How Should We Quantify Bilingual Vocabulary Knowledge?

Jennifer M. Weber (jennifer.m.ellis@colorado.edu)
Department of Psychology and Neuroscience, 345 UCB
Boulder, CO 80309 USA

Pui Fong Kan (puifong.kan@colorado.edu)
Department of Speech, Language, and Hearing Sciences, 409 UCB
Boulder, CO 80309 USA

Eliana Colunga (eliana.colunga@colorado.edu)
Department of Psychology and Neuroscience, 345 UCB
Boulder, CO 80309 USA

Abstract

Dual language learners (DLLs) constitute a large portion of the population, but relatively little is known about the best ways in which to assess their vocabulary knowledge. Past research has used both conceptual vocabulary knowledge, assessing whether a child knows a word in either language, as well as total vocabulary knowledge, assessing what words a child knows in each language separately. The present work uses neural networks to predict specific word learning for individual Cantonese-English DLLs. As its input, The model utilizes word2vec embeddings that either represent children's' conceptual word knowledge or total word knowledge. We find that using total word knowledge results in higher predictive accuracy, suggesting that knowing what specific words DLLs know in each of their languages provides the most accurate picture of DLLs' vocabulary knowledge. The present work has many implications for both identification of at-risk individuals and the creation of learning materials for DLL populations.

Keywords: neural networks; dual language learners, vocabulary representation

Background

Understanding vocabulary development is necessary for helping identify children at-risk for current and future language difficulties. Many current assessment tools use estimates of vocabulary size as one way to pinpoint those falling behind, but some recent work points to weaknesses in this general approach (Rescorla, 2011; Heilmann et al., 2005). Some suggest that also understanding the types of words children know or the structure of their vocabularies may lead to better insights into each child's individual trajectories (Beckage et al., 2020; Hills et al., 2010; MacRoy-Higgins et al., 2016). However, using individual word knowledge to help model language development is further complicated by the fact that about 60% of the world's population speaks two or more languages (ilanguages.org, 2018). How should we take into account lexical knowledge in multiple languages? The current work investigates whether representing the specific words a dual language learning child knows in both languages is a better predictor of subsequent word learning than representing their conceptual word knowledge, irrespective of which language the word is known in.

DLL Vocabulary Assessment

Children growing up in bilingual households are exposed to varying amounts of each language, which directly impacts their proficiency in both languages and more generally creates a large spectrum of possible lexical knowledge (Duursma et al., 2007; Hammer et al., 2009; Thordardottir, 2011). Further, just as with monolingual children, dual language learners' (DLLs) vocabulary skills are important predictors of their later reading achievement in elementary school (Hammer et al., 2009). Based on the epidemiological data stemming from white monolingual English-speaking school-aged children, It is anticipated that about seven to 11 percent of bilingual children have a developmental language disorder (Calder et al., 2022; Nudel et al., 2023; Wu et al., 2023). Evidence suggests that DLLs are not inherently at a disadvantage or more likely to have language delays or disorders, but that we simply need different ways of evaluating their language knowledge (Mancilla-Martinez et al., 2020). A largely debated question is exactly what information language assessments should collect to give an accurate picture of DLL development.

Multiple approaches have been used for capturing DLL word knowledge, utilizing both their first, or home, language (L1), and their second language, usually that of the community (L2). Such approaches have considered completely separating word knowledge in both languages - dog counts as one word in English *or* perro counts as one word in Spanish (Schwartz, 20144; Uchikoshi, 2014), combining checklists to gather conceptual vocabulary knowledge, or marking a word known if they know it in either language - knowing either dog or perro count as one word or concept (Gross et al., 2014; Junker & Stockman, 2002; Core et al., 2013; Gonzalez-Barrero et al., 2020), or combining checklists for total vocabulary - knowing both dog and perro count as two words (Patterson, 1998; Core et al., 2013; Gonzalez-Barrero et al., 2020). Though researchers agree that looking at knowledge in both languages is better than only

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

investigating one or the other (Peña et al., 2016), there are mixed results as to whether conceptual or total vocabulary measures are better indicators of overall lexical knowledge. Furthermore, with these methods bilingual researchers still tend to simplify vocabulary knowledge into a size estimation, without considering how individual word knowledge or vocabulary structure contributes to overall performance. However, some computational modeling work with monolinguals (e.g., Beckage et al., 2020; Weber & Colunga, 2022) suggests that extending this work to DLLs by considering the individual words known in both L1 and L2 in models of vocabulary acquisition is both warranted and necessary.

