Supplementary data - Phonological Networks and Systematicity in Early Lexical

Acquisition

Supplementary data - Phonological Networks and Systematicity in Early Lexical

Acquisition

The supplementary materials below present further analyses and examples from the main study. This data is included here for transparency and further explanation, and to support replication and further analysis of this data. All scripts and data for the analysis can be found on the project's OSF page at

https://osf.io/uzrsy/?view_only=340858d2084245d087fc00fcca41b679.

The analyses below (S1:S6) show:

- S1: Demonstration of how phonological distance was calculated, with examples
- S2: Overview of how the connectivity threshold of 0.25 was established for the analysis
- S3: AoP ~ Degree connectivity across infants in the dataset
- S4: Full model output tables from the analyses in the main paper
- S5: GAMMs with INT/EXT values scaled by speaker and age
- S6: By-speaker comparisons of the data

S1: Establishing phonological distance.   Phonological distance was established following Monaghan, Christiansen, Farmer and Fitneva's (2010) approach, with some adaptations. Note that in their study, only monosyllabic words are included, and so their approach is adapted here to include multisyllabic words. Following their method, each word was first divided into a series of 'slots', according to its phonological structure. For example, the word *baby* was separated into five slots: /b-e-i-b-i/. Because vowels were not accounted for, the nucleus of each syllable - both monophthongs and diphthongs - was then replaced by a generic V slot, i.e. /b-V-b-V/. Words were then aligned by nucleus to generate a phonological distance measure between each possible word pair. All consonants at word onset and final syllables were aligned, regardless of syllable number, such that the

final /d/ of *bed* would be aligned with the final /n/ of *balloon*. This is because infants may have a tendency to produce only certain consonants word-finally, and so this approach would capture such systematicity. For the English data, the maximal word structure considered in the analysis is C-C-C-V-C-C-|C-C-C-V-|C-C-C-V-|C-C-C, where syllable boundaries are marked with a |. This accounts for complex clusters at word onset (e.g. *splash* /splæʃ/, C-C-C-V-C), coda (*plant* /plænt/, C-C-V-C-C), and across syllable boundaries (*pumpkin* /pʌmpkɪn/, C-V-C-C-|C-V-C). In French the maximal structure was C-C-C-V-C-C-|C-C-C-V-|C-C-C-V-C|C-C-C-V-C|C-C-C-V-|C-C-C. This accounted for multisyllabic target words such as *hélicoptère* ("helicopter" /elɪkɔptɛʁ/, V-|C-V-|C-V-C-|C-V-C) and *appareil photo* ("camera" /apaʁɛjfɔto/, V-|C-V-|C-V-C-|C-V-|C-V), and complex codas as in *arbre* ("tree" /aʁbʁ/, V-C-|C-C). For vowel-initial words, the C1 slot in word-initial position is empty, but all other alignments remain the same.

This maximal structure is required in an analysis of infant word production, to account for unexpected complexities such as, production of French *mettre* "to put" as [mɛʁstʁ] and *étoile* "star" as [ɛstwal]. In the infant data, it was not always easy to determine exactly where a syllable boundary should occur in complex productions, in part because this was not predictable based on the target form due to the variability in production, so would have to be done on a word-by-word basis, and in part because the syllable boundary of some productions could not be clearly established from its phonetic transcription. Instead, consonants were always assigned to the syllable-initial cluster, rather than assigning part of the cluster to the coda of the previous syllable (e.g. for the examples above, the infant production of *mettre* was coded as C-V-|C-C-C-C and *étoile* as V-|C-C-C-V-C).

Tables 1 and 2 shows examples of how phonological distance was calculated from the data.

Table 1

Table showing how phonological distance was calculated across the dataset. Here distance is calculated between baby and balloon; note that only consonants were included in the phonological distance measure. Word position relates to each consonant's position in the word; features shows values of the 11 distinctive features that were used to generate distance values between each pair of words; sum squared differences shows the squared difference in values between baby and balloon. The phonological distance of the word pair is shown on the bottom row. S1 = first syllable, S2 = second syllable; C1 = first consonant, C2 = second consonant, CF = final consonant.

