methods.* Touchstone Applied Science Associates.

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, *13*(7-8), 789-845.

Justice, L. M., Petscher, Y., Schatschneider, C., & Mashburn, A. (2011). Peer effects in preschool classrooms: Is children's language growth associated with their classmates' skills? *Child Development*, *82*(6), 1768-1777.

Kaufman, A.S., & Kaufman, N.L. (1983). Kaufman Assessment Battery for Children. Circle Pines, MN: American Guidance Service.

Kaufman, A.S., & Kaufman, N.L. (1990). *Kaufman Brief Intelligence Test*. Pearson, Inc.

Krzanowski, W. J. (2000). *Principles of Multivariate Analysis, Revised Edition*. Oxford University Press.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Landauer, T.K., Kireyev, K., & Panaccione, C. 2011. Word maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15*(1): 92–108

Lenci, A., Sahlgren, M., Jeuniaux, P., Gyllensten, A. C., & Miliani, M. (2021). *A comprehensive comparative evaluation and analysis of Distributional Semantic Models.* arXiv. 2105.09825.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics, 3*, 211-225.

Li, J., & Jurafsky, D. (2015). Do Multi-Sense Embeddings Improve Natural Language Understanding? In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1722-1732). ACL.

Lund, K., & Burgess, C. (1996) Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers 28*(2) 203-208.

Łuniewska, M., Haman, E., Armon-Lotem, S. et al. (2016). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*, *48*(3), 1154-1177.

MacWhinney, B. (1996). The CHILDES system. *American Journal of Speech-Language Pathology*, *5*(1), 5-14.

Maddux, C. D. (1999). Peabody Picture Vocabulary Test (PPVT-III). *Diagnostique, 24*(1-4), 221-228.

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.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Mikolov, T. (2013). Word2vec-toolkit [Online forum comment]. Retrieved from https://groups.google.com/forum/#!searchin/word2vec-toolkit/c-bow/word2vec-toolkit/NLvYXU99cAM/E5ld8LcDxlAJ. Accessed 10 Jan 2022.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). *Efficient estimation of word representations in vector space.* arXiv. 1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K.Q. Weinberger (Eds.), Proceeding of the 26th International Conference on Neural Information Processing Systems* (pp. 3111-3119).

Mikolov, T., Le, Q. V., & Sutskever, I. (2013c). *Exploiting similarities among languages for machine translation.* arXiv:1309.4168.

Miller, G.A., 1995. WordNet: A lexical database for English. *Communications of the ACM 38*, 11, 39–41.

Montefinese, M., Vinson, D., Vigliocco, G., & Ambrosini, E. (2019). Italian age of acquisition norms for a large set of words (ItAoA). *Frontiers in Psychology*, *10*, 278.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A. L., ... & Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*(1), 169-177.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology Section A*, *50*(3), 528-559.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, *24*(2), 237-270.

Nelson, J., Perfetti, C., Liben, D., and Liben, M., 2012. *Measures of text difficulty: Testing their predictive value for grade levels and student performance.* Council of Chief State School Officers.

Pan, B. A., Rowe, M. L., Singer, J. D., Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development, 76*, 763–782.

Perret, C. A., Johnson, A. M., McCarthy, K. S., Guerrero, T. A., & McNamara, D. S. (2017). StairStepper: An adaptive remedial iSTART module. In B. Boulay, R. Baker & E. Andre (Eds.), *Proceedings of the 18th International Conference on Artificial Intelligence in Education (AIED),* (pp.557-560), Springer.

Robinson, B. F., & Mervis, C. B. (1998). Disentangling early language development: Modeling lexical and grammatical acquisition using and extension of case-study methodology. *Developmental Psychology*, *34*(2), 363.

Rowe, M. L., Raudenbush, S. W., & Goldin-Meadow, S. (2012). The pace of vocabulary growth helps predict later vocabulary skill. *Child Development*, *83*(2), 508-525.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences, 112*(41), 12663-12668.

Ruas, T., Grosky, W., & Aizawa, A. (2019). Multi-sense embeddings through a word sense disambiguation process. *Expert Systems with Applications*, *136*, 288-303.

Schönemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika, 31*(1), 1-10.

Shabani, K., Khatib, M., & Ebadi, S. (2010). Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. *English Language Teaching, 3*(4), 237-248.

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

Snefjella, B., & Blank, I. (2020). Semantic Norm Extrapolation is a Missing Data Problem. *ArXiv preprint.* https://doi.org/10.31234/osf.io/y2gav

Stadthagen-Gonzalez, Hans, & Davis, C. J. (2006). The Bristol Norms for Age of Acquisition, Imageability and Familiarity. *Behavior Research Methods, 38*, 598–605.

Teng, F. (2019). The effects of context and word exposure frequency on incidental vocabulary acquisition and retention through reading. *The Language Learning Journal*, *47*(2), 145-158.

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics, 71*, 249-267.

Webb, N. M. (1991). Task related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education, 22*, 366–389.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143-2152.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In: *The 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).* Vancouver, Canada.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics, 35*(5), 2173-2192.