Chunks of phonological knowledge play a significant role in children's word learning and explain effects of neighborhood size, phonotactic probability, word frequency and word length

Gary Jones[1]

Francesco Cabiddu[1]

Mark Andrews[1]

Caroline Rowland[2]

The model code and data that support the findings of this study are available from the Open Science Framework at the following address: https://osf.io/z8sa2/

[1] Department of Psychology, Nottingham Trent University

[2] Max Planck Institute for Psycholinguistics, Nijmegen and Donders Institute for Brain, Cognition and Behaviour, Radboud University

Corresponding author: Gary Jones, Department of Psychology, Nottingham Trent University, 50 Shakespeare Street, Nottingham NG1 4FQ (UK). Telephone: +44 (115) 848 2422; E-mail: gary.jones@ntu.ac.uk.

Keywords: Word learning; phonotactic probability; neighborhood density; statistical learning; child development; CLASSIC.

Abstract

A key omission from many accounts of children's early word learning is the linguistic knowledge that the child has acquired up to the point when learning occurs. We simulate this knowledge using a computational model that learns phoneme and word sequence knowledge from naturalistic language corpora. We show how this simple model is able to account for effects of word length, word frequency, neighborhood density and phonotactic probability on children's early word learning. Moreover, we show how effects of neighborhood density and phonotactic probability on word learning are largely influenced by word length, with our model being able to capture all effects. We then use predictions from the model to show how the ease by which a child learns a new word from maternal input is directly influenced by the phonological knowledge that the child has acquired from other words up to the point of encountering the new word. There are major implications of this work: models and theories of early word learning need to incorporate existing sublexical and lexical knowledge in explaining developmental change while well-established indices of word learning are rejected in favor of phonological knowledge of varying grain sizes.

Introduction

The input from which children acquire their cognitive representations often comes via rapid and transient signals (e.g. actions, speech, events), which must be processed quickly, in the moment, in order to be utilized by the child's learning mechanisms. The amount and type of knowledge already acquired at the point of learning is likely to affect how this input is processed, and thus, how new knowledge can be extracted from this input, and integrated into representations in long-term memory (Karmiloff-Smith, 1998). For example, words that are presented in familiar sentence frames are processed faster (Fernald & Hurtado, 2006), complex grammatical constructions that share structural properties with prior learned constructions are acquired more easily (Abbot-Smith & Behrens, 2006), and non-words that contain familiar, wordlike phoneme sequences are repeated more accurately (Gathercole, 1995).

Despite this, many major debates in developmental cognition have tended to ignore the role of the accumulating knowledge in driving developmental change, focusing instead on debating the role of the input and of innate constraints/knowledge in the learning process. For example, in the word learning literature, the traditional debate has focused on whether a simple associative learning mechanism, equipped with low level abilities like cross-situational statistical learning, can extract all the information it needs to learn words from the input (Smith & Yu, 2008), or whether we also need to build higher level socio-cognitive (e.g. Tomasello, 2003) or linguistic (e.g. Markman, 1989) constraints into the learning mechanisms. In these theories, the role played by the child's existing linguistic knowledge at the precise moment that she learns a new word is static across development.

The implications of this omission are profound, because it means that traditional theories effectively assume that the way in which a child learns a word remains the same

across development. This is unlikely to be the case. For example, we know that different input characteristics have different effects on vocabulary acquisition at different stages of development; 18-month olds seem to learn best from input that is high in repetition, in which a few words are repeated often, but 30-month olds benefit instead from an input that is high in variation, in which many different words are produced but rarely repeated (Rowe, 2012). Traditional associative, socio-cognitive and linguistic models, in which each word is learned in the same way regardless of the model's developmental stage, cannot explain this effect. Only models where learning occurs over developmental time and in a way that is influenced by current knowledge at that time can simulate effects of this type.

Nowadays, there is increasingly more work that factors in the child's knowledge at the point of learning a new word. Broadly speaking, this research falls under two categories: the influence of semantic characteristics and/or phonological characteristics. For semantic characteristics, research has primarily focused on network analyses, where nodes represent a word and links between nodes indicate some form of semantic relatedness between words. Hills et al. (2009) used nouns (N = 130) from the MacArthur Bates Communicative Developmental Inventory (Dale & Fenson, 1996) together with their age of first production (from 16 to 30 months). They examined how the nouns clustered together semantically, creating a network for each consecutive month that involved all nouns produced up to that point. Nouns that entered the child's productive vocabulary at a particular timepoint were best explained by their semantic relatedness to all possible nouns that could be learned across all timepoints rather than semantic relatedness restricted to nouns produced by the child up to the timepoint in question. The same was found when using a similar procedure but expanding word classes beyond nouns (N = 532) (Hills et al., 2010). Interestingly, in both studies, phonological characteristics were also examined: word frequency; neighborhood density (ND, or the number of neighbors a word has, with neighbour defined as the addition, deletion

or substitution of one sound e.g., *cat* has *cot*, *cats*, *at* etc. as neighbors); and for Hills et al. (2009), phonotactic probability (PP, typically defined as how often a biphone – or two sounds that occur sequentially – occur in a language corpus relative to all combinations of biphones). Word frequency and ND both contributed to word learning in addition to semantic factors. The network analysis approach suggests that the influence of semantic relatedness on word learning may not be dependent on those words known by children at the point of learning but that phonological characteristics also significantly contribute to word learning. However, more recent empirical work by Borovsky et al. (2016) examining semantic relatedness based on 24-month-old children's own productive vocabularies found that current semantic knowledge did influence subsequent processing of semantically related words.

Network analyses that describe phonological associations across words also use nodes to represent words, but links between nodes are typically based on satisfying the traditional definition of ND described above (e.g., *bat* and *cat* would be linked but not *bat* and *dog*). Vitevitch (2008) for example compared his network to one produced at random, showing how the adult lexicon represented a 'small world network': the number of links between words was comparable but the ND network had larger clusters of linked words, suggesting potential processing efficiencies based on ND (e.g., lexical competitors would exist within the cluster rather than being disparate). Siew (2013) used the same network but with different outcome measures, showing how each cluster of words differed in terms of such things as ND, PP, word length and word frequency (e.g., some clusters mainly involved high ND words and others involved low ND words).

Work on semantic influences on word learning have primarily focused on such network analyses. While network analyses are useful in testing competing hypotheses about word learning and for describing how words may fruitfully cluster together to create processing efficiencies, these networks do not actually implement learning. Instead, they describe

associations across words at particular points in time. Although this is also the case for work on phonological influences on word learning, there are also specific hypotheses and models that propose how current phonological knowledge may influence subsequent word learning.

The lexical restructuring hypothesis (e.g., Metsala & Walley, 1998; Walley, 1993; Walley, Metsala & Garlock, 2003) suggests that infants implicitly attune to the phonological segments of speech but quickly begin to represent words as unanalysed wholes in line with the vocabulary spurt that occurs around 18 months. Thereafter, phonological analysis of holistic word forms is primarily driven by the need to differentiate similar sounding words. The hypothesis suggests not only that the child holds greater segmental detail for words having dense neighborhoods but also that this process is gradual, being dependent upon exposure to similar sounding words. Such a view predicts that children will learn novel words more easily when the novel words have high NDs due to the greater supporting knowledge that exists for such novel words. This prediction has been upheld by numerous studies (e.g., Hollich, Jusczyk & Luce, 2002; Storkel, 2009).

However, the direct method for examining the effect of current knowledge on subsequent learning is to use computational models that learn over time as more and more input is presented. Vitevitch and Storkel (2013) used a connectionist model to operationalize a similar view to that of the lexical restructuring hypothesis. Eighteen input units were used to represent consonant-vowel-consonant (CVC) word inputs, with 6 units per phoneme; 6 hidden units were connected to all input units to capture word-level knowledge associated with the input as training progressed; these then connected to 18 output units that mirrored the input units in order to test the extent to which the network had learned the trained words and also how the network performed for novel CVC words. Trained words and novel words comprised ones from both dense neighborhoods (sharing two phonemes) and sparse neighborhoods (sharing one or zero phonemes). For both trained and novel words, the model

was more able to represent those from dense neighborhoods than sparse neighborhoods. The same was true when the training set was altered to be more developmentally plausible (i.e., where a small set of words began training with more words introduced to the set over time). Interestingly, a direct consequence of the particular word sets meant that dense neighbourhood words inevitably had biphones that were high frequency, or high PP, since they appeared across several words. As such, the findings do not concur with other work showing how young children's vocabularies mainly contain words having low PP biphones, leading the authors to suggest that PP effects require another level of representation.

The dominant factor in these views and analyses of phonological word learning is neighborhood, and neighborhood also plays a significant role in the semantic networks discussed earlier. Yet there are problems in explaining word learning in terms of the neighborhood characteristics of a particular to-be-learned word. First, ND shows significant positive correlations with PP and word frequency, while there is a particularly strong negative correlation with word length (e.g., Storkel, 2009): as word length increases, ND dramatically reduces. Second, the majority of words have either no neighbors or very few neighbors. This issue is evident in Vitevitch (2008) and Siew (2013) where out of a total of 19,340 words, 53% had no neighbors (e.g., *obtuse*) and 13% had few localized neighbors (e.g., *converse*, *converge*, *convert*), with none of these words contributing to their analyses. Third, there may be additional factors (or different factors) involved in phonological word learning that drive the perceived effect of neighborhood. It has already been acknowledged above that PP may be one of those factors, but even PP is not ideal since it traditionally focuses on biphones (Auer & Luce, 2005).