Word position	Baby		Balloon		Sum sq. diffs
	Cons.	Features	Cons.	Features	
S1C1	b	-1,1,0,-1,1,1,0,-1,1,0,0	b	-1,1,0,-1,1,1,0,-1,1,0,0	$0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0$
S1C2	-	0	-	0	-
S2C1	b	-1,1,0,-1,1,1,0,-1,1,0,0	l	0.5,0,1,0,-1,-1,1,-1,1,1,0	$2.25 + 1 + 1 + 1 + 4 + 4 + 1 + 0 + 4 + 1 + 0$
S2CF	-	0	n	0,0,1,1,1,-1,1,-1,1,1,0	$0 + 0 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0$
Phonological Distance ($\sum \sqrt{\text{total}}$)					7.21590931844225

Table 2

Table showing how phonological distance was calculated across the dataset. Here distance is calculated between baby and sky; note that only consonants were included in the phonological distance measure. Word position relates to each consonant's position in the word; features shows values of the 11 distinctive features that were used to generate distance values between each pair of words; sum squared differences shows the squared difference in each of the values between baby and sky. The phonological distance of the word pair is shown on the bottom row. S1 = first syllable, S2 = second syllable; C1 = first consonant, C2 = second consonant, CF = final consonant.

| Word position | Baby | | Sky | | |
	Cons.	Features	Cons.	Features	Sum sq. diffs
S1C1	b	-1,1,0,-1,1,1,0,-1,1,0,0	s	-0.5,1,-1,-1,0,-1,1,-1,-1,1,1,0	0.25 + 0 + 1 + 0 + 1 + 4 + 1 + 0 + 4 + 1 + 0
S1C2	-	0	k	-1,1,-1,-1,1,-1,-1,-1,-1,-1,0	1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0
S2C1	b	-1,1,0,-1,1,1,0,-1,1,0,0	-	0	1 + 1 + 0 + 1 + 1 + 1 + 0 + 1 + 1 + 0 + 0
S2CF	-	0	-	0	-
Phonological Distance ($\sum \sqrt{\text{total}}$)					9.30802897123297

S2: Establishing connectivity within the networks. To determine the 'cut-off' for connectivity in the networks (i.e., at what distance threshold are two words assumed to be connected within the network?), two approaches were taken:

1) Spearman's correlation coefficients were generated between AoP and Degree for each word in the network along a series of 1000 connectivity thresholds between 0.01 and

.99, following Amatuni and Bergelson (2017). This measure allows an overview of all potential connectivity in the networks; we expect a negative AoP~Degree correlation, since as the age of first production increases, a word should be less likely to connect to other nodes in the network. Figure 1 below shows how this relationship changes at different threshold limits: for thresholds upto around .15, there is no predictable change in the AoP~Degree correlation as threshold increases, and indeed some of the correlations are not significant in this range (shown by open circles in Figure 1). On the other hand, from thresholds of around 0.48 we start to see the relationship between AoP and Degree become less negative (i.e. rho moves closer to 0) and then plateau, such that connectivity in the network has reached its maximum potential connectivity, i.e. almost all nodes are connected. We can see that this happens on or shortly after a threshold of 0.5 for both datasets. Figure 2 plots the distribution of scaled distances between all words in each infant's network. Here we can see that there are only few words in the dataset that have a distance of more than 0.5. A threshold between 0.15 and 0.5 is thus likely to capture the most appropriate representation of connectivity within the network.

2) Taking this into account, as reported in the main paper, a value one SD below the mean was accepted to be a suitable threshold for connectivity in the dataset. A threshold of 0.25 sits at the upper end of the lowest quartiles across each of the four datasets (French and English, each with Actual and Target data; range = 0.18-0.24).

