Developmental Science

Developmental Science 19:6 (2016), pp 901–917 DOI: 10.1111/desc.12341

PAPER

Development of infants' segmentation of words from native speech: a meta-analytic approach

Christina Bergmann and Alejandrina Cristia

Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, Paris, France

Abstract

Infants start learning words, the building blocks of language, at least by 6 months. To do so, they must be able to extract the phonological form of words from running speech. A rich literature has investigated this process, termed word segmentation. We addressed the fundamental question of how infants of different ages segment words from their native language using a meta-analytic approach. Based on previous popular theoretical and experimental work, we expected infants to display familiarity preferences early on, with a switch to novelty preferences as infants become more proficient at processing and segmenting native speech. We also considered the possibility that this switch may occur at different points in time as a function of infants' native language and took into account the impact of various task- and stimulus-related factors that might affect difficulty. The combined results from 168 experiments reporting on data gathered from 3774 infants revealed a persistent familiarity preference across all ages. There was no significant effect of additional factors, including native language and experiment design. Further analyses revealed no sign of selective data collection or reporting. We conclude that models of infant information processing that are frequently cited in this domain may not, in fact, apply in the case of segmenting words from native speech.

Research highlights

- We present a meta-analysis of infants' segmentation of words from fluent speech in their native language.
- There is a significant preference for familiar over novel test stimuli.
- This overall familiarity preference is not modulated by age, even when controlling for methodological factors and language background.
- The absence of a preference for novel stimuli invites a revision of popular theories linking age to behavioral preferences.

Introduction

Words can be viewed as the building blocks of language. The process of finding and memorizing these foundational units starts very early during language acquisition: Even 6-month-olds can relate wordforms, the acoustic realization of a word, to their visually presented referents (e.g. Tincoff & Jusczyk, 2012; Bergelson & Swingley, 2012). This is remarkable, since less than 10% of word in infants' input occur in isolation (e.g. Brent & Siskind, 2001; van de Weijer, 1998). The ability to extract wordforms from running speech, a process commonly referred to as *word segmentation*, is therefore thought to play a major role in infant word learning. In this paper, we quantitatively integrate 20 years of laboratory evidence on word segmentation from natural, native speech in order to shed light on the role of a key variable of high theoretical and practical relevance: infants' age, an (admittedly imprecise) proxy of maturation and native language experience often invoked in the infant speech perception literature. Before introducing our predictions and approach, we briefly summarize the literature.

Address for correspondence: Christina Bergmann, Laboratoire de Sciences Cognitives et Psycholinguistique, Pavillon Jardin, 29, rue d'Ulm, 75005 Paris, France; e-mail: chbergma@gmail.com

A brief overview of research on infant word segmentation

The first laboratory evidence that infants can recognize words across isolated and sentential presentations in their native language came from Jusczyk and Aslin (1995). In three (of four) experiments, infants heard two words spoken in isolation until they accrued a previously set amount of experience (Jusczyk & Aslin used 30 seconds per word). Subsequently, participants heard sentences containing the target word and others that contained foils, words that were familiar to other infants in counterbalanced conditions. Differences in listening times, and thus an emergent preference for one of the two types of test trials, indicated successful segmentation. In a fourth experiment infants were first familiarized with sentences and then tested with isolated words. The reverse order yielded comparable results to what has now become the standard experimental sequence. In addition, 6-month-old American English learners were found to fail in the words-then-passages type of experiment, whereas at 7.5 month of age infants succeeded, suggesting a time frame for the emergence of segmentation skills.

Subsequent research has modulated and extended the early conclusions of Jusczyk and Aslin (1995), while continuing to acknowledge the key impact of infant age. For example, some have argued that younger infants are not completely unable to match words in isolation and in sentences, but they may require that the task be simplified. Indeed, 6-month-olds succeed in similar segmentation in their native language if the target word is placed next to the infant's own name or the highly familiar word 'Mommy' (Bortfeld, Morgan, Golinkoff & Rathburn, 2005). Age, as a proxy for cognitive maturation, is further thought to play a role in the format of representation. Jusczyk and Aslin (1995) had reported that the representation of familiarized words was detailed enough that infants would not confuse the familiar word with a one-segment mispronunciation, such as 'tup' versus 'cup'. Subsequent research using the same paradigm has gone on to suggest that young infants' representation is *too* detailed, as infants fail if the familiarization and test stimuli differ along indexical dimensions, such as speaker gender (Houston & Jusczyk, 2000, 2003; but see van Heugten & Johnson, 2012) or affect (Singh, Morgan & White, 2004). As infants grow older, their segmentation abilities become more robust to indexical variation, again suggesting that they become better at segmenting their native language.

Cross-linguistic differences

The studies briefly summarized above all tested infants acquiring American English and to this day the majority of segmentation studies have been carried out on this population. Studies on other languages are emerging, with diverse and at times even surprising results. Dutch infants, for example, succeed in the same paradigm later than their American English-speaking peers, at the age of 9 months (Houston, Jusczyk, Kuijpers, Coolen & Cutler, 2000). Even greater delays were reported for French infants, who do not succeed in the exact same segmentation task as Jusczyk and Aslin (1995) until after their first birthday (e.g. Nazzi, Mersad, Sundara, Iakimova & Polka, 2014). Catalan- and Spanish-speaking infants, in contrast, already segment words from naturally spoken sentences at 6 months (Bosch, Figueras, Teixidó & Ramon-Casas, 2013). What could cause such differences in pace? Nazzi and colleagues have argued that, to a certain extent, languages may differ in their reliance on those cues that are more accessible to a very young learner. Indeed, there is a sizable literature on cross-linguistic differences in adult use of cues for word segmentation (to give just one example, Tyler & Cutler, 2009, studied the differential use of suprasegmental cues by French, English, and Dutch native listeners). Further, infants are still acquiring the appropriate cue-weighting scheme for their native language (Jusczyk, Houston & Newsome, 1999). Even if there is a slightly different pace of development across languages, it remains a common assumption in this line of research that infants, regardless of their linguistic background, are becoming more skilled at segmenting words in their mother tongue as they get older.

Using a meta-analysis to measure the effect of age on word segmentation

Our brief review of the segmentation literature paints a mostly coherent picture, but it is based on a small selection of studies. The full picture is in fact much more complex and difficult to attain using a qualitative approach. Therefore, we turned to meta-analytic tools for an unbiased integration of previous findings. Why focus on infant age? In the word segmentation literature, infant age is a proxy thought to reflect changes in both overall cognitive maturation and linguistic proficiency. Maturation could play a major role because it correlates with changes in working memory and attention (Ruff & Rothbart, 2001), which in turn have an impact on many tasks including linguistic processing (e.g. Gathercole & Baddeley, 1993). Budding language-specific knowledge at all levels (such as the attunement to native sound categories, the acquisition of language-specific segmentation strategies, and an increased lexicon size) will influence word segmentation from native speech. For example, infants can use well-known words to extract

other words (Bortfeld *et al*., 2005). Determining the precise contribution of each of these factors requires further experiments. Nevertheless, and regardless of which specific mechanism explains a correlation between age and performance, *the expectation in all cases is that infants become more skilled at segmenting words in their native language as they age.*

How might improved performance be evident in a lab setting? Researchers working on word segmentation studies (e.g. Babineau & Shi, 2011; Bosch *et al*., 2013; Seidl & Johnson, 2008; Singh *et al*., 2004) have thought that marked increases in performance would be reflected in switches from a familiarity preference, longer listening times to familiar stimuli, to a novelty preference, as proposed by Hunter and Ames (1988) among others. Although this is not the only or even the first model to interpret infant behavior in a lab setting, it is nonetheless the most frequently cited in this literature.[1] We therefore centered our attention on it when drawing predictions for our meta-analysis. We will evaluate the accuracy and suitability of this model, related ones, and competing accounts in the Discussion.

As just mentioned, Hunter and Ames (1988) is the central citation in the word segmentation literature when improvement with age is discussed. We therefore lay out the specifics of the original model, which attempted to predict familiarity, novelty, and null preferences. The basic idea is that infants exposed to a given stimulus will attempt to encode it, and will continue to explore it (for example, by attending to it) until they have completely encoded it. Thus, if given a choice between two initially neutral stimuli, one that has been partially encoded and a second that is completely novel, the child will prefer the more familiar one. Once they have fully processed the familiarized stimulus, infants are ready to begin encoding novel information and switch to a novelty preference. The function describing the direction of preferences can be approximated by a polynomial of at least degree 3 (cubic function) that is not defined for negative familiarization times, with a null preference at the onset of familiarization and a peak in familiarity preferences before novelty preferences become the stable outcome at some point, as exemplified in Figure 1.

