Modeling vocabulary growth in autistic and non-autistic children

Eileen K. Haebig
Department of Communication Sciences and Disorders

Stanley H. West
Department of Psychology

Christopher R. Cox
Department of Psychology

Louisiana State University
Baton Rouge, LA 70803 USA

Abstract

We assessed the goodness of fit of three models of vocabulary growth, with varying sensitivity to the structure of the environment and the learner's internal state, to estimated vocabulary growth trajectories in autistic and non-autistic children. We first computed word-level acquisition norms that indicate the vocabulary size at which individual words tend to be learned by each group. We then evaluated how well network growth models based on natural language co-occurrence structure and word associations account for variance in the autistic and non-autistic acquisition norms. In addition to replicating key observations from prior work and observing that the growth models explained similar amounts of variance in each group, we found that autistic vocabulary growth also exhibits growth consistent with "the lure of the associates" model. Thus, both groups leverage semantic structure in the learning environment for vocabulary development, but autistic vocabulary growth is also strongly influenced by existing vocabulary knowledge.

Keywords: autism; vocabulary growth; network modeling

Introduction

Children on the autism spectrum have core features of restricted interests and repetitive behaviors and have challenges with social communication (Association, 2013). Although not a diagnostic criterion, the majority of autistic children have delays in early vocabulary development (Charman et al., 2003; Ellis Weismer & Kover, 2015; Luyster et al., 2007). In fact, it is estimated that the average age of autistic children producing their first words is 23 months, which is approximately 11 months later than typically developing infants (Mayo et al., 2013). These early language delays are significant because language skills are predictive of child and adult outcomes in autistic individuals (Anderson et al., 2009; Hedvall et al., 2015; LeGrand et al., 2021). Given this importance of early language, a great deal of research has been dedicated to characterizing the vocabulary abilities of autistic children. The current study aims to apply a semantic network modeling approach to not only characterize lexical acquisition of young autistic children at the word level, but to identify key insights into the learning mechanisms that drive vocabulary development in autistic children.

Autistic children's early vocabulary development

Although the late onset and slow progression of vocabulary has been well-documented in young autistic children, the current literature provides scarce insight into when individual words tend to enter an autistic child's vocabulary. Haebig et al. (2021) reported the ten most frequently reported words produced by minimally speaking autistic children. Nine of them are also among the top ten most frequently reported words in a vocabulary matched group of non-autistic children, indicating that autistic children broadly learn the same first words that non-autistic children learn. Although this is interesting, it provides only a small glimpse into the order of word learning at the earliest point in the acquisition process.

In contrast, the WordBank public repository (Frank et al., 2017), which is a data repository that contains word-level data from the MacAurthur-Bates Communicative Development Inventory (Fenson et al., 2007) for over 12,000 toddlers who speak American English—and several tens of thousands of toddlers who speak other languages—contains sufficient word-level data to derive reliable age of acquisition (AoA) data. Compared to adults reflecting on their own childhood to estimate the age when they learned specific words (Kuperman et al., 2012), parents reporting on the current abilities of their child are more time-sensitive and objective data for estimating AoA (Łuniewska et al., 2016; Smolík & Filip, 2022).

Modeling vocabulary growth as a network

Why would vocabulary norms be useful for understanding autistic vocabulary development? Network growth modeling (Hills et al., 2009; Steyvers & Tenenbaum, 2005) allows for a detailed study of the semantic structure of a child's vocabulary and its relationships to the semantic environment they are developing within (Jiménez & Hills, 2022; Wojcik, 2018). Analyzing a vocabulary as a network of semantic associations provides insight both into its conceptual organization and into mechanisms contributing to its growth. Three models of network growth have been explored:

1. *Preferential attachment* (Steyvers & Tenenbaum, 2005): unknown words that are associated with known words with the most connections to other known words are more likely to be acquired. This emphasizes the structure of the child's current vocabulary in guiding growth.

2. *Preferential acquisition* (Hills et al., 2009): unknown words with the most connections to other words—both known and unknown—are more likely to be acquired. This ignores the structure of the child's vocabulary; only the environment matters.

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.*

3. *Lure of the associates* (Hills et al., 2009): unknown words with the most connections to known words are more likely to be acquired. This emphasizes the associations at the intersection of what is known and unknown but ignores structure more generally.

