From idiosyncratic narratives to social scripts: Mapping the semantic space of self-descriptions in children and adults

Karla E. Perez[1], **Claire Augusta Bergey**[2], **Robert D. Hawkins**[2] **& Hyowon Gweon**[1]
{perezke, cbergey, hawkins, gweon}@stanford.edu
[1]Department of Psychology, [2]Department of Linguistics, Stanford University, Stanford, CA

Abstract

Self-description is a routine part of daily life, but choosing what to reveal about ourselves is a cognitively sophisticated communicative act. How do children talk about themselves, and how do they differ from adults? To characterize developmental changes in self-description, we asked 3- to 4-year-old children and adults to share 12 things about themselves. We analyzed over 3,000 self-descriptions using qualitative annotations and sentence embeddings to quantify their semantic structure. Children generated small, idiosyncratic "islands" of self-description, sampling narrowly from a specific topic, whereas adults' responses spanned a broader semantic range, reflecting a more "scripted" approach. While preliminary and descriptive in nature, our work demonstrates how advances in natural language processing can be leveraged to quantify and visualize developmental changes in self-description. The results are consistent with the possibility that self-description evolves from idiosyncratic narratives to a normative social act, raising new questions about the cognitive underpinnings of this development.

Keywords: cognitive development; self-disclosure; social cognition; natural language processing; semantic embeddings

Introduction

Humans are motivated to talk about themselves (Tamir & Mitchell, 2012), but figuring out what to reveal about oneself is a sophisticated cognitive feat. Unlike in describing something about the world (e.g., describing one's car), where the referent has properties that are directly observable and verifiable by others, describing something about the self requires retrieving from a store of knowledge in which much of the relevant content is not directly observable or verifiable. Yet, as adults, we have an intuitive sense of what information is worth sharing about us from an internal repertoire of experiences, traits, thoughts, and preferences. How does this ability develop?

Longstanding theories in linguistics and cognitive science characterize communication as a process by which two agents share information by thinking about each others' minds (e.g., Clark, 1996; Grice, 1975; Sperber & Wilson, 1986). This idea has fueled a large body of work formalizing communication and examining its development (Bonawitz et al., 2011; Goodman & Frank, 2016; Gweon, 2021; Ho et al., 2017; Shafto et al., 2014), but it has primarily focused on how humans communicate about the external world rather than about themselves. Here, we take a first step towards computationally characterizing how adults and children communicate about themselves and how this act might change over development.

Much prior work has focused on *why* people talk about themselves. Beyond the idea that people have a drive for self-disclosure (Carbone & Loewenstein, 2023; Tamir & Mitchell, 2012), existing work has explored the role of self-disclosure in building intimacy (Laurenceau et al., 1998) and how reciprocal sharing and content valence affect how people talk about themselves (e.g., Altman & Taylor, 1972; Gable et al., 2004). However, much remains unknown about the semantic structure of self-description and the cognitive mechanisms underlying *what* people choose to share about themselves.

Imagine that you just met someone; what would you say about yourself? In principle, you can say anything: your cultural and demographic background; your preferences, traits, and qualities; even your inner fears or aspirations. From this repertoire of self-knowledge, you must choose what to include or omit based on what you want others to know about you. For instance, it is acceptable (and perhaps advisable) to mention a peculiar recurring rash to your doctor, but perhaps best to avoid it in a self-introduction to your new boss. As adults, we may have some rehearsed sets of self-descriptions for different social contexts; we routinely mention our research interests in academic settings, but instead mention hobbies and music tastes in other contexts. Intuitively, we can entertain at least three key components that underlie the development of systematic, flexible self-description: source content, sampling strategy, and social scripts.

The first component is the **source content**—one's knowledge about the self—providing the "database" for any self-description. In fact, prior work analyzed self-descriptions of adults (Bugental & Zelen, 1950; Kuhn & McPartland, 1954) and children (Harter, 2012; Montemayor & Eisen, 1977) as a window into their self-concept, finding a shift from observable, concrete attributes such as physical traits among children to more abstract, psychological, and relational traits among adults. While some of this shift may indeed reflect a genuine change in what children know about themselves, the content of self-descriptions is further modulated and constrained by how we draw from the store of knowledge.

