## Literature Search Protocol

We conducted our literature search following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist [CITE].In Google Scholar, e conducted a keyword search with "Syntactic Bootstrapping" and a forward search on papers that cited the seminal paper, Naigles (1990) (retrieved between May, 2020 and July, 2020). Additional papers were identified by consulting the references section of recent literature review [CITE] and the experts in the field. We included works that are published journal articles, conference proceedings, doctorate dissertaions, and unpublished manuscripts. They will be collectively referred to as "papers" in the following sections. All papers include reports of experimental studies. Each paper may includes multiple conditions, and thus provide multiple effect sizes for the final analysis.

To be included in our final sample, the paper must satisfy the following criteria: First, the experimental paradigms must involve a two alterantive forced-choice situation, in which the participants were instructed to identify the scenes that match the linguistic stimuli. Second, the visual scenes are all displayed on a monitor, and the media can be either video reocording or animation clips. Third, the linguistic stimuli used in the experiments must include at least one novel verb embedded in a syntactically informative frame For example, "Look, Kradding!" does not provide relevant syntactic information to interpret the novel verb "kradding". In contrast, "Look, it's kradding!" embeds the novel verb in an intransitive syntactic frame. If the linguistic stimuli used in the experiments only include the former example, then the experiments would be excluded. But as long as the participants are exposed to one exemplar of the latter example, the experiments would be included. Finally, we decided to focus on English-speaking, typically-developing children. Experimental conditions that tested on other population were excluded from the sample.

Our final sample included data from `r different\_infants` unique infants, reported in `r n\_effect\_sizes` individual effect sizes from `r n\_papers` individual papers. Out of the `r n\_effect\_sizes` there were [NUMBER OF HIGH CONFIDENCE EFFECT SIZES] conditions labeled as "classical approach", i.e., the set-up and procedure is most similar to the seminal paper (Naigles, 1990). There were [NUMBER OF DEVIATING STUDIES] conditions labeled as "non-classical approach". We judged these experimental conditions to meet all of the inclusion criteria, but they deviated from the classical approach in some non-trivial aspects (LIST OUT THE DEVIATION). To preserve the homogeneity of the experiments without reducing the sample sizes, we conducted our statsitics analysis both on the entire sample and on the experimental conditions labeled as "classical approach".

???Should I explain the division??

???after cleaning only two adult effect sizes included, and one must be excluded because SD=0

??? numer of paper retrieved method?

??? coded some prepositional dative as transitive? how do I describe it in more details??

??? explain why we coded the things we coded??

## Coded Variables

For each experimental condition, we coded two types of variables: participants-related variables, the stimuli related variables, and the experimental procedure related variables The variables were either retrieved from the papers' methods sections or obtained via contacting the authors. Allinformation was either retrieved from the paper's methods sections and the figures illustrating the procedure, or obtained by contacting the authors.

Participants-related variables included infants' mean age by days and the mean productive mean productive vocabulary measured by MacArthur-Bates Communicative Development Inventories (CDI) Words and Sentences. Most papers reported mean age by months, so we converted it to mean age by days by multiplying the reported statistics with 30.44, the average number of days in a month. All conditions (N=??) have the age variable availble. [???] conditions have the vocabulary information available.

Stimuli related variables included features of both the linguistics stimuli and the visual stimuli. For linguistic stimuli during the training phase, we focused on the structure of the sentences infants heard. The sentences were categorized as transitive(N =???), if and only if when the novel verb in the sentence was preceded by a subject and followed by an object. Other sentences were categorized as intransitive (N=???). We also coded the types of words used in the agent argument and the patient argument. [N=??] experimental conditions [EXPLAIN?? the arument]. For linguistic sstimuli during the testing phase, we coded whether the infants were prompted to identify the action ("an example", N=) or the actors ("an example, N=). For visual stimuli, we coded both the types of media used (video or animation) and the types of protagonists in the events (person or non-person). [N=??] conditions used video recordings of human actors (N=??) or human actors in animal suits (N=??). The other [N=??] conditions used clips of animation with ??? as the protagonists of the events. We also coded how the onset of the linguistic stimuli aligned with the visual stimuli. The procedure was coded as "simultaenous" if the very first training sentence was presented along with the visual stimuli depicting relevant action (N=??). It was coded as "immediate after" if the first training sentence was presented along with an attention-getter or a blank screen, immediately followed by the relevant action (N=??). Finally, some experimental conditions first presented the linguistic stimuli paired with irrelevant visual scenes (for example, a person on the phone talking). The relevant visual stimuli were not shown until the training phase is over. For experimental conditions using this procedure, they were coded as "asynchronous"(N=??).

Finally, experimental procedure related variables included the type of response elicited from the participants: whether the infants were explicitly prompted to point or their eye gaze duration were measured as they heard the linguistic stimuli. ( Pointing: N = [??]; Looking: N=[???]). Three characteristics of the experimental procedures were coded as categorical variables: the inclusion of practice phase (Yes: N= ; No: N=), the inclusion of character identification phase (Yes: N=: No:), and the distribution of the training and the testing trials. A procedure was categorized as "mass" if and only if the infants were trained exclusively on one novel verb and tested on the very same verb (N=??). It was "distributed" if the infants were given multiple train and test pairs on multiple novel verbs (N=??) To better characterize the experience infants had prior to testing, we also coded how many train-test pair the infants were given (for mass procedure it was always 1), how many trials during the test phase infants were given, how many times the visual stimuli showing the relevant actions were presented, and how many timeseach novel verb was spoken in a syntactically-informative way. If the prompt question provided relevant syntactic context (EXAMPLE), then it was also counted.