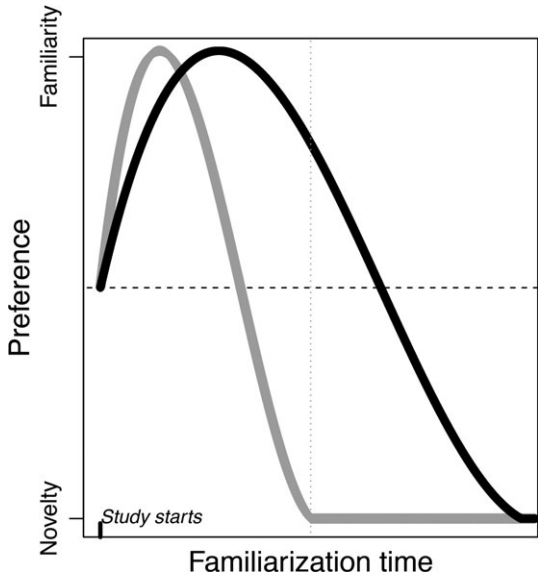

Figure 1 *Infants' preference (along the ordinate) as a function of familiarization or exposure time (along the abscissa) according to the Hunter and Ames (1988) model. The dashed horizontal line represents no preference (chance looking), and familiarity is plotted up. The two curves represent predicted age differences, with behavioral preference for older infants in gray and that for younger ones in black. Whereas the overall shape is the same, changes in behavior are slower in younger infants. Thus, at a fixed familiarization time, there must be a pair of ages at which younger infants will exhibit a significant familiarity preference, whereas older infants will show a significant novelty preference. In this representation, this predicted event is indicated by the vertical dotted line.*

Hunter and Ames (1988) explore two predictions related to experimental design. First, all else being equal, longer familiarization times should eventually lead to a novelty preference. Second, and again when all other aspects are held constant, decreasing task complexity should push the switch earlier in the timecourse of a given experiment. Hunter and Ames (specifically pp. 76f.) also discuss the effects of age: Older infants should become more efficient due both to an expansion in their cognitive capacities with maturation and an accumulation of experience with complex and diverse input in their daily lives, if the experiment uses comparable stimuli. This results in a compressed curve, such that testing two age groups with the exact same methodology (i.e. keeping the length of familiarization and task complexity the same) should yield a developmental switch from familiarity to novelty preference – provided the appropriate ages are chosen. Figure 1 exemplifies this shift with a black curve for younger infants and a grey curve for older infants.

[1] Comparing citations both in our database (51 papers, of which seven link direction of preference to models of stimulus preferences) and on scholar.google.com (30 April 2015): Hunter and Ames (1988) have six (database) and 349 (scholar) citations; Rose, Gottfried, Melloy-Carminar and Bridger (1982) two and 204; Roder, Bushnell and Sasseville (2000): one and 144; and Wagner and Sakovits (1986): one and 83. We further note that the authors citing Rose *et al*. and Wagner and Sakovits also invoke Hunter and Ames.

The data from which Hunter and Ames (1988) originally drew support (see pp. 83–86, and citations therein) were varied in terms of infant age (from 6 weeks to 12 months), in design (ranging from single sessions at the lab with exposures as short as 15 seconds, to daily 30-minute presentations over 8 weeks), and in stimuli and procedure (e.g. presenting pairs of photographs and measuring visual fixations; presenting arrays of toys and measuring focused manipulation). Although none of these studies included linguistic stimuli, we (and other researchers interested in word segmentation – see footnote 1) have interpreted this model as portraying general information processing characteristics in infants, and therefore potentially applicable to word segmentation tasks. In fact, models of infant preferences have been claimed to be modality-independent (see e.g. Wagner & Sakovits, 1986).

Other models of infant preferences in a laboratory setting make the same predictions as Hunter and Ames (1988) on the effect of age: A novelty preference must emerge at some point when all other factors are held constant. For example, Roder *et al*. (2000) disagreed with Hunter and Ames about the necessity of observing a familiarity preference phase, which they argue might be absent in some cases (see also Houston-Price & Nakai, 2004). Sirois and Mareschal (2002) place Hunter and Ames' model within the larger context of habituation/dishabituation phenomena, but, again, without contesting the original claim that older infants will prefer novel stimuli.

In sum, the predictions of the model by Hunter and Ames (1988) as well as those drawn from related accounts align in terms of how age relates to laboratory performance: If infants become more skilled at segmenting their native language with age, then, all else being equal, younger infants will show familiarity preferences whereas older infants should exhibit novelty preferences. We will assess this prediction through our meta-analysis, taking into account three experimental factors that modulate task difficulty, as well as infants' linguistic background.

The first experimental factor we considered is familiarization time, a key factor according to Hunter and Ames (1988). Second, Nazzi and colleagues (2014) observed that Parisian infants display segmentation abilities earlier when familiarized with passages instead of isolated words (see also van Heugten & Johnson, 2012). Third, characteristics of the actual stimuli might influence segmentation difficulty. To quantify task difficulty, we followed suggestions of existing reports that infants either rely on certain factors or are distracted by their presence. Consequently, we encoded the following factors: sentence edge alignment of the target word (Seidl & Johnson, 2006); match with the predominant stress pattern of the native language in multisyllabic words (e.g. Jusczyk *et al*., 1999); and changes in speaker identity (Houston & Jusczyk, 2000, 2003) or affect (Singh *et al*., 2004) between familiarization and test phase. The fourth factor we incorporated into our analyses was infants' native language, following some reports in the literature suggesting relative cross-linguistic advantages or delays.

Before proceeding, it is important to mention that a neighboring body of literature studies speech segmentation using miniature artificial languages (e.g. Saffran, Aslin & Newport, 1996). We do not include it for a number of reasons (see Supplementary Materials for a detailed account), the most important one being that the predictions of performance as a function of age need not be the same for an artificial language: In broad terms, there is no reason why further exposure to the infants' native language (as it occurs naturally with age) should lead to improved performance when segmenting artificial syllable streams. Thus, it is appropriate, and possibly even necessary, to analyze only studies using infants' native language as spoken naturally, and to leave work using artificial syllable streams to future meta-analysts. Thus, when we speak of *word segmentation* in this article, we refer specifically to the process involved in recognizing words spoken naturally, be it in isolation or embedded in a sentence frame, in the infant's native language; in short, studies that are conceptual and in many cases also methodological replications of Jusczyk and Aslin (1995).

Summary of goals and scientific approach

The present paper has three main goals. First, we use a meta-analytic approach to systematically summarize the current evidence on infant word segmentation abilities and test a key prediction of the most frequently cited model linking infant abilities to behavior by harnessing the power of 20 years of research. To anticipate our results, we will find that our data do not fit the predictions made by Hunter and Ames (1988) and related models regarding how age will affect the direction of infant preferences. Therefore, our findings invite a revision of popular theories of infant preferences. In addition, and contrary to recent reports, we find that neither methodological factors nor language background explain a significant amount of variance in effect sizes.

Our second goal is to promote discussions on scientific practices in infant speech perception research. To this end, we explain in this Introduction why our meta-analytic approach is worthwhile, as meta-analyses are relatively rare in infant speech perception research (but see e.g. Galle & McMurray, 2014; Tsuji & Cristia, 2013, for two recent examples). Aggregating experimental data

in one analysis allows for an important step in scientific progress: estimating the size of an effect and its variance. Whereas an experiment demonstrates that a phenomenon can be observed in a specific situation, an effect size drawn through meta-analytic methods speaks to the robustness of the effect, and this estimation is more reliable (the variance can be reduced) by grouping together comparable experiments. Meta-analyses rely on a crucial assumption, namely that the phenomenon under study is replicable. If we believe in the scientific method, then the data issuing from similar studies aiming to tap into the same underlying process, particularly when very similar methodology is used, should be comparable, even when collected at different points in time, by different people, and in order to study specific follow-up research questions. In other words, most experimental studies building on Jusczyk and Aslin (1995) are both conceptual and methodological replications of that initial report, since they purport to study the same process and follow essentially the same protocol as Jusczyk and Aslin's seminal experiments.

Meta-analyses can go beyond examining the size and variance of an effect by assessing the impact of factors of interest, including methodological variants (e.g. using a words-to-sentences or sentences-to-words order), and participant descriptors (infant age and native language). An experiment can manipulate these factors and provide one single observation of their effect on the phenomenon of interest. But no single experiment is a window on true and unbiased reality. By integrating results across multiple experiments quantitatively, meta-analyses can provide a more accurate measure of *how much* a factor modulates the effect, and *how accurate* our estimation is.

In addition to serving these two immediate goals (estimating the size of an effect, and its modulation by moderator variables), meta-analyses can be valuable for *future* research. For example, meta-analyses can guide experimentalists who want to determine the necessary number of participants in prospective power analyses, or who want to ensure methodological comparability with previous research. This leads us to the third and final goal of the present paper. In the process of answering our theoretical research question, we put together a database. We have followed recent recommendations by rendering our data publicly accessible and open to updates, i.e. we have built a Community-Augmented Meta-Analysis (for a summary of the benefits for a research community, see Tsuji, Bergmann & Cristia, 2014; see also Mills-Smith, Spangler, Panneton & Fritz, 2015, for a discussion of the importance, in the field of infant developmental science, to incorporate effect size into our interpretations). Here, we limit ourselves to a report on the meta-analysis; all relevant information on the database and extensive supplementary materials can be found on the companion website (http://inworddb.acristia.org).