Hills et al. (2009) observed that preferential acquisition and lure of the associates, but not preferential attachment, explained significant variance in vocabulary growth in typically developing children, with preferential acquisition being the best predictor. Importantly, a model including lure of the associates as a predictor could be improved by adding preferential acquisition, but not vice versa. More recently, Jiménez and Hills (2022) examined vocabulary growth models for non-autistic children with typical language development and those who were late talkers. In the analyses, they included data from receptive and expressive vocabulary knowledge indexed by AoA. They found that vocabulary growth in both groups was most consistent with growth by preferential acquisition relative to the other models. No previous study has applied semantic network modeling to examine vocabulary learning in autistic children, presumably because AoA norms do not exist for this group.

Vocabulary size of acquisition

Any attempt to construct vocabulary growth trajectories for autistic children confronts a fundamental problem. Compared to non-autistic children with typically developing language, autistic children are far more heterogenous and achieve language milestones at a wide range of ages. The assumption of AoA is that there is a typical age at which words are acquired, but age is not a reliable indicator of any given autistic child's level of language development.

This does not mean that characterizing the typical vocabulary growth of autistic children is impossible or ill advised. Rather than estimating the probability of producing each word as a function of age to obtain an expected AoA, these probabilities can be estimated as a function of vocabulary size to obtain a *vocabulary size of acquisition* (VSOA) for each word. VSOA estimates can then be used in place of AoA to describe the most common trajectory of language development, whenever that development begins and the rate at which it proceeds. By selecting non-autistic children to closely match the distribution of vocabulary sizes in the autistic sample, VSOA can facilitate comparisons of vocabulary composition in units of development that are aligned across groups.

Current study

Current models of vocabulary development highlight the importance of semantic structure—both the structure of the lexicosemantic environment that the child is developing in, and the structure of the child's own vocabulary relative to that environment. Thus, the trajectory of vocabulary growth in autism may contain valuable information about how they perceive and relate to the structure of their environment. Divergence from non-autistic growth trajectories, reflected in different patterns of alignment with models of vocabulary growth, may imply cognitive differences relevant to word acquisition. The current study assesses the goodness of fit of three models of vocabulary growth, each differing in its sensitivity to the structure of the environment and the learner's internal state, to estimated vocabulary growth trajectories, measured in intervals of VSOA, over words appearing on the McArthur-Bates Communicative Development Inventory (CDI; Fenson et al., 2007) separately for autistic and non-autistic children.

Methods

Vocabulary data

Word-level data that were measured using the McArthur-Bates Communicative Development Inventory (CDI; Fenson et al., 2007) were analyzed for the current study. The CDI is a vocabulary checklist for caregivers to complete to report a child's expressive vocabulary. Though the CDI was intended for and has normative data for children aged 8 to 30 months, it is appropriate to use with older children with language delays and disorders (Fenson et al., 2007). As such, clinicians and researchers frequently use the CDI to characterize vocabulary knowledge in autistic children (e.g., Arunachalam et al., 2022; Charman et al., 2003; Luyster et al., 2007).

The CDI data were gathered from two databases: the National Database for Autism Research (NDAR; National Institute of Mental Health; Payakachat et al., 2016) and WordBank public repository (Frank et al., 2017). NDAR contains data from multiple studies, each with different protocols. There is variability in the data that is available for each child, and some children participated multiple times at different ages. We included all children with a confirmed autism diagnosis—typically a record of the gold-standard Autism Diagnostic Observation Schedule (Lord et al., 2012) with documentation of an autism spectrum classification—and at least one time point of data with word-level CDI data for the child. This resulted in data from 250 autistic children and 472 unique CDI data entries. The average number of unique CDI data entries that the autistic group contributed was 1.89.

Data from the non-autistic group came from the WordBank American English data (Frank et al., 2017). WordBank is comprised of item-level CDI data from young children who have participated in various studies and who are thought to have typical development. We attempted to ensure that the children in our non-autistic sample are all developing language typically by excluding children who scored below the 15th percentile according to the CDI normative data (Fenson et al., 2007). We analyzed 1,416 unique CDI data entries from WordBank.

