


How do the kids speak? Improving educational use of text mining with child-directed language models


Peter Organisciak
University of Denver
Denver, CO, USA

peter.organisciak@du.edu

 0000-0002-9058-2280


Michele Newman
University of Washington
Seattle, WA, USA

mmn13@uw.edu

 0000-0002-5293-7992


David Eby
University of Illinois at Urbana-
Champaign
Champaign, IL, USA

davidme2@illinois.edu

 0000-0002-9722-0215

Selcuk Acar
University of North Texas
Denton, TX, USA

selcuk.acar@unt.edu

 0000-0003-4044-985X

Denis Dumas
University of Georgia
Athens, GA, USA

Denis.Dumas@uga.edu

 0000-0002-8446-4720

ABSTRACT

Purpose

Most educational assessments tend to be constructed in a close-ended format, which is easier to score consistently and more affordable. However, recent work has leveraged computation text methods from the information sciences to make open-ended measurement more effective and reliable for older students. This study asks whether such text applications need to be adapted when used with samples of elementary-aged children.

Design/methodology/approach

This study introduces domain-adapted semantic models for child-specific text analysis, to allow better elementary-aged educational assessment. A corpus compiled from a multi-modal mix of spoken and written child-directed sources is presented, used to train a children's language model, and evaluated against standard non-age-specific semantic models.

Findings

Child-oriented language is found to differ in vocabulary and word sense use from general English, while exhibiting lower gender and race biases. The model is evaluated in an educational application of divergent thinking measurement and shown to improve on generalized English models.

Originality

Research in computational measurement of open-ended responses has thus far used models of language trained on general English sources or domain-specific sources such as textbooks. This paper is the first to study age-specific language models for educational assessment. Additionally, while there have been several targeted, high-quality corpora of child-created or child-directed speech, the corpus presented here is the first developed with the breadth and scale required for large-scale text modeling.

Research limitations/implications

The findings demonstrate the need for age-specific language models in the growing domain of automated divergent thinking and strongly encourage the same for other educational uses of computation text analysis by showing a measurable difference in the language of children.

Social implications

Understanding children's language more representatively in automated educational assessment allows for more fair and equitable testing. Further, child-specific language models have fewer gender and race biases.

KEYWORDS

Educational data mining, text mining, learning, assessment, language modeling, divergent thinking

1 INTRODUCTION

Recent advancements in natural language processing are enabling a key advancement in education: the ability to parse open-ended measurement responses reliably and consistently. Doing so greatly opens the ability to measure knowledge and abilities that have traditionally not been well suited to close-ended testing, such as measures of originality and divergent thinking, which have been costly and uneven in the past (Acar et al 2021; Dumas & Dunbar 2014; Dumas et al., 2020). However, realizing the possibility of computational methods for improving assessment, particularly in contexts that involve children, requires tools that meet the needs of educational domains while fulfilling expectations of transparency and interpretability.

In this work, a corpus of child-directed language is developed and modeled to better understand how the linguistic profile of children’s language differs from general English. The MOTES Corpus, emergent from the Measurement of Original Thinking in Elementary Students project, differs from past children’s corpora by focusing on scale, which allows it to be used in modern computation text mining applications. This scale is achieved by focusing on child-directed resources, which are more comprehensively available than child-spoken or child-produced corpora. The child-directed model is compared to a general-language model in three ways: through an empirical linguistic analysis, where notable differences in language use are observed; through a bias analysis, where the children’s corpus is found to lead to lower race and gender stereotyping; and in an applied context, on tests of children’s divergent thinking from the MOTES project, where child-focused text models are found to outperform tradition models.

Applications of natural language processing often rely on models of relationships in general English. Entirely different words may mean very similar or identical things, and the relatedness of words needs to be represented for systems to understand that people talk or write about similar concepts in varying ways. In other words, models of relationships between words help applications focus on latent meanings rather than the specific words used in conveying those meanings. The best models are learned by observing great deals of text, so it has become commonplace for scholars and practitioners to use pre-trained, openly distributed models in a process called *transfer learning* (Zhuang et al. 2020). Doing so aids reproducibility while avoiding the complex task of corpus building and model training.

In computational approaches to education and learning, there is reason to expect that out-of-the-box models are insufficient, particularly when working with younger children. It has been shown that the domain of the texts on which a model is trained can affect the model significantly (Brunet et al., 2019). For example, one popular model was trained on a corpus from Google News (Mikolov et al., 2013), and the biases in the word relationships that it encodes are characteristic of news articles, including problematic biases, such as a tendency to over-represent male voices for various professions (Bolukbasi et al., 2016; Manzini et al., 2019). Further, while large models of general language have represented a big improvement for natural language processing applications in recent years, it has been shown that further pre-training those models on domain specific texts, as is done with children’s language here, improves their performance (Gururangan et al 2020; Konlea & Jannidisa 2020). In applied use, it is imperative that underlying models of language represent the language of the people they affect.

In one area of educational measurement, creativity and original thinking, models trained on a generic web text have performed best in automated scoring of adults (Dumas et al., 2020). Given the sensitivity of language models to the underlying domain of texts, there is concern that building similar systems for children with similarly generic, predominantly adult language may confound the measurement of creativity with a child’s vocabulary and the various developmental, educational, and socioeconomic

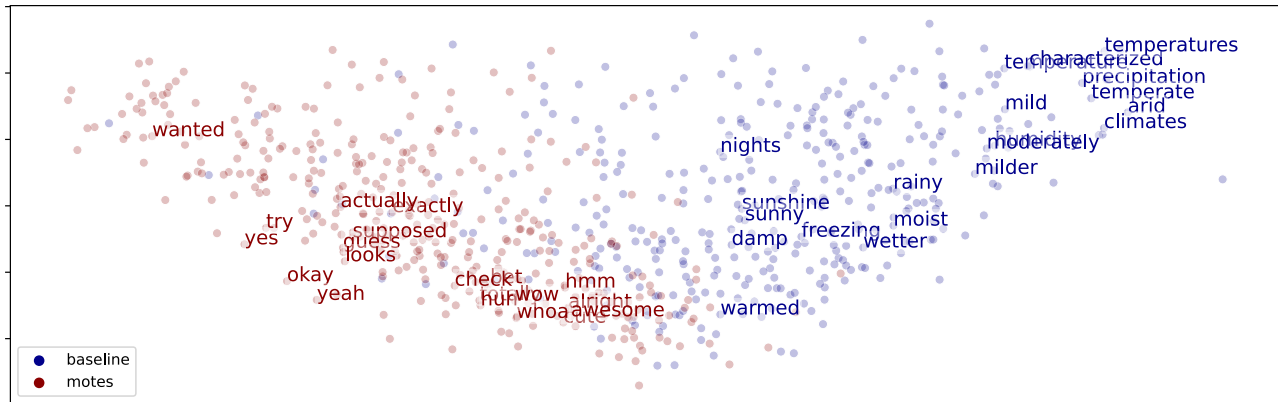


Figure 1: Nearest neighbors of the word 'cool' in the child-specific MOTES model (left/red) and a general model (right/blue), shown in two dimensions with Principal Component Analysis.

factors that affect it. The Corpus was developed to support a measure called MOTES (Measurement of Original Thinking in Elementary Students), aimed at measuring originality in late-elementary-aged (e.g., grades 3-5) children. However, it has a more general value in text mining applications which aim to accurately understand children's language, allowing for better representation of that language without sacrificing the convenience of out-of-the-box, transferable models.

This paper investigates the language profile of a set of child-oriented media: (a) in videos for children posted to the video website YouTube, (b) in the language of children's television shows, (c) in children's books, and (d) on the child and English-language-learner-focused Simple English Wikipedia. The mix of sources is intended to represent the language that children encounter, both aurally and textually. In comparison with a baseline model, the children's corpus model is found to exhibit differences in word use that better align with children's language. One such example is seen in Figure 1, which displays the tendency of a general English corpus to use *cool* in relation to weather and temperature, where the child-specific corpus uses it as an intensifier.

This work is notable for several reasons. First, it contributes a child-oriented model for computational psychometric approaches in education, built on a children's corpus of unprecedented scale. Second, it studies the language profile of child-directed media in a novel way, through the lens of a trained word embedding or neural language model. The process of modeling aims to accurately learn and represent relationships between words within a language, making it ideal for studying the biases and tendencies of different corpora separate from quirks of term counts (Gonen et al., 2020; Hamilton et al., 2016). Notably, child-oriented models – and in turn the texts they are informed by – exhibit lower gender and race biases. Finally, the efficacy of such a tailored model is demonstrated in an educational assessment context, originality scoring, where this paper finds it outperforms a general language model in scoring children's tests.