Methods

We have followed a standard meta-analytic protocol: We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement to organize our research project and ensure that all key aspects of our meta-analysis were reported on, including selection criteria (see next section) and assessment of data quality by checking for possible publication biases (Moher, Liberati, Tetzlaff, Altman & PRISMA Group, 2009). Further, we consulted the reference work by Lipsey and Wilson (2001) for specifics on effect size calculations, and employed the R (R Core Team, 2013) packages meta (Schwarzer, 2007) and metafor (Viechtbauer, 2010) to conduct our analyses. Below, we give a streamlined overview of the specific steps taken, which are illustrated in Figure 2. For an in-depth introduction to systematic reviews and meta-analyses, refer for example to Durlak (2009), Lakens (2013), and Lipsey and Wilson (2001).

Study selection

We first generated a list of potentially relevant items to be included in our meta-analysis using the google scholar search engine, with the broad search term 'infant word segmentation' (following Gehanno, Rollin & Darmoni, 2013). This search was carried out on 27 November 2012 and we inspected the first 1000 results. Fifteen additional items were included based on recommendations and by scanning references of included papers. After removing duplicates, we screened the title and abstract of each remaining item and identified 231 items for full-text inspection using the following inclusion criteria: (1) original data were reported; (2) the stimulus material was continuous natural speech spoken in the participants' native language; (3) the dependent measure was looking time (LT) at a neutral visual target (i.e. not a possible referent of one set of stimuli); (4) infants were normally developing. The final sample consisted of 38 journal articles, seven proceedings papers, one book chapter, one thesis, and one unpublished report (available from an institutional website). We will refer to these 51 items collectively as *papers*. Table 1 provides an overview of all papers.

Data entry

In the next step, we entered the 51 papers into our database, which can be conceptualized as a large table.

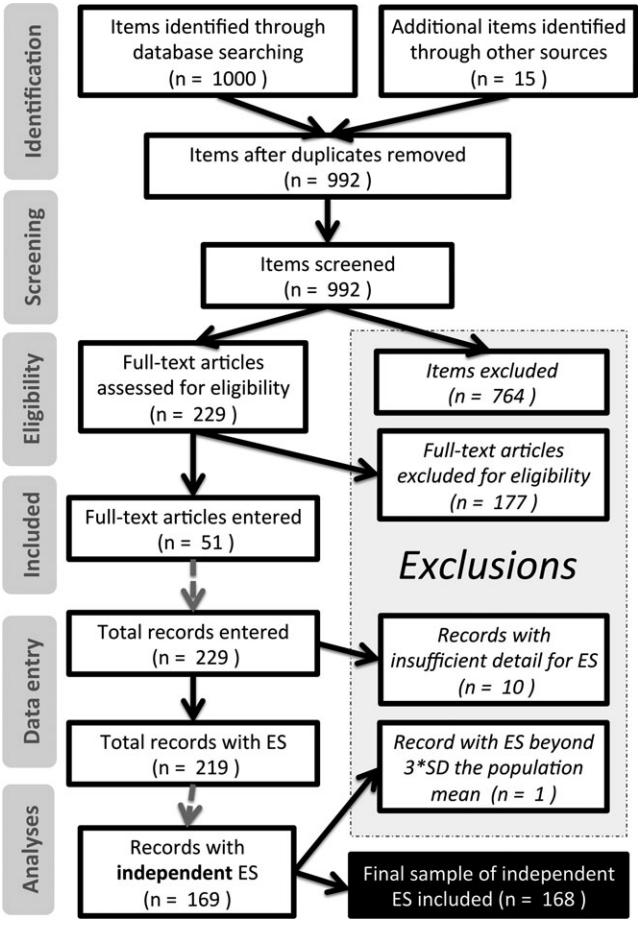

Figure 2 *Flowchart indicating data exclusion at each stage, adapted from the PRISMA flowchart. Dashed lines indicate where the number of data points changes in ways that cannot be captured in the flowchart, because multiple data points are reported in a single paper, and some of them are not mutually independent. See main text for details.*

Each column refers to a unique characteristic in a given experiment, such as mean participant age. We will refer to the columns as *fields*. An exhaustive list of coded fields is available in the Supplementary Materials and on the companion website (http://inworddb.acristia.org). For the present analyses, the relevant fields are:

1 Mean age reported per group of infants, in days;
2 Native language, including the native accent (e.g. Canadian vs. Parisian French);
3 Direction of test was either words-to-passages or passages-to-words;
4 Familiarization criterion, the number of seconds per target word;
5 Method was either headturn preference procedure or other (including central fixation);
6 A difficulty score was calculated as the sum of: linguistic difficulty (0 if the phonological form of the target word was identical across familiarization and test, 1 otherwise – e.g. *ham-hamlet* receives a score of 1); sentence alignment (0 if the target word was always aligned with a sentence edge, 1 otherwise); stress alignment (words that did not follow the predominant stress pattern for content words in the relevant language – English, German, and Dutch are largely stress-initial, Turkish stress-final – were given a score of 1, otherwise the score was 0); and indexical properties (0 if they were purposefully matched; 2 if they were purposefully changed between familiarization and test; and 1 if they were not controlled).

Each row in the database is a set of data that could give rise to an effect size. A given paper could appear in more than one row. This happened when there were multiple experiments, when there were multiple age groups within a single experiment, and when a unique group of infants was exposed to more than one condition (e.g. happy versus neutral affect; Singh, 2008) – provided that results were broken down by experiment/infant group/condition in the source paper. In other words, we coded all outcomes that had been reported separately, and which had unique infant participants and/or method characteristics. For example, if the same infants were tested on two conditions and these were reported separately, then we entered two different rows (even though there was only one group of infants), with their methodological features and unique outcomes. In such cases, we also documented that it was a repeated measure. We will call each row a *record*. In all, there were 229 records.

Effect size calculation

For each record, we attempted to code the information necessary to estimate an effect size, which in turn can be directly related to familiarity/novelty preferences. When information was missing, we contacted the authors, and many were able to provide us with further details. We did not have sufficient information for 10 records, leaving 219 records for which effect sizes could be calculated. The process and formulas we used are represented in Figure 3. All effect sizes were based on infants' looking times (LT) to a neutral visual target in response to the two types of test stimuli: familiar (stimuli including the target word presented during familiarization) and novel (stimuli not presented in familiarization). Mean LT for each test type (novel and familiar) and their respective standard deviations were available for 178 records. The effect size Cohen's *d* was calculated as the difference in

Table 1 *Overview of the included studies (identified by author names, and publication year), along with the number of independent records contributed to the meta-analysis (numbers in brackets indicate the records for which no effect size could be calculated), the age groups in months, native language of the participants, and whether children heard words during the familiarization phase and passages in the test phase*