To further explore the extent to which the results are dependent on this threshold value, the data was re-analysed at a series of additional thresholds, representing the lower and upper limits of the analysable range discussed above (0.15-0.5), as well as each of the lower quartile values for the four individual datasets, and the overall mean connectivity in the data (0.34). The full dataset was regenerated seven times with each of these seven thresholds as a cut-off for connectivity (i.e. two nodes in the network were said to be
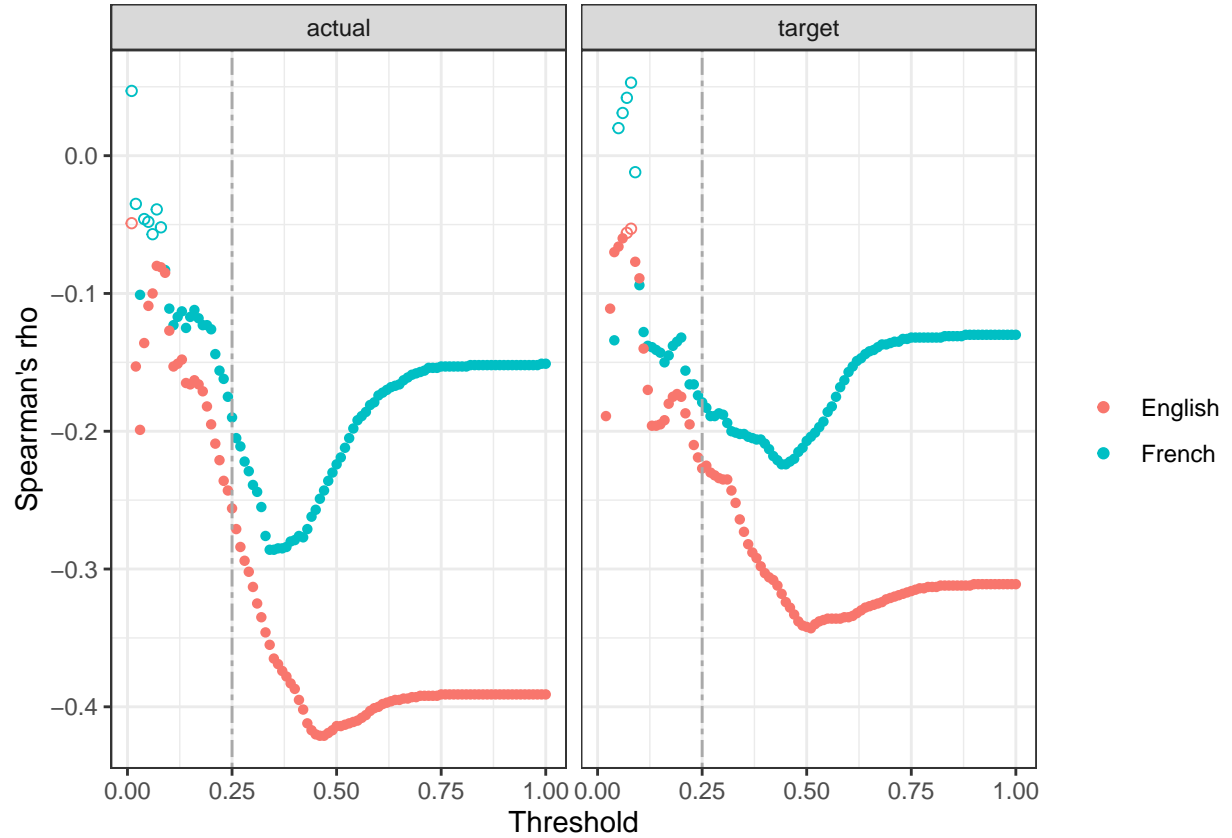
Figure 1. Correlation coefficient (Spearman's rho) between age of production (AoP) and Degree across the dataset, at connectivity thresholds ranging from 0.01-0.99. Filled circles represent rho values that were signficant at p<.05; empty circles represent non-significant correlations. Dashed line shows selected connectivity threshold of 0.25.