The second component is **sampling strategy**: the process of retrieving information from source content. Sampling from one's self-knowledge is a form of memory search (Hills et al., 2012), which changes across development (Nelson & Kosslyn, 1975; Storm, 1980), and can be adapted to the specific context and listener by choosing what is most informative or relevant (Clark & Murphy, 1982); such tailoring likely relies

on our ability to reason about others' mental states (Theory of Mind; see Rakoczy, 2022) and the expected utility of information (Goodman & Frank, 2016; Gweon, 2021), as well as our desire to manage our reputation via communication (Asaba & Gweon, 2022; Baumeister, 1982). As these social-cognitive capacities undergo substantial changes in early childhood, developmental shifts in how children talk about themselves may also reflect changes in their sampling strategy.

A third component is what people consider as the **social norms or 'scripts'** about self-description (e.g., what information is typically included in self-introduction). Such knowledge can be acquired through repeated experience or observation across different settings. Research suggests that through repeated social interactions, individuals internalize autobiographical "schemas" that standardize how a life story or identity is presented (Habermas & Bluck, 2000; Nelson & Fivush, 2004). These scripts may provide templates that help speakers gauge what is typical, appropriate, or relevant to share in a given context. Thus, the development of self-description may further reflect the gradual alignment between one's internal sense of the self and their broader cultural environment.

In addition to these key components, the ability to verbally express one's thoughts also undergoes substantial development (e.g., vocabulary size, see Frank et al., 2021). Thus, linguistic competence is yet another factor that constrains the content of self-description. All in all, characterizing the cognitive processes that underlie self-description in humans is an ambitious undertaking that requires an understanding of the complex interplay between these components and beyond.

With this broader goal in mind, the current work takes a key first step towards building a such unified understanding: Characterizing the developmental shift itself. Only after knowing how children's self-descriptions differ from those of adults can we start generating hypotheses about their underlying cognitive mechanisms and their development. While some past work has analyzed self-descriptions in adults (Bugental & Zelen, 1950; Kuhn & McPartland, 1954) and children (Harter, 2012; Montemayor & Eisen, 1977), they primarily relied on qualitative methods and classifying the contents into discrete categories. Thus, it remains difficult to distinguish whether observed changes in self-descriptions reflect genuine changes in the self-concept or development of other factors such as sampling strategies, social scripts, or linguistic competence.

Recent advances in computational linguistics and natural language processing (NLP) have made it possible to quantify, compare, and visualize the semantic content of multifarious utterances. Beyond classification-based approaches to identify latent signals of personality (Park et al., 2014) or mental health status (Coppersmith et al., 2014; Eichstaedt et al., 2018), some recent work has used sentence embeddings to characterize the semantic structure of conversations and its change over time (e.g., see Hawkins et al., 2020; Schmidt et al., 2025). This approach allows researchers to analyze how people "explore" a latent semantic space across multiple utterances, characterizing not just their semantic contents but also their trajectory in semantic space. Thus, language embeddings can serve as a powerful tool to precisely quantify changes in how people describe themselves across development. While language embeddings in general have been productively applied to child-produced language corpora (e.g., Charlesworth et al., 2021), no prior work to our knowledge has used sentence embeddings to reveal developmental trends in how people explore a latent semantic space.

The current study is exploratory and descriptive in nature, aiming to use qualitative annotations and sentence embeddings to identify both well-established and novel developmental changes. To maximize the range of developmental change, we start with the youngest possible children who can understand the task and produce verbal responses: 3- to 4-year-olds. In Part 1, we describe the behavioral paradigm for children and adults. In Part 2, we report results from qualitative analyses using manual annotation. In Part 3, we present our quantitative analyses using sentence embeddings.

Part 1: Behavioral Paradigm

In this section, we describe the behavioral paradigms used to elicit self-disclosure in children and adults.

Participants

Children: Thirty-four 3- and 4-year-old children were recruited from a preschool in the United States, where English was the primary language of instruction ($M(SD)_{age}$ = 4.03(0.56) yrs, Range: 3.12 - 5.01 yrs).
Adults: Two hundred and fifty adults were recruited on Prolific ($M(SD)_{age}$ = 41.63(13.44) yrs, Range: 18-82 yrs) and paid $1.25 for participation. Only participants who had an approval rate of 95% or higher, spoke English as a first language, were based in the United States, were eligible to participate.