2 BACKGROUND

Research in educational assessment has a history of using language models to score students' open-ended writing (Miller, 2003; Millis et al., 2006), an area that has been growing in the emerging domain of *computational psychometrics* (von Davier, 2017). One common tool used in the domain is a *word embedding model*, which seeks to represent words in a lower-dimensional linear space. Geometric distance in such a model indicates how related or unrelated the words are. Word embedding models and their predecessors (Deerwester et al., 1990; Mikolov et al., 2013; Pennington et al., 2014) are also more

explainable than the newest large language models, which operate on weights of millions of neural network parameters rather as distance in a geometric space. This explainability is an important factor in educational applications.

An area of interest in word embedding models is in original thinking measurement (Acar & Runco 2014; Dumas & Dunbar 2014; Dumas et al., 2020; Forthmann et al., 2019), which is historically difficult to measure since measuring novel ideation does not lend itself well to close-ended items. Essentially, if student originality and creativity is the focus of an assessment, then students must provide open-ended and ill-structured responses. The need to use an open-ended response format has severely limited the application and interpretation of creativity assessments, this is because traditionally available scoring systems for these responses (e.g., human-judges) are costly, time-consuming, and inconsistent.

The limitations surrounding creativity assessment also influence what abilities schools can focus on supporting within their students. Requirements on school systems mean that if public funds are spent to support a particular ability in students, students' concomitant growth and improvement in that ability must be measured and tracked (Lissitz & Jiao, 2014). Therefore, without cost-effective, reliable, and valid creativity assessments, schools have historically struggled to situate creative thinking within their curricula and measurement.

There has been emerging success applying computational methods to scoring original thinking in adults (Acar & Runco 2014; Dumas & Dunbar 2014; Forthmann et al., 2019). Compared to human graders, a well-performing computational instrument can be performed with more consistency and at lower cost. It also has the benefit of offering a more manifest view of its biases, allowing it to be studied and addressed in ways that are difficult for traditional graders. While various forms of language models import their own biases from the corpora on which they are trained, these may be studied, laid bare, and potentially corrected (Caliskan et al., 2017, Badilla et al., 2020; Sweeney & Najafian, 2019; Bolukbasi et al. 2016, Zhao et al. 2018).

However, positive results with adults do not mean that all the necessary means are available to conduct the same procedure with younger students. It has been shown that even with high-performing general models of the English language, domain-adaptive pre-training – fine-tuning the models on texts from the usage domain – consistently improves performance (Gururangan et al., 2020). Vocabulary and style may vary in different contexts. This change is particularly notable in language directed and used by children, which is not specific to a single topical domain but often adapted in style and vocabulary to better support language development.

In the remainder of this work, the literature on language modeling and on the differentiation of children's language from adult language is reviewed. Next, a domain-specific corpus is collected, and incorporated into a model of child-directed language. The model's deviations from general English are subsequently studied, including a bias analysis, and measuring in an originality scoring application.

3 RELATED WORK

Prior work has endeavored to develop child-oriented text corpora. An early project is CHILDES (MacWhinney, 2000), a part of the TalkBank project (MacWhinney, 2007). CHILDES is a corpus of children's spoken language built to study audio-based conversational interaction of children, collecting transcripts from numerous past studies. Similar corpora include WordBank (Frank et al., 2017), a collection of child vocabulary development data, and the CMU Kids Corpus (Eskenazi et al., 1997), an audio-based corpus of children reading predetermined sentences. Such corpora generally focus on

understanding semantics and grammar development within children (Rabkina et al., 2019). However, they tend to be more focused and smaller, due to their focus on traditional study and the effort in transcribing and cleaning such data. Automated transcription is also more challenges for child speech (Lileikyte et al 2022). The MOTES Corpus differs in its focus on learning a robust and transferable model for text mining applications, focusing on larger quantities of text and in turn, emphasizing child-directed media rather than child-produced text.

The form of language adapted to communicating with children is referred to as *child-directed speech*, or CDS. CDS entails both syntactic considerations – of word use and grammatical construction (e.g., Cameron-Faulkner et al., 2010; Cristia, 2019) – and verbal or prosodic construction – the rhythm and pitch at which it is spoken (e.g., Biersack et al., 2005; Zahner, 2016). The study of CDS often aims to understand how we adapt our language in interactions with children and how the linguistic environment for children shapes their learning. For example, work has found the importance of specifically child-directed versus overheard speech on children’s vocabulary development (Weisleder and Fernald 2013). Other work focuses on how infants and children acquire word segmentation ability in early learning (e.g., Saffran et al. 1996). It is also possible to study the presence and style of use for specific concepts, such as Pleyer’s look at the use of the term ‘pretend’ in children’s environments (2020). The present study pursues somewhat different goals, seeking to effectively model the in-the-wild syntactic and semantic diversity of children’s language for text mining applications, agnostic of the role of that language in children’s development.

Text-mining has emerged as an approach to analyze big data in education as part of the larger movement of Educational Data Mining, or EDM (Aldowah et al., 2019; Sachin & Vijay, 2012, Rodrigues et al., 2018). EDM is applied for a variety of formative and evaluative purposes, including assessment, and a systematic review identified 4.75% of EDM applications used text mining techniques (Aldowah et al., 2019). With the growth of online learning, researchers have paid special attention to finding patterns in large-scale educational data sets, to better support online learning and curriculum development with the assistance of machine learning (Romero & Ventura, 2013). These advancements in text and data mining allow for improvements in understanding learning, cognition, and assessment (Sachin & Vijay, 2012).

One area where text mining is increasingly applied to improving education is in measures of divergent and original thinking, which by their nature rely on testing with open-ended responses. Advanced semantic and grammatical capabilities have been a marker for the identification of gifted children for decades, and divergent thinking is found to correlate with more traditional intelligence measures (Preckel et al., 2006). The process of text modeling language, which seeks to represent related words and concepts closer together than divergent concepts, aligns well with the goals of divergent thinking research, which similarly aims to measure whether an idea is predictable or original. In divergent thinking research, using the distance between concepts in a semantic model as a proxy for divergent thinking has been validated for respondent across multiple responses (Dumas and Dunbar 2014; Acar and Runco 2014; Beaty and Johnson 2021, Dumas et al., 2020). However, the field is still addressing various challenges, such as that semantic distance is primarily a measure of novelty, without much regard to usefulness (Beaty and Johnson 2021), and is partially confounded by elaboration, referring to the detail of the response (Forthmann et al., 2019) and can be susceptible to adversarial nonsense responses, which ‘cheat’ by using semantically distant responses which are not valid uses (Dumas et al 2020). To date, work has focused on adults, using general English language models or models of language in textbooks. The present paper evaluates the need for tailored models for children’s applications.

Prior to text mining methods, originality was primarily scored with human judges. In applied school-based contexts, the Torrance Test of Creative Thinking (TTCT; Torrance, 1966) has been the most common test. It uses multiple tasks, such as a version of the alternate uses task (called the unusual uses task in their context), and a task for generating consequences to outlandish situations. Scoring is done based on a list of commonly produced ideas known as the zero-originality list, which has been generated on past tests, and a response is considered original if it is not on the zero-originality list. In research contexts, non-automated scoring is typically done by multiple trained judges on a graded scale (e.g., Silvia et al 2008) or based on the frequency of individual responses within the entire pool of responses produced by the study sample where only unique (ideas that were only presented by a single participant) or infrequent ideas (e.g., below a certain threshold such as 5%) are counted (Forthmann et al., 2019; Runco and Acar, 2012). Recent work (Forthmann and Doebler 2022) has rediscovered early automated scoring methods by Paulus (1970), which used word and grammar features for scoring TTCT tests.

Prior work has found value in adopting text-mining for measurement beyond divergent thinking scoring. For example, Huang et al. (2006) mapped responses of high school students to an educational film through the parsing of their text responses (Huang et al., 2006). In the social sciences, methods such as content analysis (Neuendorf & Kumar, 2015) and lexical analysis (Monroe et al., 2008) increasingly embrace computational methods. Word embedding models have been consistently used in assessment. Early work used *latent semantic analysis* (LSA), an approach that builds sparse term-document occurrence matrices, and factorizes them to a dense, low-dimension space (Deerwester et al., 1990). Landauer and Dumais (1997) have argued that models like LSA more closely align with how the mind works, arguing that the networks of latent concepts undergirding each word provided a path toward understanding language learning. More recent work continues in that tradition, and the question of robustness of word embedding models, particularly LSA, remains an active domain of study in education (Crossley et al., 2017; Cai et al., 2018).

A related type of model is the neural-network based large language model, which improves on word embedding models by including positional context – e.g., ‘bank’ is represented differently depending on whether it is in the context of ‘river’ or ‘teller’ – and using more sophisticated training methods to learn that context from large corpora (Vaswani et al 2017). These models, which include BERT (Devlin et al 2018), RoBERTa (Liu et al 2019), T5 (Raffel et al 2020), and GPT-3 (Brown et al., 2020), usually perform more robustly than word embedding models, other than cases where linear, geometrically comparable embeddings are desired (Reimers et al 2019). However, their lower explainability and interpretability of model output make them more difficult to apply in educational contexts. Especially when working with children, it is vital to maintain transparency in how a model decision is made in response to a given input.