In this paper we leave aside the influence of semantic information on word learning and focus instead on phonological characteristics that network analyses, modeling work and empirical work all suggest is influential for word learning. Our goal is to examine the extent

to which word learning can be explained by appealing to phonological information; but importantly we also consider whether the observed effects of ND and PP (together with effects of word length and word frequency) can be explained by a simple chunking account of phonological word learning that pays no attention to definitions of ND and PP. On our view, chunks are created incrementally in line with the frequency of encounter of a particular phoneme sequence in the input (e.g., for *hand -> h*, *ha*, *han*, *hand*) such that long phoneme sequences that occur often are likely to be represented as large chunks. This view has fruitfully been applied to word segmentation – locating word boundaries within continuous speech, a feat typically achieved by the developing infant between the ages of around 0;6-1;6. For example, BootLex (Batchelder, 2002) parses continuous speech into potential words by a combination of knowledge of optimal word length and selection of the (incrementally chunked) phoneme sequences having the highest combined frequency; while TRACX (French, Addyman & Mareschal, 2011) shows over a series of studies how recognition of previous frequently encountered phoneme sequences is able to mimic behavior in studies of segmentation. Similar to TRACX, our view records no frequency information; rather, frequently encountered phoneme sequences form larger and larger chunks. However, learning phoneme sequence chunks can be facilitated if there is prior knowledge of other chunks that share sequential phonological information (e.g., *ha* from *hat*, and *nd* from *stand*; *ha + nd -> hand*). Word learning is the learning of a chunked phonological sequence that corresponds to a word. This view is supported by Goldstein and Vitevitch (2014). In their study, novel words were controlled for ND and other measures but differed in their clustering coefficient (the extent to which neighbors of a word are also neighbors of each other; e.g., *badge -> ban, bath* etc. vs. *log -> dog, lawn* etc.). Novel words with a higher clustering coefficient were learned more easily, giving credence to the notion that shared sublexical phoneme sequences across many words facilitates learning.

Under our view, the concept of ND is fluid, since it relates to both the shared phoneme sequences between known words and a to-be-learned word and the size of the existing chunks that contain those phoneme sequences. The concept of PP is equally fluid since it is not related to the raw frequency of a phoneme or biphone but rather the size of the chunks in which phoneme sequences appear and whether they can be applied to the to-be-learned word. We also examine word length and word frequency effects in the model. For word length, all things being equal, longer words will require more exposure to be learned; however, this is mediated by the constituent phoneme sequences since if these are shared by many other words, it is likely to lead to the word being learned more quickly. For word frequency, highly frequent words are likely to be learned quickly but these in turn may facilitate the learning of words that rarely occur, based on shared phoneme sequences.

The role of this kind of sublexical knowledge has been particularly neglected in the word learning literature. We already know that children and adults store sublexical chunks because they use them in non-word repetition tests, in which they have to repeat sequences of syllables (e.g. *dop, te-vack, ver-zer-dut*). Wordlike non-words, which contain phoneme sequences that are frequent in real words, are repeated much more accurately than non-wordlike non-words, which do not contain these sequences (Gathercole, 1995; Jones et al., 2010). This shows that both children and adults' lexicons store sequences of phonemes (sublexical chunks), which can then be retrieved to help them parse, encode, and thus repeat non-words. Given this, it seems likely that children and adults also use such sequences to parse, encode and learn new real words.

Our approach has been used to illustrate how children's early vocabularies are shaped by current knowledge of chunked phoneme sequences. In Jones and Rowland (2017), the model initially benefitted from high levels of repetition in the input, because this allowed it to quickly build up a store of frequently occurring lexical and sublexical (phoneme sequence)

patterns in its lexicon. Once it had learned most of the frequently occurring patterns though, repetition yielded no additional advantage. Instead, the model started to benefit from a more lexically diverse input, which allowed it to learn a large number of sublexical sequences which could be used to then code the input utterances (and the lexical items therein) more efficiently. In other words, the definition of optimal input changed over development, depending on the linguistic knowledge that the model possessed, producing effects that mirrored those of Rowe (2012) outlined above (see Bohannon & Hirsh-Pasek, 1984, for similar arguments).

The aim of the present paper was to test whether existing phonological knowledge plays a role in new word learning. We implemented this hypothesis in the same model used by Jones and Rowland (2017; CLASSIC), described above, to model the changing role of the input throughout early development. The key parameters of the model are a) a chunk-based learning mechanism that learns by gradually chunking sequential information in the model's internal representational system on the basis of incoming input and b) a probabilistic processing constraint that means that, on average, only a certain number of chunks can be encoded for any given input.

Here we use the model to test whether existing knowledge at the point of learning, particularly sublexical knowledge in the form of stored sequences of phonemes (chunks), plays a role in new word learning. We first test whether the model successfully (and developmentally) simulates four well-established properties of word learning in human children: the effects of a word's a) length, b) frequency, c) PP and d) ND on the likelihood of it being learned; the latter two of which cannot be explained by models and verbal theories that ignore the role of existing sublexical knowledge in the learning process. We then provide an even more robust test of the model's ability to simulate the mechanism of word learning,

by explicitly testing a new prediction of the model, derived from the model's learning behavior, on children's early productions.

CLASSIC: A computational model that learns phonological, word, and multi-word knowledge

CLASSIC (e.g., Jones, 2016; Jones & Macken, 2015; Jones & Rowland, 2017) implements a chunk-based learning mechanism that has become established as one of the key mechanisms of human cognition (Gobet et al., 2001). CLASSIC begins with one chunk for each of the phonemes in standard British English. It takes as input a corpus of phonemically-transcribed utterances, processing each utterance one at a time. For each utterance, the model learns via two simple processes. First, it encodes the input utterance into as few chunks as possible based on chunks that have been learned thus far. Second, it learns new chunks by joining together adjacent chunks from the encoded utterance. For example, if 'Hello!' is repeated three times as the first three utterances in the corpus, on the first presentation, the model would access the four phonemes that make up the word /h/, /e/, /l/, and /ow/, and chunk up adjacent chunks to learn three new chunks: /he/, /el/, and /low/. The second utterance in the corpus is then processed, and this can now be encoded using only two chunks: /he/ and /low/. A new chunk will then be learned by joining these adjacent chunks to form /helow/. The third presentation of 'hello' would therefore be encoded using only one chunk as the word has now been learned.

Note that the phonemic input is word-delimited. This is a constraint that we have imposed in most of our modeling work examining both early language development and nonsense word repetition performance using CLASSIC (e.g., Jones, 2016; Jones & Rowland, 2017; Jones & Macken, 2018). While other models of word learning also use word-delimited inputs (e.g., Vitevitch & Storkel, 2013), as we saw above, models of word segmentation do

not. Our rationale here is that we compare model performance to children of mean age 1;10-2;10, ages at which the word segmentation process is well-established (children are capable of determining word boundaries via a range of phonetic, phonological and distributional cues by their first birthday, see Rowland, 2014, for a review). Previous derivatives of CLASSIC that examined nonsense word repetition have shown that the inclusion/exclusion of word boundaries only moved the fit between model and child data from within 7% to within 8% (Jones, 2016). For current purposes, removal of the word boundary information would necessitate a mechanistic account of segmentation in order to determine whether or not a learned chunk constitutes a word, and this is beyond the scope of the paper.

There are also two constraints on learning. There is a limit on the number of chunks that can be encoded from an input utterance (on average 4.5 chunks) favoring those at the end of the utterance in line with well-established recency effects (e.g., Grenfell-Essem & Ward, 2012). In addition, word boundaries are not crossed unless the chunks themselves are words, at which point a chunk corresponding to a multi-word sequence would be learned. Other than a 1.00 probability of learning (i.e., learning at every opportunity) and encoding only 4.5 chunks of an utterance on average, there are no further parameters to the model. Learning occurs at every opportunity because the input to the model will be very small compared to the developing child who hears up to half a million utterances in a 3-week period (Swingley, 2007). As one can see in the example (learning the new chunk /helow/ from the encoded chunks /he/ and /low/), for a word to be learned, that word must be encoded using only two chunks. A detailed description of the model can be found in the Supplementary Materials.

Note that *phonological knowledge* in the model acts as a proxy for a range of performance benefits as linguistic exposure increases for the developing child. For example, accurate production of CC and CCC word-onsets increases with the input frequency of those onsets in 0;11-4;0 infants (Ota & Green, 2013) while 18-month-old infants are more able to

detect close mispronunciations of a target referent if they have encountered the referent frequently as opposed to infrequently (Swingley, 2007). These performance differences are captured within the model in terms of learning increasingly larger chunks as exposure to the same phoneme sequence increases. Over time, this means a word and the utterance in which the word appears can both be represented using fewer chunks, with fewer chunks equating to better performance.

Our data

The mother (model input) and child data came from the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001) on CHILDES (MacWhinney, 2000). The corpus consists of 12 mother-child dyads, each containing 34-hour transcriptions from audiotaped interactions between mothers and children, taken twice every three weeks over a period of one year. At the beginning of the study, child mean age was 1;10 (range 1;8-2;0). The children were 6 males and 6 females, and were all first born monolinguals from middle-class families, half from Manchester and half from Nottingham (both UK). Since the data is from CHILDES transcripts, speech errors were at the interpretation of the original transcriber. All utterances were converted to their phonemic equivalent using the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). Utterances containing words that did not exist in the dictionary were omitted. Information on word types and word tokens for each set of utterances is given in the supplementary materials. Each transcript was divided into 20 stages (approximately 1.7 hours of transcript for each stage) so that developmental progression could be examined. For each of the 20 stages, unique word types were extracted (i.e., word types that had not been produced in a previous stage). In line with other corpus analyses (e.g., Maekawa & Storkel, 2006), child acquisition of a word type was assumed to be when the word was first produced in the child's utterances. The Supplementary Materials

provide details of the vocabularies of mothers, children, and models, showing not only that the models more closely match the children than the mothers on which they are trained, but also how both models and children show vocabulary effects that are consistent with previous findings.

Analysis strategy

One run of the model was undertaken for each individual set of maternal utterances[1], resulting in a total of 12 model simulations. The same modeling environment was used for each simulation therefore any effects seen are solely attributable to differences in the linguistic input. The model's vocabulary learning was examined after every 5% increment of the input to enable direct comparison to the 20 stages of the child utterances.

Data manipulation and analysis were carried out using the R programming language (version 3.4.3; R Core Team, 2017) and RStudio IDE for R (version 1.1.423; RStudio Team, 2016). At each stage, for the mother and child utterances, word types that had not appeared in any previous stage were recorded. For the model, words learned were recorded at every 5% increment of the input. As is often the case when dealing with large input samples (e.g., Hart & Risley, 2003; van Heuven et al., 2014), every unique word was treated as a word type (e.g., *Bert, Bertie, ring, ringing* were treated as separate word types) except for plural nouns which were ignored (e.g., *balls*, *chickens*)[2].

Word frequency, ND, and PP were computed using two different corpora: (a) the spoken part of the British National Corpus (BNC, 2007), transcriptions of unscripted informal conversations balanced for age, region, context and social class, containing approximately 10

[1] Since all the inputs for the model come from the mothers of the children, who were also their primary caregivers, we use maternal and mother throughout the article. This is not intended to imply that all caregiver speech comes from the mother.