Authors	Year	Number of independent records	Infant age groups (in months)	Native language	Words during familiarization, Passages during test
Altvater-Mackensen & Mani	2013	3	7	German	yes
Babineau & Shi	2011	3	20, 24	Canadian French	no
Barker & Newman	2004	2	7.5	American English	yes
Bartels, Darcy, & Höhle	2009	2	8.5	German	yes
Bortfeld, Morgan, Golinkoff, & Rathburn	2005	3	6	American English	no
Bosch, Figueras, Teixidó, & Ramon-Casas	2013	6	6, 8	Spanish and Catalan	no
Gonzalez-Gomez & Nazzi	2013	4	10.5, 13.5	Parisian French	no
Gout, Christophe, & Morgan	2004	2	10, 13	American English	yes
Höhle & Weissenborn	2003	2	6, 8	German	yes
Hollich, Newman, & Jusczyk	2005	4	7.5	American English	no
Houston & Jusczyk	2000	4	7.5, 10.15	American English	yes
Houston & Jusczyk	2003	4	7.5, 10.15	American English	yes
Houston, Jusczyk, Kuijpers, Coolen, & Cutler	2000	1	9	Dutch	yes
Houston, Santelmann, & Jusczyk	2004	6	7.5	American English	yes
Johnson	2005	2	10.5	American English	no
Johnson	2008	1	12	American English	no
Johnson, Jusczyk, Cutler, & Norris	2003	4	12	American English	yes
Jusczyk & Aslin	1995	4	6, 7.5	American English	yes (3), no (1)
Jusczyk, Houston, & Newsome	1999	15	7.5, 9, 10.5	American English	yes (8), no (7)
Katz-Gershon	2007	2	8	American English	no
Kuijpers, Coolen, Houston, & Cutler	1998	1	7.5	Dutch	yes
Marquis & Shi	2008	2	8, 11	Canadian French	yes
Marquis & Shi	2009	2	11	Canadian French	yes
Marquis & Shi	2012	3	11	Canadian French	yes
Mason-Apps, Stojanovik, & Houston-Price	2011	1	10, 19	British English	yes
Mattys & Jusczyk	2001a	8	8.5, 10.5, 13, 16	American English	yes
Nazzi, Dilley, Jusczyk, Shattuck-Hufnagel, & Jusczyk	2005	4 (2)	10.5, (13.5,) 16.5	American English	yes
Nazzi, Iakimova, Bertoncini, Frédonie, & Alcantara	2006	8	8, 12, 16	Parisian French	yes
Nazzi, Mersad, Sundara, Iakimova, & Polka	2014	5	8, 12, 16	Parisian French	no
Newman & Jusczyk	1996	4	7.5	American English	yes (3), no (1)
Polka & Sundara	2011	5	8	Canadian French	yes
Schmale & Seidl	2009	1	9, 13	American English	yes
Seidl & Johnson	2006	2	8	American English	no
Seidl & Johnson	2008	3	11	American English	no
Shi	2007	1	8	Canadian French	yes
Shi, Cutler, Werker, & Cruickshank	2006	4	8, 11	American English	Not applicable: words and word-groups
Shi & Lepage	2008	2	8	Canadian French	Not applicable: words and word-groups
Shi, Marquis, & Gauthier	2006	4	6, 8	Canadian French	Not applicable: words and word-groups
Singh	2008	8	7.5	American English	yes
Singh & Foong	2012	3	7.5, 9, 11	Mandarin and English	yes
Singh, Morgan, & White	2004	6	7.5, 10.5	American English	yes
Singh, Nestor, & Bortfeld	2008	4	7.5, 10.5	American English	yes
Singh, Nestor, Parikh, & Yull	2009	2	7.5	American English	yes
Singh, Reznik, & Xuehua	2012	1	7.5	American English	yes
Singh, White, & Morgan	2008	4	7.5, 9	American English	yes
Tsay & Jusczyk	2003	1	7.5	Mandarin	yes
van Heugten & Johnson	2012	2	7.5	Canadian English	no
Willits, Seidenberg, & Saffran	2009	3	7.5, 9.5	American English	no
van Kampen, Parmaksiz, van de Vijver, & Höhle	2007	0 (1)	9	Turkish	Not applicable: words and word-groups
Jusczyk, Hohne, & Bauman	1999	0 (4)	9, 10.5	American English	yes
Mattys & Jusczyk	2001b	0 (3)	9	American English	no

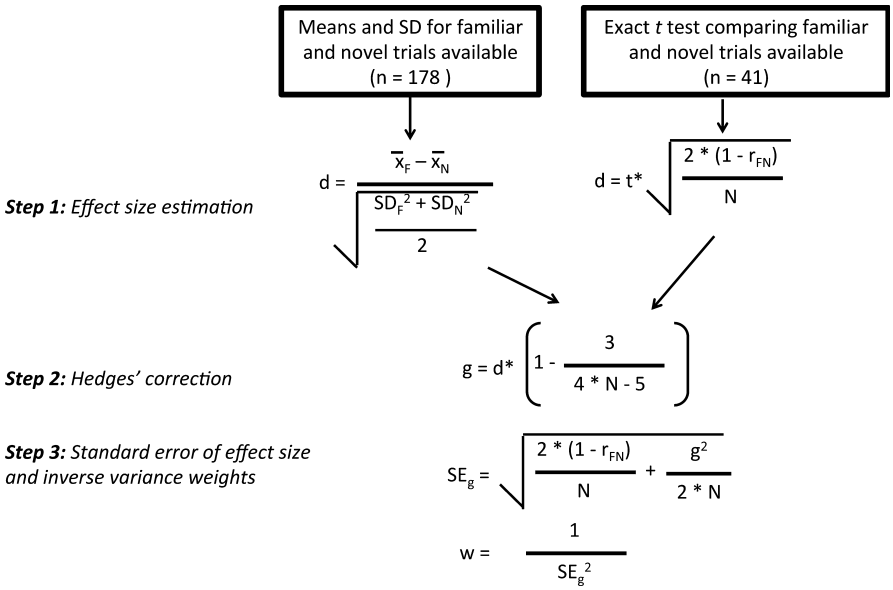

Figure 3 *Process and formulae through which effect sizes, standard error, and weights were calculated. Here, x_F means LT to familiar test trials, x_N is LT to novel test trials, $SD_{F/N}$ are the respective standard deviations, N signifies number of participants in a given experiment, t is the exact t value, r_{FN} stands for Pearson correlation coefficient of familiar and novel test trials, d is Cohen's d and g is Hedges' g, SE means standard error and w is the inverse variance weight.*

looking times to familiar minus novel trials divided by the pooled standard deviation (Lipsey & Wilson, 2001, p. 44; Figure 3, Step 1). In 41 additional records, the exact *t*-value from a paired *t*-test comparing LT was available whereas the raw looking data were not. Here, we estimated effect size using an approximation formula (Dunlap, Cortina, Vaslow & Burke, 1996, p. 171). The resulting 219 effect sizes were then unbiased to yield Hedges' *g* (Morris, 2010, p. 21; see also Figure 3, Step 2). Hedges' *g* is recommended, because it provides a more conservative estimate for small sample sizes by introducing a correction factor; for large sample sizes Hedges' *g* is virtually identical to Cohen's *d*. In general, Hedges' *g* can be evaluated using the same criteria put forward by Cohen (1988). The standard errors of effect sizes were estimated using the appropriate formula for repeated measures (Lipsey & Wilson, 2001, p. 44), as were the inverse variance weights (Lipsey & Wilson, 2001, p. 44; Figure 3, Step 3).

Both the estimation of Cohen's *d* from a *t*-value and the calculation of the effect size's standard error require the Pearson correlation coefficient of individual LTs across the two test trial types (familiar and novel), because these effect sizes are within-participant repeated measures (see Figure 3). However, this information is never reported. As it was not feasible to recover correlations for all papers, we only asked for correlations from authors whom we contacted for other information (usually because their reports lacked critical information for the calculation of effect size). Correlations were returned to us for 50 records; we randomly selected values from those for the remaining records using the *impute* function of the Hmisc package in R (Harrell, 2013). This function randomly samples from the available data and thus fills in all gaps where correlations were not available with values observed in actual studies.

As mentioned above, some studies reported multiple outcomes for the same infant. Effect sizes emerging from these multiple outcomes are not mutually independent. In these cases, we combined outcomes across studies by estimating the median among the repeated measures (for the effect size, its standard error, and its weight). This resulted in 169 independent effect sizes. One of them was more than 3 standard deviations away from the effect size mean, and was excluded following standard meta-analytic practice. The final dataset thus included 168 independent effect sizes from 3774 unique infants.

We followed a standard pipeline of analysis. First, we estimated the weighted mean effect size and heterogeneity in the sample. The test statistic for unexplained variance is the Q statistic, which tests the null hypothesis that the range of observed effect sizes can be explained by sampling error alone. If the null hypothesis is rejected, there is unaccounted heterogeneity in the sample. Critical values for Q follow the chi-square distribution. QM is the Q statistic for moderator analyses, testing whether specific factors account for a significant proportion of variance. As is standard procedure in meta-analyses, we

also inspected a funnel plot and a forest plot. The funnel plot will be introduced in the Results section (see Figure 4). Forest plots display all effect sizes. With 168 entries, ours was too large to be included in the present report; it is available from the companion website.

Continuing with our analysis pipeline, we observed that heterogeneity was significant, which invited an analysis for potential moderators. In line with our research question, we investigated the moderating role of infant age, taking into account native language and differences in task difficulty, as explained above. We also carried out targeted analyses that are not typical of meta-analysis, and which will be introduced below.

Results

Preliminary analyses

The weighted mean effect size over the whole dataset was significantly above zero ($p < .0001$), with Hedges' $g = 0.22$, $SE = 0.025$. Since effect sizes are derived from LT to familiar minus novel test trials, positive values indicate longer listening to familiar stimuli and negative values relate to longer listening to novel stimuli. In other words, the median effect size in infant word segmentation is small in size (according to the criteria set by Cohen, 1988) but significantly above zero, which is consistent with a *familiarity* preference.

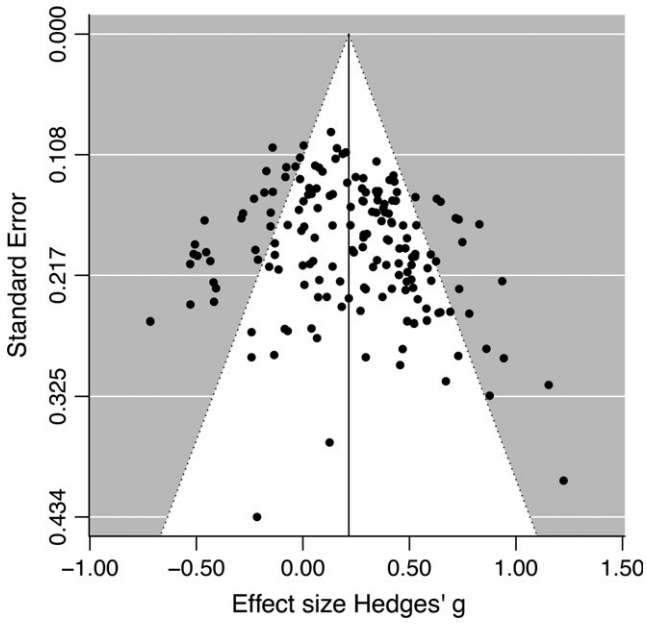

Figure 4 *Funnel plot, showing standard error of the effect size Hedges' g as a function of effect size.*

Funnel plots depict standard error of the effect size as a function of effect size. The smaller the error, the closer the effect size is expected to be to the true population mean; and the larger the error, the further away from the population mean an effect size can be. If all results are published, then studies will deviate from the population mean in either direction, whereas if a field of research systematically ignores a certain direction, then this plot can be asymmetrical. Figure 4 shows the funnel plot for our data: There is no salient evidence of bias, as the data are spread symmetrically around the mean.