Modeling Vocabulary Acquisition

Predicting vocabulary growth using computational models is a fruitful avenue to both help us understand acquisition mechanisms and develop tools for assessment and intervention. These methods have looked at vocabulary structure, or how different words contribute to overall vocabulary knowledge, rather than more general size estimations. In fact, one recent analysis suggests using individual word knowledge produces better fitting models than those that only take into account vocabulary size and other demographic information, such as sex and age (Beckage et al., 2020). Though such methods have modeled growth in different ways, for example using age of acquisition norms (Alhama et al., 2020), different growth algorithms such as preferential attachment (Beckage & Colunga, 2019; Hills et al., 2010), or neural networks (Beckage et al., 2020), they have come to similar conclusions about the importance of understanding vocabulary structure over and above vocabulary size. Further, modeling work with monolinguals suggests sources of data that more accurately represent the knowledge of the population of interest can impact model accuracy and subsequent conclusions (Weber & Colunga, 2022). This work likely extends to DLLs, in that utilizing data that accurately represents knowledge in both languages will result in more accurate models than those representing knowledge in just one or the other.

That said, modeling endeavors have rarely tackled the question of bilingual development. Bilson et al. (2015) constructed semantic network representations of bilingual and monolingual children's vocabulary over multiple timepoints and found that learning words in one language does facilitate word learning in the other, suggesting that we cannot simply look at bilingual development as two separate monolingual trajectories. The question remains though, what is the best way to quantify a bilingual's vocabulary when modeling their growth?

Current Study

In the present analyses we aimed to understand whether representing DLL vocabulary knowledge using conceptual vocabulary (L1 and L2 combined) or total vocabulary (L1 and L2 separately) leads to more accurate word learning predictions. We further wished to do so using neural network models which take into account individual word knowledge, rather than simply using overall vocabulary size. We created two identical neural network architectures with the only difference being whether the input consisted of semantic information about word knowledge in each language separately, or combined into conceptual knowledge. We hypothesized that retaining information about vocabulary knowledge in each language separately, arguably providing more information overall, would lead to better model performance than using conceptual vocabulary knowledge as the input.

Method

The Vocabulary Data

To compare our neural network models, we utilized longitudinal data collected over three timepoints from Cantonese-English DLLs in a Head Start program located in San Francisco, California. All children were typically developing and learned Cantonese (L1) at home and English (L2) in school. The majority of children using Head Start services are at or below the federal poverty level or are under certain disadvantageous circumstances. Data from 125 children were included in the analyses; 44 children were assessed at two consecutive timepoints and 81 at all three timepoints. At the initial visit children ranged in age from 37 to 62 months of age (M = 49.14, SD = 6.70). The average time between timepoints was 2.84 months.

Children's vocabularies were assessed in both their first (L1) and second (L2) language on separate occasions for each timepoint, to avoid interference. In other words, children were visited two-four times over a short period (typically within a week), to gather all the relevant vocabulary data, and this constitutes one timepoint. Among other measures that will not be analyzed here, each child completed the Kai Ming Vocabulary Test (KMVT; Kan et al., 2020). The KMVT was specifically developed for Cantonese-English bilingual preschool children, with words chosen to represent a range of lexical knowledge relevant for this population (Kan et al., 2020). The task consists of a total of 194 items, 91 of which are queried receptively and 103 which are queried expressively. The same 194 words were queried in both languages. Target words include animals (elephant, horse), foods (cookie, noodles, soy sauce), household items (TV, wok, toothpaste), and other nouns, as well as a few verbs (kiss, cry) and adjectives (short, bitter). In past work, this task has shown a strong correlation with other language sample measures (Kan et al., 2020).