[2] For example, *ball* and *balls* were treated as one word type (*ball*). This is only for our analysis – the input to the model still distinguished plurals from singulars.

million word tokens; and (b) the maternal corpus outlined above. As per the mother and child data, all BNC utterances were converted into their phonemic equivalent using the CMU Pronouncing Dictionary. We used two corpora because previous research examining effects of word frequency, ND and PP have used both adult corpora (e.g., Storkel, 2009) and caregiver corpora (e.g., Swingley & Humphrey, 2018).

For both corpora, word frequency, ND, and PP were computed only for words that existed in the dictionary. ND and PP were computed using the same formulae used in the Irvine Phonotactic Online Dictionary (version 2.0; IPhOD, Vaden, Halpin, & Hickok, 2009). ND was consistent with numerous other studies (e.g., Storkel, 2009; Swingley & Humphrey, 2018) and defined as unstressed phonological ND, referring to the number of words that differ from a given word by one phoneme (i.e., by the addition, deletion, or substitution of one sound, e.g. *pin/spin, bit/bat*). PP was unstressed word-average biphone probability, consistent with numerous previous studies (e.g., Storkel, 2003; Storkel & Lee, 2011; Vitevich & Luce, 1998, 1999) and defined as the weighted likelihood of occurrence of ordered phoneme pairs that are present in a given word (i.e., accounting for the frequency of the biphones).

Experiment 1

The aim of experiment 1 was to determine whether the model can simulate four robust properties of human word learning; frequency, word length, ND and PP. The first two are well-established effects that many models simulate successfully, but are important to establish the plausibility of our simulation of word learning. With regard to **frequency**, all else being equal, young children tend to acquire high frequency members of a word category before low frequency members, and these effects are seen across a number of categories

(verbs, nouns, adjectives, closed-class words etc.; Huttenlocher et al., 1991; Naigles & Hoff-Ginsberg, 1998; Goodman, Li & Dale, 2008; Swingley & Humphrey, 2018; for a review, see Ambridge et al., 2015). With regard to **word length**, shorter words tend to be learned before longer words (Maekawa & Storkel, 2006; though note that there can be large variations in phonemic length even for monosyllabic words; Coady & Aslin, 2003). Furthermore, word length effects are seen in tasks that bear strong relationships with vocabulary learning, such as non-word repetition tasks where, even in 2 to 4 year old children, there is a significant decline in performance as length increases (Roy & Chiat, 2004). The same pattern is observed in almost every study, though most involve older children (e.g., Gathercole & Baddeley, 1989; Jones et al., 2010).

The latter two effects constitute strong tests of our hypothesis that children use existing sublexical material stored in the lexicon to learn new words (effects that cannot be simulated by models and verbal theories that do not posit a role for linguistic knowledge at the sublexical level)[3]. For **ND,** infants are more likely to learn words with high NDs (words having a high number of phonologically similar words) than those with low NDs (Hollich, Jusczyk and Luce, 2002; Storkel, 2009). We first determined what ND effects need to be explained in our data (i.e., do the children in our sample learn high or low ND words first) and then tested whether our model can simulate these data. In addition, since words with high ND tend to be shorter than words with low ND (Storkel, 2009; Vitevitch & Luce, 2016) we tested both the model's ability to simulate raw ND effects, and its ability to simulate the ND effects that emerge when we control word length, to ensure we are not simply modeling ND

[3] Effects of PP require sublexical representations beyond individual phonemes because differences are seen between words containing relatively familiar sound sequences versus relatively unfamiliar sound sequences. We argue that effects of ND are also likely to require sublexical representations beyond individual phonemes. While one could determine neighbor words via a phoneme-by-phoneme comparison across words, this method would be highly inefficient. While there are some models (e.g., TRACE, McClelland & Elman, 1986) that simulate PP and ND effects while only explicitly encoding phonemes and words, there has to be some form of (implicit or explicit) knowledge of phoneme sequences (e.g., for TRACE, "each [phoneme/word] unit stands for a hypothesis about a particular perceptual object occurring at a particular point in time defined relative to the beginning of the utterance" (p. 8).

effects as a by-product of an ability to simulate word length effects. (Effects of ND when controlling for frequency can also be found in Supplementary Materials).

Finally, we modeled **PP** effects. Interestingly, the literature is not clear-cut on whether low or high PP words are learned first. Very young infants (M = 38 weeks, range 37-40) prefer to listen to nonsense words comprised of high PP CVC sound sequences (words with sound sequences that are high frequency in the language; Jusczyk, Luce & Charles-Luce, 1994), suggesting a high PP word advantage. However, the early productive vocabularies of infants (M = 1;11, range 1;4-2;6) tend to be comprised of low PP words (Storkel, 2009). We additionally examined PP when controlling for word length to be consistent with the ND data (see Supplementary Materials for PP effects when controlling for frequency).

Prior to plotting the data, we first determined whether growth patterns for children and models for each of the four measures were better explained by either a linear or logarithmic fit using linear mixed-effect modelling. As the different proportions of word types (outcome) must sum to 1, this variable was recoded. We took one of the levels (e.g., three-syllable group in the word length analysis) as a reference level, calculating the natural logarithm of each remaining proportion relative to the proportion of the reference level. Taking the word length analysis as an example, the recoded predictor variable was computed as:

$$ln_1 = log\ (\frac{Pr(\text{one syllable})}{Pr(\text{three syllable})}) \qquad ln_2 = log\ (\frac{Pr(\text{two syllable})}{Pr(\text{three syllable})})$$

To examine whether the data would be better represented by a linear or logarithmic fit, we compared two multilevel linear models using the bootstrap 95% confidence interval around the difference in Adjusted R^2 ($\Delta AdjR^2$). The confidence intervals were based on 1,000 iterations and adjusted for multiple comparisons using Holm's correction. If a confidence interval contained 0, we then concluded that no significant difference between linear and logarithmic fit was found, indicating that a parsimonious linear fit better represented children and models' growth patterns.

Each multilevel linear model predicted ln_k as a function of a) *stage* or *log(stage)*, b) *k* (where *k* is 1 or 2), c) whether the data is from the child or model (*set*), and d) the interactions between these three predictors. We included random intercept and slope for child *id*, as influenced by *stage* /*log(stage)*, *set* and *k*. As shown in Table 1, a linear fit better represented the data only in the case of PP, and only when this measure was based on the Spoken BNC.

Table 1. Difference in Adjusted R^2 between linear and logarithmic fit, for each of the four measures and for each corpus that the measures were computed from.

Linear vs. Logarithmic	Corpus	$\Delta AdjR^2$
Length	BNC	-0.049 [-0.083, -0.025]
Frequency	BNC	-0.034 [-0.047, -0.026]
Neighbourhood density	BNC	-0.053 [-0.07, -0.039]
Phonotactic probability	BNC	-0.006 [-0.02, **0.004**]
Length	Maternal	-0.049 [-0.084, -0.024]
Frequency	Maternal	-0.016 [-0.03, -0.005]
Neighbourhood density	Maternal	-0.018 [-0.028, -0.014]
Phonotactic probability	Maternal	-0.018 [-0.038, -0.001]

Figure 1 shows the results for children and models for all four properties when measures are based on the spoken BNC and Figure 2 is when measures are based on the maternal corpus. Despite the four measures being computed using two different corpora, the children and models are very similar to one another in both Figures. To examine whether growth patterns for children and models differ statistically from one another, we were specifically interested in two interaction terms of the full model previously described: the two-term interaction between *k* and *set*, and the three-term interaction between *log(stage)/stage*, *k* and *set*. Taking the word length analysis as an example, the first

interaction term tells us whether, at each length, children and models' proportional means of word types differ statistically. In other words, this term indicates whether, at each length, children and models' curves differ in terms of average height on the y axis. Further, the second interaction term tells us whether children and models' curves differ in slope.

We compared a full model which excluded the above interaction terms to a full model, using the bootstrap 95% confidence interval around the differences in AIC (ΔAIC) and BIC (ΔBIC). Confidence intervals were adjusted using Holm's correction. If the confidence interval contained 0, we concluded that only a negligible change in model fit was found between child and model data. We chose to rely on confidence intervals rather than *p* values of Chi-square difference tests applied to the nested models because the latter are directly affected by sample size (for large samples trivial differences may become significant).

As shown in Table 2, the confidence intervals for ΔAIC and ΔBIC contained 0 for every measure when the Spoken BNC is used, indicating that adding the two interaction terms of interest did not provide a significantly better fit to the data. In other words, children and models' growth patterns did not differ in average height and slope of the fitted curves/lines. The same is true when the maternal corpus is used, except for ND – in this instance, further examination (see Supplementary Materials) shows that children and models differ only for the lowest ND tertile, suggesting that the significant difference found is probably of low practical significance.