Heterogeneity was significant according to a meta-analytic linear model ($Q(167) = 506$, $p < .0001$; total heterogeneity $I^2 = 69.45\%$), which means that the sample contains unexplained variance leading to significant differences across studies. In view of this, we turned to our key question. To test the influence of age on infants' segmentation abilities, we carried out three analyses. First, we tested for non-linear effects of age, following the proposal by Hunter and Ames (1988). Second, we introduced additional variables to take into account differences in task difficulty, which might mask an influence of age on effect size. Finally, we limited our analyses to studies that tested two age groups with identical designs.

Modulation of word segmentation as a function of age

The model by Hunter and Ames (1988) predicts a shift from familiarity to novelty preferences (see Figure 1). The square and cubic functions are appropriate to model such a development. The test for centered age, its square, and its cube as moderators was not significant ($N = 168$, $QM(3) = 0.68$, $p = .87$), lending no support for a non-linear change with age. Importantly, the square and cube functions had no explanatory value for the data, as they were both estimated to be zero. In addition, the estimate for age was *positive*, which is consistent with increases in familiarity preferences with age, but it was not significantly different from zero (ß = .0003, $SE = .0005$; see Figure 5).

Age revisited: taking into account task difficulty and native language

The effects of age may be obscured if experimenters present more difficult tasks to older infants. To control for such differences statistically, we included three methodological factors that modulate task difficulty according to previous work: familiarization criterion (the number of seconds infants had to attend to the familiarization stimuli, a continuous variable, centered to obtain a distribution around 0), direction of test (words

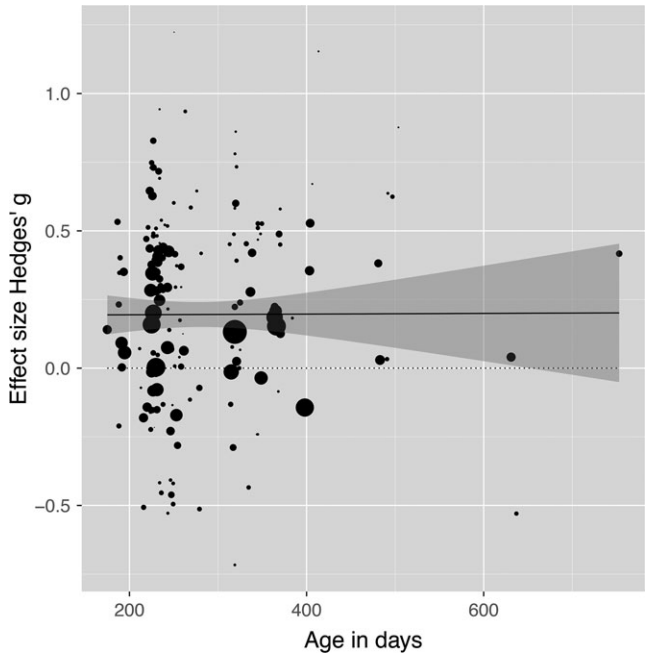

Figure 5 *Effect size Hedges' g as a function of age (in days). Each point corresponds to a unique experiment and the size to the inverse variance weight. The dotted line indicates no effect and the solid line shows the weighted linear regression.*

during familiarization and passages during test or vice versa, binary), and linguistic difficulty (a composite score of different linguistic manipulations, see Methods, ordinal).

Infants' native language was a further factor, in addition to age and task difficulty. As shown in Table 1, there are many languages for which only one or a handful of data points were available, and thus we could not fit regressors reliably with such few data points. Therefore, we created larger categories, on the basis of language families when possible, to have at least 10 independent effect sizes per category. This yielded the following groups: American English (105 samples), Canadian French (21), European French (17), and other Germanic languages (15, containing German, Dutch, British English, and Canadian English). Mandarin (four) and Catalan and Spanish (six) could not be included in this analysis.

Since there was no reason to assume that methodological factors interact with each other, with language, or age, and in the interests of maximizing our power, only age and language were allowed to interact. The resulting formula was Effect Size ~ Infant Age * Native Language + Difficulty + Familiarization Criterion + Passages In Test. Results showed no significant impact of the moderators as a whole (QM(14) = 11.18, p = .67). The estimate for age in this model was zero (ß = 0; SE = .0005).

Age revisited II: controlling for task difficulty by focusing on paired observations

As a second approach to achieving the goal of keeping 'all else equal', we carried out an analysis limited to studies where the *exact same experiment* was presented to younger and older infants. We identified 20 papers, containing 64 records, which complied with this criterion. When more than two age groups were tested with the same materials, we only considered the youngest and the oldest. The resulting analysis with age group (younger or older, ordinal) as a moderator showed again no significant impact (QM(1) = 1.6, p = .20). The estimate for age group in this model was positive, but not significantly different from zero (ß = 0.17; SE = 0.13). In addition, we assessed whether the median effect size for the younger and older groups was significantly above zero to determine whether the lack of a switch to novelty might have been due to the data points coming from the first section of the predicted curve (familiarity preference to zero, Figure 1). Both age groups are significantly above 0, with the younger group yielding Hedges' g = 0.15 (SE = 0.05), z = 2.81, p < .0001; and the older group Hedges' g = 0.28 (SE = 0.07), with z = 3.73, p < .0001. In sum, both age groups are above 0 and, while the direct comparison yielded no significant outcome, we observe a numerical increase of effect size.

Interim summary

Taking our three analyses on the influence of age on infants' segmentation performance together, we find that (1) there is, if anything, a linear and positive, albeit non-significant, effect of age (evident both in the general model, and in the one using paired observations); (2) task-dependent factors do not significantly account for any variance and in this analysis age was estimated with 0. Although these results are not clear on whether familiarity becomes stronger with age, they show with certainty that there is no shift towards a novelty preference as infants grow older.

Post-hoc analyses: publication biases

A key step in a meta-analysis is to assess for the presence of bias, which is further necessary in the present case where results fail to support popular theoretical models. We carried out several post-hoc analyses, of which we report two (see the companion website for additional analyses, and Figure 4 for a funnel plot, which is further

explained in the Preliminary Analyses section). We first examined how reports of novelty preferences are distributed over the last 20 years. If missing results are due to biases leading to data selection, this should increase over time, such that initially novelty preferences are reported and their proportion decreases as evidence for a familiarity preference continues to accumulate. To test this, we looked at the history of appearances of at least small negative effect sizes (according to Cohen's criteria, below −0.1; recall that negative effect sizes directly reflect novelty preferences). There were 31 such records (12 were reported to be statistically significant).[2] To assess whether there was a historical pattern, we calculated the percentage of published reports that contain such negative effect sizes out of the total number of published studies for every year from 1995 to 2013. This percentage correlated positively with year of publication (Spearman correlation coefficient $\rho_S(177) = 0.73$, $p = .001$; when only considering negative effect sizes associated with a p-value below alpha level: $\rho_S(260) = 0.61$, $p = .01$). These results point to an *increased acceptance* of novelty preferences as outcomes of segmentation studies using natural speech. Notably, only two studies bore on infants older than 12 months.

We further estimated the number of missing data points necessary to observe a shift in the oldest age group typically tested for their segmentation abilities, following the predictions by Hunter and Ames (1988). To calculate how many studies showing a novelty preference would be needed for a significant negative overall effect in older infants, we isolated the records on infants older than 12 months (23 records, median effect size Hedges' $g = 0.28$, $SE = 0.07$, above 0 with $p < .0001$; 3 negative effect sizes, 2 below −0.1). We expanded this dataset by adding a new data point drawn at random from the set of effect sizes in the original dataset, but with an inverted sign (i.e. negative if it was initially positive, and vice versa), until a negative effect size reached significance. With 1000 repetitions of the simulation, we had to add on average 41.5 ($SD = 12$) data points, of which 38.5 ($SD = 9.6$) were negative to reach a significantly negative effect (median Hedges' $g = -0.10$; $p = .043$). Compared to the number of negative effect sizes we started with, namely three, the amount of unpublished novelty preferences would have to be about 10 times higher than the published record to show the expected switch from a familiarity to a novelty preference as infants mature.

[2] A non-significant effect might be associated with a moderate effect size, depending on the sample size.