All testing was completed one-on-one with a trained research assistant in a quiet space inside the Head Start facility. The receptive task asked the child to point to the named (either in English or in Cantonese) item from among four images, whereas the expressive task asked children to name (either in English or in Cantonese) the depicted image on the page. Children's accuracy in both languages on the KMVT were combined to create a child's "vocabulary": 164 words were modeled in both languages, for a total vocabulary

of 328 possible known words. Children learned on average, 28.47% (*SD* = 12.22%) of the words they did not know in their L1 between visits, and 29.99% (*SD* = 12.62%) of the words they had left to learn in their L2.

The Neural Networks

The goal of the neural networks was to predict which words the child knew at the subsequent visit, based on their current vocabulary. To this end, each network utilizes vocabulary data from 2 timepoints, constituting a pre and a post measure. A child's vocabulary from the pre-test forms the input to the network, with the network attempting to produce the post-test results as its output. The child's actual post-test data is used as the gold standard to train the network. Each pre/post pair was treated separately by the network, for a total of 368 data points. That is, if a child participated in three visits, this constituted two separate pre/post groupings, visits one (input) and two (output), and visits two (input) and three (output). As in previous work, (Beckage et al., 2020; Weber & Colunga, 2022) data was normalized so that new words could only be learned and none could be "forgotten". That is, if a child was reported to know a word at the input visit, they continued to know that word at the output visit regardless of how they performed on the output visit's vocabulary tests.

The models were created and trained using the keras package in python, with a training/validation/test split of 60/20/20. The split was created by dividing up the pre/post pairing described above. The overall architecture was optimized to consist of an input layer, two 100-unit hidden layers, and an output layer. The hidden layers utilized the ReLu activation function with a dropout rate of 0.1. Each network was optimized using a version of stochastic gradient descent and a learning rate of 0.001. Networks were trained using a batch size of 10 over 200 epochs, or 100 passes through the pre/post pairings in the training set.

To represent a child's vocabulary knowledge, we used a pre-trained Wikipedia2Vec model trained on an English Wikipedia corpus (Yamada et al., 2018). Wikipedia2Vec uses a skip-gram model, or the Word2Vec algorithm, to learn vector representations for all words in the corpus. The resulting embeddings take into account both words' co-occurrence in the corpus, as well as when words appear in similar contexts. The Wikipedia2Vec pretrained embeddings have been shown to be useful for a range of tasks, including entity linking, question answering, and text classification (Yamada et al., 2016; Yamada et al., 2018; Poerner et al., 2019). Wikipedia2Vec was chosen specifically because it also provides pre-trained embeddings in 11 other languages, making it optimal for future cross-linguistic and bilingual research. Further, preliminary searches found no other available corpora in Cantonese or representing children, especially Cantonese-English children, that contained enough of the vocabulary words in the KMVT to be able to model word-learning with neural networks. These models were trained using a window size of 5 and an iteration value of 10, and here we used the 300- dimension vectors. That is, each word is represented by 300 different values. To use the Wikipedia2Vec embeddings as the input for our neural network, the vector for each word the child was reported to know at the current visit was summed together, for an input 300 units in length. This method of summing was found to be optimal in prior analyses using Word2Vec methods (Beckage et al., 2020). Rather than averaging known-word vectors, which produces an input that is size-invariant, summing the vectors together keeps information about vocabulary size.

