Journal of Memory and Language 92 (2017) 265–280

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases

Inbal Arnon [a,*], Stewart M. McCauley [b], Morten H. Christiansen [b,c]

[a] Department of Psychology, Hebrew University, Jerusalem 91905, Israel
[b] Department of Psychology, Cornell University, Ithaca, NY 14853, USA
[c] The Interacting Minds Centre, Aarhus University, 8000 Aarhus C, Denmark

ARTICLE INFO

Article history:
Received 21 December 2015
revision received 3 July 2016
Available online 25 July 2016

Keywords:
Age-of-Acquisition
Multiword units
Language learning

ABSTRACT

Words are often seen as the core representational units of language use, and the basic building blocks of language learning. Here, we provide novel empirical evidence for the role of *multiword* sequences in language learning by showing that, like words, multiword phrases show age-of-acquisition (AoA) effects. Words that are acquired earlier in childhood show processing advantages in adults on a variety of tasks. AoA effects highlight the role of words in the developing language system and illustrate the lasting impact of early-learned material on adult processing. Here, we show that such effects are not limited to single words: multiword phrases that are learned earlier in childhood are also easier to process in adulthood. In two reaction time studies, we show that adults respond faster to early-acquired phrases (categorized using corpus measures and subjective ratings) compared to later-acquired ones. The effect is not reducible to adult frequencies, plausibility, or lexical AoA. Like words, early-acquired phrases enjoy a privileged status in the adult language system. These findings further highlight the parallels between words and larger patterns, demonstrate the role of multiword units in learning, and provide novel support for models of language where units of varying sizes serve as building blocks for language.

Introduction

Traditionally, words are seen as the basic building blocks of language learning and processing (e.g., Chomsky, 1965; Pinker, 1991). Recent years, however, have seen a shift away from this perspective. There is increasing theoretical emphasis on, and empirical evidence for, the idea that multiword units, like words, are integral building blocks for language. This idea is found in linguistic approaches that emphasize the role of constructions in language (Culicover & Jackendoff, 2005; Goldberg, 2006; Langacker, 1987) and is advocated in single-system models of language which posit that all linguistic material – whether it is words or larger sequences – is processed by the same cognitive mechanisms (Bybee, 1998; Christiansen & Chater, 2016b; Elman, 2009; McClelland, 2010). The role of multiword units in language is also highlighted in usage-based approaches to language learning, which have been gaining prominence in recent years (Bannard, Lieven, & Tomasello, 2009; Christiansen & Chater, 2016a; Lieven & Tomasello, 2008; Tomasello, 2003). In such models, language is learned by abstracting over stored exemplars of various sizes and levels of abstraction (from syllables through words to constructions). Multiword units are predicted to play a role in learning by providing children with information about the distributional and structural relations that hold between words (Abbot-Smith & Tomasello, 2006; Bod, 2006, 2009; McCauley & Christiansen, 2014). Children are

[*] Corresponding author.
E-mail address: inbal.arnon@mail.huji.ac.il (I. Arnon).

http://dx.doi.org/10.1016/j.jml.2016.07.004
0749-596X/© 2016 Elsevier Inc. All rights reserved.

expected to draw on both words and multiword units in the process of learning.

Accordingly, there is growing developmental and psycholinguistic evidence that children and adults are sensitive to the properties of multiword sequences and draw on such information in learning, production, and comprehension (e.g., Arnon & Cohen Priva, 2013, 2014; Arnon & Snider, 2010; Bannard, 2006; Bannard & Matthews, 2008; Bybee & Schiebman, 1999; Janssen & Barber, 2012; Jolsvai, McCauley, & Christiansen, 2013; Reali & Christiansen, 2007; Tremblay & Tucker, 2011). Adult speakers, for instance, are faster to recognize and produce higher frequency four-word phrases (Arnon & Cohen Priva, 2013; Arnon & Snider, 2010) and show better memory of them (Tremblay, Derwing, Libben, & Westbury, 2011), an effect that is not reducible to the frequency of individual substrings. This sensitivity is evident early on; young children (two- and three-year-olds) are faster and more accurate at producing higher frequency phrases (Bannard & Matthews, 2008), while four-year-olds show better production of irregular plurals inside frequent frames (e.g., *Brush your – teeth*, Arnon & Clark, 2011). Analyses of early child language also support the role of multiword chunks in early learning: up to 50% of children's early multiword utterances include 'frozen' chunks (sequences that are not used productively, Lieven, Behrens, Speares, & Tomasello, 2003; Lieven, Salomo, & Tomasello, 2009), a pattern that is also found in computational simulations of early child language (Bannard et al., 2009; Borensztajn, Zuidema, & Bod, 2009; McCauley & Christiansen, 2011; McCauley & Christiansen, 2014).