Procedure & Data Preparation

To elicit self-descriptions in a context-neutral way, we asked participants to share 12 things about themselves. This was inspired by the Twenty Questions Test used in prior work (Bugental & Zelen, 1950; Kuhn & McPartland, 1954; Montemayor & Eisen, 1977) where participants were asked to provide 20 different answers to the question "Who am I?".

Child Paradigm Children were tested individually in a quiet room at their preschool. The experimenter first identified herself as a new 'game-room teacher'–a designation familiar to the children at this site. The experimenter then introduced the task, framed as a game where the child earned a uniquely colored 'star' (laminated paper with Velcro backing) for each piece of self-information shared, to attach on a felt-covered board (approx. 12 x 12 in.). The experimenter provided a scripted prompt: "For every one thing that you share about yourself, I am going to give you a star. You can place the star anywhere on the board."

To establish her baseline knowledge about the child and initiate the task, the experimenter acknowledged the child's

name and school affiliation and expressed her interest in getting to know more about the child: "I know that your name is X and that you go to [school name]. What else can I learn about you?" This initial cover story and open-ended prompt provided a naturalistic conversational context to facilitate children's self-disclosure. After each response, the child received a star to place on the board and was prompted for another response (e.g., "Thank you for sharing, here's a star. What else would you like to say about yourself?"). The session continued until the child opted to stop or collected all 12 stars.

Adult Paradigm Adults completed a custom jsPsych experiment after passing a Google reCAPTCHA v2 verification to screen for bots. Participants were instructed to imagine that they had just met someone with whom they had just exchanged names, and then asked to answer the question "What are some other things you'd like to say about yourself?" across 12 free-response trials. To mimic the task for children, adults were prompted with variations of "What['s] [the first/another/else] [thing/would] you['d] like to [say/share] about yourself?" above a text box. Participants were required to provide text input before advancing to the subsequent trial. Similar to the stars in the child version of the task, a progress indicator (e.g., "x out of 12") was displayed throughout the task.

Data pre-processing We used a two-step pre-processing pipeline to prepare this dataset for subsequent analyses.

1. **Transcription and extraction:** Child sessions were transcribed using WhisperX (Bain et al., 2023). We then hand-extracted and de-capitalized each discrete response. Some responses were extracted over a few exchanges. Consider the following exchange, for instance, between the child (**C**) and the experimenter (**E**): (C) "I have a big bobby house"; (E) "A bobby house?"; (C) "No, a BOBBY house, like with dolls" (E) "Oh, a Barbie house!". This exchange was coded as "i have a big barbie house".
2. **Standardization:** To prevent confounds in the semantic analyses, we equated orthography between the child and adult data by de-capitalizing all text responses and removing any punctuation that was not a period or apostrophe.

This pre-processing procedure yielded a dataset containing a total of 3,339 discrete responses across all participants.

Part 2: Qualitative analysis

In this section, we describe our qualitative approach to characterizing and comparing the contents of responses provided by children and adults. First, in light of prior work that suggests increasing levels of abstraction in children's self-concept (Harter, 2012), we asked whether children tend to generate responses that are more concrete and episodic in nature, which may become more abstract with age. We then looked at the specific content of these responses to analyze the number of topics children and adults mention in their self-descriptions.

The level of generality and abstractness

Our first analysis examined the level of generality and abstractness in children and adults' responses. We developed a 7-point coding scheme to assess generality (from episodic facts to general traits or qualities) and abstractness (from concrete or perceptual facts to abstract qualities) in participants' self-descriptions: 0 = No response; 1 = irrelevant (i.e., not about the self); 2 = self-relevant but impossible/not true (e.g., "I can fly"); 3 = self-relevant, concrete, episodic, retrieval of past events (e.g., "I ate yogurt this morning"); 4 = self-relevant, concrete, episodic, future-oriented (e.g., "I will walk my dog this evening"); 5 = self-relevant, concrete or perceptual traits and qualities (e.g., "I have brown eyes"); 6 = self-relevant, abstract/conceptual traits and qualities (e.g., "I like to dance").

We found that children's tendency to mention more abstract traits and qualities was significantly different than adults; their average abstractness score ($M(SD)$ = 3.67(2.04)) was significantly lower than the average score in adults ($M(SD)$ = 5.52(0.75)), W = 476, $p < .001$, Wilcoxon rank sum test). Within child participants, however, age was not a significant predictor of abstractness score (β = .35, p = .29).