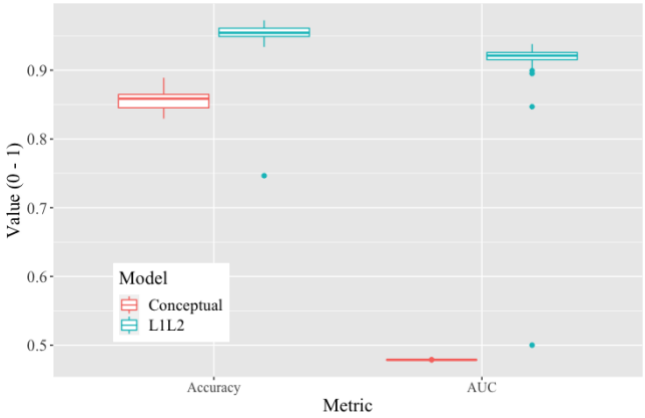
We wanted to compare whether including vocabulary information from both languages was better than understanding conceptual vocabulary knowledge using the same architecture. That is, is knowing the specific words a child knows in each language separately more informative than simply knowing which words they knew in either language? To compare these knowledge representations, we created a neural network with an input of 600 units. To investigate each language separately, we gathered the embeddings of the words a child knew in their first language (L1), and separately the embeddings they knew in their second language (L2). Three hundred input units were used for the summed vectors in their L1, and 300 were used for the summed vectors in their L2. To represent conceptual vocabulary knowledge, we gathered the embeddings the child knew in at least one of their two languages. They did not have to know the word in both languages to be considered conceptually known. We summed the embeddings and that same conceptual vector was duplicated for two identical 300-unit inputs. It is possible for the two representations to be identical (a child's conceptual input and the input of both language separately) if the child knew the same exact words in both their L1 and L2. However, this never occurred. The output consisted of 328 units representing the 328 words a child could know at the subsequent visit, 164 words from the vocabulary tests for both languages. The output layer used a sigmoid activation function to predict, for each output unit, whether the word was unknown at the subsequent visit (output value of 0) or known (value of 1). Fifty runs of each network were trained and tested, in order to compare the two representation methods statistically.

The two methods were compared on three metrics: loss, AUC, and calculated model accuracy. Overall loss was measured using mean squared error. The model seeks to minimize loss as much as possible during training, with smaller losses representing better model performance. AUC, or area under the curve, measures the tradeoff between sensitivity (true positives: the model predicted the child knew the word when they did) and specificity (true negatives: the model predicted the child did not know the word when they in fact did not). Finally, accuracy was calculated by first taking the output prediction for each node and binarizing it; for each node, if it predicted a value above or equal to 0.5, it was predicted to be known or have a value of 1, and for each predicted value below 0.5, it was changed to be 0 or unknown. These were then compared to the actual values, and accuracy was measured as the average number of nodes correctly predicted by the model.

Results

We tested whether representing vocabulary knowledge in both languages separately (L1L2) resulted in more accurate word-learning predictions than representing conceptual vocabulary (Conceptual). We conducted independent samples t-tests comparing the test results for the 50 runs of each of the two network types. We found that the L1L2 network performed better across our three evaluation metrics. The L1L2 network (M = 0.0452, SD = 0.0310) minimized loss to a greater extent than the Conceptual network did (M = 0.1436, SD = 0.0142), t(98) = 20.40, p<.001. The AUC was greater for the L1L2 network (M = 0.9113, SD = 0.0610) compared to the Conceptual network (M = 0.4787, SD = 0.0), t(98) = 50.12, p<.001. Finally, the calculated accuracy was also greater for the L1L2 network (M = 0.9512, SD = 0.0307) compared to the Conceptual network (M = 0.8564, SD = 0.1416), t(98) = 19.82, p<.001. Figure 1 shows the accuracy and AUC results. To note, though the L1L2 network performed significantly better, the Conceptual network still had decent performance, as measured by accuracy and loss.

Figure 1: Sliding Window Network Comparison



Discussion

The present analyses provide evidence that modeling DLL growth by using word-level vocabulary data from each language separately enhances predictive performance compared to using a combined measure representing conceptual language knowledge. The present results corroborate past research suggesting that vocabulary size measures that take into account total word knowledge rather than conceptual word knowledge may be better indicators for language assessment (Core et al., 2013; Gonzalez-Barrero et al., 2020). This work also expands on past work showing that vocabulary growth is well captured by representing individual word knowledge in monolinguals, over other measures using only vocabulary size (Beckage et al., 2020; Weber & Colunga, 2022). These analyses bolster the idea that clinicians should consider more than just raw vocabulary size or even a normed percentile when making diagnoses and creating treatment plans for young children, whether they are bilingual or not. Understanding a child's vocabulary structure in addition to percentile rankings provides a fuller picture of an individual child's needs. The more information we have about the structure of a child's vocabulary in each language, the better we are able to model and predict future lexical growth.