Generation of acquisition norms

Normative vocabulary size of acquisition (VSOA) scores for autistic and non-autistic children were estimated using logistic regression following an established protocol for estimating

AoA (Goodman et al., 2008). The probability of each word existing in a child's vocabulary is estimated as a function of vocabulary size, group, and their interaction. The vocabulary size at which the model for a word predicts a probability of .5 for each group is the word's expected VSOA. It is the vocabulary size at which roughly half of the group is expected to produce the word. As such, common first words (e.g., mommy, daddy, ball, bye) that tend to have young AoAs will have small VSOAs and words that typically have older AoAs will have large VSOAs. Out of 680 words on the CDI, we obtained VSOAs for 661 words.

Language environment and features

It is well-established that children learn language through exposure to the language that they are learning (Hart & Risley, 1995). Furthermore, the structure—including the semantic structure—of the language input that children receive is linked to vocabulary developmental trajectories (Hills et al., 2009; Jiménez & Hills, 2022). Therefore, to test theories of vocabulary learning, well-established data that estimate the semantic relationships among early-acquired words are needed.

Free association norms provide one perspective on the relationships among words and have been found to predict lexical growth in typically developing toddlers (Hills et al., 2009). Natural language co-occurrence statistics provide a different, complementary perspective on these relationships. The Child Language Data Exchange system (CHILDES; MacWhinney, 2000) provides multiple corpora containing transcripts of child-adult verbal interactions that have been used estimate word co-occurrence structure in especially relevant contexts (e.g., Jiménez & Hills, 2022). The current study leverages both natural language co-occurrence structure from CHILDES corpora and a recently validated and publicly available child-oriented word association database (Cox & Haebig, 2023) to estimate separate directed networks of word associations.

Additionally, psycholinguistic features such as word frequency and phonological complexity influence word learning in young children with typical and atypical development (Kover & Ellis Weismer, 2014; Schneider et al., 2015). Therefore, we fit a baseline model that included the following psycholinguistic variables: number of phonemes, word frequency (derived from the CHILDES corpus; Bååth, 2010), and phonotactic probability and phonological neighborhood density (estimated using the Vitevitch and Luce (2004, 2016) calculator).

Network definitions

The CHILDES natural language network was defined by sliding a five-word window over the sequence of words in each processed transcript[1] involving a 3 – 60-month-old child and an adult caretaker. Windows span utterances but not transcripts. The directional relationships between the earliest word in the window to the four that follow were tabulated. The resulting asymmetric co-occurrence matrix was subset to include just the 581 words on the CDI for which we had VSOA, estimated non-autistic AoA ≤ 30 (the oldest typically-developing child for which the CDI is validated for use) and that are not homonyms, idiosyncratic proper nouns ("babysitter's name", "child's own name", "pet's name") or the four short phrases ("give me five!", "gonna get you!", "so big!", "this little piggy"). An unweighted, directed edge (connection) was drawn between each pair of nodes (words) if the receiving node followed the sending node at least once in the corpus to form the natural language network.

This yields a very dense network ($\approx 40\%$ of possible edges exist), but increasing this threshold created subnetworks with no paths between them. Despite the density, this natural language network explains substantial variance in typical vocabulary development in 16- to 30-month-olds (Cox & Haebig, 2023).

The word association network was derived from child-oriented word associations cued by all CDI items (except the four phrases and idiosyncratic proper nouns mentioned above; Cox & Haebig, 2023). Referencing responses to the 581 cue words described above, an unweighted, directed edge was drawn between each pair nodes if the receiving (response) node was generated in response to the sending (cue) node at least once. This yields a single sparse network ($\approx 4\%$ of possible edges exist) with paths between all nodes. These child-oriented responses explain more variance in typical vocabulary development in 16- to 30-month-olds than unconstrained responses, and explained substantial variance left unexplained by the natural language network (Cox & Haebig, 2023).

Growth values

The three growth models can each be implemented as a function that takes two arguments: 1) a network defining associations among the set of words that can be learned and 2) a list containing a subset of those words (i.e., those that exist in the vocabulary). Each function returns a growth value for every word not in the vocabulary. An unknown word's growth value is defined by each growth model as:

1. *Preferential attachment*: the average indegree of the known words it is associated with.

2. *Preferential acquisition*: its own indegree in the context of the full network, irrespective of what words are known.

3. *Lure of the associates*: the number of known words that are associated with it.

Note that indegree refers to the number of edges that are directed toward a node. Words assigned large values by a growth model should be more likely to be acquired, if that model accurately characterizes vocabulary growth.

[1] Processing involved regularizing spelling, tokenizing phrases (e.g., "thank you", "go potty"), flagging proper nouns and nonwords (e.g., babbling, invented words), and reducing morphosyntactic variability (lemmatization) over 4.5 million tokens.