The breadth of topics

Our second analysis examines the number of "topics" that child and adult participants mentioned across their 12 responses. As a preliminary pass at categorizing the topic of each response, the first author manually coded the responses for unique topics within each group. Then, we counted the number of unique topics mentioned (e.g., favorite activities, physical attributes, etc.) within each group.

We found that, rather than selecting key facts from a diverse range of topics (as adults do), children tend to construct one or two ad-hoc domains on-the-fly (e.g., favorite activities) and keep generating closely thematically related facts; the difference between children and adults in their average topic count was statistically significant (W = 406.5, $p < .001$, Wilcoxon rank sum test). Age was a significant predictor of

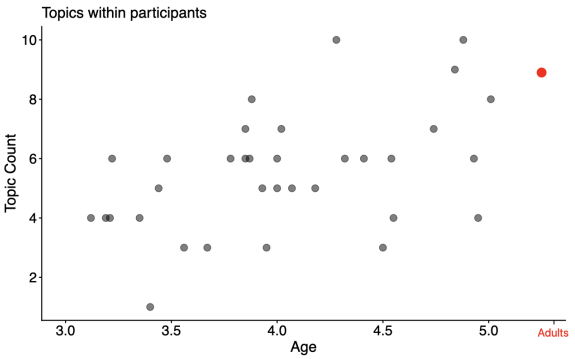

Figure 1: The number of topics mentioned by each child participant (n = 34 total) with the mean topic count for a random subset of adult participants (n = 14).

topic count (β = 1.68, p = .005; see Figure 2), suggesting a clear developmental trajectory toward a more diverse range of topics with age.

Part 3: Characterizing the semantic space of self-descriptions

In this section, we leverage language embeddings to quantify how children and adults explore semantic space as they describe themselves. Language embedding models allow us to precisely locate each self-description in a multidimensional semantic space, making it possible to conduct quantified comparisons of each person's descriptions to their counterparts (Hawkins et al., 2020). We use a sentence transformer model, which represents the meaning of each word in its sentential context and achieves high alignment with humans' ratings of sentence similarities (Reimers & Gurevych, 2019). This class of models has been fruitfully used to examine how humans explore semantic space in dialogues (Schmidt et al., 2025). By representing the semantic content of participants' utterances as high-dimensional embeddings, we can quantify the semantic similarity between utterances both within and across individuals and precisely compare adult and child groups.

To quantify the semantic content of children's and adults' self-descriptions, we used the Sentence-BERT (SBERT) framework (Reimers & Gurevych, 2019) (specifically the `all-distilroberta-v1` model hosted on the Hugging Face) to extract a 768-dimensional vector for each item produced by the participants. Full vectors were used for analyses; we used UMAP, a dimensionality reduction algorithm, to reduce representations to two dimensions for visualization in plots (McInnes et al., 2018).

We started by comparing the mean verbosity (word count per response) between groups to see whether semantic structural differences might be driven by response length. Children produced shorter responses than adults on average (M_{child} = 6.68 words, M_{adult} = 8.31, Welch's $t(60) = 2.54, p = .01$, Cohen's d = .37, 95% $CI[0.08, 0.67]$). To assess whether the semantic differences reported below are driven by this verbosity difference, we control for mean word count in subsequent analyses.

Internal cohesion in children vs. adults

Our primary finding from the qualitative topic coding was that the content of self-descriptions spans more topics in adults than children, suggesting that the semantic content of self-descriptions becomes more varied with age. In the present analysis, we tried to quantify the breadth of topics using semantic similarity, which represents the tightness, or coherence, of a participant's self-descriptions in a semantic space.

To assess within-participant semantic similarity, we calculated the mean pairwise cosine similarity between all possible pairs of responses for each participant. Children exhibited significantly higher internal similarity than adults ([M = 0.92 vs. M = 0.90], Welch's $t(44.54) = -8.25$, p < .001). To verify that this difference was not driven by the verbosity gap reported above, we conducted a linear regression including each participant's mean word count as a covariate. The age effect remained large and significant ($\beta = .023, SE = .003, t(280) = 8.08, p < .001, \eta_p^2 = .21$, 95% $CI[0.14, 1.00]$), with the full model explaining 24.9% of variance ($F(2, 280) = 47.62, p < .001$). This indicates that a given child produces a more semantically constrained set of descriptions independent of how many words they produce, whereas a given adult covers a wider semantic space.