The accuracy of the total vocabulary model (L1 and L2 separately) supports and builds on prior findings that knowing the specific words a child knows is useful for predicting vocabulary growth over time (Beckage et al., 2020). The model taking into account word-level knowledge in both languages was over 95% accurate in predicting which words that child would know roughly three months later. That is, having in-depth knowledge of vocabulary structure, particularly when this knowledge was for both languages separately, resulted in an accurate description of a DLL's vocabulary trajectory. Having the predictive power to understand what a child's vocabulary structure may look like in the future has many implications for both identification of at-risk individuals and the creation of intervention materials. Such models could be used to select target vocabulary words for intervention services, or even for general enrichment for children with typically developing language skills. Specifically, these models predict which words are learned in each language, which may be particularly helpful for sequential bilinguals who may have more vocabulary deficits in one language (their L2) than the other. Future work with predictive models should further investigate the relationship between the predicted words in each language, or how initial knowledge distribution across languages impact predictions, in order to understand how best to use these models for applied work.

Model Accuracy

One important note about the accuracy of these models: they are predicting what words are known at the next timepoint. So, the networks are partially learning that any "words" known at the initial visit should still be known, and also predicting which new words should be learned as well. But, we do not explicitly give the model the exact words the child knew individually, but give them the sum of all the vectors of words known. This means knowing the exact words known at the initial visit requires a decomposition of the sum of the vectors.

Though the results indicated that the model utilizing conceptual vocabulary information performed worse across metrics, interestingly the gap between the two models was substantially larger for the AUC than for calculated accuracy. Though both models were decently accurate at predicting, across all 328 possible known words, which were known in the future based on current semantic knowledge, the conceptual model was unable to optimize the tradeoff between sensitivity and specificity, resulting in an AUC of less than 50%. The question is, why would information about word knowledge in both languages enhance the tradeoff between sensitivity and specificity more so than general accuracy? or stated another way, why does conceptual knowledge information help general accuracy more so than AUC. The conceptual model had an AUC of less than 0.5,

indicating that this model is in fact worse than a random classifier when looking across nodes. Because of a likely imbalance in classes (e.g., less known (1) and more unknown (0) words) across most children, it may be more pertinent to look more so at precision rather than tradeoff between both precision (true positive) and specificity (true negatives). This means our measure of calculated accuracy may be a better indicator of model performance between the two models types. Although both accuracy and AUC indicate the L1L2 model is significantly better, we still see admirable performance on accuracy for both the L1L2 and conceptual models.

Limitations and Future Directions

A future analysis should further pick apart AUC under different conditions, specifically when children know very few or many of the tested words. As suggested above, this is because the imbalance in known and unknown words could greatly impact the calculated AUC for the different models. When a child initially knows and learns very few words, the imbalance between known (1) and unknown (0) words is quite large, but as children learn more words this imbalance decreases, and the utility in understanding the tradeoff between sensitivity and specificity also decreases. Though there is an interesting model component to understand (i.e. whether the conceptual or L1L2 model perform better on AUC), it will also be fruitful to understand for what children and under what circumstances this AUC measure is best used. Overall model accuracy or an Receiver-Operating Characteristic (ROC) curve may be better utilized when the number of known and unknown words are more evenly distributed.

Many DLLs residing in the United States constitute an interesting population, as they are exposed to both American culture as well as the culture of their other language. This culture may impact their word knowledge as well. Despite many similarities in vocabulary acquisition, children of different languages and cultures may learn different words and word types at different rates early in vocabulary development (Caselli et al., 1995; Frank et al., 2017; Choi & Gopnik, 1995). Further, there is evidence that the amount of similarity in the two languages of a DLL can impact this order and rate of development in each language (Barac Bialystok, 2012). In order to best model children with multiple cultural influences, we need to choose resources that reflect this varied experience. However, this is no easy feat, and has likely impeded bilingual modeling work. First, the amount of influence each individual culture has on a DLL's development is hard to quantify. Second, finding resources to accurately model multiple languages or cultures adds another layer of complexity. Many resources are only available in one language, usually being English, and don't have comparable counterparts in other languages. Further, some languages are wholly underrepresented in available resources, making it difficult to even begin to study them. Cantonese-English bilinguals are one such population that may be hard to study given a dearth of Cantonese resources, and the likelihood of complex cultural influences, given differences between Eastern (Cantonese) and Western (English) culture (Chang, 2001; Nisbett, 2007).