Comparing autistic and non-autistic vocabulary growth trajectories

Relative to the patterns of vocabulary growth characterized in ostensibly neurotypical children, autistic children may be found to differ in at least three ways. First, we may observe poorer model fits to the autistic vocabularies in general, with no evidence of superior fits. This would imply that their vocabulary growth is less influenced by semantic structure. Second, we may observe similar model fits overall between groups, with a different profile of fits across models. This would imply that autistic children are sensitive to different aspects of the semantic environment or utilize that information differently. Third, we may observe superior model fits to the autistic vocabularies by one or more models. Note that being more or less in line with a particular growth model is not indicative of "better" or more efficient learning, and it would not be inconsistent for autistic children to be more influenced by the semantic environment while language delayed.

We first grouped the full set of 581 words, W, into $N = 31$ intervals according to their VSOA: $\Pi = \{(-\infty, 20], (20, 40] \ldots (600, \infty]\}$. We will refer to the sets of known words accumulated over interval steps as Ψ_k, $k \in 1..N$, where $\Psi_1 = \Pi_1$, $\Psi_2 = \Pi_1 \cup \Pi_2$, $\Psi_3 = \Pi_1 \cup \Pi_2 \cup \Pi_3$, etc. The corresponding sets of unknown words will be denoted $\Omega_k = W \setminus \Psi_k$, where $\setminus$ is the operator for taking the difference of two sets. The model begins by "knowing" all words in Ψ_1, and growth values can be estimated for all words in Ω_1 within the networks of word associations and child-directed speech. Growth values derived from each network are standardized to have zero mean and unit variance, then stored. This procedure is repeated for $\Omega_1 \ldots \Omega_N$. The result is a matrix, X, where rows correspond to $\Omega_1 \cup \cdots \cup \Omega_N$ and each column contains growth values derived from a particular growth model applied to a particular network. We then add columns to X that contain the psycholinguistic baseline variables that characterize the word represented in each row.

The probability of learning each unknown word in Ω_k, here interpretted as a subset of rows in X that correspond to a selection of unknown words as defined above, is estimated based on a ratio of strengths:

$$p_y = \frac{\exp(\beta y^\top)}{\sum_{x \in \Omega_k} \exp(\beta x^\top)},\ y \in \Omega_k,\ \Omega_k \subset X \tag{1}$$

Which variables are included in X is manipulated to determine the significance of including or excluding growth values from each of the three growth models; baseline models will include only the psycholinguistic baseline variables. Simply put, the model estimates the probability of acquiring unknown word $y \in \Omega_k$ as the weighted sum of the evidence for acquiring word y, $\exp(\beta y^\top)$, divided by the sum total of evidence over all unknown words $x \in \Omega_k$, $\sum_x \exp(\beta x^\top)$. The weights in β are tuned to minimize the negative log likelihood of the model using the `stats::optim` function in `R` (R Core Team, 2024), where log likelihood is defined as the sum of log-transformed probabilities for words that are learned:

$$\log \theta(\beta) = \sum_k^N \sum_{y \in \Delta} \log p_y,\ \Delta = \Omega_k \setminus \Omega_{k+1} \tag{2}$$

In Equation 2, Δ denotes the set difference between the words that are currently unknown and the words that will remain unknown after the words in Π_{k+1} are acquired (i.e., the words that are expected to be learned in the next vocabulary growth step).

Nested models can be compared using a likelihood-ratio test—the difference of log likelihoods follows a χ^2 distribution. In equations 3 and 4, θ_0 and θ_1 denote the likelihood of the restricted and full models in the nested pair, respectively:

$$-2\left(\log \theta_1 - \log \theta_0\right) \sim \chi^2 \tag{3}$$

We compute the Bayesian Information Criterion (BIC) associated with the addition of variables over a restricted model as follows, where p is the total number of model parameters and n is the total number of datapoints the model is trained on:

$$\text{BIC} = 2\log\left(\theta_1 - \theta_0\right) - p \log n \tag{4}$$

Using this modeling framework, we first fit a model attempting to predict vocabulary growth using just the psycholinguistic baseline variables: number of phonemes, word frequency (derived from the CHILDES corpus; Bååth, 2010), and phonotactic probability and phonological neighborhood density estimated using the calculator provided by Vitevitch and Luce (2004, 2016). Next, models are fit that also include the growth values derived from each growth model in turn (both those derived from the natural language network and the word association network, cf. Cox & Haebig, 2023). These are compared to the psycholinguistic baseline model. Finally, using each of these models including growth values from one growth model as a baseline, we fit a set of models including growth values from one additional growth model. This allows us to test whether growth models explain unique variance relative to each other and the psycholinguistic baseline. All models are fit to autistic and non-autistic vocabularies separately.