connected at a threshold of $<n$, where n is the threshold being tested) and both the AoP~Degree correlations and the full GLMER models were run for both Actual and Target data, as in the main paper. Results are shown in Tables 3 and 4.
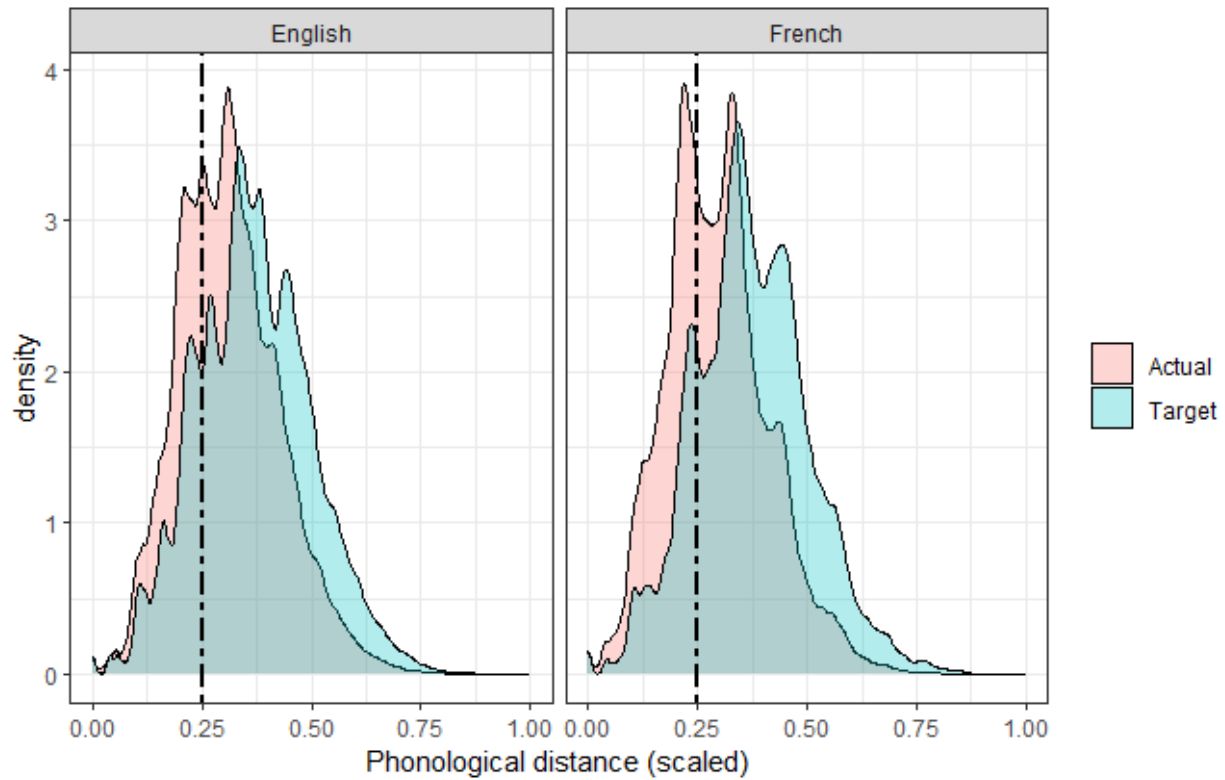
Figure 2. Density plot showing distribution of scaled phonological distance between word pairs in each infant's data. Plot is faceted by corpus; Actual data is shown in pink and Target data is shown in blue. The dashed line represents the connectivity threshold of 0.25, as applied in the analysis.

Table 3

Summary of key variables in the Actual dataset across seven different connectivity thresholds. At each threshold, the number of words that do not connect to any other word in the network is shown, alongisde Spearman's rho values for AoP Degree correlation tests and model coefficients (estimates, p values and 95% confidence intervals) for GLMER models testing the effect of INT and EXT on word learning. GLMERs were the same as those reported in the main paper, varying only by connectivity threshold.

			INT			EXT		
Threshold	Unconnected words	Spearman's rho	Estimate	p	95% CI	Estimate	p	95% CI
0.15	329	-0.15	0.44	<0.001	[0.33,0.55]	1.06	0.03	[0.12,2.01]

0.18	210	-0.15	0.42	<0.001	[0.31,0.53]	0.25	0.38	[-0.31,0.82]
0.19	173	-0.16	0.38	<0.001	[0.27,0.49]	0.07	0.78	[-0.44,0.58]
0.21	109	-0.19	0.42	<0.001	[0.31,0.53]	0.12	0.58	[-0.3,0.53]
0.24	59	-0.22	0.45	<0.001	[0.34,0.57]	0.03	0.85	[-0.27,0.33]
0.34	8	-0.33	0.42	<0.001	[0.28,0.55]	-0.07	0.46	[-0.25,0.11]
0.50	0	-0.33	0.45	<0.001	[0.29,0.62]	0.00	0.96	[-0.16,0.17]

Table 4

Summary of key variables in the Target dataset across seven different connectivity thresholds. At each threshold, the number of words that do not connect to any other word in the network is shown, alongisde Spearman's rho values for AoP Degree correlation tests and model coefficients (estimates, p values and 95% confidence intervals) for GLMER models testing the effect of INT and EXT on word learning. GLMERs were the same as those reported in the main paper, varying only by connectivity threshold.