Such findings highlight the parallels in processing words and larger sequences, and undermine a strict representational distinction between words and phrases. However, the existing findings do not provide conclusive evidence for the role of multiword units in learning. Finding that higher frequency phrases are easier to process means that adult speakers are sensitive to distributional information about multiword sequences, but does not attest to their role in learning. Similarly, the presence of multiword chunks in children's production does not necessarily mean such units were used as building blocks for learning, especially since most of children's early productions are single words and not multiword sequences. Moreover, since children's receptive vocabulary is typically much larger than their productive one (Clark & Hecht, 1983; Grimm et al., 2011) it is hard to identify early linguistic representations based on their early productions (e.g., children show a preference for sentences with grammatical forms even when such morphemes are omitted in their own speech; Shi et al., 2006). A similar comprehension-production asymmetry has also been observed in a computational model that uses multiword sequences as its building blocks (Chater, McCauley, & Christiansen, 2016; McCauley & Christiansen, 2013).

In this paper, we address the challenge of identifying children's early linguistic units by turning to adult processing as a window onto the early units of learning. We provide novel evidence for the prediction that multiword units serve as building blocks for language learning by showing that, like words, multiword phrases show age-of-acquisition (AoA) effects: multiword phrases that were acquired earlier in childhood show processing advantages in adult speakers, after controlling for adult usage patterns. The finding that AoA effects are not limited to single words has consequences beyond the role of larger units in learning: such a finding provides additional evidence for the parallels in processing and representation between words and larger phrases, and expands our understanding of the linguistic information speakers are sensitive to.

Lexical Age-Of-Acquisition effects

Words that are acquired earlier in childhood show processing advantages for adult speakers in a variety of lexical and semantic tasks, including lexical decision, picture naming, word naming, sentence processing, and more (Ellis & Morrison, 1998; Juhasz & Rayner, 2006; Morrison & Ellis, 1995). Early-acquired words tend to be responded to faster than later-acquired ones, after controlling for adult usage patterns (the frequency of the word in adult language). For instance, despite having similar frequency in adult language, adults would be faster to recognize the early-acquired *bell* compared to the later-acquired *wife* (AoA and frequency taken from Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). These AoA effects have been found in numerous studies across different languages and tasks (see Johnston & Barry, 2006; Juhasz, 2005, for reviews). One of the major challenges in studying the effect of AoA on processing is separating the effect of order of acquisition from that of other factors that are naturally correlated with it, like cumulative frequency (early-acquired words have been known longer), frequency trajectory (early-acquired words tend to have a high-to-low frequency trajectory across the life span), concreteness (early-acquired words tend to be more concrete), and length (early-acquired words tend to be shorter).

While the precise mechanism that gives rise to AoA effects is still debated (e.g., Ghyselinck, Lewis, & Brysbaert, 2004; Marmillod et al., 2012), there is substantial evidence that AoA does affect processing and is not just a proxy for other factors, or a frequency effect in disguise. AoA effects are found after controlling for other factors known to affect lexical processing (e.g., Brysbaert & Ghyselinck, 2006). They are particularly robust in tasks such as picture naming or lexical decision where such effects persist after controlling for frequency, cumulative frequency (Ghyselinck et al., 2004; Moore & Valentine, 1998), and frequency trajectory (Perez, 2007; Maermillod, Bonin, Meot, Ferrand, & Paindavoine, 2012). For instance, AoA effects are found even when adult frequencies are higher for the late-acquired words, as in the comparison between high-frequency/late-acquired words like *cognition* (for psychologists) and low-frequency/early-acquired words like *pony* (Stadthagen-Gonzalez et al., 2004). More importantly, AoA effects do not increase with age, as would be expected if they simply reflected cumulative frequency (Kuperman et al., 2012; Morrison, Hirsh, Chappell, & Ellis, 2002; but see Catling, South, & Dent, 2013), and are also found in artificial language learning, where both frequency and cumulative frequency (as well as other word properties) can be tightly controlled

(Catling, Dent, Preece, & Johnston, 2013; Izura et al., 2011; Stewart & Ellis, 2008). Taken together, the converging evidence suggests that the order-of-acquisition of words has an independent effect on adult processing.