The MOTES Corpus builds on the tradition of word embedding models and extends it, combining at scale both verbal and textual works intended for children.

4 DATA

For data collection, four sources of information were employed: YouTube subtitles for children's videos, closed captioning from children's television shows, a full dump of the Simple English Wikipedia, and the text of scanned children's books. The motivation was to span a variety of mediums and topics. The size of each of these components is shown in Table I.

Table I: Corpus sizes

Corpus	Number of Words
--------	-----------------

YouTube Made for Kids video captions	9.2 million
US Children’s Shows	12.4 million
Children’s books 1850-2010	582 million
Simple English Wikipedia	27 million

4.1 *YouTube Collection*

YouTube is a video sharing platform launched in 2005. It offers content primarily developed by users, but also content from larger entertainment entities. Due to its accessibility on a variety of devices, from televisions to smartphones, it is a prime source for entertainment for children. According to Pew, 81% of parents with children 11 or younger allow their child to watch YouTube, and 34% do so regularly (Smith & Kessel, 2018). Other research has found similar usage among children aged 0-7, who spend an average of nearly 1.5 hours per day on the service (Neumann & Herodotou, 2020a, 200b).

For data collection, a snowball sampling approach was employed to find videos that are made for kids. First, videos were searched in various topic categories using keywords such as 'kids' or 'children', filtered specifically to videos that have closed captioning and are relevant to English language searchers. Since Jan 2020, videos that are intended for children need to be marked as 'madeForKids' on YouTube (Alexander, 2020). From the sample pool, the videos tagged as 'madeForKids' were analyzed to identify other keywords that had a high log odds-ratio of occurring in children's videos, which in turn motivated new queries. Additionally, notable children’s channels were identified from the sample of videos and collected. English-language captions for all sampled videos were collected for the sample.

4.2 *Children’s Television Shows*

The corpus also pulled from a variety of subtitles from children’s shows across a variety of genres and networks that are aimed toward kids including Disney Channel, Nickelodeon, and Cartoon Network. The inclusion list for television shows was derived from the “List of Children's Shows” on Wikipedia. Subsequently, subtitles were collected from a subtitle website, addic7ed.org.

4.3 *Children’s Books*

For a corpus of children's books, scanned and automatically transcribed materials from the HathiTrust Digital Library were used. The HathiTrust is a non-profit consortium of institutions that collects scanned books from a consortium of institutions around the world (York, 2010). Its digital library comprises over 17 million volumes, about half of which are in English. Book data was acquired from the HathiTrust Research Center's Extracted Features Dataset, specifically to allow use of in-copyright works (Organisciak, 2017). The trade-off for access at such a scale is that this dataset only includes word counts, not information on their order. This requires a loosening of some of the assumptions in building a semantic model, which is discussed later.

For the MOTES Corpus, library metadata records were used to identify known children's books in the HathiTrust corpus, identifying books cataloged as [E] or [Fic] in the Dewey Decimal System – children up to age 8 and young adults, respectively – or with Library of Congress Classification numbers beginning with PZ7, *General juvenile belles lettres* 1870-. This resulted in a collection of 14345 books, or 2.2 million pages. Though the corpus skews modern (Figure 2), the corpus was split by date so that more recent books may be given stronger representation in modeling.

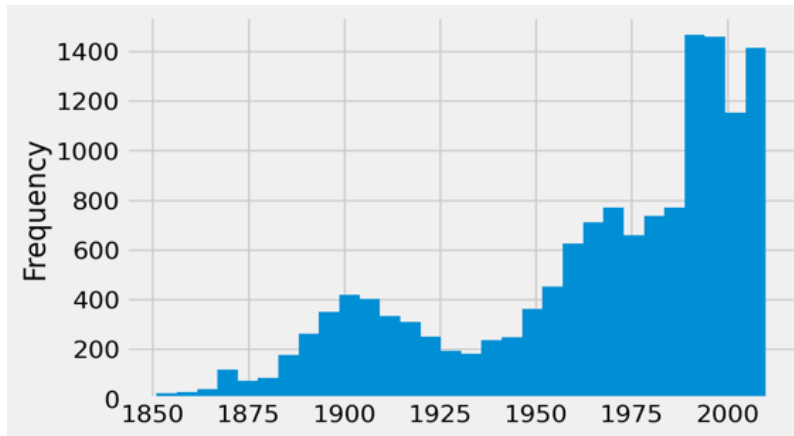


Figure 2: Date Distribution of Children's Books.

4.4 Simple English Wikipedia

Simple English Wikipedia is a collaborative online encyclopedia that focuses on using basic and clear English. Started in 2001, its audience includes "children and adults who are learning English" (*Simple Wikipedia Homepage*, n.d.) and boasts 170 thousand articles.

5 METHODS

5.1 Modeling

To understand the notable differences in language, word embeddings were trained for MOTES using the GloVe architecture (Pennington et al., 2014). GloVe is a form of word embedding model which has been used in educational contexts (Dumas et al. 2020). It seeks to learn a semantic distance vector space from word co-occurrences, where related concepts are close together. Functionally, a GloVe-trained model is similar to models such as Word2Vec (Mikolov et al., 2013) and LSA (Deerwester et al., 1990). While the outputs are similar for various semantic models, GloVe was chosen because its training procedure lends itself well to this study's goals of understanding how language differs in children's contexts. Here, we use modeling for analysis, inspecting the characteristics of a corpus through the lens of how it looked when modeled. GloVe is trained on a co-occurrence matrix, which makes it possible to stack difference subcorpora for training to see their effect on the model. That co-occurrence matrix is also a word-to-word matrix, counting when words occur within a window of proximity, which scale better for large corpora than a document-term matrix, like with LSA (Deerwester et al., 1990), which counts word co-occurrence in documents. Finally, as a semantic model, it can be inspected more directly than a large language model, such as BERT (Devlin et al 2018), which have millions of internal parameters.

For the purposes of this study, seeking to understand how language differs in children's contexts, GloVe operates directly on the cooccurrence matrix between words. This project makes use of that property to easily train on different mixes of corpora.

5.2 Preprocessing

For preparation, texts are tokenized and lowercased, though not modified through stemming or lemmatization. Stoplisting and other term removal is not done at the preprocessing stage, though is performed in comparable manners at the training and analysis stage. For each sub-corpus, a $n \times n$ training matrix is initialized, where $n = 400000$, aligned to the 400k word vocabulary used in an earlier GloVe model, the English Wikipedia+Gigaword model (Pennington et al. 2014). The vocabulary was

chosen simply as a broad English-language vocabulary. For each target word of each document, cooccurrences with a given context window w are saved in the training matrix. Here, $w = 10$ is used, a context window previously shown to be effective (see Pennington et al., 2014). A variable weight is assigned to the context words, $weight = \frac{w}{c_i}$, where c_i is the context distance from the target word, such that words that are further from the target word are weighted down. Counts below .2 in the final co-occurrence matrix are dropped, to remove noisy data and improve training performance. With the models reported here, the vocabulary was truncated to $n = 100000$, which preserves over 96% of the weighted counts seen in the full 400k token matrix.

5.3 Training on Bags of Words

The children's book sub-corpus presents a challenge to training, because it does not include positional word information – a so-called *bag of words* – and therefore, the context words around a given word are not known. This was addressed with adjustments that loosen the assumptions about context. Primarily, while the immediate context of a word is not known, there is signal in the fact that two words occurred together on the same page. Thus, an entire page is treated as a 'context'. This is different than a term-document matrix factorization approach, as it still does away with the concept of a document in training, and only counts words in proximity to each other.

One concern about this approach of using a full page as a context window is that pages have variable length, so longer pages can exert unfair influence on the co-occurrence patterns of the corpus. This is addressed by assigning all cooccurrences on a page a weight of $1/n(\text{context words})$. Thus, two words that occurred on a page of 11 words counts receive a weight of 0.1, whereas two words on a page of 101 only counts as 0.01.

Since the children's books in the dataset are derived from scanned texts, there may be issues with optical character recognition (OCR). Such issues are unevenly applied, stemming from challenges such as poor scans, old books, outdated OCR software, hard-to-OCR content such as tabular data, or content misidentified as text, such as illustrations. To limit the influence of any one scan, pages were truncated to a maximum of 400 randomly selected words. Removing the context window means co-occurrences are likely to be noisier and require more raw text to find a signal. This was aided by the scale of the book corpus: 582.2 pages. On the other end, to separate the noise from the signal, more aggressive filtering was performed for this corpus, dropping any target-context count below 1; with the length-based weighting, this is equivalent to 100 co-occurrence instances on 100-word pages.