|			INT			EXT		
Threshold	Unconnected words	Spearman's rho	Estimate	p	95% CI	Estimate	p	95% CI
0.15	388	-0.18	0.65	<0.001	[0.44,0.86]	1.97	<0.001	[0.88,3.05]
0.18	298	-0.17	0.60	<0.001	[0.39,0.8]	0.59	0.056	[-0.01,1.2]
0.19	269	-0.17	0.57	<0.001	[0.37,0.78]	0.48	0.089	[-0.07,1.02]
0.21	179	-0.19	0.64	<0.001	[0.44,0.85]	0.09	0.685	[-0.35,0.53]
0.24	74	-0.21	0.64	<0.001	[0.43,0.86]	-0.07	0.655	[-0.39,0.24]
0.34	7	-0.25	0.76	<0.001	[0.52,1]	-0.05	0.584	[-0.24,0.13]
0.50	0	-0.29	0.58	<0.001	[0.28,0.88]	0.06	0.478	[-0.11,0.23]

S3: Age of production (AoP) ~ connectivity.

Table 5

*Outputs (rho and p values) of AoP-degree Spearman's correlation tests for each infant in
the dataset.*

		Actual		Target	
Speaker	Corpus	Spearman's rho	p	Spearman's rho	p
Alex	English	-0.27	<0.001	-0.22	<0.001
Lily	English	-0.28	<0.001	-0.26	<0.001
Naima	English	-0.34	<0.001	-0.22	<0.001
Violet	English	-0.28	<0.001	-0.25	<0.001
William	English	-0.28	<0.001	-0.20	<0.001
Anais	French	0.02	0.769	-0.09	0.131
Marie	French	-0.22	<0.001	-0.22	<0.001
Nathan	French	-0.16	0.036	-0.04	0.580
Tim	French	-0.34	<0.001	-0.18	<0.001

Table 5 and Figure 3 show Spearman's correlation coefficients and corresponding p
values, as well as visualized data, for AoP~Degree correlations at a threshold of 0.25 for
each infant in the dataset.
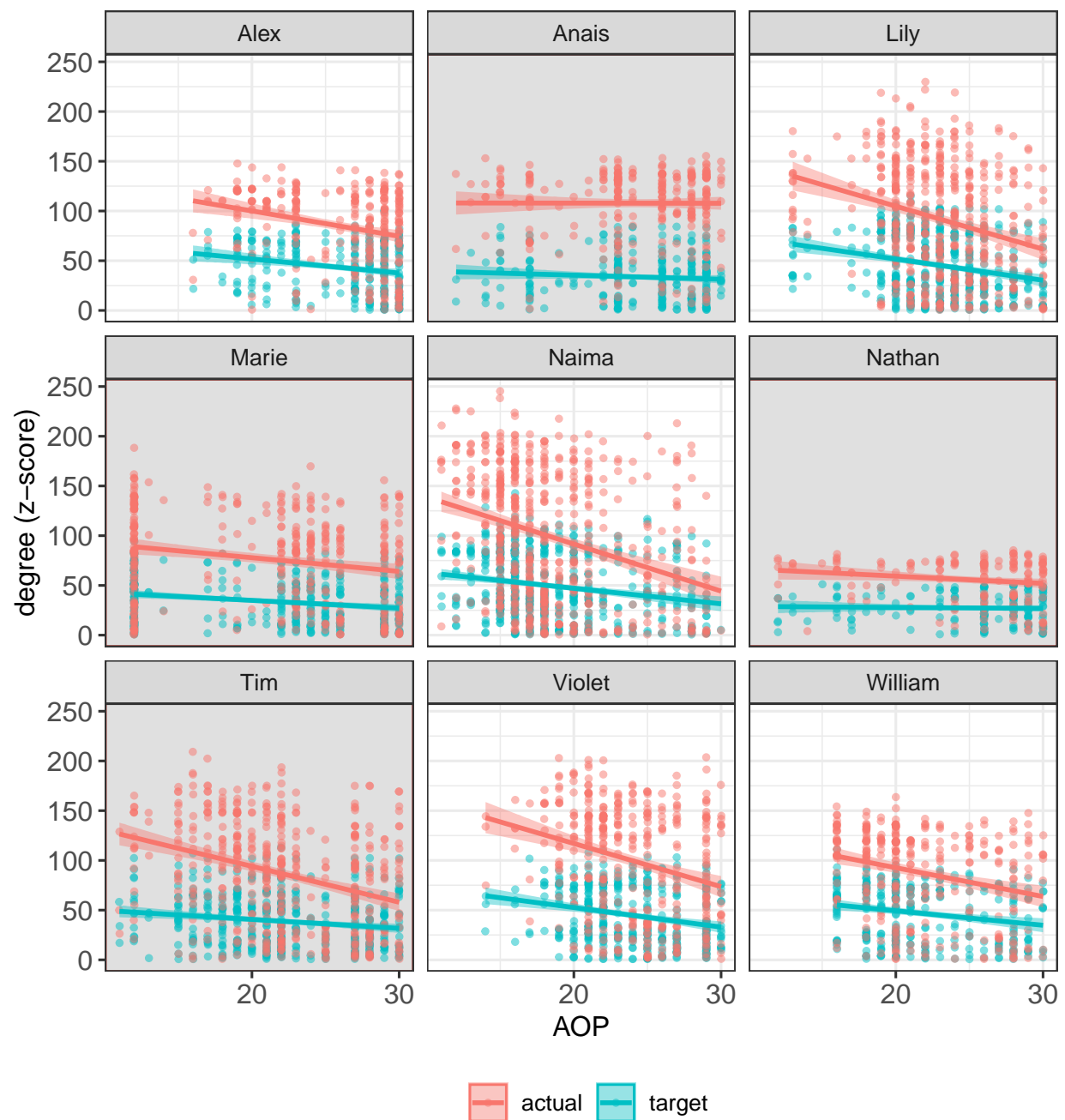
Figure 3. Age of production in relation to degree (z-score) of each word in the data. Individual points show individual word types produced in each month, with regression lines showing correlations for each infant. Data type (Actual vs. Target) is indexed by colour, with shaded colours showing 95% confidence intervals; shaded boxes show French data, un-shaded show English data.