Summary of results

We carried out several analyses with one coherent result: Age does *not* lead to novelty preferences, contrary to our expectations based on previous piecemeal results and theoretical accounts. This result persisted even when taking into account experimental factors and infants' linguistic background, both explicitly in our regressions and through a targeted analysis of paired observations, and post-hoc analyses suggested that it was unlikely that a reporting bias would explain this result. A large amount of variance (almost 70%) remains unexplained.

Discussion

Garnering the power of 168 experiments, with data from 3774 unique infants, we sought to shed light on a key question: Does the ability to segment words from sentences change as a function of infant age (and consequently maturation and linguistic experience)? We expected to observe a switch in infants' stimulus preference, from familiarity (encoded in positive effect sizes) to novelty (negative effect sizes), indicating improvement following previous work on infant information processing (e.g. Colombo, 2002; Houston-Price & Nakai, 2004; Hunter & Ames, 1988; Roder *et al*., 2000; Sirois & Mareschal, 2002) and in line with current assumptions in the field (see e.g. Babineau & Shi, 2011, p. 35; Bosch *et al*., 2013, p. 9; Seidl & Johnson, 2008, pp. 15f.; Singh *et al*., 2004, pp. 184f.). We can confidently state that this pattern is *not* present in the existing data. Instead, we observe an overall familiarity preference, which is not significantly affected by age.

Revising mainstream assumptions

We are certain, based on our careful investigation, that we are facing a true result; there is no switch to novelty with age in word segmentation tasks. This is a surprising result given the two mainstream assumptions laid out in the Introduction: (1) skilled processors should display novelty preferences (all else being equal); and (2) infants' segmentation of words from their native speech improves with age. Each of these assumptions could be incorrect.

As for the first assumption, Hunter and Ames' (1988) model may not be the most appropriate theoretical framework for interpreting preferences in word segmentation from natural, native speech. In fact, Hunter and Ames' proposal and related models have not been free from criticism, because an initial period of stable familiarity preference is not always observed (e.g. Roder *et al*., 2000), but also because models of stimulus preferences are mostly unable to account for differences

in preference that do not relate to the timeline of processing but rather to other key cognitive features of the task. For instance, Kidd, Piantadosi and Aslin (2012) document that entropy in visual stimulation modulates attention through a U-shaped function, as infants attended longest to visual displays with moderate difficulty, and less to both overly simple and overly complex ones. Similarly, and somewhat closer to the skill discussed here, Gerken, Balcomb and Minton (2011) document that infants look longer when they are listening to stimuli from which a rule can be discovered compared to very similar sequences where no such rule is present. Put more generally, in the context of word segmentation from native speech, infants' attention may be better predicted by other factors, such as the characteristics of the stimuli (see also Aslin, 2007; Bergmann, ten Bosch, Fikkert & Boves, 2013).

In view of these previous results, one could propose an interpretation whereby a sustained familiarity preference, unchanged (or even increased) by age, is desirable when segmenting words from native speech. Indeed, an anonymous reviewer suggested that even toddlers may continue to attend to a familiar wordform because they can use it as an anchor to learn other valuable pieces of information, such as the linguistic context (e.g. other words the target co-occurs with; Bortfeld *et al*., 2005), and a post-hoc analysis of our database suggested that this was an interesting possibility to explore in further work.[3]

It remains clear that age *per se* does not lead to a switch to novelty, either because 'completeness of processing' is not the most important factor affecting infants' preferences, and/or because infants and toddlers continue attempting to derive information from familiarized words. In other words, Hunter and Ames' (1988) may not be the most relevant model, and authors in this literature should be cautious when interpreting a novelty outcome as merely the result of age.

Could our second mainstream assumption also be false? Indeed, authors (including us) frequently use age as a proxy for maturation, which should lead to increases in attentional and memory resources. However, we know of no work convincingly documenting a link between variation in attention and memory, on the one hand, and specifically word segmentation skills, on the other, within infancy. The second way in which age should correlate with increases of skill is if infants adopt segmentation strategies that are appropriate for the native language, and acquire other helpful linguistic skills (e.g. knowing more words). Here as well, we have to point out that there is no research directly demonstrating that infants from different language backgrounds use different word segmentation strategies. Thus, we cannot at present even be certain of this fundamental assumption.

Alternative explanations

One may wonder whether a novelty preference does emerge at an age greater than the ones we considered. This does not appear to be likely, since a novelty preference has also been observed in English learners, for example at 7.5 months (Singh, 2008) and at around 11 months (Seidl & Johnson, 2008). In fact, the majority of novelty preferences in the present meta-analysis stem from infants younger than 12 months (29 out of 31). Importantly, these studies suffice to show that novelty preferences *can* be elicited in younger infants segmenting native speech. It remains unclear what distinguishes these studies from others in the same age group that showed the predominant familiarity preference.

A second alternative explanation is that we are drawing from a biased literature. Ioannidis (2005, p. 0700) famously stated that 'the claimed effect sizes [may be] measuring nothing else but the net bias that has been involved in the generation of [a given] scientific literature'. In the present case, the bias would have been generated with the initial study by Jusczyk and Aslin (1995), which yielded familiarity preferences. Thereafter, the bias would be reinforced through data selection at the submission and/or publication stages (i.e. authors confine their results to the file-drawer when seeing a novelty preference, or their study is rejected by skeptical reviewers). The post-hoc analyses reported above (and in the Supplementary Materials) revealed no evidence of publication biases. We expect that a continuous accumulation of evidence will allow us to inspect the distribution of emerging and previously unpublished studies as they are added, but at present selective reporting cannot explain the lack of a switch to novelty preferences.

Additional methodological considerations

Based on our investigation of possible publication biases, we can further conclude that there are sound and reliable

[3] If infants continue to attend to familiar stimuli for the purposes of extracting information, they may be more likely to disengage from the familiar word (and explore a novel competitor) when there is little to be gained by continuing to attend to it; in other words, novelty preferences should be more common among studies using the passage-to-word design, where no linguistic context is available in the test phase. As reported in the Results section, we included this factor when accounting for task difficulty, but the test for moderators was overall not significant, not warranting further statistical exploration. Nonetheless, a direct comparison reveals a trend towards more novelty preferences in passage-to-word studies ($\chi^2(1) = 3.33$, $p = .06$), but age plays no role in this analysis.

reporting practices in place within the research domain of word segmentation from real speech. This finding, while not the core aim of our endeavor, is reassuring for a field characterized by small sample sizes, large individual variability, and ensuing low study power. We have also observed that methodologies are mostly stable, and that no gross statistical errors or misdemeanors were evident. To give an example that readers can verify using the publicly available data that we published on a companion website: Effect sizes are not related to other indices of data selection, such as proportion of infants excluded and number of test trials (the original study by Jusczyk & Aslin, 1995, had 16, and some papers report 8 or 12, which may reflect selection of early blocks of trials). This should reinforce our trust in data gathered in this domain, which, at least on paper, adheres to excellent practices.

Nonetheless, we take this opportunity to point out possible improvements. First, we strongly suggest that authors report the Pearson correlation coefficient between test conditions to facilitate future meta-analyses and prospective power calculations that authors might want to conduct before gathering experimental data. For within-participant designs, the Pearson correlation coefficient is necessary to estimate the effect size (based on *t*-values) and its standard error. Further, the Pearson correlation coefficient can serve as an indicator of the systematicity of infant behavior, allowing readers to evaluate experiments on information beyond (non)significant *p*-values. In addition, two-tailed tests are the norm, with only two exceptions in 75 *t*-tests reported in the 51 papers analyzed. While familiarity preferences constitute a large majority, novelty preferences emerge across languages and age groups and remain a possible outcome. Our careful meta-analytic investigations did not lead to any specific factor that modulated effect sizes significantly, and consequently we have currently no way of predicting whether a given outcome will be one of the rare novelty cases. In this context of uncertainty, two-tailed tests remain appropriate.

A related question is whether any significant preference is equally acceptable irrespective of the direction, as suggested for example by Houston-Price and Nakai (2004, p. 344, see also Aslin, 2007, p. 51). While this may be a reasonable approach the very first time one uses a paradigm, science is built on the cornerstone of replicability when using comparable methods. In the present meta-analysis, we purposefully confined ourselves to studies testing infants in their native language following the paradigm set out in Jusczyk and Aslin (1995), which are thus both conceptual and methodological replications. If results are *not* comparable when such similar methods are used, then one of two conclusions must ensue: Either we postulate that this line of research is not scientific, or we accept that there are variables that have not captured researchers' (and therefore, these meta-analysts') attention. We favor the latter conclusion, and hope that our complementary community-augmented website and database helps the community discover other variables that may account for the large proportion of variance in our data which remains unexplained (about 70%).