The training adjustment employed here is entirely novel, to our knowledge, and may be used for other information science contexts where copyright limits access to full-text materials. For example, earlier work seeking to measure bias in K-12 books using the HathiTrust Digital Library was limited by the availability of full-text books (Mohan 2021).

5.4 Training

The GloVe training objective is used for training (Pennington et al., 2014). Each reported model has 300 dimensions and is trained with 100 epochs. This means that the resulting models place each of their vocabulary words into a 300 number address that puts it close (in a geometrically measurable sense) to synonymous or related words. The hyperparameters were kept constant in the interest of comparability between models.

For the baseline model, a 2020 dump of the English-language Wikipedia is used. Wikipedia is a common corpus for language modeling, because it is large, easy to work with, and comprehensive in

topics discussed. This work uses 317000 articles for training, and drop co-occurrence counts below 100 prior to training, which account for 3% of the counts.

The MOTES Corpus models are built on top of the baseline ‘general-language’ model. This is done by addition of co-occurrence matrices, effectively overlaying them. The choice to stack the children’s corpora on the base corpus is to ensure that the model still understands concepts outside the domain of children’s literature. It is inspired by fine-tuning in transfer learning for large language models, such as is commonly done with BERT (Vaswani et al., 2017). With fine-tuning, training for a large standard model is picked up and resumed with text in the domain of the target problem. This ensures that a sound foundation for the language is being used, while nudging the final model to the specific quirks of the language for the context at hand. Fine-tuning is less common for word embedding models like GloVe and pre-trained models generally are not distributed with enough information to resume training, though methods have been proposed for doing so (Dingwall & Potts, 2018). The method in this paper follows the same goal, of ensuring a solid foundation underlying a domain-specific corpus, though does it by fine-tuning the input data, prior to training rather than after.

Without model stacking on the baseline, the model would be poor at understanding words that are very uncommon in children’s texts. The words that differ most from the baseline in a model trained without stacking demonstrate this, including *spokesman*, *reporters*, *thursday*, *eu*, *investors*, *trading*, and *analysts*. They tend to be words that are underrepresented in the MOTES Corpus – and thus, poorly modeled – rather than indicators of actual differences in word usage.

When stacking, sub-corpora are scaled according to a scaling factor, s_i . The scaling factor is meant to account for the varying corpus sizes, regarding number of words, so that the linguistic features of that subcorpus are not overwhelmed by a larger corpus that it is stacked on. The scaling factor is selected based on the total sum of occurrences in the top 10 words in the baseline model divided by the target model. Additionally, a weighting factor, w_i , is manually set to specify how much representation each sub-corpus has in the final model. Thus, for each sub-corpus, b of the full set C , the training co-occurrence matrix is constructed as:

$$X_{children} = X_{base} + \sum_{b \in C} X_b s_b w_b$$

For the full MOTES Corpus model, the weighting factor of YouTube, Simple English Wikipedia, TV shows, and books 2005-2010 are set to 0.5. From the book corpus, older books are weighted down, with weights of 0.4, 0.3, 0.2, and 0.1 for ranges 2000-2014, 1990-1999, 1980-1989, and 1850-1979, respectively. When comparing individual sub-corpora, a weighting factor of 1 is used.

Models are trained without stoplisting, but for analysis, words with fewer than 500 post-weighting co-occurrences are excluded, as well as words from the `spacy` stoplist (Honnibal & Montani, 2017), akin to the exclusion performed by (Gonen et al., 2020).

6 RESULTS

This section reports how children’s language differs from general English, as seen through the MOTES Corpus models. The MOTES Corpus model and code is available at <https://osf.io/pwvda>.

6.1 Word Distribution

Prior to comparing modelled language use, it is worth understanding the underlying vocabulary use in each sub-corpus.

Table II shows the log-odds-ratio for each corpus relative to the baseline. Log odds ratio is a symmetrical measure of how likely a word is to occur in corpus A relative to corpus B. Where n_j is the total count of words in corpus j and the proportion of a word i 's frequency w_i in that corpus is $p_{i,j} = \frac{w_{i,j}}{n_j}$, it's odds are $O_{i,j} = p_{i,j}/(1 - p_{i,j})$. Subsequently, the odds ratio between two corpora is $OR_{i,A,B} = O_{i,A}/O_{i,B}$. Log-odds-ratio is sensitive to bias from very rare words (Monroe et al., 2008), so only words from the top 20k of the vocabulary are reported.

The words seen in Table II reflect a mix of domain-specific context and medium. *Mom* and *dad* repeat frequently, as well as simple affirmatives and exclamations. In the children's books, there is a heavy emphasis on literary style, such as verbs like *smiled*, *shrugged*, and *laughed*. Likewise on YouTube and in T.V. shows, the more common words are more spoken-language terms as well as verbs that may be artifacts of descriptive captioning. Finally, the Simple English Wikipedia terms does not appear to offer any meaningful reflection of the language, and instead emphasizes various historical news topics. This serves as a reminder that while the Simple English language is intended to be accessible to children, it still maintains a large topical breadth. The inverse situation, of words notable in the baseline relative to the children's corpora, represented more niche concepts rarely addressed by children's material, such as *gastropod* and *nkorea*.

Overall, the medium is a large factor in the vocabulary of the sub-corpora. To better understand if and how children's language differs from general English, we need to look at how it is used.

Table II: Words with highest log-odds ratio for being represented in the noted corpus relative to the baseline corpus.

Corpus	Highly Represented Words (log-odds-ratio)
Full Children's Corpus	okay, huh, yeah, smiled, goo, mom, uh, hey, dad, shrugged, sorry, oh, maybe, guys, wow
Simple English	websites, kostunica, izetbegovic, mladic, paralympics, denktash, communes, djindjic, kph, blackhawks, walesa, deh, pages, ipod, Ocalan
YouTube	okay, goo, yeah, huh, mom, wow, awesome, oh, dad, hey, uh, guys, thank, applause, um
Books	smiled, shrugged, okay, mom, laughed, staring, leaned, dad, shook, yelled, hugged, cried, wondered, huh, wo
T.V. Shows	huh, okay, yeah, uh, hey, sorry, oh, guys, duh, um, wow, nah, awesome, maybe, thank

6.2 Different Language Usage in MOTES Corpus

Statistics related to term frequencies offers some insight into a text and have long been performed in text analysis (e.g., Burrows 1987). However, it does not necessarily inform how that language is *used*. For example, how often 'cool' is used would not indicate the different manner of use between general English and child-directed language seen in Figure 1. For that, it is useful to train and compare models.

Here, the approach from Gonen et al. (2020) is employed to compare word usage differences between the baseline and MOTES Corpus. This paper will refer to that approach as the *nearest neighbor intersection score*. For each token t_i of the shared vocabulary between models, the top k nearest neighbors to the word are taken, and a score is derived from the count of intersecting words. For more interpretable results, this paper divided the score from by k to bind it between $[0,1]$ – zero signifying no overlap between top- k nearest neighbors and one signifying no difference between models.

Rather than focusing on which words are used more, the primary focus here is *how* words are used differently. Table III compares the full MOTES Children's Corpus to the baseline, listing the top words representative of a divergence in language. For interpretive clarity, words from among the top 10k are presented.

Table III: Words with most notable usage shift in full children’s corpus, scored by nearest neighbor intersection.

Words
bucks, lula, affirmative, yah, nah, corrections, responsibilities, skipper, rein, toughest, tally, sparks, ing, sec, dax, rescuers, trailed, booming, looming, umbrella, slaying, birdie, sidelines, xavier, hah, woo, spill, blames, checking, shrinking, taped, tumbled, barney, hopefully, fortunately, newt, kofi, blueprint, starters, grounded, pose, specter, toss, lott, separatists, losers, youngsters, slash, roundup, pence

The nearest neighbors of these words may be viewed to better understand the reason for the difference in word usage. Table IV shows the 15 nearest neighbors for a selection of low intersect words, as suggested by Gonen et al. (2020). For example, for *bucks* the baseline model discusses sports words, motivated particularly by the *Milwaukee Bucks*, where the children's corpus considered bucks primarily as a synonym for dollars or money. Similarly, lower use of sports-based language was seen with words such as *dodgers* and *foul*. The word *affirmative* discusses affirmative consent and affirmative action-related words in the baseline but appears to be a science fiction-related word for kids, for instance as a word that a fictional robot may say. Names, such as *lula* and *dax*, tend to be less likely to fixate on a single person association in the children's corpus.