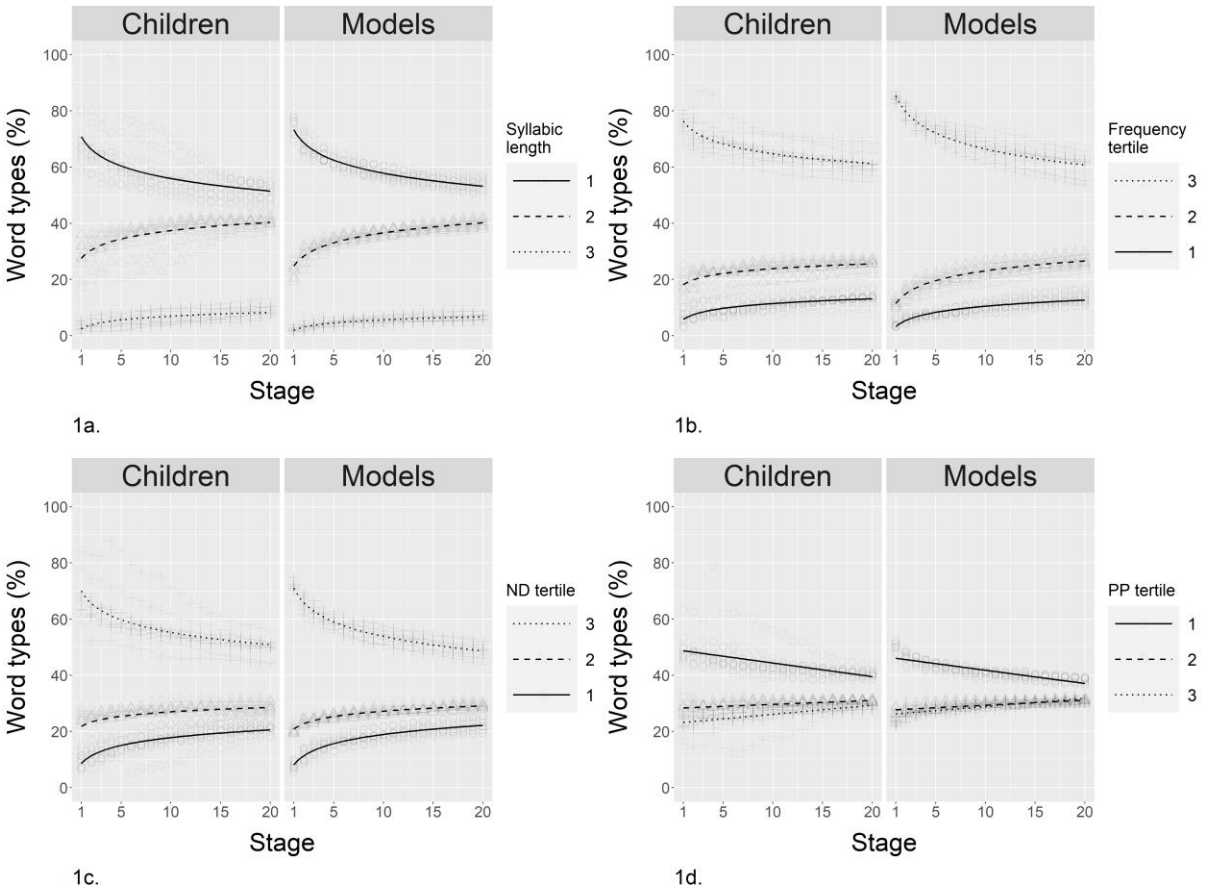


Figure 1a-d. Children and model comparisons when measures are computed using the spoken BNC. Each graph shows the proportion of new (unique) words produced at each stage (i.e., those words produced at a stage that were not present in any previous stage) of the child transcriptions and model's learning, by a) syllabic length, b) frequency tertile, c) ND tertile and d) PP tertile across the two groups. For tertiles, 1 = lowest 33%, 2 = 33-67%, and 3 = highest 33%. Each datapoint represents a different child/model, with regression lines provided as summaries for each variable and group. Proportions are calculated over the overall number of words produced by each subject/model at each stage. Logarithmic curves better represented length, frequency and ND growth patterns. A linear fit better represented the PP growth patterns.

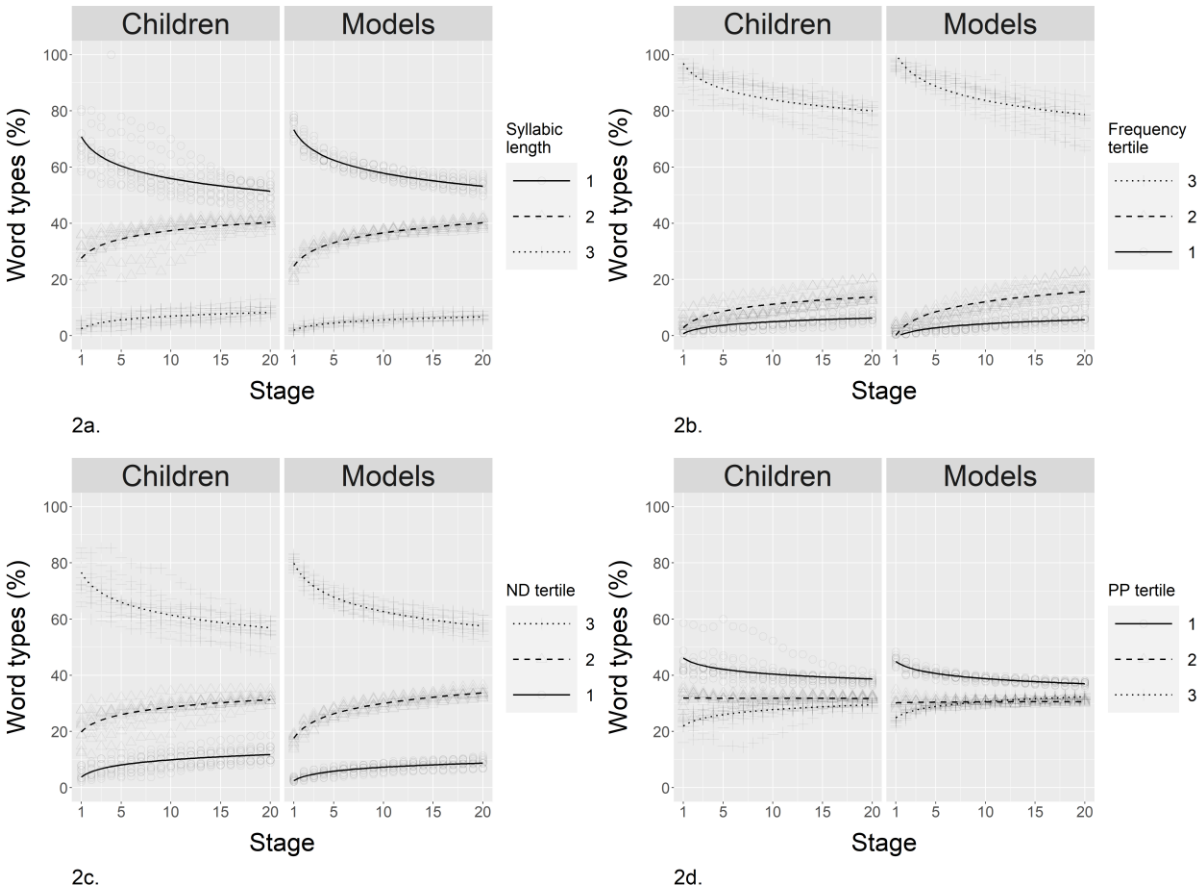


Figure 2a-d. Children and model comparisons when measures are computed using the maternal corpus. Each graph shows the proportion of new (unique) words produced at each stage (i.e., those words produced at a stage that were not present in any previous stage) for the child transcriptions and model's learning, by a) syllabic length, b) frequency tertile, c) ND tertile and d) PP tertile across the two groups. For tertiles, 1 = lowest 33%, 2 = 33-67%, and 3 = highest 33%. Each datapoint represents a different child/model, with regression lines provided as summaries for each variable and group. Proportions are calculated over the overall number of words produced by each subject/model at each stage. Logarithmic curves better represented length, frequency, ND and PP growth patterns.

Table 2. Model and child comparison of growth patterns for each of the four measures and for each corpus (BNC or maternal corpus), using the best fit (linear or logarithmic) as determined in Table 1.

Child vs Model	Type	Fit	ΔAIC	ΔBIC	ΔDeviance	$\Delta AdjR^2$
Length	BNC	Log	9 [-3, 25]	0 [-13, 15.025]	13 [1, 29]	-0.001 [-0.003, 0]
Frequency	BNC	Log	21 [-3, 88.538]	11 [-13, 79.05]	25 [1, 92.538]	-0.001 [-0.005, 0]
Neighborhood density	BNC	Log	21 [-3, 113.025]	11 [-13, 103.025]	25 [1, 117.025]	-0.001 [-0.008, 0]
Phonotactic probability	BNC	Lin	22 [-3, 96.513]	13 [-13, 86.756]	26 [1, 100.513]	-0.002 [-0.014, 0]
Length	Maternal	Log	9 [-3, 27.513]	0 [-13, 17.513]	13 [1, 31.513]	-0.001 [-0.004, 0]
Frequency	Maternal	Log	127 [9.325, 203.05]	117 [-0.675, 193.375]	131 [13.325, 207.05]	-0.006 [-0.017, -0.001]
Neighborhood density	Maternal	Log	**442 [116.438, 535.294]**	**432 [107.194, 526.294]**	**446 [120.438, 539.294]**	-0.012 [-0.023, -0.005]
Phonotactic probability	Maternal	Log	-2 [-4, 43]	-12 [-14, 33]	2 [0, 47]	0 [-0.006, 0]

Having established that there is almost no difference in growth patterns between children and models across all four measures – regardless of whether the measures are computed using the spoken BNC or maternal input – we now examine what these growth patterns tell us. Since Figures 1 and 2 show overall trends across the graphs that are broadly similar across all four measures, hereafter we use the spoken BNC data (Figure 1).

Figure 1a illustrates the effect of length: the proportion of new (unique) words produced at each stage (e.g., those words produced at stage 3 that were not present in any previous stage) that are one-syllable, two-syllable, and three-syllable at each stage of the child transcriptions and model's learning. Figure 1b shows the same proportions by frequency tertile, calculated by grouping all word types produced by children and models, deriving the frequency of each word from the spoken BNC, and then splitting the words equally into tertiles on the basis of their frequency (1 = lowest 33%, 2 = 33-67%, 3 = highest 33%).

Both the models and the children show the predicted effect. Early on in development, the children produce and the models learn, more short word types than long word types, and more high frequency than low frequency words (e.g. at stage 1, 68% of children's words are

one syllable vs. 74% of the model's). For an analysis of length in number of phonemes, which yields similar results, see Supplementary Materials.

Figures 1c and 1d show the proportion of new (unique) words produced at each stage by ND tertile (1c) and PP tertile (1d). To calculate the ND and PP tertiles, we grouped all word types produced by children and models, derived NDs and PPs for each word by applying the IPhod formulae to the spoken BNC and then split words equally into tertiles on the basis of their ND or PP scores (1 = lowest 33%, 2 = 33-67%, 3 = highest 33%). Children produce a greater proportion of high ND than low ND word types, and a greater proportion of low PP than high PP word types (consistent with Storkel, 2009, and Storkel & Lee, 2011).

The fact that the model simulates the proliferation of low PP word types produced by the children even though learning is incremental and based on exposure to the maternal input, is surprising. Although the model does not record word frequency per se (once a word is learned, the only further learning is for multi-word sequences involving the word), more phonological knowledge will be learned for those phoneme sequences that appear often, or across numerous words, in the input. This suggests the reverse effect should be seen – words should be more easily learned when they contain high PP sequences because those sequences appear often (and/or across different words). We thus examined why we simulated the low PP advantage by looking at how PP changes with phonemic word length.