Conclusions

The present study sought to shed light on a key prediction derived from a common assumption in the word segmentation literature (infants' word segmentation skills improve with age) and popular models tying skill to behavioral performance in lab-based tasks (more skilled participants show novelty preferences). The combined results from 168 experiments revealed that there is no shift to novelty with age, and that this could not be explained by either infants' native language and differences between experiments or by publication biases or data selection. Our results invite the revision of the frequent assumption that novelty preferences must occur at some age, and a thorough evaluation of the applicability of current models of stimulus preferences in language processing tasks, specifically to word segmentation. The present research has also led to the establishment of a public and open database, which facilitates the integration of unpublished and future results.

Acknowledgements

The work was supported by a Fondation Pierre-Gilles de Gennes grant to Christina Bergmann, institutional support from ANR-10-LABX-0087 and ANR-10-IDEX-0001-02, and grant ANR-14-CE30-0003 MechE-Lex to Alejandrina Cristia. The authors wish to thank Amanda Seidl for her invaluable contribution to the database that forms the basis for this meta-analysis. Many thanks also to Laura Bosch, Anne Christophe, Marieke van Heugten, Elizabeth Johnson, Rochelle Newman, Jon Willits, Jenny Saffran, Rushen Shi, Leher Singh, and Emily Mason-Apps, whose responses enriched the database. Attendees of a dedicated lunch discussion at WILD 2013 and of a presentation at BUCLD 2014 as well as colleagues at the Max Planck Institute for Psycholinguistics and the LSCP were critical, encouraging, and helpful discussion partners. Finally, the editors and anonymous reviewers provided very insightful and constructive feedback on earlier versions of this paper.

References

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*, **16** (6), 980–990. doi: 10.1111/desc.12071

Aslin, R.N. (2007). What's in a look? *Developmental Science*, **10** (1), 48–53. doi: 10.1111/j.1467-7687.2007.00563.x

Babineau, M., & Shi, R. (2011). Processing of French liaisons in toddlers. In N. Davis, K. Mesh & H. Sung (Eds.), *Proceedings of the 35th Boston University Conference on Language Development* (pp. 25–37). Somerville, MA: Cascadilla Press.

Barker, B.A., & Newman, R.S. (2004). Listen to your mother! The role of talker familiarity in infant streaming. *Cognition*, **94**, B45–B53. doi:10.1016/j.cognition.2004.06.001

Bartels, S., Darcy, I., & Höhle, B. (2009). Schwa syllables facilitate word segmentation for 9-month-old German-learning infants. In J. Chandlee, M. Franchini, S. Lord & G.-M. Rheiner (Eds.), *Proceedings of the 33rd Boston University Conference on Language Development* (pp. 73–84). Somerville, MA: Cascadilla Press.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, USA*, **109** (9), 3253–3258. doi: 10.1073/pnas.1113380109

Bergmann, C., ten Bosch, L., Fikkert, P., & Boves, L. (2013). A computational model to investigate assumptions in the headturn preference procedure. *Frontiers in Psychology*, **4**, 676. doi: 10.3389/fpsyg.2013.00676

Bortfeld, H., Morgan, J.L., Golinkoff, R.M., & Rathbun, K. (2005). Mommy and me: familiar names help launch babies into speech-stream segmentation. *Psychological Science*, **16** (4), 298–304. doi: 10.1111/j.0956-7976.2005.01531.x

Bosch, L., Figueras, M., Teixidó, M., & Ramon-Casas, M. (2013). Rapid gains in segmenting fluent speech when words match the rhythmic unit: evidence from infants acquiring syllable-timed languages. *Frontiers in Psychology*, **4**, 106. doi: 10.3389/fpsyg.2013.00106

Brent, M.R., & Siskind, J.M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, **81** (2), B33–B44. doi: 10.1016/S0010-0277(01)00122-6

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn.). Hillsdale, NJ: Erlbaum.

Colombo, J. (2002). Infant attention grows up: the emergence of a developmental cognitive neuroscience perspective. *Current Directions in Psychological Science*, **11** (6), 196–200. doi: 10.1111/1467-8721.00199

Dunlap, W.P., Cortina, J.M., Vaslow, J.B., & Burke, M.J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, **1** (2), 170–177. doi: 10.1037/1082-989X.1.2.170

Durlak, J.A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, **34** (9), 917–928. doi: 10.1093/jpepsy/jsp004

Galle, M.E., & McMurray, B. (2014). The development of voicing categories: a quantitative review of over 40 years of infant speech perception research. *Psychonomic Bulletin & Review*, **21** (4), 884–906. doi: 10.3758/s13423-013-0569-y

Gathercole, S.E., & Baddeley, A.D. (1993). *Working memory and language*. Hove, East Sussex: Lawrence Erlbaum Associates.

Gehanno, J.F., Rollin, L., & Darmoni, S. (2013). Is the coverage of Google Scholar enough to be used alone for systematic reviews. *BMC Medical Informatics and Decision Making*, **13** (1), 7. doi: 10.1186/1472-6947-13-7

Gerken, L., Balcomb, F.K., & Minton, J.L. (2011). Infants avoid 'labouring in vain' by attending more to learnable than unlearnable linguistic patterns. *Developmental Science*, **14** (5), 972–979. doi: 10.1111/j.1467-7687.2011.01046.x

Gonzalez-Gomez, N., & Nazzi, T. (2013). Effects of prior phonotactic knowledge on infant word segmentation: the case of nonadjacent dependencies. *Journal of Speech, Language, and Hearing Research*, **56** (3), 840–849. doi: 0.1044/1092-4388(2012/12-0138)

Gout, A., Christophe, A., & Morgan, J.L. (2004). Phonological phrase boundaries constrain lexical access II: Infant data. *Journal of Memory and Language*, **51** (4), 548–567. doi: 10.1016/j.jml.2004.07.002

Harrell, F.E. (2013). Harrell Miscellaneous. R package. Version 3.14-6. Available at: http://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf

Höhle, B., & Weissenborn, J. (2003). German-learning infants' ability to detect unstressed closed-class elements in continuous speech. *Developmental Science*, **6** (2), 122–127. doi: 10.1111/1467-7687.00261

Hollich, G., Newman, R.S., & Jusczyk, P.W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development*, **76** (3), 598–613. doi: 10.1111/j.1467-8624.2005.00866.x

Houston, D.M., & Jusczyk, P.W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, **26** (5), 1570–1582. doi: 10.1037/0096-1523.26.5.1570

Houston, D.M., & Jusczyk, P.W. (2003). Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance*, **29** (6), 1143–1154. doi: 10.1037/0096-1523.29.6.1143

Houston, D.M., Jusczyk, P.W., Kuijpers, C., Coolen, R., & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin & Review*, **7** (3), 504–509. doi: 10.3758/BF03214363

Houston, D.M., Santelmann, L., & Jusczyk, P.W. (2004). English-learning infants' segmentation of trisyllabic words from fluent speech. *Language and Cognitive Processes*, **19** (1), 97–136. doi: 10.1080/01690960344000143

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, **13** (4), 341–348. doi: 10.1002/icd.364

Hunter, M.A., & Ames, E.W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, **5**, 69–95.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine*, **2** (8), e124. doi: 10.1371/journal.pmed.0020124

Johnson, E.K. (2005). English-learning infants' representations of word forms with iambic stress. *Infancy*, **7** (1), 99–109. doi: 10.1207/s15327078in0701_8

Johnson, E.K. (2008). Infants use prosodically conditioned acoustic-phonetic cues to extract words from speech. *Journal of the Acoustical Society of America*, **123** (6), EL144–EL148. doi: 10.1121/1.2908407

Johnson, E.K., Jusczyk, P.W., Cutler, A., & Norris, D. (2003). Lexical viability constraints on speech segmentation by infants. *Cognitive Psychology*, **46** (1), 65–97. doi: 10.1016/S0010-0285(02)00507-8

Jusczyk, P.W., & Aslin, R.N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, **29** (1), 1–23. doi: 10.1006/cogp.1995.1010

Jusczyk, P.W., Hohne, E.A., & Bauman, A. (1999a). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, **61** (8), 1465–1476. doi: 10.3758/BF03213111

Jusczyk, P.W., Houston, D.M., & Newsome, M. (1999b). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, **39** (3), 159–207. doi: 10.1006/cogp.1999.0716

Katz-Gershon, S. (2007). Word extraction in infant and adult directed speech: Does dialect matter? Doctoral dissertation, Wayne State University.

Kidd, C., Piantadosi, S.T., & Aslin, R.N. (2012). The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE*, **7** (5), e36399. doi: 10.1371/journal.pone.0036399

Kuijpers, C.T.L., Coolen, R., Houston, D.M., & Cutler, A. (1998). Using the head-turning technique to explore cross-linguistic performance differences. In C. Rovee-Collier, L. Lipsitt & H. Hayne (Eds.), *Advances in infancy research* (Vol. 12, pp. 205–220). Stamford, CT: Ablex.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t* tests and ANOVAs. *Frontiers in Psychology*, **4**, 863. doi:10.3389/fpsyg.2013.00863

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Marquis, A., & Shi, R. (2008). Segmentation of verb forms in preverbal infants. *Journal of the Acoustical Society of America*, **123** (4), EL105–EL110. doi: 10.1121/1.2884082

Marquis, A., & Shi, R. (2009). The recognition of verb roots and bound morphemes when vowel alternations are at play. In J. Chandlee, M. Franchini, S. Lord & M. Rheiner (Eds.), *A supplement to the Proceedings of the 33rd Boston University Conference on Language Development*.