Table IV: Nearest neighbors for low-intersect words in general language (baseline) and children’s corpus models

Word	Nearest Neighbors
bucks	GENERAL: milwaukee, colts, braves, browns, knicks, phillies, royals, northampton, suns, bulls CHILD: dollars, owe, worth, cents, thousand, paid, pay, tickets, buy, dollar
affirmative	GENERAL: nudity, discrimination, fairness, imperative, liar, questionable, abort, prejudice, proposition, hibernation CHILD: holographic, asap, heatwave, vert, activating, siphon, martians, shutting, treadmill, bot
fantastic	GENERAL: incredible, marvel, avengers, kirby, strange, hulk, weird, superheroes, mister CHILD: wonderful, job, good, want, awesome, seeing, great, talented, terrific BOTH: amazing
anxious	GENERAL: uncomfortable, frustrated, unhappy, eager, impatient, preoccupied, wanting, worried, compelled, unwilling CHILD: puzzled, glance, remarked, exclaimed, uneasy, felt, manner, scarcely, spite, eagerly
batch	GENERAL: prototypes, deliveries, shipped, prototype, stored, locomotives, initial, processing, process, commence CHILD: buttered, chocolate, cookies, sliced, potatoes, bacon, decided, dropped, butter, toast
foul	GENERAL: batter, touching, ball, outfield, throws, thrown, throw, bounced, touches, hitting CHILD: creep, smelling, instinct, scent, stench, cue, mouthed, shrieking, knives, burners

Removing the focus on most common words above, the top words with different usage in the full vocabulary are *delacorte* and *pew*. The former is the name of a press, featured heavily in book frontispieces next to copyright information. *Pew* is used as onomatopoeia for laser shots by children (*playfully, blasting, dun*), while it more commonly refers the Pew Research Center elsewhere (*gallup, surveys, survey, research, study*). Such in-depth investigation of long-tail words, although interesting, would be too out-of-scope here. However, these examples again reinforce the overall model shift toward language more qualitatively representative of children.

Is there an overall relationship to which types of words change? To better interpret the divergence between models, the word embedding model space is projected to two dimensions with principal component analysis (Figure 3), and the nearest neighbor words are shown for two low intersect words (*bucks* and *detention*), and two high intersect words (*highway* and *Africa*). Words are marked by whether they are nearest neighbors in the MOTES model, baseline model, or both. Here it can be seen that the neighbors of low-intersect words occupy different parts of the language space. That is, they group into entirely different groups of semantic meaning.

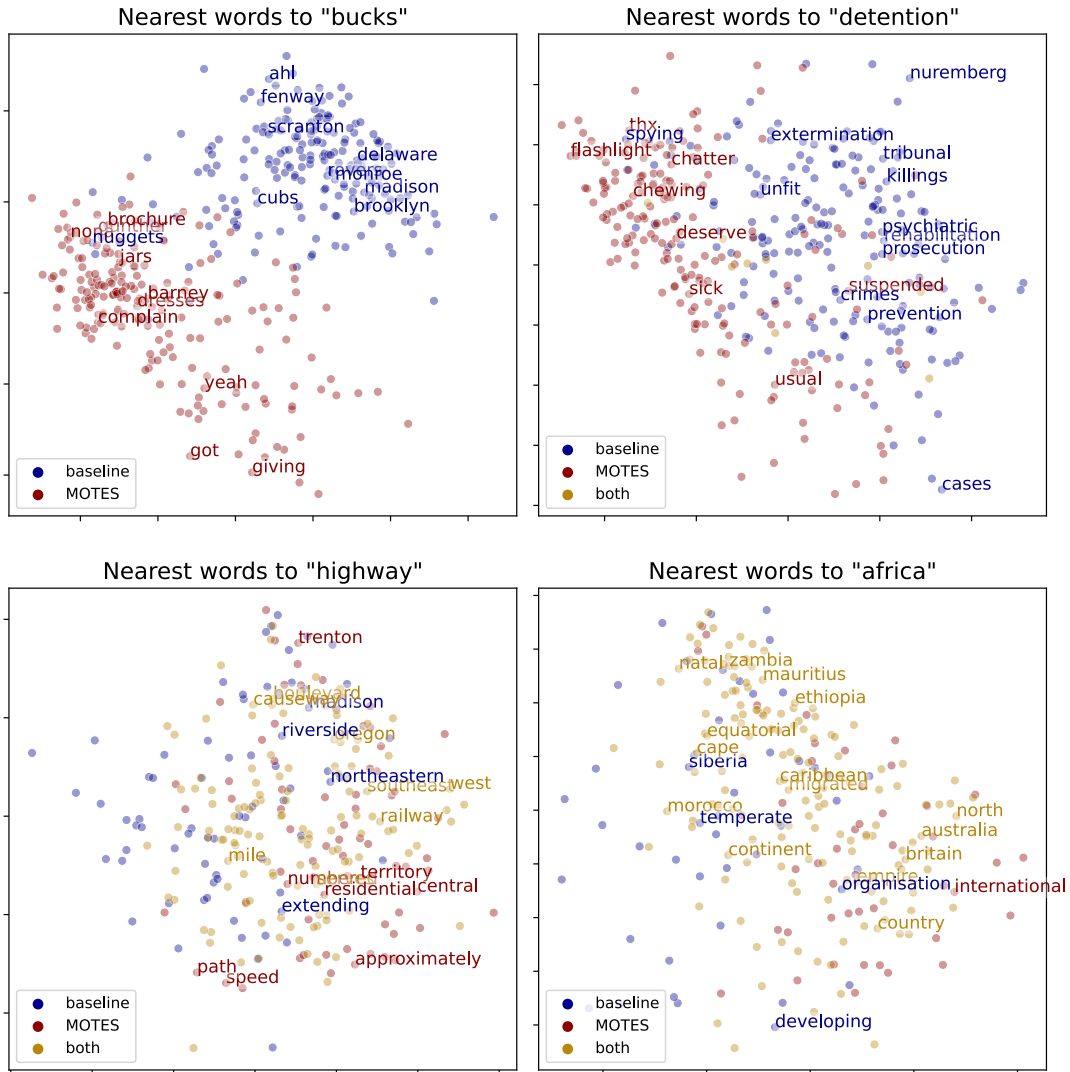


Figure 3: Nearest neighbors of baseline and children's corpus models for example low-intersect words (above) and high intersect words (below), shown in two dimensions with Principal Component Analysis.

6.3 Bias Analysis

Language models in their various forms seek to model the input corpora as accurately as possible, which means that they can import biased associations on matters such as race and gender. To understand biases within the MOTES Corpus, the Word-Embedding Association Test, or WEAT (Caliskan et al., 2017) was applied to the baseline and MOTES model that was trained on top of it. Additionally, the Glove 6B word pre-trained model is included for comparison (Pennington et al., 2014), which is a common out-of-the-box model used for divergent thinking scoring (e.g., Dumas et al 2020). For a fair assessment, the final model was truncated to the vocabulary used by the others.

The WEAT tests work by comparing a given target words in a variety of categories and with two sets of attribute words aimed at identifying how closely related these words are to the target words. The first two WEAT tests show an association that is acceptable: that flowers are considered more pleasant than insects, and instruments are more pleasant than weapons. The remaining tests look at biases that are more pernicious: favorability of European American vs African American names (using different name and attribute lists in WEAT 3-5), gendering of professions, math and science, stigmatization of different illness, and age bias. The target and attribute lists used by WEAT originate from studies of implicit biases in people (Bertrand et al., 2004; Greenwald et al., 1998; Nosek et al. 2002). Results are shown in Figure 4, included an average of the main association tests on race and gender. Positive values show a greater bias toward the first category; after the first two tests, the ideal is a score of zero.