Semantic dispersion in children vs. adults

The fact that individual children respond in tightly-clustered areas of semantic space is consistent with at least two possibilities. One is that children all respond in a small shared region of semantic space; another is that each child samples from a small but idiosyncratic region of semantic space. Here, we test dispersion across individuals: As a group, how much of semantic space do children cover compared to adults?

We analyzed the semantic distances between participants' response centroids within each age group. Each participant's centroid was defined as the mean vector of their response embeddings, which we then L2-normalized. For each participant, we computed the mean cosine distance from their centroid to every other same-group participant's centroid, yielding a per-participant idiosyncrasy score representing how far each participant sat from their group's typical position. We then fit a linear regression model predicting this score using age group (children vs. adults) and mean word count.

Children's centroids were significantly further from same-group peers than adults', even after controlling for verbosity (mean word count); $\beta = .006, SE = .001, t(281) = 3.53, p < .001, \eta_p^2 = .03$, 95% $CI[0.01, 1.00]$). Verbosity, however, was also an independent and somewhat larger predictor, as it accounted for more variance, of inter-participant distance ($\beta < .001, SE < .001, t(281) = 4.69, p < .001, \eta_p^2 = .07$, 95% $CI[0.03, 1.00]$). The model including verbosity accounted for 10.0% of total variance in inter-participant distance ($F(2, 281) = 15.65, p < .001$). Complementing the difference between children and adults, a non-parametric permutation test of multivariate dispersion (PERMDISP) also found greater dispersion among children than adults, $F(1, 282) = 4.21, p = .047$.

Together with the cohesion result, these findings suggest that children's self-descriptions cluster in distinct regions of semantic space. It is notable that dispersion is larger in children than adults despite the fact that individual adults' responses cover a wider space than individual children's responses. This raises the possibility that although adults sample broadly to provide a representative list of self-descriptions, their contents are overall more aligned to each other than those of children.

Sequential leap distance in children vs. adults

The difference between children and adults—tightly clustered but idiosyncratic regions in children versus widely spread but overlapping (across individuals) regions in semantic space