These findings have psycholinguistic and developmental implications. From a psycholinguistic perspective, they highlight the richness of information that adult speakers are sensitive to (e.g., Elman, 2009): not only the frequency with which words are used, but also their order of acquisition. More importantly, lexical AoA effects illuminate the process of language acquisition: they illustrate the lasting impact of early-learned material on subsequent representation, and show that early-learned words play an important part in shaping the adult language system. Put differently, AoA effects offer a *window* into the process of language learning: we can look at adult processing to identify early units of learning and assess their impact on the adult system.

The current study

If multiword units serve as building blocks for language learning, they should also exhibit AoA effects. In the present study, we test this prediction and go beyond existing findings to show that AoA effects are not limited to single words, but are also found for multiword phrases (three-word sequences). We show that early-acquired phrases, like early-acquired words, show processing advantages in adult processing. Such findings add a novel dimension to what speakers know – not only the properties of words but also of multiword sequences; reveal further parallels in the processing of words and larger phrases, and most importantly, provide novel empirical evidence for the prediction about the role of larger units in language learning.

A major challenge in testing this prediction lies in identifying the AoA of multiword sequences: how can we know when (or rather, in which order) multiword phrases were acquired? We turn to the lexical AoA literature, which was faced with a similar challenge. In the lexical AoA literature, the most commonly used method for determining AoA is simply asking participants to estimate the age (in years) when they learned a word. These subjective ratings provide the relative order-of-acquisition of words and are used to classify items into early and later acquired. These ratings have been used in multiple studies and have been validated as reliable estimates of AoA in several ways. First, they predict reaction times on a variety of tasks (see Juhasz, 2005, for a review): subjective AoA ratings from one sample of participants predicts reaction times collected from a different sample. Second, subjective ratings are correlated with actual naming data collected from children: they accurately reflect the age at which most children (over 75%) understand a word (Morrison, Chappell, & Ellis, 1997). Finally, subjective ratings are consistent across participants: they result in similar rankings across different samples of speakers (Kuperman et al., 2012; Stadthagen-Gonzalez & Davis, 2006). In sum, speakers seem to be able to estimate the age at which words were acquired (or at least their relative order).

However, it is not clear that this ability can scale up to three-word sequences, which are less concrete by nature. Because we did not want to assume what we are trying to test, mainly, that speakers *are* sensitive to multiword AoA, we decided to use a combination of corpus-based measures and subjective ratings to create our early- and later-acquired items. As a first step, we used a large-scale corpus of child-directed speech to extract trigrams (three-word sequences) that appeared frequently in speech directed to children under the age of three. We used those frequent trigrams as our early-acquired candidates. We then matched each of these trigrams with another trigram that differed by only one word, but rarely appeared in the same child corpus: we extracted pairs of trigrams that differed in frequency in child-directed speech (e.g., high-frequency*: take them off*, vs. *take time off*, which did not appear in the child-directed corpus). The logic behind this is that children are unlikely to acquire forms they are never (or rarely) exposed to. We only selected trigrams whose words were early-acquired (based on established norms, Kuperman et al., 2012) to control for the effect of lexical AoA on processing. We then ensured that the two trigrams had a similar distribution in adult language by only selecting pairs where the two trigrams had similar unigram, bigram, and trigram frequencies in adult speech (estimated using two large-scale adult corpora), meaning that any difference in response times between them would not reflect adult usage patterns. We ended up with a set of trigrams pairs that were matched on all adult frequencies (based on a large adult corpus) but differed in their frequency in child-directed speech. To ensure that any difference in reaction time is not due to adult usage patterns, we conducted additional corpus simulations (see Methods for details) to show that our frequency estimates are reliable and do now show 'burstiness' (the tendency of words or phrases to occur in bursts throughout a corpus, e.g. Katz, 1996; Pierrehumbert, 2012). The resulting set of items was rated by a different set of participants for plausibility to control for possible differences between the trigrams.