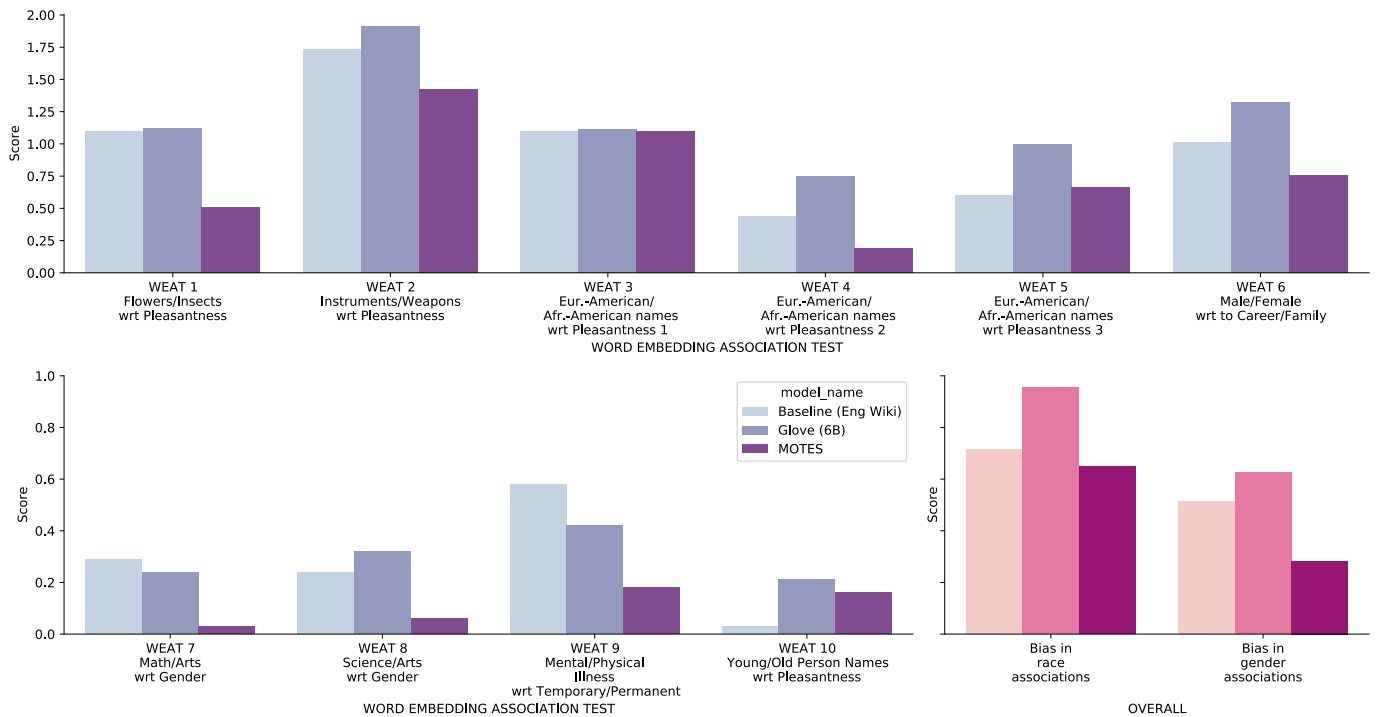


Figure 4: Bias analysis with the Word Embeddings Associations Test. Individual tests and average of race and gender tests shown.

7 ORIGINALITY SCORING

Finally, the MOTES Corpus model is evaluated in an applied context: scoring original thinking in children. Specifically, the model is piloted through an assessment of originality administered to elementary school-aged children.

According to contemporary curriculum models (Anderson et al., 2001; Krathwohl, 2002), creativity is the highest order thinking that must be fostered in education and the core component of creativity is originality. Evaluation of original thinking is a domain which is already applying language models toward addressing measurement challenges and offers us a grounded setting for evaluating whether those models may better serve children with child-directed language customizations.

An assessment which captures open-ended responses related to divergent thinking was created as part of the MOTES Corpus. In the present evaluation, the MOTES Corpus model is used for scoring a cognitive pilot of a child-specific divergent thinking test, applied to 35 elementary-aged US-based students. The MOTES cognitive pilot consists of four games, comprising 29 questions, totaling 963 valid responses. Game one asks for alternate uses for everyday items e.g., “What is a surprising use for a ball?”. Known as the Alternative Uses Task, it is the most common divergent thinking test (Acar et al., 2020; Gerwig et al., 2021; Runco et al., 2016). Game two asks for surprising examples of an adjective, game three focuses on situations, and game four asks students to complete a given sentence. Table VI shows example questions from each. Ages of respondents range from 8 to 11 years old.

Scoring of tests was done by measuring the distance between a prompt and response in the vector space of the given model. Both the baseline and MOTES Corpus are word embedding models which operate at a word-level, so projecting a single prompt or response to the vector space requires a method for aggregating the words of that texts into a single point. For the results reported here, that was done by using a weighted average, where words were weighted with inverse document frequency based on their relative prevalence. Less discriminatory words, ones that occur in a wide array of English documents, are afforded less weight than more unique words. Additionally, a basic stop list is employed to exclude words that say little about the topical content of the phrase, such as *the*, *and*, or *as*. Once the prompt and child’s response are embedded into the model’s vector space, the cosine distance between the two is adopted as the measure of originality.

Ground truth was derived from a three-coder human evaluation, with coders scoring the originality of a response on a 7-point scale. Coders were graduate students in a School of Education, trained by a domain expert. To ensure that they properly understood children’s references, coders also met to discuss ambiguous responses. Table V presents the Pearson correlation between the machine-graded and human-graded responses. The results are presented by game. Past work has validated automated scoring on a multiple-response, test-level (Dumas et al 2020, Beaty and Johnson, 2021). On individual responses, automated scoring is noisier, but provides us a more nuanced view of the difference between the MOTES Corpus and the baseline. The baseline performance on this elementary-aged testing is poor relative to similar tests done with adults, particularly the game 1-comparable Alternate Uses Task (Dumas et al., 2020), and is improved upon by the MOTES model in every instance. For Games 1-3, the MOTES model is a statistically significant improvement over the baseline, at $p < 0.01$, for each test reported by Diederhoefen and Musch (2015).

Table V: Results on divergent thinking test (N=1008, * denotes one-tailed significance at $p < 0.01$)

Game		Example Question	Baseline	MOTES
1	Uses	What is a surprising use for a ball?	.207	.388*
2	Instances	What is a surprising example of something that is red?	.315	.427*
3	Consequences	What would happen if people could travel through time?	.222	.357*
4	Complete the Sentence	When I got on the school bus, I saw...	.153	.162

A look at the effect of each subcorpus on model performance is presented in Table VI. In most cases, just the single sub-corpus improved on the baseline, though the spoken word corpora appeared to have the most

notable effect. The results suggest that the stacking approach taken here, where corpora were weighed up and stacked on top of the baseline, overcomes the need for extraordinary corpus scale.

Table VI: Individual subcorpus performance on divergent thinking test ($N=1008$, *denotes one-tailed significance at $p < 0.05$, ** at $p < 0.01$)

Game	Baseline	YouTube	Kids' Books	Simple Wiki	TV Shows
Uses	.207	.341**	.266**	.318**	.308**
Instances	.315	.483**	.310	.178	.458**
Consequences	.222	.389**	.287**	.301**	.432**
Complete the Sentence	.153	.212*	.210*	.199	.107

Finally, it is worth considering whether the MINIs performance is significant because it better matches the language of children, or because it is simply a stronger, less-biased model. To answer that, we re-analyzed adult data of an alternate uses task from Dumas et al. (2020). There, the difference was not statistically significant, where the MOTES Corpus had an $r=.372$ correlation with judges, versus a baseline performance of $r=.374$.

This example context uses established methods from the divergent thinking community, applied with and without child-specific models. Recent work has found that scoring individual responses can be greatly improved with large language models (Organisciak et al. 2022). These results, as well as general work on pre-training large language models (Gururangan et al 2020), suggest that such work could also be further improved with pre-training on child-specific corpora, though that remains to be seen.

8 DISCUSSION

There are two primary contributions of this work. First, we consider whether text mining applications which deal with children should consider child-specific modeling. We find that children’s language diverges notably from general English and, in the one application considered here, using a child-specific model outperforms an established method which does not. The second contribution is the MOTES Corpus itself, which may be used for creating child-specific models that can be applied to additional child-related research areas. Here, the discussion considers the significance and challenges of those differences.

8.1 Shifts in Language Use and the Benefits of a Child-Specific Corpus

Interpreting models is complicated by the fact that at some point, a qualitative assessment may be required, an interpretation that may be skewed by expectation. However, in this paper’s results, several notable shifts may be noted. First, it is clear the language in the child-specific corpora is not identical to a general corpus. By inspecting the *model* rather than the *corpus*, this study does not look for what is used *more*, but *how* it is used, and a notable portion of the words in the model shift their meaning.

In quantifying the word’s shifts using nearest neighbor intersect (Gonen et al., 2020), the shifts are partially topical, partially related to a more colloquial language, and partially reflective of the spoken and literary language represented in medium of some of the sub-corpora. However, avoiding the issue seen in the earlier analysis of word frequencies, it appears that the medium does not contribute to an overly strong effect in considering modeled word usage. For example, between the sub-corpora of books and television subtitles, 72% of their 100 highest usage shift words overlap. What is more difficult to infer is the quality of the shifts. In a qualitative consideration of high-shift words, the shifts in language in the child-directed corpora aligns with children’s topics. For example, when a child is using the word *pentagon*, it is more

sensible to assume discussion of shapes, rather than military and defense; when they use *umbrella*, a system parsing their language would be more accurate is assuming they are not discussing umbrella organizations.

What is clear is that in application to elementary-aged assessment in this paper, the general-language baseline performed poorly, whereas the child-specific MOTES Corpus model notably outperformed it. This shows the applied value of a children's corpus and model, and it is worth further study on other elementary applications.