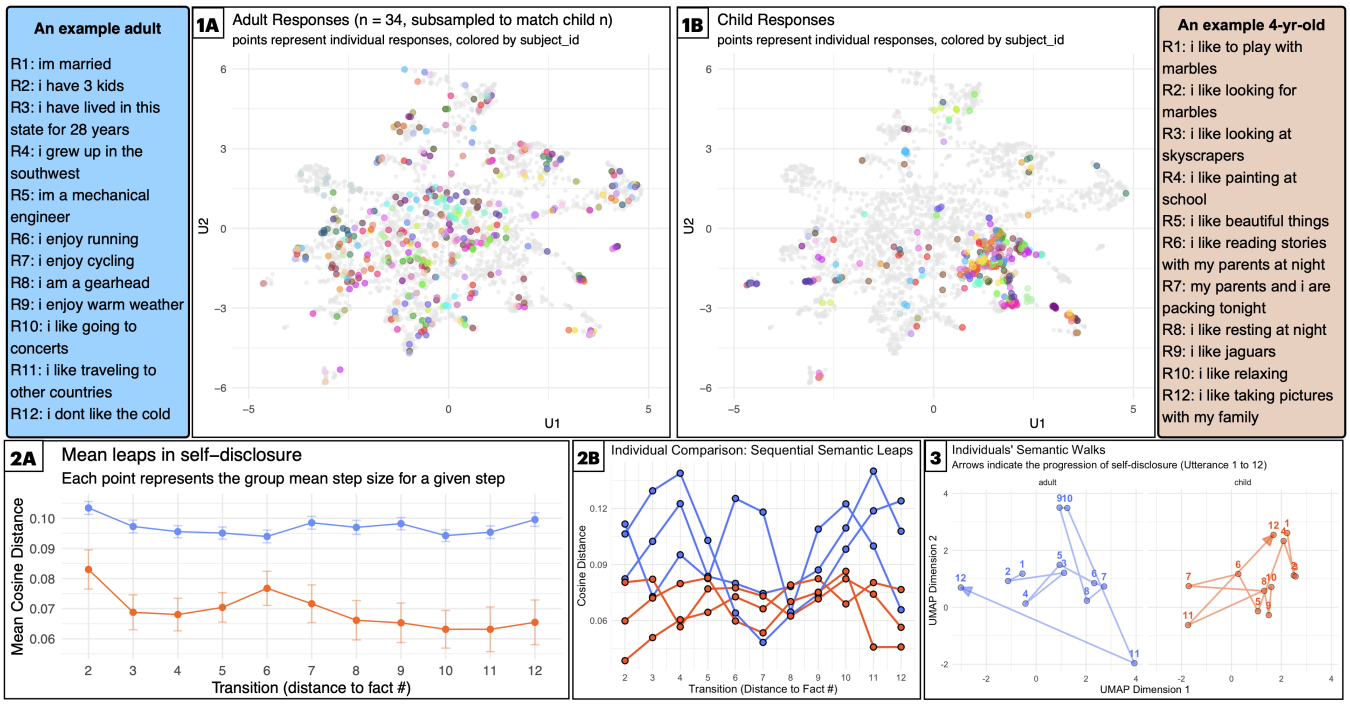

Figure 2: 1A & 1B: UMAP visualization of participants' responses, with responses from an example participant for each group. Each point in the UMAP represents a single response, colored by Subject ID for adults (1A; $n = 34$), randomly sub-sampled to match the sample size for children (1B; $n = 34$). Children's responses are both more clustered within participants than adults and more dispersed across participants over semantic space than adults. 2A & 2B: Cosine distance between consecutive participant responses: Mean leap distance between transitions (2A) and leap distance in individual trajectories (2B) showing 3 adults and 3 children as examples. 3: UMAP visualization of individual participants' trajectory (1 adult, 1 child)

in adults—could result from different patterns of sequential sampling. Thus, we further sought to examine how each group explores the semantic space as they sample one self-description after another, moment-to-moment.

To compute the step length between responses, we calculated the cosine distance between each set of consecutive responses within individuals. We then ran a linear mixed-effects model to determine whether age group predicts the average distance between individuals' sequential semantic steps, with a random intercept for subject identity to control for individual differences in baseline disclosure style. This model revealed a significant main effect of age groups: Children took significantly smaller steps in semantic space compared to adults ($\beta = -0.025$, $SE = .003$, $t(300) = -7.98$, $p < .001$).

To assess the effect size of age group on sequential semantic distance, we calculated the marginal and conditional R^2 values for the mixed-effects model. The fixed effect of age group explained a 6% of the total variance (Marginal $R^2 = .06$) in step-distance, and the full model—including random intercepts for individual participants—accounted for 22.7% of the variance (Conditional $R^2 = .23$). The Intraclass Correlation Coefficient (ICC) was .18, indicating that approximately 18% of the variance in semantic step size was attributable to stable individual differences between participants, which further justifies our use of a multilevel modeling approach.

General Discussion

Using both qualitative annotation and semantic embeddings, we examined how young children and adults talk about themselves. In particular, our sentence embeddings approach identified differences between adults' and children's semantic trajectories that would be difficult to precisely characterize using traditional qualitative methods.

We found that children's self-descriptions were internally cohesive in semantic space while also exhibiting striking idiosyncrasy: While adults' responses covered larger areas of semantic space that overlap with those of other adults, children tended to pick a distinctive region of semantic space and sample narrowly. As a result, each child's responses formed semantic 'islands' that were distinct from those of their peers, but as a group, their responses were dispersed over a larger semantic space than those of adults.

While these findings are descriptive in nature, we can speculate on how these patterns bear on potential developmental changes in the key components of self-description: source content, sampling strategy, and social script.

First, it is notable that even children as young as 3 were able to talk about themselves, suggesting they already have a form of self-knowledge—source content—from which to generate these responses. But the fact that their responses are clustered around fewer topics suggests that this source content continues

to grow and become more refined in development.

Second, children and adults may also differ in their sampling strategy; adults have better pragmatic competence and more experience retrieving from their memory, leading them to sample broadly to form a representative list of self-descriptions; in contrast, children might forage rather ineffectively through their self-knowledge (Abbott et al., 2015; Hills et al., 2012)—retrieving and assembling self-relevant facts on the fly—without the communicative goal to form a representative description of the self.