S4: Network growth models: Full model outputs. Table 6 shows the full model outputs from the logistic mixed effects regression models run in the main paper, this time with word category included in the table.

Table 6

Results from maximal logistic regression model (model 3) testing the effects of network growth values, corpus (English as baseline), input frequency, comprehensive AoA, word category (nouns as baseline) and word length to predict word acquisition. All variables were scaled and centred. Word category was defined according to word categories on the McArthur Bates CDI (Fenson et al., 1994).

	Actual					Target				
Effect	beta	SE	z	p	95% CI	beta	SE	z	p	95% CI
Intercept	-3.35	0.27	-12.51	<0.001	[-3.87,-2.82]	-2.96	0.38	-7.79	<0.001	[-3.71,-2.22]
INT value	0.88	0.07	12.50	<0.001	[0.75,1.02]	0.47	0.06	8.08	<0.001	[0.35,0.58]
EXT value	0.02	0.05	0.46	0.648	[-0.07,0.12]	0.00	0.05	0.04	0.964	[-0.1,0.1]
Age	0.93	0.14	6.57	<0.001	[0.65,1.2]	1.25	0.13	9.77	<0.001	[1,1.5]
AoA	-0.22	0.04	-6.13	<0.001	[-0.29,-0.15]	-0.21	0.04	-5.79	<0.001	[-0.29,-0.14]
Length	-0.07	0.06	-1.29	0.198	[-0.19,0.04]	-0.14	0.06	-2.32	0.021	[-0.25,-0.02]
Input frequency	0.17	0.05	3.31	0.001	[0.07,0.27]	0.19	0.05	3.71	<0.001	[0.09,0.29]
Corpus	0.52	0.31	1.65	0.099	[-0.1,1.13]	0.93	0.47	1.99	0.047	[0.01,1.84]
Category: verbs	-0.53	0.08	-6.69	<0.001	[-0.69,-0.38]	-0.61	0.08	-7.59	<0.001	[-0.77,-0.45]
Category: adjectives	-0.15	0.14	-1.05	0.292	[-0.43,0.13]	-0.13	0.15	-0.89	0.373	[-0.42,0.16]
Category: games/routines	-0.16	0.16	-0.99	0.321	[-0.47,0.16]	-0.14	0.16	-0.83	0.405	[-0.46,0.18]
Category: prepositions	-0.06	0.42	-0.13	0.896	[-0.88,0.77]	-0.17	0.46	-0.37	0.710	[-1.08,0.73]
Category: pronouns	-0.29	0.40	-0.73	0.467	[-1.09,0.5]	-0.40	0.39	-1.01	0.310	[-1.17,0.37]
Category: quantifiers	0.87	0.68	1.28	0.200	[-0.46,2.21]	0.42	0.68	0.62	0.537	[-0.91,1.75]
Category: onomatopoeia	0.29	0.17	1.70	0.089	[-0.04,0.62]	0.17	0.17	1.01	0.313	[-0.16,0.5]
Category: time	-1.67	0.62	-2.72	0.007	[-2.88,-0.47]	-1.95	0.61	-3.20	0.001	[-3.14,-0.76]
Category: locations	0.34	0.26	1.33	0.182	[-0.16,0.85]	0.23	0.26	0.89	0.371	[-0.28,0.74]
Age x INT	0.12	0.05	2.27	0.023	[0.02,0.23]	-0.07	0.04	-1.62	0.106	[-0.16,0.01]
Age x EXT	-0.02	0.04	-0.42	0.673	[-0.1,0.07]	-0.03	0.05	-0.56	0.575	[-0.12,0.06]
Age x AoA	0.11	0.03	3.92	<0.001	[0.06,0.17]	0.10	0.03	3.28	0.001	[0.04,0.16]
Age x Length	0.04	0.05	0.76	0.449	[-0.06,0.13]	0.01	0.05	0.22	0.825	[-0.09,0.11]
Age x Input frequency	-0.06	0.03	-1.86	0.063	[-0.12,0]	-0.06	0.03	-1.92	0.055	[-0.13,0]