Figure 3a shows the PP tertile breakdown across phonemic length (we focus on two to six phonemes because children's productive vocabularies do not contain many words of seven phonemes or more). Crucially, here, the proportion of word types is calculated by grouping all words together regardless of length, partitioning them into PP tertiles (T1 = lowest 33%, T3 = highest 33%), and then calculating the proportion of low/medium/high PP words at each phonemic length. It is clear that the vast majority of low PP word types are three or fewer phonemes in length, which explains their ease of learning in the model. Since

the majority of early word types produced are short (see Figure 1a) and the majority of short words are of low PP, the model learns low PP words first, not because they are low PP but because the model learns short words more quickly than long words. This explains why the model appears to learn low PP words first, even though its learning algorithm favors high PP words over low PP words.

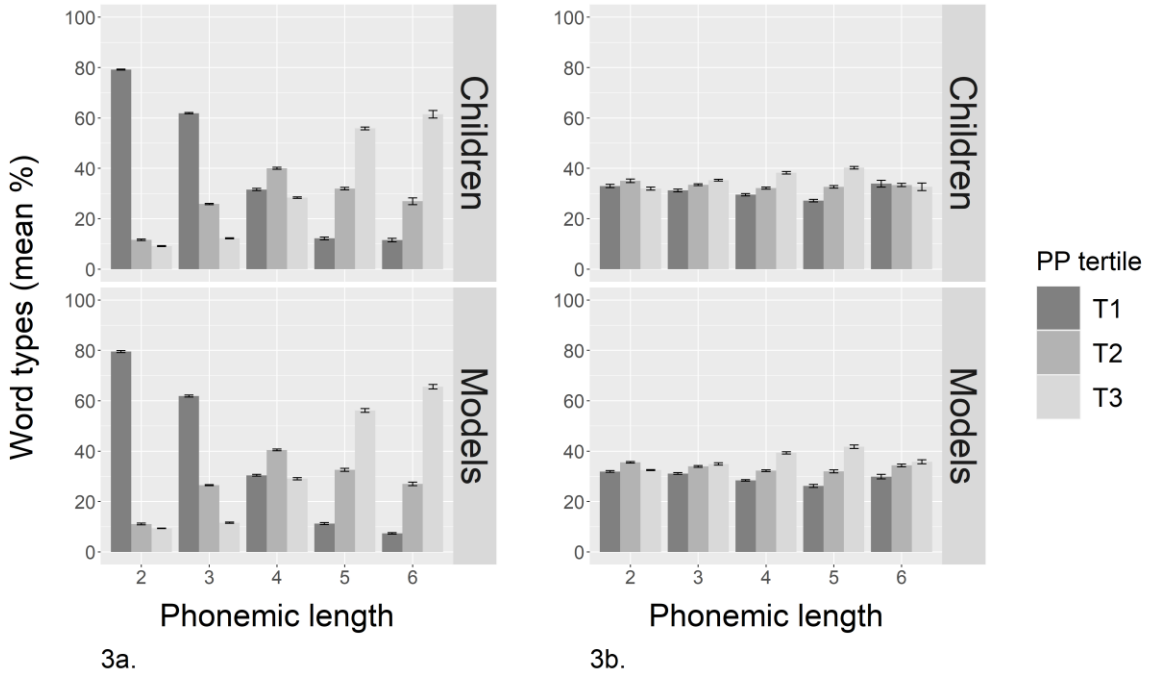
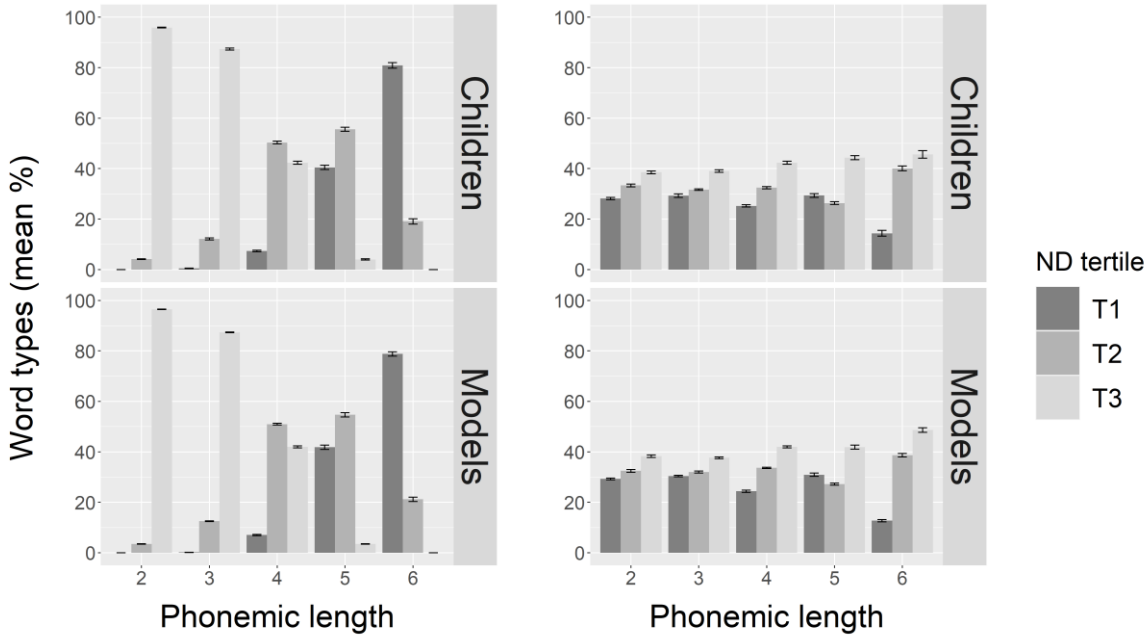


Figure 3. Mean proportion of word types produced by children (top) and learned by models (bottom) for each PP tertile by phonemic length. In figure 3a) Tertiles are defined by grouping all words together regardless of length and then partitioning them into tertiles. In figure 3b) Tertiles are defined by grouping all words of a particular phonemic length and then partitioning them into tertiles. In both figures T1 = lowest 33%, T3 = highest 33%. Error bars indicate the Standard Error of the mean.

When we repeated the analysis but this time controlling for length by considering only words of a particular length when computing the proportion of low/medium/high PP words (e.g., considering only three phoneme words when calculating tertiles for three phoneme

words), a different pattern emerges (figure 3b). Now we find a high PP advantage for both children and models for the majority of phonemic lengths. That is, the models tend to learn, and the children tend to produce, more high PP words than low PP words once word length is taken into account.

Figure 4 illustrates children's word productions and words learned by the models for different ND tertiles split into different word length tertiles; both before and after controlling for word length. As with PP above, we see an interaction between ND and word length (figure 4a). For both children and models, short words tend to be from dense neighborhoods and long words tend to be from sparse neighborhoods. Figure 4b shows that when we control for word length (e.g., considering only four phoneme words when computing tertiles for four phoneme words), there is a consistent advantage for high ND words, in both the children and the models.



4a. 4b.

Figure 4. Mean proportion of words produced by children (top) and learned by models (bottom), for each ND tertile by phonemic length. In figure 4a) tertiles are defined by grouping all words together regardless of length and then partitioning them into tertiles. In figure 4b) tertiles are defined by grouping all words of a particular phonemic length and then partitioning them into tertiles. In both figures T1 = lowest 33%, T3 = highest 33%. Error bars indicate the Standard Error of the mean.

In sum, the aim of study 1 was to determine whether the model could simulate four properties of human infant word learning, two of which (PP and ND) cannot be simulated by models or verbal theories that do not implement a role for sublexical knowledge in new word learning. The model simulated all four effects, including an unexpected effect in which both children and models seemed to favor low PP words over high PP words early in development. However, this was due to the fact that low PP words also tend to be short (figure 3a). When we controlled for word length, there tended to be an advantage for high PP words in both children and models. There was also a change in patterns for ND: ND declines as length increases, but once length is controlled (i.e., separating ND tertiles when only considering words of a particular length), there is a general advantage for words having high NDs. In the Supplementary Materials we also show how the models simulate children's word learning for interactions of PP and frequency and ND and frequency.

Thus, the effects of word length, word frequency, PP and ND, together with interactions between PP/ND and word length and between PP/ND and word frequency, are simulated by a model that assumes children gain greater phonological knowledge from their experience with words, and use this knowledge in new word learning. All four effects can be explained in terms of how phonological knowledge is represented as chunks of sound sequences of varying lengths in the model.

Experiment 2

In the CLASSIC model, learning is faster for those words that can be represented using a few long chunks (reflecting prior learned knowledge of long sound sequences) as opposed to many short chunks (reflecting only knowledge of short sound sequences), even when the words are of the same length. This is because there are fewer chunks involved in ultimately learning the whole word. Thus, the model predicts that children will find it easier to learn new words that can be represented by long sound sequences that already exist in the child's phonological repertoire. Work on nonsense word repetition supports this view, at least for individual sounds, where performance is superior for nonsense words that comprise sounds that have been attested in children's prior productions versus those that do not (e.g., Keren-Portnoy et al., 2010; Schwartz & Leonard, 1982).

In experiment 2, we therefore tested whether the size of the phoneme sequences that have been previously attested in the child's productions influence how likely they are to learn, and thus produce, a new word. The maternal input contains words that children subsequently produce (i.e., there is definitive proof they have been learned) and also words that children do not produce (i.e., from the perspective of the transcripts alone, we have no evidence that they have been learned). Our hypothesis is that the probability of a child producing a newly presented word in their input increases in step with the size of the phoneme sequences that are shared between the word and other words that have been produced by the child; in other words, longer previously produced phoneme sequences should be better able to "scaffold" the production of new words. Our prediction, therefore, is that those words that children produce from their maternal input will share larger phoneme

sequences with other words in the child's productive vocabulary than those words from the maternal input that are not produced by the child.

We used the child data described in experiment 1 above, together with the mother data that was previously used to train the models. We first extracted all nouns that were used in the maternal input. We selected nouns because they are the most common word type produced by the children and also the most concrete word category (e.g., relating to objects), and because they constitute a strong test of our hypothesis because other aspects of the communication exchange (e.g., gestures, eye movements, semantics) are very likely to play a role in learning. We split the monosyllabic, bisyllabic and trisyllabic nouns[4] produced in the maternal input into those that the child also produced (hereafter *produced-nouns*) and those that the child did not produce (hereafter *not-produced-nouns*). Since produced-nouns also tend to be of higher frequency in the input than non-produced-nouns, we removed all high frequency produced-nouns: starting from the most frequent, nouns were excluded from the produced set until the produced set was equal in mean frequency to the set of not-produced-nouns. In all analyses, noun plurals were excluded (e.g., *ball* and *balls* were analyzed as one word type, *ball*).