Marquis, A., & Shi, R. (2012). Initial morphological learning in preverbal infants. *Cognition*, **122** (1), 61–66. doi: 10.1016/j.cognition.2011.07.004

Mason-Apps, E., Stojanovik, V., & Houston-Price, C. (2011). Early word segmentation in typically developing infants and infants with Down syndrome: a preliminary study. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences*, 1334–1337. Available at: www.icphs2011.hk

Mattys, S.L., & Jusczyk, P.W. (2001a). Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance*, **27** (3), 644. doi: 10.1037/0096-1523.27.3.644

Mattys, S.L., & Jusczyk, P.W. (2001b). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, **78** (2), 91–121. doi: 10.1016/S0010-0277(00)00109-8

Mersad, K., & Nazzi, T. (2012). When Mommy comes to the rescue of statistics: infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development*, **8** (3), 303–315. doi: 10.1080/15475441.2011.609106

Mills-Smith, L., Spangler, D.P., Panneton, R., & Fritz, M.S. (2015). A missed opportunity for clarity: problems in the reporting of effect size estimates in infant developmental science. *Infancy*. Early view. doi: 10.1111/infa.12078

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D.G., & the PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, **151** (4), 264–269. doi: 10.7326/0003-4819-151-4-200908180-00135

Morris, S.B. (2010). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, **53**, 17–29. doi: 10.1348/000711000159150

Nazzi, T., Dilley, L.C., Jusczyk, A.M., Shattuck-Hufnagel, S., & Jusczyk, P.W. (2005). English-learning infants' segmentation of verbs from fluent speech. *Language and Speech*, **48** (3), 279–298. doi: 10.1177/00238309050480030201

Nazzi, T., Iakimova, G., Bertoncini, J., Frédonie, S., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: emerging evidence for crosslinguistic differences. *Journal of Memory and Language*, **54** (3), 283–299. doi: 10.1016/j.jml.2005.10.004

Nazzi, T., Mersad, K., Sundara, M., Iakimova, G., & Polka, L. (2014). Early word segmentation in infants acquiring Parisian French: task-dependent and dialect-specific aspects. *Journal of Child Language*, **41** (3), 600–633. doi: 10.1017/S0305000913000111

Newman, R.S., & Jusczyk, P.W. (1996). The cocktail party effect in infants. *Perception & Psychophysics*, **58** (8), 1145–1156. doi: 10.3758/BF03207548

Polka, L., & Sundara, M. (2012). Word segmentation in monolingual infants acquiring Canadian English and Canadian French: native language, cross-dialect, and cross-language comparisons. *Infancy*, **17** (2), 198–232. doi: 10.1111/j.1532-7078.2011.00075.x

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Roder, B.J., Bushnell, E.W., & Sasseville, A.M. (2000). Infants' preferences for familiarity and novelty during the course of visual processing. *Infancy*, **1** (4), 491–507. doi: 10.1207/S15327078IN0104_9

Rose, S.A., Gottfried, A.W., Melloy-Carminar, P., & Bridger, W.H. (1982). Familiarity and novelty preferences in infant recognition memory: implications for information processing. *Developmental Psychology*, **18** (5), 704–713. doi: 10.1037/0012-1649.18.5.704

Ruff, H.A., & Rothbart, M.K. (2001). *Attention in early development: Themes and variations*. Oxford: Oxford University Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, **274** (5294), 1926–1928.

Schmale, R., & Seidl, A. (2009). Accommodating variability in voice and foreign accent: flexibility of early word representations. *Developmental Science*, **12** (4), 583–601. doi: 10.1111/j.1467-7687.2009.00809.x

Schwarzer, G. (2007). Meta: an R package for meta-analyses. *R News*, **7**, 40–45.

Seidl, A., & Johnson, E.K. (2006). Infant word segmentation revisited: edge alignment facilitates target extraction. *Developmental Science*, **9** (6), 565–573. doi: 10.1111/j.1467-7687.2006.00534.x

Seidl, A., & Johnson, E.K. (2008). Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *Journal of Child Language*, **35** (01), 1–24. doi: 10.1017/S0305000907008215

Seidl, A., Tincoff, R., Baker, C., & Cristia, A. (2015). Why the body comes first: effects of experimenter touch on infants' word finding. *Developmental Science*, **18** (1), 155–164. doi: 10.1111/desc.12182

Shi, R. (2007). Infants' recognition of function words in continuous speech. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences*, 1541–1544. Available at: www.icphs2011.hk

Shi, R., Cutler, A., Werker, J., & Cruickshank, M. (2006a). Frequency and form as determinants of functor sensitivity in English-acquiring infants. *Journal of the Acoustical Society of America*, **119** (6), EL61–EL67. doi: 10.1121/1.2198947

Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science*, **11** (3), 407–413. doi: 10.1111/j.1467-7687.2008.00685.x

Shi, R., Marquis, A., & Gauthier, B. (2006b). Segmentation and representation of function words in preverbal French-learning infants. In D. Bamman, T. Magnitskaia & C. Zaller (Eds.), *Proceedings of the 30th Annual Boston University Conference on Language Development* (pp. 549–560). Sommerville, MA: Cascadilla Press.

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, **106** (2), 833–870. doi: 10.1016/j.cognition.2007.05.002

Singh, L., & Foong, J. (2012). Influences of lexical tone and pitch on word recognition in bilingual infants. *Cognition*, **124** (2), 128–142. doi: 10.1016/j.cognition.2012.05.008

Singh, L., Morgan, J.L., & White, K.S. (2004). Preference and processing: the role of speech affect in early spoken word recognition. *Journal of Memory and Language*, **51** (2), 173–189. doi: 10.1016/j.jml.2004.04.004

Singh, L., Nestor, S.S., & Bortfeld, H. (2008a). Overcoming the effects of variation in infant speech segmentation: influences of word familiarity. *Infancy*, **13** (1), 57–74. doi: 10.1080/15250000701779386

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, **14** (6), 654–666. doi: 10.1080/15250000903263973

Singh, L., Reznick, J.S., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: a longitudinal analysis. *Developmental Science*, **15** (4), 482–495. doi: 10.1111/j.1467-7687.2012.01141.x

Singh, L., White, K.S., & Morgan, J.L. (2008b). Building a word-form lexicon in the face of variable input: influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, **4** (2), 157–178. doi: 10.1080/15475440801922131

Sirois, S., & Mareschal, D. (2002). Models of habituation in infancy. *Trends in Cognitive Sciences*, **6** (7), 293–298. doi: 10.1016/S1364-6613(02)01926-5

Tincoff, R., & Jusczyk, P.W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, **17** (4), 432–444. doi: 10.1111/j.1532-7078.2011.00084.x

Tsay, J., & Jusczyk, P.W. (2003). Detection of words in fluent Chinese by English-acquiring and Chinese-acquiring infants. In D. Houston, A. Seidl, G. Hollich, E. Johnson & A. Jusczyk (Eds.), *Jusczyk Lab Final Report*. Retrieved from: http://hincapie.psych.purdue.edu/Jusczyk

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: toward cumulative data assessment. *Perspectives on Psychological Science*, **9** (6), 661–665. doi: 10.1177/1745691614552498

Tsuji, S., & Cristia, A. (2013). Perceptual attunement in vowels: a meta-analysis. *Developmental Psychobiology*, **56** (2), 179–191. doi: 10.1002/dev.21179

Tyler, M.D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America*, **126** (1), 367–376.

van de Weijer, J. (1998). Language input for word discovery. Doctoral dissertation, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

van Heugten, M., & Johnson, E.K. (2012). Infants exposed to fluent natural speech succeed at cross-gender word recognition. *Journal of Speech, Language, and Hearing Research*, **55** (2), 554–560. doi: 10.1044/1092-4388(2011/10-0347)

van Kampen, A., Parmaksiz, G., van de Vijver, R., & Höhle, B. (2008). Metrical and statistical cues for word segmentation: the use of vowel harmony and word stress as a cue to word boundaries by 6- and 9-month-old Turkish learners. In A. Gavarró & M.J. Freitas (Eds.), *Language acquisition and development* (pp. 313–324.) Newcastle: Cambridge Scholars Publishing.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, **36** (3), 1–48.

Wagner, S.H., & Sakovits, L.J. (1986). A process analysis of infant visual and cross-modal recognition memory: implications for an amodal code. *Advances in Infancy Research*, **4**, 195–245.

Willits, J.A., Seidenberg, M.S., & Saffran, J.R. (2009). Verbs are lookING good in early language acquisition. In N. Taatgen, van Rijn H., L. Schomaker & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2570–2575). Austin, TX: Cognitive Science Society.

Received: 22 December 2014
Accepted: 1 June 2015

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Supplementary Material.