S5: GAMMs with INT/EXT values scaled by speaker and age. In the original paper, GAMMs testing differences in INT values by Age (H2) and Data type (H3) drawn on INT and EXT values scaled only by Speaker. This is inconsistent with the logistic regression models testing H1, which draw on INT/EXT values scaled by both Speaker and Age. Different scaling was used across models to improve visual clarity in the smooth plots for the GAMMs. However, results were consistent when scaling by Age was also applied, as shown in Table 7.

Table 7

Outputs from nested model comparisons of GAMMs testing the effect of age on INT and EXT values in Actual and Target data (Models 1 and 2), and the effect of Data type on INT values (Model 3). Model comparisons compared full models against those without parametric and smooth terms that included the variable being tested.

	Model	Actual			Target		
		Df	Chi Sq	p	Df	Chi Sq	p
1	INT:Age	14.000	17.886	0.001	14.000	33.594	<.001
2	EXT:Age	14.000	6.088	0.592	14.000	10.084	0.125
3	INT:Data type	7.000	496.560	<.001			

S6: By-speaker comparisons. Between-child differences are well established in the phonological development literature (Maekawa & Storkel, 2006; Vihman, Ferguson, & Elbert, 1986). There is thus reason to expect that the nine infants in the current dataset may have differential paths to word production, and so results may be variable when considered across speakers. Figure 4 below shows a by-infant breakdown of the INT value trajectories visualised in Figure 4 of the main paper. Descriptively, in