This selection process is based on several assumptions, all of which are motivated by existing findings. Using corpus frequencies as a proxy for order of acquisition is motivated by several lines of research. First, there is a large literature showing that more frequent elements (sounds, words, constructions) tend to be acquired earlier (see Ambridge, Kidd, Rowland, & Theakston, 2015; Diessel, 2007; Lieven, 2010, for reviews). It seems reasonable to assume that phrases that were used often in the input may be acquired earlier than ones that occur rarely. Second, words that appear often in child-directed speech do seem to be acquired earlier: input frequencies in child-directed speech are correlated with age of acquisition as assessed using the MacArthur-Bates Communicative Development Inventory which provides norming data for vocabulary acquisition (Goodman et al., 2008). Together, the findings provide some support for the postulated relation between multiword frequency in child-directed speech and order of acquisition.

A second assumption is that while child and adult usage patterns are correlated - in the sense that many items that are frequent in child-directed speech will also be frequent in adult-to-adult speech - there are also meaningful differences in the way language is used with children and adults.[1] These differences stem from the different situations experienced by young children and adults, as well as the unique communicative and social settings. Unfortunately, while many studies examine the unique properties of child-directed speech (see Soderstrom, 2007, for a review), very few compare the distributional properties of child-directed and adult-to-adult speech. One study, however, compared verb use in child-directed and adult-to-adult speech (Buttery & Korhonan, 2005) and found both overlap and distinct patterns. For instance, action verbs like *play*, *eat*, *and put* were much more frequent in child-directed speech while mental state verbs like *know*, *mean*, *and feel* were more frequent in adult-to-adult speech. As our item selection will demonstrate, it is possible to find items that are highly frequent in child-directed speech but not in adult conversations. For instance, the phrase *a good girl* is much more frequent than the phrase *a good dad* in child-directed speech, but both are similarly infrequent in adult-to-adult conversations.

Finally, we collected subjective ratings for all our item pairs. We asked a new set of participants (that did not take part in the experiments or in the plausibility ratings) to estimate the age (in years) when they first understood the trigram, using a rating method identical to the one used to assess lexical AoA (Kuperman et al., 2012). We did this for two reasons. First, we wanted to validate our corpus-based classification and see if the trigrams we defined as early-acquired (based on corpus frequencies) were also rated as having a lower AoA. Second, the ratings provide another way to ask if speakers are sensitive to multiword AoA. If they are, then the ratings should predict reaction times (for a separate sample of participants), as they do for words.

To reiterate, the current study has several goals. First, we wish to determine if adult participants are sensitive to the relative order-of-acquisition of multiword phrases. Such a finding would further support the parallels in processing words and larger patterns and provide novel support for the idea multiword phrases serve as building blocks for language learning. Second, we ask if participants are capable of estimating the AoA of multiword phrases, and if those ratings predict reaction times as they do for individual words. If so, this would both provide a replicable way of assessing multiword AoA and further support the idea that speakers are sensitive to the order-of-acquisition of larger patterns. We test these predictions in two reaction time studies with adult participants using two different sets of items, with the second study having a more stringently controlled set of items in terms of lexical AoA. This was done to increase the reliability and validity of the results and ensure they are not confined to a particular item set, and are not driven by adult usage patterns or differences in lexical AoA.

Experiment 1

Methods

Participants

Seventy undergraduate students from Cornell University participated in the study in exchange for course credit (mean age: 20.6, range 19–25; 37 females and 33 males). All participants were native English speakers, did not have any language or learning disabilities, and reported normal or corrected-to-normal vision. Since this is the first study to look at multiword AoA effects, we did not have a priori estimates of the expected effect size and therefore of the appropriate sample size. As a result, data collection was done for a predetermined duration (three weeks before the end of the semester). At the end of this period there were seventy participants. The data was analysed only after that date had passed.