Results

VSOA findings

When examining non-autistic VSOAs, words with early AoAs are found to have small VSOAs and non-autistic VSOA scores correlate strongly with AoA (Kendall's $\tau_b = .91$, $n = 598$). Autistic VSOA scores are much less correlated with non-autistic AoA estimates (Kendall's $\tau_b = .64$, $n = 598$), indicating that the order in which autistic and non-autistic children learn words diverge. Though the VSOAs for most words were similar between groups (median absolute difference = 53 words), there were also many words with large

Table 1: Differences in VSOA for individual words

Word	Autistic	Non-autistic	Difference
Skate	202	554	-352
This little piggy	107	411	-304
Paint	113	409	-303
Hide	143	383	-240
Brother	254	473	-219
Cut	198	414	-216
Pour	246	458	-212
Daddy	453	1	452
Mommy	320	1	319
Have	645	402	243
Home	416	182	234
Bib	487	255	232
Peek-a-boo	337	110	227
Baby	253	38	215

Note. The 14 individual words with the largest differences in VSOA. Words autistic children tend to learn earlier in vocabulary development are presented above the midline. "Pet's name" and "child's own name" also have significantly higher VSOAs in autistic children, but we excluded proper nouns from this table.

Table 2: Nested model comparisons.

Autistic children				
θ_0	θ_1	χ^2	BIC	p
Baseline	Att.	3.00	−9.72	.224
Baseline	Acq.	72.61	59.89	<.001
Baseline	LOA	41.54	28.82	<.001
Acq.	Acq.+LOA	14.88	2.16	.001
LOA	LOA+Acq.	45.95	33.23	<.001
Non-autistic children				
θ_0	θ_1	χ^2	BIC	p
Baseline	Att.	1.63	−11.06	.444
Baseline	Acq.	77.05	64.36	<.001
Baseline	LOA	41.45	28.77	<.001
Acq.	Acq.+LOA	3.23	−9.46	.199
LOA	LOA+Acq.	38.83	26.14	<.001

Note. All model comparisons are on two degrees of freedom. Baseline models include only psycholinguistic baseline variables (see Methods). Att.: Preferential Attachment; Acq.: preferential acquisition; LOA: lure of the associates.

VSOA differences in both directions.[2] Table 1 provides examples of words that had larger VSOAs for autistic children relative to non-autistic children (i.e., words that autistic children tend to learn later in vocabulary development) and words that had had smaller VSOAs for autistic children relative to non-autistic children (i.e., words that autistic children tend to learn earlier in vocabulary development).

Growth model comparisons

Our second aim was to use each group's word-level VSOA data to test whether the three growth models (preferential attachment, preferential acquisition, and lure of the associates) predicted vocabulary growth within each group. Note that growth values derived from child-oriented word associations and child-directed speech (CHILDES) are always included together for a given growth model. Thus, all nested model comparisons have two degrees of freedom.

We first replicated and extended earlier work (Cox & Haebig, 2023; Hills et al., 2009) by demonstrating that preferential acquisition and lure of the associates, but not preferential attachment, predict non-autistic vocabulary growth beyond the psycholinguistic baseline model when defining expected vocabulary compositions with VSOA intervals rather than AoA (Table 2, bottom panel, $\theta_0 =$ Baseline). The same pattern was also found using the autistic VSOA data (Table 2, top panel, $\theta_0 =$ Baseline). These findings indicate that the structure of the semantic environment (preferential acquisition) and the child's existing vocabulary knowledge (lure of the associates) influence vocabulary growth in both groups. Because preferential attachment was not predictive of growth in either group, it is excluded from subsequent analyses.