Anais, Nathan and Alex's data, INT values increase exponentially over time, suggesting that INT is an increasingly influential factor in their early production. On the other hand, INT values plateau relatively early in the data for William, Naima and Lily, suggesting that other factors may have a stronger influence in these infants' word production and acquisition after a certain time-point.
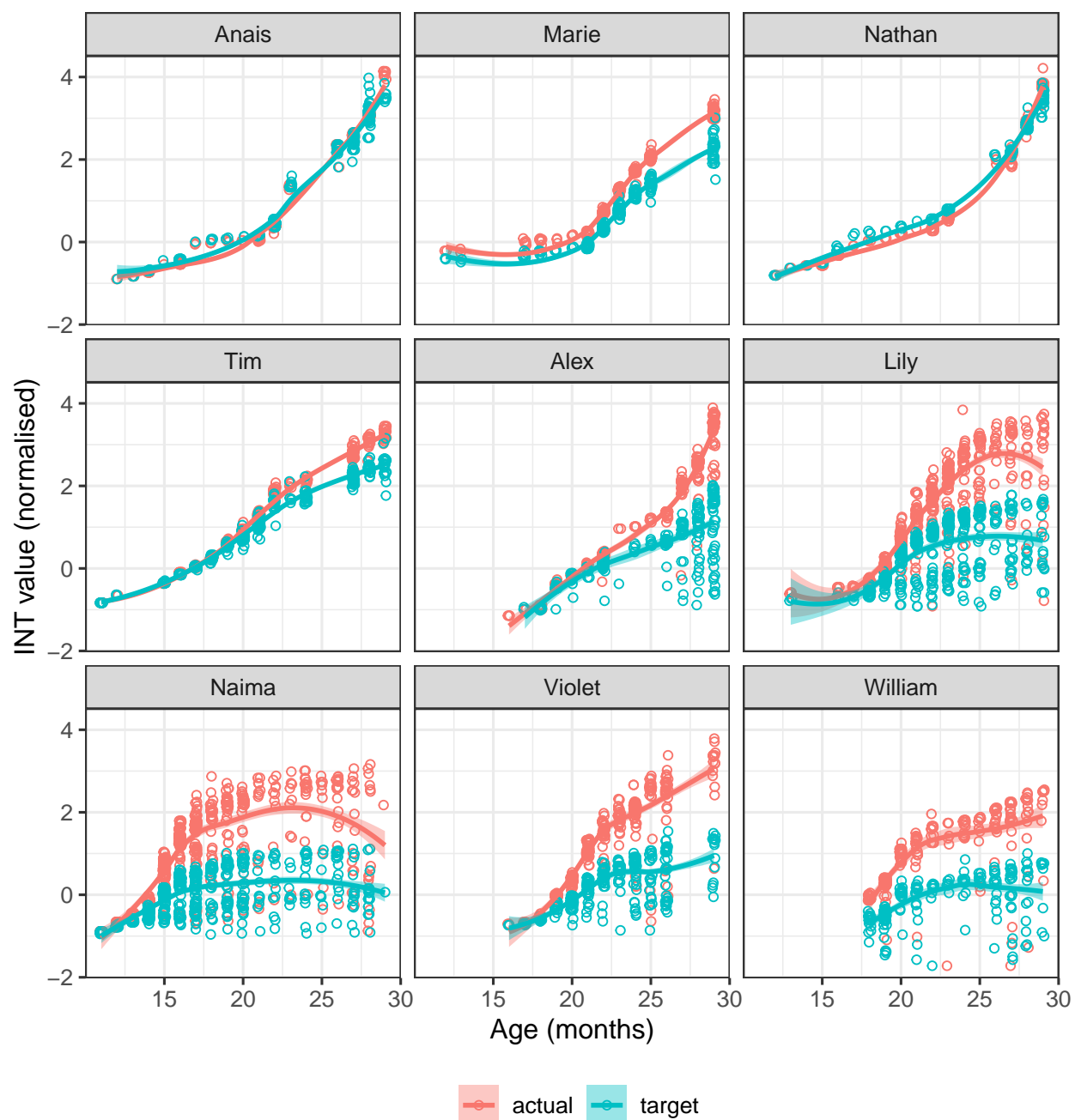
Figure 4. Smooth plot showing scaled INT values in Actual vs. Target forms. Shaded areas show 95% confidence intervals, lines indicate mean trajectories over time, coloured circles represent individual datapoints, jittered for visual clarity. Plot is faceted by infant for comparison across subjects.

Amatuni, A., & Bergelson, E. (2017). *Semantic networks generated from early linguistic input* [Preprint]. Animal Behavior; Cognition. https://doi.org/10.1101/157701

Maekawa, J., & Storkel, H. L. (2006). Individual differences in the influence of phonological characteristics on expressive vocabulary development by young children. *Journal of Child Language*, *33*(3), 439–459. https://doi.org/10.1017/S0305000906007458

Monaghan, P., Christiansen, M. H., Farmer, T. A., & Fitneva, S. A. (2010). Measures of phonological typicality. *The Mental Lexicon*, *5*(3), 281–299. https://doi.org/10.1075/ml.5.3.02mon

Vihman, M., Ferguson, C., & Elbert, M. (1986). Phonological development from babbling to speech: Common tendencies and individual differences. *Applied Psycholinguistics*, *7*, 3–40.