8.2 *Bias in the Corpus*

Algorithms learn from the material they are given, which is inherently biased. Therefore, understanding undesired associations is paramount to understanding the how the model will affect underserved groups in applied use.

The word embedding association test that was applied in this study was modeled after implicit association tests, meant for measuring embedded biases in people, and they found that the biases in popular pre-trained word embedding models aligned very similarly to the biases in people (Caliskan et al., 2017). Caliskan et al. note that such associations are rooted in a complex mix of sources, not always ill-intentioned: for example, they note that the degree to which a career is gendered in GloVe correlates highly with the actual gender makeup of that profession.

The associations test is notable in for what it demonstrates about child-directed media. In nearly every category the MOTES Corpus had a less biased set of internal associations, with gender bias lowered to a surprising degree. There are various possible interpretations for this effect. For example, authors and producers of child-directed media may be more cognizant of inherent gender biases and work to avoid them or may simply avoid topics which produce biased language. The corpus also performs better on negative race associations, though less drastically, pointing to less progressive countering of race biases in the children's media sources used here.

Since the MOTES Corpus improves on common human biases, it raises another potentially valuable use of text mining approaches to open-ended assessment: the ability to avoid bias which human graders may have. More study would be needed to understand the strength of implicit associations in teachers and other graders; however, corpora like MOTES can be further adapted to attempt to remove the negative associations altogether. Noting the importance of corpus composition, for example, a future direction for MOTES and similar projects is to include more diverse sources, actively seeking out more texts that reflect black culture as well as other communities of color. Further, there are additional tests for reviewing bias associations in word embedding models (Badilla et al., 2020; Sweeney & Najafian, 2019), as well as methods for debiasing models after they've been trained (Bolukbasi et al. 2016, Zhao et al. 2018). These can be used in concert to remove negative associations; still, it has been argued that such techniques are not entirely successful (Gonan and Goldberg 2019), which makes the inherently less biased MOTES Corpus a good starting point. Acknowledging human bias also calls into question the ground truth that we use. In a use case such as the example one presented in divergent thinking scoring, aiming to perform similarly to human judges could present a possible opportunity for bias reinforcement, and requires care in judge training and auditing of their work.

8.3 *Challenges*

There are some limitations to using word embedding models in applications with children. Foremost is the issue of out of vocabulary errors. A term-term word embedding model with a fixed vocabulary does not account for the reality that children misspell. While common misspellings show up in the MOTES large

vocabulary, there is no guarantee that they are seen enough to be modeled accurately. Further, to compile a corpus at scale, this project focuses on language that children encounter, not that they write themselves.

There are various potential solutions to out of vocabulary errors related to misspellings, as well as made-up words. In education testing, responses may be administered and written down by adults. Spelling correction may also be used. On the model side, a potential solution is with subword vector embeddings. These models treat words by their component parts. For example, with *fastText* (Bojanowski et al., 2017) the model uses character ngrams as its tokens and represents each word as a sum of its characters' vectors (e.g., *apple* may be seen as a mix of *ap*, *pp*, *pl*, and *le*). This allows models to learn semantic relationships but also syntactic relationships, including those between commonly misspelled character sequences. Other approaches include variable-length subwords, such as with *byte-pair encoding* (Sennrich et al., 2016; Gage, 1994).

In education, the use of word embedding models is common because they are easier to interpret than large language models. When applied in testing that may affect students, it is important to be able to explain the choices a computational approach makes. That does present challenges for growth in the field, given that neural language modeling approaches generally outperform word embedding models in applied use, including recent reinterpretations of divergent thinking scoring (Organisciak et al. 2022). Nevertheless, the MOTES Corpus may be similarly applied to pretraining BERT-like large language models (Gururangan et al 2020), as it was done with GloVe in this work. The exception is in the children's book portion of the corpus, which is only available as page-level bag-of-words representations, though recent work has suggested that large language models are still robust with bags of words (Gupta et al. 2021, Hessel and Schofield, 2021).

9 CONCLUSION

Word embedding models are common in text mining applications because of their ability to represent *latent* meaning underlying word use. They present tremendous potential in education and learning, offering consistent and valid ways to understand how people learn or think (Dumas and Dunbar, 2014; Dumas et al., 2020; Forthmann et al., 2019; Landauer and Dumais, 1997). Natural language processing also offers a pathway from close-ended measurement, a boon for certain sub-domains that struggle to be measured in that form, warranted further consideration of its applications in education. However, the performance of word embedding models depends on a sound underlying sense of how language is used. This presents a challenge to applications involving children. As we show, a word embedding model build on general sources of English text may not accurately represent children's language.

In this paper, we prepare and analyze a word embedding model for children's language, based on a multi-domain corpus of texts from multi-modal child-oriented sources. By employing an approach which stacks source text on top of large baseline corpus, we find that the resulting model does not lose its ability to represent words that are uncommon in child-facing sources – words like *administrative*, *republic*, *empire* but effectively nudge words like *bucks*, *pentagon* or *fantastic* closer to how children use them. We quantify the language with the greatest change and qualitatively review their linguistic shifts, seeing indication that the child-facing corpus consistently performs more appropriately to children's language. The primary contribution of this work is demonstrating the semantic difference of children's language from general language texts and showing that child-tailored text models are stronger in an established educational context.

ACKNOWLEDGEMENTS

Thank you to Kelly Berthiaume, Maggie Ryan, and the full MOTES team for additions contributions and advice.

FUNDING SOURCES

This study was funded by the Institute of Education Sciences (IES) (Grant No. R305A200519).

RESEARCH DATA

The MOTES Corpus model as well as code for reproducing data collection and modeling are available at <https://osf.io/pwvda> .

REFERENCES

- Acar, S., Berthiaume, K., Grajzel, K., Dumas, D., Flemister, C. “Tedd,” & Organisciak, P. (2021). Applying Automated Originality Scoring to the Verbal Form of Torrance Tests of Creative Thinking. *Gifted Child Quarterly*. <https://doi.org/10.1177/00169862211061874>
- Acar, S., & Runco, M. A. (2014). Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal*, 26(2), 229–238. <https://doi.org/10.1080/10400419.2014.901095>
- Acar, S., Runco, M. A., & Park, H. (2020). What should people be told when they take a divergent thinking test? A meta-analytic review of explicit instructions for divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, 14(1), 39–49. <https://doi.org/10.1037/aca0000256>
- Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 37(4), 13–49. <https://doi.org/10.1016/j.tele.2019.01.007>
- Alexander, J. (2020, January 6). *YouTube officially rolls out changes to children’s content following FTC settlement*. The Verge. <https://www.theverge.com/2020/1/6/21051465/youtube-coppa-children-content-gaming-toys-monetization-ads>
- Anderson, L.W. and Krathwohl, D.R. (2001) *A taxonomy for learning, teaching, and assessing : a revision of Bloom’s taxonomy of educational objectives*. Available at: <https://eduq.info/xmlui/handle/11515/18345>.
- Badilla, P., Bravo-Marquez, F., & Pérez, J. (2020). *WEFE: The Word Embeddings Fairness Evaluation Framework*. 1, 430–436. <https://doi.org/10.24963/ijcai.2020/60>
- Beaty, R.E., Johnson, D.R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research* 53, 757–780. <https://doi.org/10.3758/s13428-020-01453-w>
- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4), 991–1013. <https://doi.org/10.1257/0002828042002561>
- Biersack, S., Kempe, V., & Knapton, L. (2005). Fine-tuning speech registers: A comparison of the prosodic features of child-directed and foreigner-directed speech. *Ninth European Conference on Speech Communication and Technology*.
- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *ArXiv:1607.06520 [Cs, Stat]*. <http://arxiv.org/abs/1607.06520>
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. <https://doi.org/10/gfw9cs>
- Bordia, S., & Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. *ArXiv:1904.03035 [Cs]*. <http://arxiv.org/abs/1904.03035>
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” *ArXiv:2005.14165 [Cs]*, July. <http://arxiv.org/abs/2005.14165>.
- Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019). Understanding the origins of bias in word embeddings. *ArXiv:1810.03611 [Cs, Stat]*. <http://arxiv.org/abs/1810.03611>
- Burrows, J. F. (1987). Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2(2), 61–70. <https://doi.org/10.1093/lc/2.2.61>
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. <https://doi.org/10/f93cpf>

- Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, 27(6), 843–873. https://doi.org/10.1207/s15516709cog2706_2
- Cristia, A., Dupoux, E., Ratner, N. B., & Soderstrom, M. (2019). Segmentability Differences Between Child-Directed and Adult-Directed Speech: A Systematic Test with an Ecologically Valid Corpus. *Open Mind*, 3, 13–22. https://doi.org/10.1162/opmi_a_00022
- Cai, Z., Graesser, A. C., Windsor, L., Cheng, Q., Shaffer, D. W., & Hu, X. (2018). Impact of corpus size and dimensionality of LSA spaces from Wikipedia articles on AutoTutor answer evaluation. *Journal of Educational Data Mining*.
- Crossley, S., Dascalu, M., & McNamara, D. (2017). How important is size? An investigation of corpus size and meaning in both latent semantic analysis and latent Dirichlet allocation. *The Thirtieth International Flairs Conference*.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407. <https://doi.org/10/db4ft5>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://arxiv.org/abs/1810.04805v2>
- Dingwall, N., & Potts, C. (2018). Mittens: An extension of GloVe for learning domain-specialized representations. *ArXiv:1803.09901 [Cs]*. <http://arxiv.org/abs/1803.09901>
- Dumas, D., & Dunbar, K. N. (2014). Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity*, 14, 56–67. <https://doi.org/10/f6wb79>
- Dumas, D., Organisciak, P., & Doherty, M. D. (2020). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*. <https://doi.org/10/ghcsqq>
- Eskenazi, M., Mostow, J., & Graff, D. (1997). The CMU Kids Corpus LDC97S63. *Web Download*. Philadelphia: Linguistic Data Consortium.
- Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior*, 53(4), 559–575. <https://doi.org/10/ghcsqk>
- Forthmann, B. and Doeblner, P. (2022) Fifty years later and still working: Rediscovering Paulus et al.'s (1970) automated scoring of divergent thinking tests. Pre-print. <http://dx.doi.org/10.1037/aca0000518>
- Forthmann, B., Paek, S. H., Dumas, D., Barbot, B., & Holling, H. (2020). Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *British Journal of Educational Psychology*, 90(3), 683–699.
- 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. <https://doi.org/10/f96wfg>
- Gage, P. (1994). A new algorithm for data compression. *C Users Journal*, 12(2), 23–38.
- Gerwig, A., Miroshnik, K., Forthmann, B., Benedek, M., Karwowski, M., & Holling, H. (2021). The Relationship between Intelligence and Divergent Thinking—A Meta-Analytic Update. *Journal of Intelligence*, 9(2), 23. <https://doi.org/10.3390/jintelligence9020023>
- Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. NAACL 2019.
- Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *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), 609–614. <https://doi.org/10.18653/v1/N19-1061>

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480. <https://doi.org/10.1037//0022-3514.74.6.1464>

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360. <https://doi.org/10.18653/v1/2020.acl-main.740>

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *ArXiv Preprint ArXiv:1605.09096*.

Hessel, J., Schofield, A. (2021). How effective is BERT without word ordering? Implications for language understanding and data privacy, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Presented at the ACL-IJCNLP 2021, Association for Computational Linguistics, Online, pp. 204–211. <https://doi.org/10.18653/v1/2021.acl-short.27>

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To Appear*, 7(1).

Huang, C.-J., Tsai, P.-H., Hsu, C.-L., & Pan, R.-C. (2006). Exploring cognitive difference in instructional outcomes using text mining technology. *2006 IEEE International Conference on Systems, Man and Cybernetics*, 3, 2116–2120. <https://doi.org/10/bvg2g5>

Krathwohl, D.R. (2002) 'A revision of Bloom's taxonomy: An overview', *Theory into practice*, 41(4), pp. 212–218.

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. <https://doi.org/10/dcpw35>

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv:1907.11692 [Cs]*. <http://arxiv.org/abs/1907.11692>

Lileikyte, R., Irvin, D. and Hansen, J.H.L. (2022) 'Assessing child communication engagement and statistical speech patterns for American English via speech recognition in naturalistic active learning spaces', *Speech Communication*, 140, pp. 98–108. Available at: <https://doi.org/10.1016/j.specom.2022.01.006>.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed* (pp. xi, 366). Lawrence Erlbaum Associates Publishers.

MacWhinney, B. (2007). The Talkbank project. In *Creating and digitizing language corpora* (pp. 163–180). Springer.

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. *ArXiv:1904.04047 [Cs, Stat]*. <http://arxiv.org/abs/1904.04047>

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

Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research*, 29(4), 495–512. <https://doi.org/10/c8wp3j>

- Millis, K., Magliano, J., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading, 10*(3), 225–240. <https://doi.org/10/bzzw3f>
- Millis, K., Magliano, J., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading, 10*(3), 225–240. <https://doi.org/10/bzzw3f>
- Mohan, P. (2021). *An Analysis of Gender Bias in K-12 Assigned Literature Through Comparison of Non-Contextual Word Embedding Models* [University of Washington]. <https://www.proquest.com/docview/2495370013>
- Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis, 16*(4), 372–403. <https://doi.org/10/cb486t>
- Neumann, M. M., & Herodotou, C. (2020). Evaluating YouTube videos for young children. *Education and Information Technologies, 1*–17.
- Neumann, M. M., & Herodotou, C. (2020). Young children and YouTube: A global phenomenon. *Childhood Education, 96*(4), 72–77. <https://doi.org/10/ghbkc7>
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6*(1), 101.
- Organisciak, P., Capitanu, B., Underwood, T., & Downie, J. S. (2017). Access to Billions of Pages for Large-Scale Text Analysis. *iConference 2017 Proceedings Vol. 2*. iConference 2017, Wuhan, China. <http://hdl.handle.net/2142/98873>
- Organisciak, P., Acar, S., Dumas, D., Berthiaume, K. Beyond Semantic Distance: Automated Scoring of Divergent Thinking Greatly Improves with Large Language Models. Pre-print. <http://dx.doi.org/10.13140/RG.2.2.32393.31840>
- Paulus, D.H. (1970). Computer Simulation of Human Ratings of Creativity. Final Report.
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. <https://doi.org/10/gfshwg>
- Pleyer, M. (2020). *The Everyday Use of pretend in Child Language and Child-Directed Speech: A Corpus Study* [Dissertation]. <https://doi.org/10.11588/heidok.00028873>
- Preckel, F., Holling, H., & Wiese, M. (2006). Relationship of intelligence and creativity in gifted and non-gifted students: An investigation of threshold theory. *Personality and Individual Differences, 40*(1), 159–170. <https://doi.org/10/fhxk8t>
- Rabkina, I., Nakos, C., & Forbus, K. D. (2019). Children's sentential complement use leads the theory of mind development period: Evidence from the CHILDES Corpus. *CogSci, 2434*–2639.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.
- Rodrigues, M. W., Isotani, S., & Zárata, L. E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. *Telematics and Informatics, 35*(6), 1701–1717. <https://doi.org/10.1016/j.tele.2018.04.015>
- Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs data mining and knowledge discovery, 3*(1), 12–27. <https://doi.org/10/cz59>
- Runco, M. A., Abdulla, A. M., Paek, S. H., Al-Jasim, F. A., & Alsuwaidi, H. N. (2016). Which test of divergent thinking is best. *Creativity. Theories–Research–Applications, 3*(1), 4-18. Doi: 10.1515/ctra-2016-0001
- Sachin, R. B., & Vijay, M. S. (2012). A Survey and Future Vision of Data Mining in Educational Field. *2012 Second International Conference on Advanced Computing Communication Technologies, 96*–100. <https://doi.org/10/ghcsqp>

- Saffran, J.R., Newport, E.L. and Aslin, R.N. (1996) ‘Word Segmentation: The Role of Distributional Cues’, *Journal of Memory and Language*, 35(4), pp. 606–621. Available at: <https://doi.org/10.1006/jmla.1996.0032>.
- Simple Wikipedia Homepage*. (n.d.). Simple Wikipedia. https://simple.wikipedia.org/wiki/Main_Page
- Smith, A., & Kessel, P. van. (2018). *Many Turn to YouTube for Children’s Content, News, How-To Lessons* (Internet & Technology). Pew Research Center. <https://www.pewresearch.org/internet/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/>
- Sweeney, C., & Najafian, M. (2019). A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1662–1667. <https://doi.org/10.18653/v1/P19-1162>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. <http://arxiv.org/abs/1706.03762>
- von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, 54(1), 3–11. <https://doi.org/10/gcphmd>
- Weisleder, A. and Fernald, A. (2013) ‘Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary’, *Psychological Science*, 24(11), pp. 2143–2152. Available at: <https://doi.org/10.1177/0956797613488145>.
- York, J. (2010). Building a future by preserving our past: The preservation infrastructure of HathiTrust digital library. *World Library and Information Congress: 76th IFLA General Conference and Assembly*, 10–15.
- Zahner, K., Schönhuber, M., Grijzenhout, J., & Braun, B. (2016). *Konstanz prosodically annotated infant-directed speech corpus (KIDS corpus)*. 562–566. <https://doi.org/10.21437/SpeechProsody.2016-115>
- Zhao, Jieyu, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. “Learning Gender-Neutral Word Embeddings.” In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4847–53. Brussels, Belgium: Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1521>.
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 1–34. <https://doi.org/10/ghcsqr>