For each of the produced-nouns, we extracted all possible phoneme sequences of lengths 2, 3, 4 and 5 phonemes, where applicable (henceforth *phoneme sequences*). We then extracted every word containing those phoneme sequences that was produced by the child on, or prior to, the stage at which the produced-noun was first produced (henceforth, *scaffolds*). For example, if Anne first produced *teddy* (/t/eh/d/iy) at stage 5, we considered as phoneme sequences /teh/, /ehd/, /diy/, /tehd/, /ehdiy/, and /tehdiy/. All words produced by Anne in stages 1-5 that contained these phoneme sequences were then included as scaffolds for *teddy*. We include the stage at which the produced-noun was first produced so that we could include

[4] There were very few nouns longer than three syllables.

all relevant scaffolding words but we exclude every instance of the produced-noun itself. Since not-produced-nouns will not have a stage at which they were produced, we considered all scaffolds that were produced on or before the average stage of the first production of nouns in the produced-nouns set.

Produced-nouns can be acquired at different stages, so we needed to compute the proportion of scaffolds relative to all available word types. We did so by dividing the number of unique scaffolds by the total number of word types produced up to the noun production stage and computing an average proportion of scaffold types across produced-nouns for each child. We calculated two dependent variables - the number of different scaffolds in which phoneme sequences appeared (type frequency) and the frequency of the scaffolds (token frequency) – to see whether the number of different words in which scaffolds were produced was a key driver and/or the raw frequency by which scaffolds were produced. Note that token frequency was not double-counted (e.g., *car* shares with *card* two different phoneme sequences of length 2 phonemes, so only one of these was used in the calculation of token frequency for phoneme sequences of length 2 phonemes).

We present data divided by word length to control for word length effects. For reference, the mean number of produced-nouns and not-produced-nouns (after matching for noun frequency) was 138 and 154 for monosyllabic nouns; 158 and 220 for bisyllabic; and 50 and 108 for trisyllabic; the mean phonemic word lengths were 3.3 and 3.4, 5.1 and 5.4, and 7.5 and 7.5 for monosyllabic, bisyllabic and trisyllabic produced-nouns and not-produced-nouns respectively.

Figure 5 shows the mean proportion of unique scaffolds (number of scaffold types) in children's vocabulary sharing a phoneme sequence of length 2 to 5 phonemes with the produced-nouns and not-produced-nouns. Table 3 shows effect sizes and statistical comparisons for these data. As one can see, the produced set of nouns is almost always larger

than the not-produced set when the noun shares a phoneme sequence of length three or more with scaffolds (Figure 5) and the effect size is large for all but one comparison (Table 3). Significant effects (noting the small sample size) are also seen when scaffolds shared at least a four phoneme-long sequence. As predicted, produced-nouns shared longer phoneme sequences with other words in the child's productive vocabulary than not-produced-nouns, and this held even for short (monosyllabic) nouns (see figure 5a).

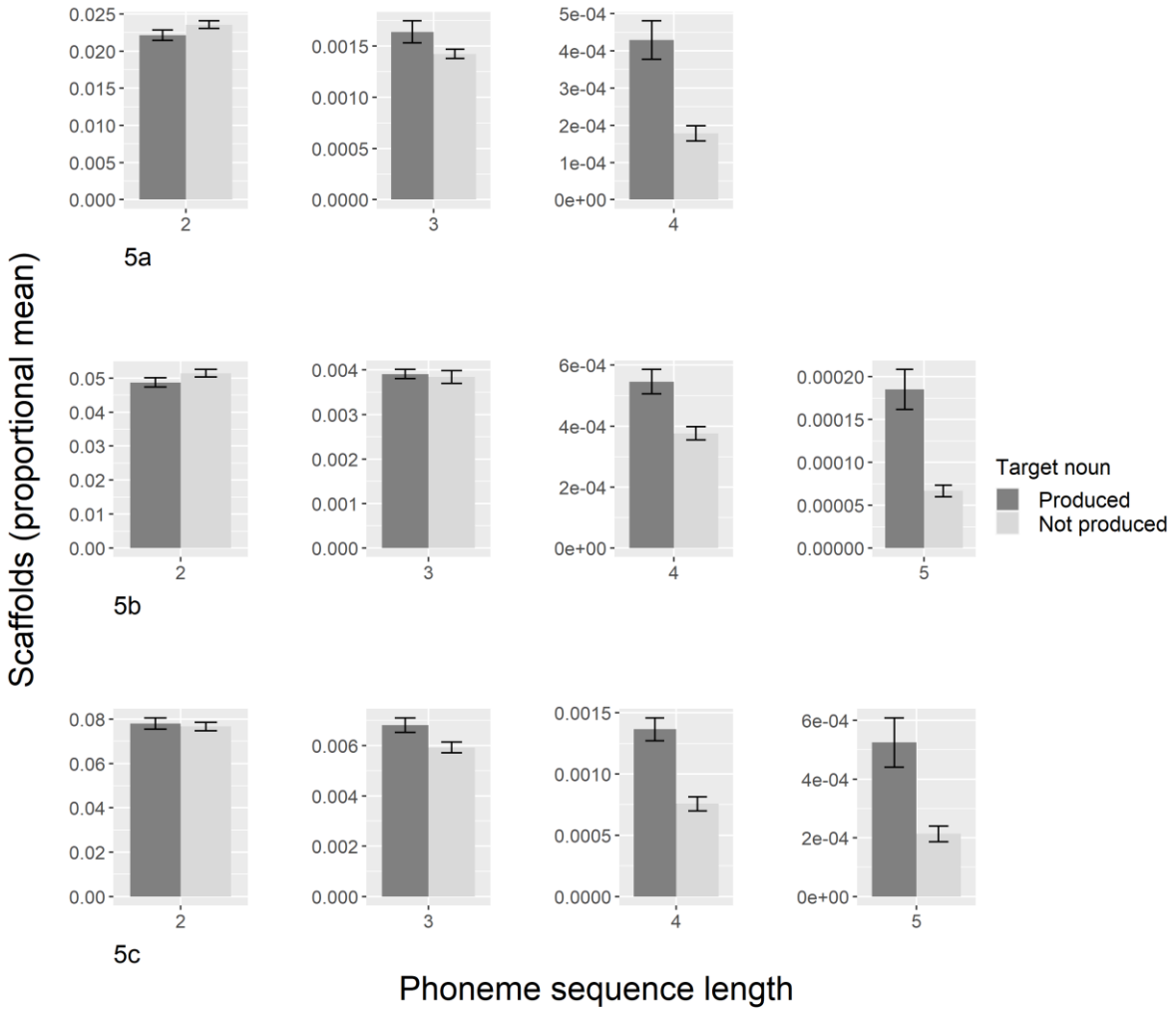


Figure 5. Mean proportion of unique scaffolds (y axis) sharing phoneme sequences of lengths 2 to 5 phonemes (x axis) with a target noun from the maternal input that was either produced or not produced by children (N = 12). Error bars show the standard error of the mean. Data is shown for a) monosyllabic, b) bisyllabic and c) trisyllabic target nouns. Note: There were too

few scaffolds containing phoneme sequences of 5 phonemes for monosyllabic target nouns (5 in the produced set and 8 in the not produced set).

Figure 6 shows the same data but calculated for scaffold tokens, with effect size and statistical comparisons shown in Table 3. Interestingly, differences between the produced and not-produced sets of nouns tend to occur at a longer phoneme sequence length than for the word type data above. Indeed, large effect sizes are only seen when the phoneme sequences shared between nouns and scaffold tokens are long (see Table 3), and while there were numerical differences for the frequency of scaffolds, all but one were not statistically significant (see Table 3). This suggests the number of *different* scaffold words sharing a phoneme sequence with a produced-noun is more critical than the raw frequency with which the scaffolding phoneme sequence has been produced by the child.