Third, adults' sampling was not only broad but also more aligned across participants, reflecting the role of a shared social script; after all, adults have years of experience introducing themselves—or observing others introduce themselves—in various social settings, which could lead to a learned set of informative and socially-expected self-descriptions. It is particularly notable that semantic dispersion was smaller in adults than children, even though individual adults occupied a larger space than children. In other words, acquisition of social script may manifest as convergence or alignment towards a group's central tendency (centroid).

These possibilities are not mutually exclusive, and the current study was not designed to adjudicate between these potential mechanisms. Nonetheless, our findings suggest that self-descriptions reflect more than one's self-concept or knowledge. Instead, the construction of the social self—the self as presented to others through self-descriptions—may be the product of a complex interplay between internal discovery, memory retrieval, and cultural alignment, enabled by linguistic competence. These findings also lay the groundwork for understanding how the mundane yet highly personal act of self-disclosure becomes conventionalized as a communicative act (see Hawkins et al., 2019).

This preliminary study provides initial evidence that sentence embeddings can be productively applied to analyze developmental data and reveal differences in how children and adults talk about themselves. One might wonder, however, whether these findings are unique to the self: Are children's idiosyncratic, ad-hoc semantic structures specific to self-descriptions, or does this reflect a broader developmental trend in how children retrieve information from memory? Although children's sampling strategy might develop throughout early childhood, given that children this age have little experience introducing themselves, this tendency may be particularly pronounced in the domain of self. One way to address this question is to examine whether children show a similar foraging dynamic when they produce descriptions in other familiar domains, such as their parent, their favorite animals, or recent activities. More generally, disentangling self-specific from domain-general retrieval processes is a meaningful direction for future work.

The current work used a simplified version of the Twenty Questions Test used in prior work (e.g., Kuhn & McPartland, 1954; Montemayor & Eisen, 1977). The use of this minimal task was a deliberate decision; by minimizing conversational dynamics such as real-time feedback, mutual disclosure, or negotiation of common ground (Schmidt et al., 2025), we could ensure that the observed child–adult differences cannot be attributed to these factors. Though unnaturalistic, we view this minimal baseline task as groundwork for follow-up studies that manipulate specific conversational features (e.g., feedback contingency, communicative goals) to measure how they modulate disclosure patterns.

However, we also note several limitations related to the task. First, it limits the generalizability of our findings; in richer, naturalistic conversations, social feedback and other factors could either encourage broader semantic exploration (Barron et al., 2018) or reinforce local cohesion (Brennan & Clark, 1996). Second, while the task requires producing multiple discrete facts, we did not have an independent measure of semantic fluency. Third, the framing of the task and response format (speech vs. text) could have been better matched between children and adults.

Additionally, there are limitations related to our sample. First, the sample size was asymmetric; the child sample was much smaller than our adult sample ($n = 34$ vs. $n = 250$), limiting the precision of our results and statistical power. Second, as we focused on recruiting the youngest age group for maximal contrast with adults, a finer-grained, continuous developmental trend between childhood and adulthood remains to be seen. Third, our samples were drawn from a single sociolinguistic context, and our child data were collected from a single university-based preschool. Given that norms around appropriate self-disclosure, and the developmental experiences that shape them, vary substantially across cultural contexts (Nelson & Fivush, 2004), the demographics of our sample constrains the generalizability of our findings. One way to address at least some of these concerns is to collect a larger sample of children spanning a wider age range—ideally from a more demographically diverse participant pool—along with additional tasks that can measure semantic fluency. In principle, our analytic approach could be applied to a much larger dataset of self-descriptions collected from different languages and cultures.

As a final point, although children in our study showed less consolidated, more idiosyncratic self-descriptions than adults, prior work has demonstrated that children are remarkably adept in informative, cooperative communication (Gweon, 2021). Thus, it is possible that despite the idiosyncrasy in their self-descriptions, they can nonetheless flexibly modulate their sampling strategy depending on the context, such as how much their communicative partner knows about them, or the understanding that communicative utterances should be informative. Experimental studies with young children might help shed light on these possibilities.

In sum, our work represents the first step in characterizing the semantic space of self-descriptions in children and adults using computational tools. These initial findings provide a rich foundation for further studying how humans come to learn and communicate about themselves.