We next consider whether the variance explained by lure of the associates is distinct from that explained by preferential acquisition in each group. For non-autistic children, preferential acquisition explains a significant amount of variance that is not explained by lure of the associates ($\chi^2(2) = 38.83$, $p < .001$), but not vice versa ($\chi^2(2) = 3.23$, *n.s.*). However, for autistic children, both model comparisons are significant: preferential acquisition explains a significant amount of variance that is not explained by lure of the associates ($\chi^2(2) = 45.95$, $p < .001$) and lure of the associates explains a significant amount of variance not explained by preferential acquisition ($\chi^2(2) = 14.88$, $p = .001$). Therefore, non-autistic children are predominantly influenced by the structure of the environment when acquiring their first words, while autistic children are additionally influenced by their existing vocabulary knowledge.

Discussion

The current study provides novel information about early vocabulary growth in autistic and non-autistic children. First, we developed a method to calculate word-level acquisition norms for each group. Up to this point, AoA data were available for non-autistic children but could not be generated for autistic children because of the wide heterogeneity in autistic children's expressive spoken vocabulary development. As such, the ages at which an autistic child is reported to produce an individual word varies widely. This is partic-

[2]Bonferroni corrected confidence intervals were bootstrapped (100,000 repetitions) around the VSOA group difference for each word. This indicates 146 statistically reliable differences, with 80 words acquired in smaller vocabularies by autistic children. This statistical analysis is imperfect; words typically among the first to be produced or rarely produced even by children with large vocabularies will have VSOAs outside the observable range with enormous standard errors. E.g., "mommy" and "daddy" have massive, reliable group differences but are not identified by this analysis.

ularly true when including children across the language endowment spectrum, principally minimally speaking autistic children. The VSOA estimates that we developed can be used to compare vocabulary composition in units of vocabulary size development rather than development in chronological age, which allows for alignment across groups. These word-level VSOA data will be useful for numerous studies focusing on vocabulary development and learning and this technique could support the study of groups that may experience differing levels of developmental delays.

The second contribution of the current study is the testing of vocabulary growth models. The current study replicated findings of previous growth model comparisons that used AoA data. As with previous work with non-autistic children that originally used AoA data to examine vocabulary trajectories, the current analyses of the non-autistic VSOA data revealed that the preferential acquisition model added significant explained variance above and beyond the lure of the associates model of vocabulary growth (Cox & Haebig, 2023; Hills et al., 2009). Additionally, our results replicated previous findings that both the lure of the associates model and the preferential acquisition model explain variance above and beyond psycholinguistic variables, while the preferential attachment model does not. This finding now extends to autistic vocabulary growth models.

Moreover, when considering structure from the child-oriented word associations and child-directed speech (CHILDES) together, the preferential acquisition growth model explains unique variance that the lure of the associates model does not in the vocabulary growth trajectories of both groups. This is further evidence that both autistic and non-autistic children are sensitive to the structure within their language environment. Though language input is associated with language growth in young autistic children (Haebig et al., 2013; Siller & Sigman, 2008; Walton & Ingersoll, 2015), researchers have suggested that the ability to learn from linguistic input may differ between autistic and non-autistic children (Arunachalam & Luyster, 2018). The current findings suggest that autistic children are able to process semantic statistics in their environment and leverage it to support word learning. This finding aligns with statistical learning studies that report that autistic children are sensitive to other types of statistical regularities within their input (e.g., phonotactic probabilities; Mayo & Eigsti, 2012; Obeid et al., 2016). Notably, intact statistical learning of phonotactic probabilities has been found in both autistic children who have typical structural language abilities and autistic children who have structural language impairments (Haebig et al., 2017).

Surprisingly, the current study found that autistic children may differ from non-autistic children by being more influenced by their existing vocabulary knowledge when learning new words. The lure of the associates growth model only explained variance not redundant with preferential acquisition when modeling autistic vocabulary growth. Although the semantic structure of the learning environment influences word learning, autistic children's existing vocabulary knowledge also appears to lure in new words to the lexicon. Future research will investigate whether potential learning bias can be observed in experimentally controlled settings, and whether it is a feature of autism or moderated by age (Kalish et al., 2015). This may require observing novel learning experiences in an age matched sample.

We estimated autistic VSOAs from aggregate data. While this is consistent with the way AoA is obtained for the non-autistic population, aggregation may obscure important variability that will be important for understanding language learning differences within the autistic population. Nevertheless, leveraging the large sample available from NDAR revealed reliable differences between groups at the resolution of individual words on the CDI. Without such a database, it would have been infeasible to compute VSOA for autistic children. Future studies should examine whether longitudinal data of child-level vocabulary sizes associated to individual words and certain child characteristics that are associated with autism (e.g., restricted interests) are associated with this novel finding.