For example, the present analyses used embeddings from a pre-trained English Wikipedia model. This specific Wikipedia2Vec model was chosen because there are comparable pre-trained models in other languages. However, the Wikipedia2Vec models in other languages still do not contain the same amount of information as the English one does, and there is no Cantonese version, only a Mandarin model. Other unpublished analyses using word2vec semantic embeddings derived from a Cantonese corpus actually suggests that learning in both Cantonese (L1) and English (L2) was better represented using the English-derived embeddings. This could be due to the cultural influence of growing up in the United States, but could also be due to the smaller size and less information available in the Cantonese corpus used. This cannot be untangled until more language resources are available for minority languages.

Similarly, though other research has suggested that models using corpora that are more reminiscent of what a typical child might hear may perform better than adult-oriented corpora such as Wikipedia (e.g., Hills et al., 2010; Weber & Colunga, 2022), many child corpora are not available in other languages or large enough to gather neural network-based distributional semantics from. Prior to the current analyses, the authors investigated gathering embeddings from CHILDES, a corpus of transcribed caregiver-child conversations (MacWhinney, 2014), but many of the vocabulary words were not found in the English CHILDES, and even fewer in the Cantonese or Mandarin CHILDES corpora. One future avenue would be a new push to collect corpora in other languages, both for adults but especially for developmental research. This includes recording language from children as well as transcribing children's books and movies from other cultures. The second avenue would be to perform both similar and new analyses on a range of bilingual and multilingual children, modeling their respective languages using embeddings gathered from corpora in that respective language. Though this initial analysis provides proof of concept, further cross- linguistic analyses are needed.

Other neural architectures and modeling techniques could also be explored. The present analyses optimized multiple parameters such as the dropout rate, number of hidden layers and units, and the learning rate, but other neural architectures besides the standard feedforward network may perform better or provide other clues to learning mechanisms. Similarly, there are other ways to create and model the language input into such networks, such as using other algorithms or metrics to gather semantic information, or using other methods to input the individual words into the model. Other modeling techniques entirely, such as preferential attachment or logistic regression, may also deepen our understanding of bilingual vocabulary development.

References

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

Beckage, N. M., & Colunga, E. (2019). Network growth modeling to capture individual lexical learning. Complexity, 2019.

Beckage, N. M., Mozer, M. C., & Colunga, E. (2019). Quantifying the role of vocabulary knowledge in predicting future word learning. IEEE Transactions on Cognitive and Developmental Systems, 12(2), 148-159.

Barac, R., & Bialystok, E. (2012). Bilingual effects on cognitive and linguistic development: Role of language, cultural background, and education. *Child development*, *83*(2), 413-422.

Bilson, S., Yoshida, H., Tran, C. D., Woods, E. A., & Hills, T. T. (2015). Semantic facilitation in bilingual first language acquisition. Cognition, 140, 122-134.

Calder, S. D., Brennan-Jones, C. G., Robinson, M., Whitehouse, A., & Hill, E. (2022). The prevalence of and potential risk factors for Developmental Language Disorder at 10 years in the Raine Study. *Journal of Paediatrics and Child Health*, *58*(11), 2044-2050.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. Cognitive Development, 10(2), 159-199.

Chang, E. C. (2001). Cultural influences on optimism and pessimism: Differences in Western and Eastern construals of the self.

Choi, S., & Gopnik, A. (1995). Early acquisition of verbs in Korean: A cross-linguistic study. Journal of child language, 22(3), 497-529.

Core, C., Hoff, E., Rumiche, R., & Señor, M. (2013). Total and conceptual vocabulary in Spanish–English bilinguals from 22 to 30 months: Implications for assessment.

Duursma, E., Romero-Contreras, S., Szuber, A., Proctor, P., Snow, C., August, D., & Calderón, M. (2007). The role of home literacy and language environment on bilinguals' English and Spanish vocabulary development. *Applied Psycholinguistics*, *28*(1), 171-190.