Materials

Corpus-based item extraction. To obtain the early-acquired trigrams, we used an aggregated corpus of American-English child-directed speech from the CHILDES database (MacWhinney, 2000) to extract three-word sequences that appeared frequently in the speech directed to children under the age of three. The aggregated corpus had 5.3 million words from 39 different CHILDES corpora. We excluded corpora that contained speech directed to multiple children of different ages to ensure the speech was directed to children below three years of age. We then matched each of the frequent trigrams with another trigram that differed only in one word and satisfied the following four constraints: first, the two variants had similar word (unigram), bigram, and trigram frequencies in adult speech (within a window of ±20%). This was done to ensure that any difference in processing between the trigrams did not reflect adult usage patterns (any remaining differences in part frequencies were controlled for in the statistical analyses, see below). We calculated adult frequencies using a 20-million word corpus created by combining the Fisher corpus (Cieri, Miller, & Walker, 2004) with the Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992). Second, the other, late-acquired trigram did not appear in the speech produced by any child in the aggregated corpus, and occurred rarely in the speech directed to children (average of less than one occurrence [0.95] in the whole corpus). There were almost 2000 trigrams pairs that fulfilled these frequency constraints. We then applied two additional constraints: Third, all of the single words in the two variants were early acquired (based on Kuperman et al., 2012). Fourth, both variants were complete intonational phrases (and not sentence fragments) and both variants had to be judged as complete syntactic constituents by an independent research assistant.

Applying these criteria to our early-acquired candidates resulted in 46 item pairs: each pair consisted of an early-acquired and late-acquired variant (see Table 1 for examples of early and late variants, and Appendix A for the full item list). The early and late items did not differ in adult

[1] There is a vast literature on the unique properties of child-directed speech. However, most of it focuses on phonological, prosodic and lexical characteristics. There are very few studies that compare lexical frequencies between child-directed and adult-to-adult speech and none (to our knowledge) that examine multiword frequencies.

unigram frequency (word$_1$: $t(90) = -0.004$, $p > .9$, word$_2$: $t(90) = 0.0001$, $p > .9$, word$_3$: $t(79) = -.067$, $p > .9$), bigram frequency (bigram$_1$: $t(90) = .0001$, $p > .9$, bigram$_2$: $t(90) = -.095$, $p > .9$), and trigram frequency ($t(90) = -.08$, $p > .9$). See Table 2 for the frequency properties of the items. Since we were interested in controlling for the effect of multiword frequency on processing (rather than testing for it), the items had relatively low trigram frequency and did not span a large trigram frequency range (mean = 0.43 per million, range 0.04–4 per million). However, the early and late items did differ in number of letters (early: 11.76, late: 12.78, $t(90) = -3.4$, $p < .01$). Also, while all the words in the trigrams had a lexical AoA of under six, the early and late items set did differ in average lexical AoA with later-acquired phrases having a slightly later lexical AoA (early: 3.84, late: 4.49, $t(90) = -3.98$, $p < .01$). This difference will be controlled for in the analyses to ensure that the effect of multiword AoA occurs after controlling for lexical AoA (a factor known to affect decision times).

To make sure that the frequency difference found between our item pairs reflects a real difference in the language used with children and adults (and is not merely the result of comparing two different corpora), we applied our item selection process to two different sets of spoken adult corpora (Switchboard vs. Fisher). We extracted all the trigrams that appeared over ten times per million the Switchboard corpus. We then looked for all the trigrams that differed in only in one word, had similar unigram, bigram and trigram frequencies in the Fisher corpus (within a 20% window but appeared under one time per million in the Switchboard corpus (the "child" corpus in this example). That is, we looked for trigram pairs where the pair had similar frequency in one corpus (Fisher) but different frequency in another (Switchboard). Using these two large corpora (larger than the ones we used for extracting the experimental items), we only found **100** such pairs (compared with 1800 when comparing child and adult speech). Of these, only **eleven** complied with the additional criteria used in our paper that all trigrams had to be syntactic constituents and form one prosodic unit.