Currently, there are limited data that point to the theories that explain how autistic children sample and learn from the input from their environment (Arunachalam & Luyster, 2016). The present study contributes important insight into learning theories that are associated with autistic vocabulary development. This work extends other studies that have examined these learning theories in the context of delayed vocabulary development that is not associated with autism (i.e., late talkers, Jiménez & Hills, 2022). Given the importance of autistic children's existing expressive vocabulary knowledge for subsequent vocabulary growth, future work should consider leveraging vocabulary comprehension data from the CDI-Words and Gestures form to examine how word knowledge within the comprehension domain may influence the trajectory of expressive vocabulary growth (see Jiménez & Hills, 2022 for a late talker application). This would be particularly interesting given that the receptive-expressive gap is smaller in young autistic children (e.g., Charman et al., 2003; Davidson & Ellis Weismer, 2017; Luyster et al., 2007).

Conclusion

Our novel technique of deriving VSOA estimates aligns well with existing AoA estimates from non-autistic toddlers and is able to be extended to autistic children. These VSOA estimates allow for insightful analyses to examine word learning that have not previously been feasible. The current study also examined theory-driven vocabulary growth models, replicating previous non-autistic vocabulary growth findings and revealing novel insights into autistic vocabulary growth. Autistic children are biased to learn words that are well connected in their learning environment like non-autistic children; however, word learning is further enhanced by autistic children's existing word knowledge.

Acknowledgments

We thank our funding source (Louisiana Board of Regents, LEQSF(2020-23)-RD-A-05; PI: Haebig). Also, we are grateful to the families and researchers that contributed data to WordBank and NDAR. Data used in the preparation of this manuscript were obtained from the NIH supported National Database for Autism Research (NDAR). NDAR is a collaborative informatics system created by the National Institutes of Health to provide a national resource to support and accelerate research in autism. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or of the Submitters submitting original data to NDAR. Dataset identifier: 10.15154/q64a-9k34.

References

Anderson, D. K., Oti, R. S., Lord, C., & Welch, K. (2009). Patterns of growth in adaptive social abilities among children with autism spectrum disorders. *Journal of Abnormal Child Psychology*, *37*, 1019–1034.

Arunachalam, S., Avtushka, V., Luyster, R. J., & Guthrie, W. (2022). Consistency and inconsistency in caregiver reporting of vocabulary. *Language Learning and Development*, *18*(1), 81–96.

Arunachalam, S., & Luyster, R. J. (2016). The integrity of lexical acquisition mechanisms in autism spectrum disorders: A research review. *Autism Research*, *9*(8), 810–828.

Arunachalam, S., & Luyster, R. J. (2018). Lexical development in young children with autism spectrum disorder (asd): How asd may affect intake from the input. *Journal of Speech, Language, and Hearing Research*, *61*(11), 2659–2672.

Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).

Bååth, R. (2010). Childfreq: An online tool to explore word frequencies in child language. *Lucs Minor*, *16*, 1–6.

Charman, T., Drew, A., Baird, C., & Baird, G. (2003). Measuring early language development in preschool children with autism spectrum disorder using the macarthur communicative development inventory (infant form). *Journal of Child Language*, *30*(1), 213–236.

Cox, C. R., & Haebig, E. (2023). Child-oriented word associations improve models of early word learning. *Behavior Research Methods*, *55*(1), 16–37.

Davidson, M. M., & Ellis Weismer, S. (2017). A discrepancy in comprehension and production in early language development in asd: Is it clinically relevant? *Journal of Autism and Developmental Disorders*, *47*, 2163–2175.

Ellis Weismer, S., & Kover, S. T. (2015). Preschool language variation, growth, and predictors in children on the autism spectrum. *Journal of Child Psychology and Psychiatry*, *56*(12), 1327–1337.

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, J. S., & Bates, E. (2007). Macarthur-bates communicative development inventories: User's guide and technical manual (2nd ed.).

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.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515–531.

Haebig, E., Jiménez, E., Cox, C. R., & Hills, T. T. (2021). Characterizing the early vocabulary profiles of preverbal and minimally verbal children with autism spectrum disorder. *Autism*, *25*(4), 958–970.