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.

Gonzalez-Barrero, A. M., Schott, E., & Byers-Heinlein, K. (2020). Bilingual adjusted vocabulary: A developmentally-informed bilingual vocabulary measure.

Gross, M., Buac, M., & Kaushanskaya, M. (2014). Conceptual scoring of receptive and expressive vocabulary measures in simultaneous and sequential bilingual children. American Journal of Speech-Language Pathology, 23(4), 574-586.

Hammer, C. S., Davison, M. D., Lawrence, F. R., & Miccio, A. W. (2009). The effect of maternal language on bilingual children's vocabulary and emergent literacy development during Head Start and kindergarten. *Scientific studies of reading*, *13*(2), 99-121.

Heilmann, J., Weismer, S. E., Evans, J., & Hollar, C. (2005). Utility of the MacArthur—Bates Communicative Development Inventory in identifying language abilities of late-talking and typically developing toddlers.

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.

Junker, D. A., & Stockman, I. J. (2002). Expressive vocabulary of German-English bilingual toddlers.

Kan, P. F., Huang, S., Winicour, E., & Yang, J. (2020). Vocabulary growth: Dual language learners at risk for language impairment. *American Journal of Speech-Language Pathology*, *29*(3), 1178-1195.

MacRoy-Higgins, M., Shafer, V. L., Fahey, K. J., & Kaden, E. R. (2016). Vocabulary of toddlers who are late talkers. *Journal of Early Intervention*, *38*(2), 118-129.

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

Mancilla-Martinez, J., Hwang, J. K., Oh, M. H., & Pokowitz, E. L. (2020). Patterns of development in Spanish–English conceptually scored vocabulary among elementary-age dual language learners. *Journal of Speech, Language, and Hearing Research*, *63*(9), 3084-3099.

Multilingual people. (2018). ilanguages.org. https://ilanguages.org/bilingual.php

Nisbett, R. E. (2007). Eastern and Western ways of perceiving the world.

Nudel, R., Christensen, R. V., Kalnak, N., Schwinn, M., Banasik, K., Dinh, K. M., ... & DBDS Genomic Consortium. (2023). Developmental language disorder–a comprehensive study of more than 46,000 individuals. *Psychiatry Research*, *323*, 115171.

Patterson, J. L. (1998). Expressive vocabulary development and word combinations of Spanish-English bilingual toddlers. American journal of speech-language pathology, 7(4), 46-56.

Peña, E. D., Bedore, L. M., & Kester, E. S. (2016). Assessment of language impairment in bilingual children using semantic tasks: Two languages classify better than one. International Journal of Language & Communication Disorders, 51(2), 192-202.

Poerner, N., Waltinger, U., & Schütze, H. (2019). E-BERT: Efficient-yet-effective entity embeddings for BERT. *arXiv preprint arXiv:1911.03681*.

Rescorla, L. (2011). Late talkers: Do good predictors of outcome exist?. Developmental disabilities research reviews, 17(2), 141-150.

Schwartz, M. (2014). The impact of the first language first model on vocabulary development among preschool bilingual children. Reading and writing, 27(4), 709-732.

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. International Journal of Bilingualism, 15(4), 426-445.

Weber, J., & Colunga, E. (2022, June). Representing the Toddler Lexicon: Do the Corpus and Semantics Matter?. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 3960-3968).

Wu, S., Zhao, J., de Villiers, J., Liu, X. L., Rolfhus, E., Sun, X., ... & Jiang, F. (2023). Prevalence, co-occurring difficulties, and risk factors of developmental language disorder: first evidence for Mandarin-speaking children in a population-based study. *The Lancet Regional Health–Western Pacific, 34*.

Uchikoshi, Y. (2014). Development of vocabulary in Spanish-speaking and Cantonese-speaking English language learners. Applied Psycholinguistics, 35(1), 119-153.

Yamada, I., Shindo, H., Takeda, H., & Takefuji, Y. (2016). Joint learning of the embedding of words and entities for named entity disambiguation. *arXiv preprint arXiv:1601.01343*.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2018). Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. *arXiv preprint arXiv:1812.06280*.