Acknowledgments

We are grateful to the staff at Bing Nursery School for supporting data collection, and to the children and families who participated in our study.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological review*, *122*(3), 558–569.

Altman, I., & Taylor, D. A. (1972). *Social penetration: The development of interpersonal relationships*. Holt Rinehart; Winston.

Asaba, M., & Gweon, H. (2022). Young children infer and manage what others think about them. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(32), e2105642119. https://doi.org/10.1073/pnas.2105642119

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *Interspeech 2023*, 4489–4493. https://doi.org/10.21437/Interspeech.2023-78

Barron, A. T. J., Huang, J., Spang, R. L., & DeDeo, S. (2018). Individuals, institutions, and innovation in the debates of the French Revolution [_eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1717729115]. *Proceedings of the National Academy of Sciences*, *115*(18), 4607–4612. https://doi.org/10.1073/pnas.1717729115

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin*, *91*(1), 3–26. https://doi.org/10.1037/0033-2909.91.1.3

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482

Bugental, J. F., & Zelen, S. L. (1950). Investigations into the'self-concept'. i. the way technique. *Journal of personality*.

Carbone, E., & Loewenstein, G. (2023). The drive to disclose. *Consumer Psychology Review*, *6*(1), 17–32.

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J. Le Ny & W. Kintsch (Eds.), *Language and comprehension* (pp. 287–299). Advances in Psychology.

Coppersmith, G., Harman, C., & Dredze, M. (2014). Measuring post traumatic stress disorder in twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*, 579–582. https://doi.org/10.1609/icwsm.v8i1.14574

Eichstaedt, J. C., et al. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, *115*(44), 11203–11208. https://doi.org/10.1073/pnas.1802331115

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

Gable, S. L., Reis, H. T., Impett, E. A., & Asher, E. R. (2004). What do you do when things go right? The intrapersonal and interpersonal benefits of sharing positive events. *Journal of Personality and Social Psychology*, *87*(2), 228–245. https://doi.org/10.1037/0022-3514.87.2.228

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Science*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Science*, *25*(10), 896–910. https://doi.org/10.1016/j.tics.2021.07.008

Habermas, T., & Bluck, S. (2000). Getting a life: The emergence of the life story in adolescence. *Psychological Bulletin*, *126*(5), 748–769. https://doi.org/10.1037/0033-2909.126.5.748

Harter, S. (2012). *The construction of the self: A developmental perspective*. Guilford Press.

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *Cognitive science*, *44*(6), e12845.

Hawkins, R. D., Goodman, N. D., & Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in cognitive sciences*, *23*(2), 158–169.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431.

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*, *167*, 91–106.

Kuhn, M. H., & McPartland, T. S. (1954). An empirical investigation of self-attitudes. *19*(1), 68–76.

Laurenceau, J.-P., Barrett, L. F., & Pietromonaco, P. R. (1998). Intimacy as an interpersonal process: The importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges. *Journal of Personality and Social Psychology*, *74*(5), 1238–1251. https://doi.org/10.1037/0022-3514.74.5.1238

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for dimension re-

duction. *arXiv preprint arXiv:1802.03426*. https://arxiv.org/abs/1802.03426

Montemayor, R., & Eisen, M. (1977). The development of self-conceptions from childhood to adolescence. *Developmental Psychology*, *13*(4), 314–319. https://doi.org/10.1037/0012-1649.13.4.314

Nelson, K., & Kosslyn, S. (1975). Semantic retrieval in children and adults. *Developmental Psychology*, *11*(6), 807.

Nelson, K., & Fivush, R. (2004). The emergence of autobiographical memory: A social cultural developmental theory. *Psychological Review*, *111*(2), 486–511. https://doi.org/10.1037/0033-295X.111.2.486

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., & Seligman, M. E. P. (2014). Automatic personality assessment through social media. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, *1*(4), 223–235. https://doi.org/10.1038/s44159-022-00037-z

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Schmidt, H., Bergey, C. A., Zhou, C., Helion, C., & Hawkins, R. (2025). Dynamics of topic exploration in conversation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Blackwell.

Storm, C. (1980). The semantic structure of animal terms: A developmental study. *International Journal of Behavioral Development*, *3*(4), 381–407.

Tamir, D. I., & Mitchell, J. P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(21), 8038–8043. https://doi.org/10.1073/pnas.1202129109