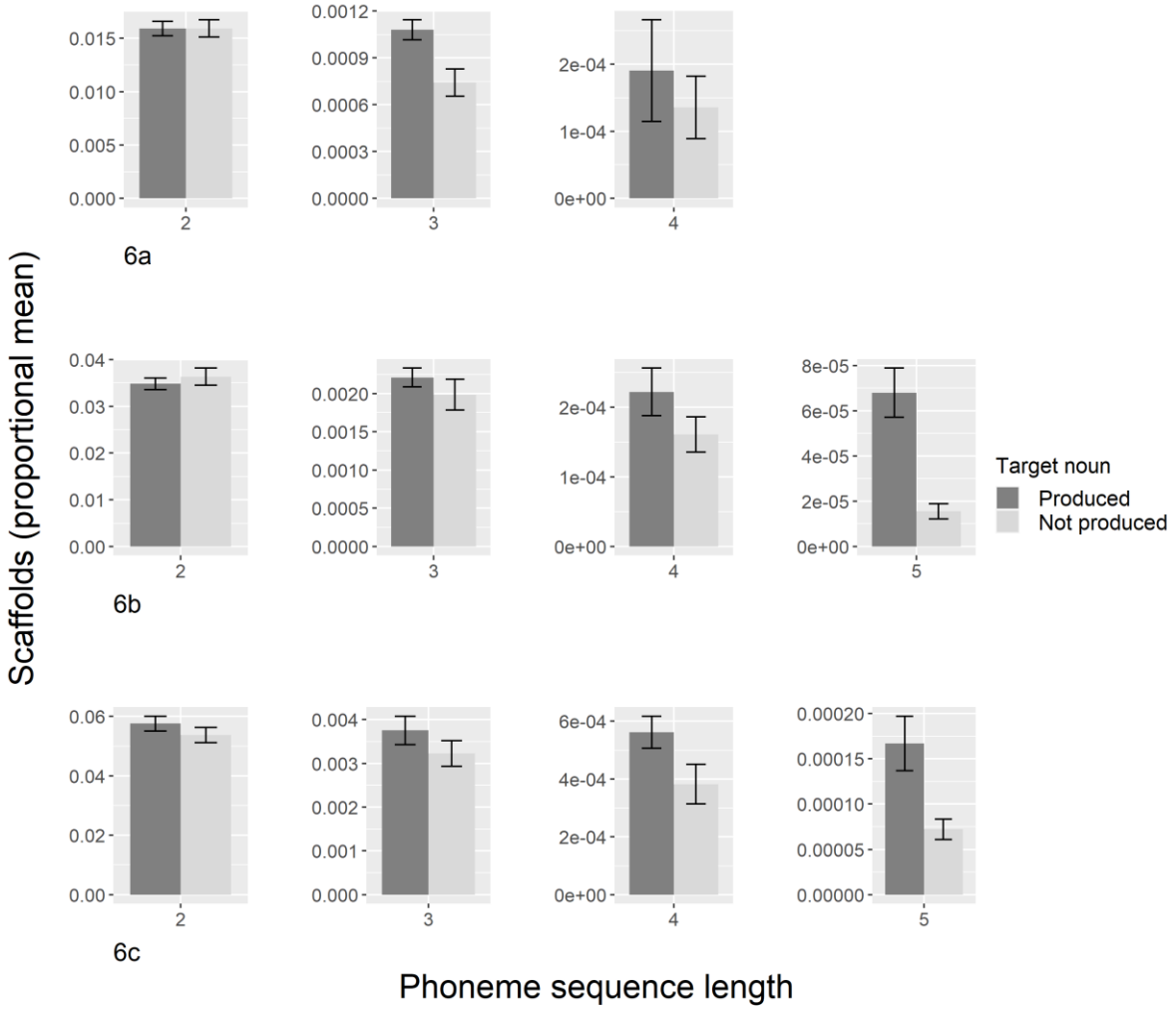


Figure 6. Mean proportion of scaffold tokens (y axis) sharing phoneme sequences of lengths 2 to 5 phonemes (x axis) with a target noun from the maternal input that was either produced or not produced by children (N = 12). Error bars show the standard error of the mean. Data is shown for a) monosyllabic, b) bisyllabic and c) trisyllabic target nouns. Note: There were too few scaffolds containing phoneme sequences of 5 phonemes for monosyllabic target nouns (5 in the produced set and 8 in the not produced set, see Supplementary Materials).

Table 3. Comparisons between produced-nouns and not-produced-nouns of different syllabic lengths and for scaffolds of 2, 3, 4, and 5 phonemes in length, for scaffold types and scaffold

tokens. A scaffold length of 5 is omitted for monosyllabic targets due to the low number of applicable scaffolds. Holm's correction for multiple comparisons was applied. Hedges' *g* was used to compute effect sizes due to our small sample size and was computed using the R package *effsize* (Torchiano, 2020). In this package, the magnitude of the effect size is assessed using the thresholds provided in Romano et al. (2006), considering the absolute value of *g*: < .15 "negligible", < .33 "small", < .47 "medium", otherwise "large").

Syllabic length	Sequence length	Scaffold types			Scaffold tokens		
		t	*p*	*G*	*t*	*p*	*G*
1	2	-1.66	.567	-.68	.00	1	.00
1	3	1.83	.521	.75	3.11	.061	1.27
1	4	4.51	.005	1.84	.62	1	.25
2	2	-1.52	.569	-.62	-.68	1	-.28
2	3	.37	1	.15	.95	1	.39
2	4	3.72	.015	1.52	1.44	1	.59
2	5	4.87	.003	1.98	4.62	.006	1.89
3	2	.43	1	.18	1.08	1	.44
3	3	2.50	.147	1.02	1.20	1	.49
3	4	5.60	< .001	2.29	2.04	.486	.83
3	5	3.54	.029	1.44	2.94	.106	1.20

Discussion

We have demonstrated that a model which learns words by building up chunks of phoneme sequences can not only model the changing role of the input throughout development (Jones & Rowland, 2017), but can also successfully simulate four key features of the word learning

process (frequency, word length, PP and ND effects). In addition, the model – to our knowledge – makes a unique prediction about which words are likely to be produced, based on the length, and number of phoneme sequences that it shares with previously produced (scaffolding) words, a prediction that was upheld in the children's data. Taken together, these results provide strong support for the hypothesis that the phoneme sequences that already exist in a child's phonological repertoire play an important part in new word learning.

These findings advance knowledge in four important ways. First, they suggest that the amount and type of knowledge already acquired at the point of learning has a major influence on how the child's input is processed, and thus how new knowledge is acquired and integrated into long-term memory (for similar ideas about developmental cascades, see Karmiloff-Smith, 1998). Thus, theories that do not incorporate a role for existing, accumulating knowledge in driving developmental change are missing an important driver of acquisition. Second, they suggest that to simulate developmental changes in word learning, we need models that store linguistic knowledge at the sublexical, as well as the lexical level. Models that store linguistic knowledge only at the whole word level are unlikely to be able to explain the PP and ND effects we simulate here. Third, the findings demonstrate that the model's learning architecture does not need to be complex to achieve these effects. CLASSIC has a relatively simple architecture; its key parameters are a chunk-based learning mechanism that learns by gradually chunking information in the model's internal representational system on the basis of incoming input and a probabilistic processing constraint such that, on average, only a certain number of chunks can be encoded for any given input. Fourth, our results question the utility of PP and ND as indices of phonological knowledge and word learning more generally, showing instead that it is the (varied) grain size of phonological knowledge at the point of learning that is important (see also Jones, 2016; Szewczyk et al., 2018).

All the effects we see here fall out of two simple processes: a sequential learning mechanism operating on natural language input and a constraint on the amount of information that can be learned at any one time. Since phonological knowledge acquisition involves the gradual accumulation of larger and larger chunks of phoneme sequences in the model, short words are more likely to be learned before long words. This accumulation proceeds more rapidly for highly frequent words despite the fact that the model does not record word frequency per se (once a word is learned, the only further learning is for multi-word sequences involving the word), because more phonological knowledge is learned for those phoneme sequences that appear often, or across numerous words, in the input. Words with high PP are learned more quickly than words with low PP, once we control for word length, because their constituent sound sequences also accumulate more rapidly with increasing frequency of encounter. Words with high ND share a greater number of sound sequences with other words than do words with low ND and therefore phonological knowledge is learned more quickly for high ND words. In addition, the model's learning, like the child's, is affected by the interaction between these effects. When we do not control for length, we see that words with low PP seem to be acquired first both by the model and the child, simply because such words tend to be short. This also provides an explanation for apparent contradictions in the literature: studies that find a high PP advantage tend to focus on words of a particular length (e.g., Jusczyk, Luce & Charles-Luce, 1994) whereas those that find a low PP advantage test words at all lengths (e.g., Storkel, 2009).

Our view is broadly consistent with several other views on children's phonological word learning, although it differs in the source of such learning. The lexical restructuring hypothesis, for example, suggests that early word learning is holistic with segmental detail emerging based on exposure to words that sound similar to one another (i.e., neighborhood). The same end point is reached by CLASSIC but from bottom-up rather than top-down

learning, since segmental detail is gradually built up over time in the form of phoneme sequence chunks, with larger chunks being learned where the model has been exposed to sequences appearing in many words (e.g., neighbors). Our view is also consistent with those models of word segmentation that give chunk-based explanations for the identification of word boundaries.

It is important to highlight that our view dispenses with traditional definitions of ND and PP[5], viewing both of these as operating at the wrong grain size (word-level and biphone-level respectively). Rather, both ND and PP largely depend upon the size of chunked phoneme sequences, which in turn depends upon the extent of exposure to such sequences, be it from one word or many words. Both Schwartz and Leonard (1982) and Keren-Portnoy et al. (2010) have shown how nonsense words are produced more easily by infants when the nonsense words comprise sounds that have been attested in the infant's productions, while work on older children shows that prior knowledge of sound *sequences* also influences production (e.g., Gathercole, 1995; Jones & Macken, 2018). In order to explore our view more fully, one could create nonsense words that hold ND, PP, and word length constant while manipulating the extent to which the nonsense words share phoneme sequences with other words. The expectation here is that the greater the number of words sharing a phoneme sequence, the easier the nonsense word will be to learn. If our view is supported, then it follows that children's vocabulary learning can be facilitated by learning particular words that are 'hubs', or words that have sound sequences shared by many other words, since learning these words would help to subsequently learn other related words.

It is also vital to acknowledge some of the limits of our model. CLASSIC implements just one of many potential models of word learning, so future work will need compare the

[5] Here we refer to the well-used definitions of a neighbor being a word that differs by the addition/substitution/deletion of one phoneme and PP being calculated based on biphones.

predictions of our model, in which the input interacts with a chunk-based learning mechanism, with models that implement other types of learning mechanisms. For example, Storkel and Lee (2011) have suggested that PP and ND effects in four-year-old children can be explained in terms of broad cognitive mechanisms of retention and retrieval, such as triggering (allocating a new representation in memory), configuration (storing form and meaning) and engagement (integrating new and existing representations); though note that the triggering hypothesis predicts that low PP words will be more easily learned than high PP words, which is not the case when you control for word length. It will also be important to see how this model fares against other models that simulate development dynamically (e.g. Samuelson, Spenser & Jenkins, 2013). Perhaps more importantly, though, CLASSIC can only simulate the acquisition of the phonological form of words, not the mapping of this form to its correct semantic representation. This is an additional complex learning task that is also likely to be influenced by the nature of the linguistic knowledge already accumulated (see Borovsky et al., 2016, for a proposal that a word's position and connectivity in developing semantic networks influences acquisition; but also see the work by Hills and colleagues discussed earlier showing that semantic influences may not be based on the current semantic knowledge held by the child). In addition, it is probable that the child's developing semantic and syntactic knowledge, as well as her phonological knowledge, will affect the acquisition of phonological forms in turn (see Dautriche et al., 2018, for evidence that children learn homophones more easily when the meanings of the two forms are made distinct via semantic and syntactic context). We also acknowledge that our analyses are based on children's productive vocabularies since the Manchester corpus does not include receptive vocabulary checklists. Unfortunately it is not straightforward to examine whether phonological knowledge also influences children's receptive vocabularies because it requires the linguistic input that the child receives during the period for which receptive vocabulary checklists are

taken and we are not aware of any corpus that includes this information. The model also focuses on how the maternal input influences children's spoken words, but ignores changes in maternal speech on the basis of the child's own spoken words, such as a subsequent re-naming of a dummy/pacifier as a 'beebee' based on the child re-naming the item.

