CLAY Report 7/1

The BabyLM project of CLAY lab over the summer of 2024 is attempting to utilize large language and vision models to search for learnable patterns in a subset of the data children receive when learning languages. One aspect of the project looks to Vong et al. 2024 to produce further analysis of the results of the work and identify if there are any key properties (physical, cognitive, or semantic) of certain lexical categories that make them harder or easier to learn the necessary word-meaning pairs.

Analysis

One analysis we attempted to quantify asymmetries in the data that might be able to allow children to make generalizations about word meanings. For example if words are highly probable to be uttered whenever the object they refer to is salient in some way (for example visible) then it might be possible to make a connection between the word and the cognitive object it refers to. In order to see if these asymmetries occur we did logistic regressions to predict whether a word occurred given relevant variables. Consider the following regressions, which both ran on the following dataset. The original saycams dataset had a row for each utterance. We filtered this by removing any utterances without an attested age or those spoken by the child or those in which the utterance is null. We then created a new dataset where each row in the original dataset corresponds to several rows in the new dataset (one row for each Konkle object). For each of these rows we recorded whether or not the word (or its plural) occurred in the utterance, whether it was salient to the utterance and what the child's age was. This created two binary columns and one column for the age. This allowed us to do the following regressions:

Regression I:

In one logistic regression we tried to predict whether a word is salient given that it is uttered and given the child’s age. We get the following results:

formula = Salient ~ Occurs * Child_Age
family = binomial

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	-5.312730	0.028460	-186.670	< 2e-16	***
Occurs	5.155733	0.094017	54.838	< 2e-16	***
Child_Age	0.013710	0.001725	7.946	1.93e-15	***
Occurs:Child_Age	-0.038126	0.006345	-6.009	1.86e-09	***

The significant positive coefficient for *occurs* indicates that (holding child age constant) if a word is uttered, it is significantly more likely for the corresponding object to be salient.

The significant positive coefficient for *child age* indicates that if a word is not uttered then the older a child is the more likely the object is to be salient. However if we take the significant negative interaction term into account, then this relationship actually reverses if a word *is* uttered. In other words if a word is uttered, then the older a child is the less likely a word is to be salient.

Regression II:

In another logistic regression we tried to predict whether a word is uttered given that it is salient and given the child's age. We get the following results:

formula = Occurs ~ Salient * Child_Age
family = binomial

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.221036	0.056972	-109.195	< 2e-16	***
Salient	5.208442	0.097228	53.569	< 2e-16	***
Child_Age	-0.026155	0.003759	-6.957	3.47e-12	***
Salient:Child_Age	-0.041889	0.006604	-6.343	2.25e-10	***

The significant positive coefficient for *salient* indicates that (holding child age constant) if an object is salient, it is significantly more likely for the corresponding word to be uttered.

The significant negative coefficient for *child age* indicates that if a word is not salient then the older a child is the less likely the word is to be uttered. And if we take the significant negative interaction term into account, then this relationship becomes even more negative.

The way to interpret the results of these regressions is that as children get older, parents become more likely to refer to objects that are not present (or otherwise salient).