Haebig, E., McDuffie, A., & Ellis Weismer, S. (2013). Brief report: Parent verbal responsiveness and language development in toddlers on the autism spectrum. *Journal of Autism and Developmental Disorders*, *43*, 2218–2227.

Haebig, E., Saffran, J. R., & Ellis Weismer, S. (2017). Statistical word learning in children with autism spectrum disorder and specific language impairment. *Journal of Child Psychology and Psychiatry*, *58*(11), 1251–1263.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children*. Paul H. Brookes Publishing.

Hedvall, Å., Westerlund, J., Fernell, E., Norrelgen, F., Kjellmer, L., Olsson, M. B., Carlsson, L. H., Eriksson, M. A., Billstedt, E., & Gillberg, C. (2015). Preschoolers with autism spectrum disorder followed for 2 years: Those who gained and those who lost the most in terms of adaptive functioning outcome. *Journal of Autism and Developmental Disorders*, *45*, 3624–3633.

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

Jiménez, E., & Hills, T. T. (2022). Semantic maturation during the comprehension-expression gap in late and typical talkers. *Child Development*, *93*(6), 1727–1743.

Kalish, C. W., Zhu, X., & Rogers, T. T. (2015). Drift in children's categories: When experienced distributions conflict with prior learning. *Developmental Science*, *18*(6), 940–956.

Kover, S. T., & Ellis Weismer, S. (2014). Lexical characteristics of expressive vocabulary in toddlers with autism spectrum disorder. *Journal of Speech, Language, and Hearing Research*, *57*(4), 1428–1441.

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

LeGrand, K. J., Weil, L. W., Lord, C., & Luyster, R. J. (2021). Identifying childhood expressive language features that best predict adult language and communication outcome in individuals with autism spectrum disorder. *Journal of Speech, Language, and Hearing Research*, *64*(6), 1977–1991.

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (2012). *Autism diagnostic*

observation schedule (2nd ed.). Western Psychological Services.

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowski, B., Southwood, F., Anđelković, D., Blom, E., Boerma, T., Chiat, S., de Abreu, P. E., et al. (2016). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*, *48*, 1154–1177.

Luyster, R., Lopez, K., & Lord, C. (2007). Characterizing communicative development in children referred for autism spectrum disorders using the macarthur-bates communicative development inventory (CDI). *Journal of Child Language*, *34*(3), 623–654.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk: Transcription format and programs*. Lawrence Erlbaum Associates Publishers.

Mayo, J., Chlebowski, C., Fein, D. A., & Eigsti, I.-M. (2013). Age of first words predicts cognitive ability and adaptive skills in children with asd. *Journal of Autism and Developmental Disorders*, *43*, 253–264.

Mayo, J., & Eigsti, I.-M. (2012). Brief report: A comparison of statistical learning in school-aged children with high functioning autism and typically developing peers. *Journal of Autism and Developmental Disorders*, *42*, 2476–2485.

Obeid, R., Brooks, P. J., Powers, K. L., Gillespie-Lynch, K., & Lum, J. A. (2016). Statistical learning in specific language impairment and autism spectrum disorder: A meta-analysis. *Frontiers in Psychology*, *7*, 205557.

Payakachat, N., Tilford, J. M., & Ungar, W. J. (2016). National database for autism research (ndar): Big data opportunities for health services research and health technology assessment. *PharmacoEconomics*, *34*(2), 127–138.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Schneider, R. M., Yurovsky, D., & Frank, M. (2015). Large-scale investigations of variability in children's first words. *Proceedings of the Cognitive Science Society*, 2110–2115.

Siller, M., & Sigman, M. (2008). Modeling longitudinal change in the language abilities of children with autism: Parent behaviors and child characteristics as predictors of change. *Developmental psychology*, *44*(6), 1691.

Smolík, F., & Filip, M. (2022). Corpus-based age of word acquisition: Does it support the validity of adult age-of-acquisition ratings? *PLoS ONE*, *17*(5), e0268504.

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

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481–487.

Vitevitch, M. S., & Luce, P. A. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics*, *2*, 75–94.

Walton, K. M., & Ingersoll, B. R. (2015). The influence of maternal language responsiveness on the expressive speech production of children with autism spectrum disorders: A microanalysis of mother–child play interactions. *Autism*, *19*(4), 421–432.

Wojcik, E. H. (2018). The development of lexical–semantic networks in infants and toddlers. *Child Development Perspectives*, *12*(1), 34–38.