In summary, our work shows that knowledge of phoneme sequences of varying sizes influence subsequent word learning. In study one, we showed how such learning is able to capture effects of children's word learning in terms of ND, PP, word length and word frequency without any need for mechanisms that capture traditional definitions of ND and without any explicit mechanism to monitor the sorts of frequency information that are involved in PP and word frequency. Rather, the model simply creates larger and larger phoneme sequence chunks for phoneme sequences that occur frequently and it is this that enables the model to capture all four effects. This not only questions the utility of measures such as ND and PP but also shows how one of the most potent predictors of word learning – that of a word's frequency – can be simulated without any monitoring of such frequency. In study two, we supported our view by showing how the words produced by two-year-old children were more likely to contain known phoneme sequences than those that were not produced by the children and that this was more likely to occur as the size of the known phoneme sequences increased. All told, this shows the critical nature of phoneme sequence knowledge in children's word learning. In order to simulate key features of child word learning, it is crucial that we build models that incorporate a role for the effect of developmental change on the learning process itself; models in which learning is influenced by the nature of the current lexical and sublexical knowledge stored in the child's developing lexicon at the fleeting moment when learning occurs.

References

Abbot-Smith, K. & Behrens, H. (2006). How known constructions influence the acquisition of other constructions: The German passive and future constructions. *Cognitive Science, 30*, 995-1026. https://doi.org/10.1207/s15516709cog0000_61.

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*, 239–273. https://doi.org/10.1017/S030500091400049X.

Auer, E. T., & Luce, P. A. (2005). 25 Probabilistic phonotactics in spoken word recognition. *The Handbook of Speech Perception*, 610–644. https://doi.org/10.1002/9780470757024.ch25.

Batchelder, E. O. (2002). Bootstrapping the lexicon: a computational model of infant speech segmentation. *Cognition*, *83*, 167–206. https://doi.org/10.1016/S0010-0277(02)00002-1.

BNC (2007). The British National Corpus, version 3 (BNC XML Edition) [on-line database]. http://www.natcorp.ox.ac.uk/.

Bohannon, J. N., & Hirsh-Pasek, K. (1984). Do children say as they're told? A new perspective on motherese. In C. G. & R. M. G. L. Feagans (Ed.), *The origins and growth of communication* (pp. 176–195). Ablex.

Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. L. (2016). Semantic structure in vocabulary knowledge interacts with lexical and sentence processing in infancy. *Child Development*, *87*, 1893–1908. https://doi.org/10.1111/cdev.12554.

Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child Language*, *30*, 441–469. https://doi.org/10.1017/S0305000903005579.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*, 125–127. https://doi.org/10.3758/BF03203646.

Dautriche, I., Fibla, L., Fievet, A.-C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology, 104*, 83-105. https://doi.org/10.1016/j.cogpsych.2018.04.001.

Fernald, A. & Hurtado, N. (2006). Names in frames: Infants interpret words in sentence frames faster than words in isolation. *Developmental Science, 9*, F33-F40. https://doi.org/10.1111/j.1467-7687.2006.00482.x.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: a recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614–636. https://doi.org/10.1037/a0025255.

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition*, *23*, 83–94. https://doi.org/10.3758/BF03210559.

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language*, *28*, 200–213. https://doi.org/10.1016/0749-596X(89)90044-2.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*, 236-243. https://doi.org/10.1016/S1364-6613(00)01662-4.

Goldstein, R., & Vitevitch, M. S. (2014). The influence of clustering coefficient on word-learning: How groups of similar sounding words facilitate acquisition. *Frontiers in Psychology*, *5*, 2009–2014. https://doi.org/10.3389/fpsyg.2014.01307.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*, 515–531. https://doi.org/10.1017/S0305000907008641.

Grenfell-Essam, R., & Ward, G. (2012). Examining the relationship between free recall and immediate serial recall: The role of list length, strategy use, and test expectancy. *Journal of Memory and Language*, *67*, 106–148. https://doi.org/10.1016/j.jml.2012.04.004.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, *27*, 1–6.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2009). Longitudinal analysis of early semantic networks. *Psychological Science*, *20*, 729–739. https://doi.org/10.1111/j.1467-9280.2009.02365.x.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative strength of language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*, 259–273. https://doi.org/10.1016/j.jml.2010.06.002.

Hollich, G., Jusczyk, P. W., & Luce, P. A. (2002). Lexical neighborhood effects in 17-month-old word learning. In *26th Annual Boston University Conference on Language Development* (pp. 314–323). Cascadilla Press.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: relation to language input and gender. *Developmental Psychology*, *27*, 236–248. https://doi.org/10.1037/0012-1649.27.2.236.

Jones, G. (2016). The influence of children's exposure to language from two to six years: The case of nonword repetition. *Cognition*, *153*, 79–88. https://doi.org/10.1016/j.cognition.2016.04.017.

Jones, G., & Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. *Cognition*, *144*, 1–13. https://doi.org/10.1016/j.cognition.2015.07.009.

Jones, G., & Macken, B. (2018). Long-term associative learning predicts verbal short-term memory performance. *Memory and Cognition*, *46*, 216–229. https://doi.org/10.3758/s13421-017-0759-3.

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1–21. https://doi.org/10.1016/j.cogpsych.2017.07.002.

Jones, G., Tamburelli, M., Watson, S. E., Gobet, F., & Pine, J. M. (2010). Lexicality and frequency in Specific Language Impairment: Accuracy and error data from two nonword repetition tests. *Journal of Speech Language and Hearing Research*, *53*, 1642–1655. https://doi.org/10.1044/1092-4388(2010/09-0222).

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645. https://doi.org/10.1006/jmla.1994.1030.

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, 2, 389-398. https://doi.org/10.1016/S1364-6613(98)01230-3.

Keren-Portnoy, T., Vihman, M. M., Depaolis, R. A., Whitaker, C. J., & Williams, N. M. (2010). The role of vocal practice working memory. *Journal of Speech, Language, and Hearing Research*, *53*, 1280–1293. https://doi.org/10.1044/1092-4388(2009/09-0003).

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Lawrence Erlbaum Associates.

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*, 439–459. https://doi.org/10.1017/S0305000906007458.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86. https://doi.org/10.1016/0010-0285(86)90015-0.

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.) *Word recognition in beginning literacy* (pp. 89-120). Lawrence Erlbaum Associates.

Naigles, L. R., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of Child Language, 25*, 95–120. https://doi.org/10.1017/S0305000997003358.

Ota, M., & Green, S. J. (2013). Input frequency and lexical variability in phonological development: A survival analysis of word-initial cluster production. *Journal of Child Language, 40*, 539–566. https://doi.org/10.1017/S0305000912000074.

Romano, J., Kromrey, J. D., Coraggio, J., & Skowronek, J. (2006). Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys. In *Annual Meeting of the Florida Association of Institutional Research* (pp. 1-33).

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development, 83,* 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x.

Rowland, C. (2014). *Understanding child language acquisition*. Abingdon: Routledge. https://doi.org/10.4324/9780203776025.

Roy, P., & Chiat, S. (2004). A prosodically controlled word and nonword repetition task for 2- to 4-year-olds: Evidence from typically developing children. *Journal of Speech, Language, and Hearing Research, 47*, 223-234. https://doi.org/10.1044/1092-4388(2004/019).

Samuelson, L. K., Spencer, J. P., & Jenkins, G. W. (2013). A dynamic neural field model of word learning. In *Theoretical and Computational Models of Word Learning: Trends in Psychology and Artificial Intelligence* (pp. 1-27). IGI global. https://doi.org/10.4018/978-1-4666-2973-8.ch001.

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language, 9*, 319–336. https://doi.org/10.1017/S0305000900004748.

Szewczyk, J. M., Marecka, M., Chiat, S., & Wodniecka, Z. (2018). Nonword repetition depends on the frequency of sublexical representations at different grain sizes: Evidence from a multi-factorial analysis. *Cognition, 179*, 23–36. https://doi.org/10.1016/j.cognition.2018.06.002.

Siew, C. S. Q. (2013). Community structure in the phonological network. *Frontiers in Psychology*, *4*, 1–17. https://doi.org/10.3389/fpsyg.2013.00553.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010.

Storkel, H. L. (2003). Learning new words II: Phonotactic probability in verb learning. *Journal of Speech, Language, and Hearing Research, 46*, 1312–1323. https://doi.org/10.1044/1092-4388(2003/102).

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language, 36*, 291–321. https://doi.org/10.1017/S030500090800891X.

Storkel, H. L., & Lee, S. (2011). The independent effects of phonotactic probability and neighborhood density on lexical acquisition by preschool children. *Language and Cognitive Processes, 26*, 191–211. https://doi.org/10.1080/01690961003787609.

Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology, 43*, 454–464. https://doi.org/10.1037/0012-1649.43.2.454.

Swingley, D., & Humphrey, C. (2018). Quantitative Linguistic Predictors of Infants' Learning of Specific English Words. *Child Development, 89*, 1247–1267. https://doi.org/10.1111/cdev.12731.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127–152. https://doi.org/10.1017/S0305000900004608.

Tomasello, M. (2003). *Constructing a Language : A Usage-based Theory of Language Acquisition*. Harvard University Press.

Torchiano M (2020). *effsize: Efficient Effect Size Computation*. https://doi.org/10.5281/zenodo.1480624), R package version 0.8.1.

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. [Data file]. http://www.iphod.com.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*, 1176–1190. https://doi.org/10.1080/17470218.2013.850521.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, *51*, 408–422. https://doi.org/10.1044/1092-4388(2008/030).

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*, 325–329. https://doi.org/10.1111/1467-9280.00064.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*, 374–408. https://doi.org/10.1006/jmla.1998.2618.

Vitevitch, M. S., & Luce, P. A. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics*, *2*, 7.1-7.20. https://doi.org/10.1146/annurev-linguistics-030514-124832.

Vitevitch, M. S., & Storkel, H. L. (2013). Examining the acquisition of phonological word forms with computational experiments. *Language and Speech*, *56*, 493–527. https://doi.org/10.1177/0023830912460513.

Walley, A. C. (1993). The role of vocabulary development in children's spoken word recognition and segmentation ability. *Developmental Review*, *13*, 286–350. https://doi.org/10.1006/drev.1993.1015.

Walley, A. C., Metsala, J. L., & Victoria, M. (2003). Spoken vocabulary growth : Its role in the development of phoneme awareness and early reading ability. *Reading and Writing*, *16*, 5–20. https://doi.org/10.1023/A:1021789804977.