To further ensure that burstiness (Katz, 1996) did not bias our material selection, we defined 100 random contiguous chunks of text (with "wraparound" at the edges of the corpus, when necessary, to avoid under-sampling at the margins), each consisting of 20% of the overall adult corpus material. We used contiguous chunks because the "burstiness" argument pertains to continuous samples of text/conversation. For each trigram, we collected mean frequencies and standard deviations across all randomly selected chunks. We then compared the Early and Late conditions to ensure that neither the standard deviation ($t = 0.64$, df = 85.32, p-value = 0.5177) nor the mean ($t = 0.0456$, df = 97.994, p-value = 0.963) of the groups differed. A significant difference in standard deviation would indicate that one of the conditions was more "bursty" than the other – such a difference was not found, suggesting that our items were well-matched in terms of adult frequencies.

Plausibility ratings. Multiword sequences that appear more frequently in child-directed speech may also refer to more plausible events. To control for this in the analyses, we used Amazon Mechanical Turk (AMT) to collect plausibility ratings for all the experimental items. AMT is a crowd-sourcing, web-based service (https://www.mturk.com) that enables the collection of responses from anonymous users. AMT is increasingly used for psycholinguistic research and norming data collected using AMT has been shown to reliably replicate lab-based findings (Gibson et al., 2011). Following Kuperman et al., 2012, we filtered non-native participants by only using responses from participants who were currently residing in the US, who entered a valid US state when asked where they lived during their first seven years of their life, and who completed the task in a predefined time. Thirty-five native English speakers (19 females and 15 males) were asked to rate the plausibility of the items on a scale from 1 to 7 (1: highly implausible – 7: highly plausible). Plausibility was defined as "describing an entity or situation that is likely to occur in the real world" (the same definition used in Arnon & Snider, 2010). In addition to the 92 experimental items, participants also rated 40 implausible filler sequences. The task took about 15 minutes to complete. While all the experimental items were judged as more plausible than the implausible fillers (experimental: 5.6,

Table 1
Examples of matched early- and later- acquired trigrams and their plausibility and frequency measures for Experiment 1.

Early-trigram	Early child-directed-freq	Early-plausibility	Early-adult-freq	Late-trigram	Late-child-directed-freq	Late-plausibility	Late-adult-freq
are you drawing	59	6.14	1	are you proud	1	6.3	1
for the baby	102	6.02	17	for the teacher	2	6.12	15
in the trash	84	6.4	30	in the hills	1	5.05	34

Table 2
Adult frequency properties in the two conditions (per million words) for items in Experiment 1.

Condition	Word1	Word2	Word3	Bigram1	Bigram2	Trigram
Early	12,075	23,360	741	1280	17	1.5
Late	12,995	23,100	746	1305	14.5	1.3

fillers: 4.3, $t(130) = 8.5$, $p < .0001$), the early acquired items were more plausible than the late acquired ones (early: 6.0, late: 5.4, $t(90) = 5.07$, $p < .001$). The plausibility rating of each item was therefore controlled for in the statistical analyses reported below.

Subjective AoA ratings. In order to validate our corpus-based classification and determine whether participants can estimate AoA for multiword sequences like they do for words, we collected subjective AoA ratings for the forty item pairs. We used AMT to collect subjective ratings from 32 native English speakers (17 females and 15 males, screened in the same way as in the plausibility rating study). We followed the same procedures and instructions used by Kuperman et al. (2012) in their large-scale AMT word AoA rating study. Participants rated all ninety-two experimental items (46 early-acquired and 46 late-acquired) as well as seventy single words taken from the Kuperman et al. norms. We included the single words to ensure that our sample provides similar AoA estimates for words as in the Kuperman et al. study. On each trial, participants saw a trigram or word on the screen and were asked to estimate the age (in years) when they first understood the item (even if they did not use it at the time). The study took around fifteen minutes to complete.

All participants completed the task suggesting they were able to estimate the AoA for multiword sequences. The results corroborated our corpus-based classification: our early-acquired items were rated as learned earlier than our later-acquired ones (early: 3;8, late: 5;3, $t(90) = -9.38$, $p < .001$). Importantly, the correlation between the lexical AoA in our participant sample and that in the large-scale lexical AoA study (Kuperman et al., 2012) was very high ($r = .96$), further confirming the validity of the sample and the reliability of the subjective rating method.

Procedure

Participants completed a phrasal decision task, modelled on the classic lexical decision task used commonly in psycholinguistic research. The phrasal decision task has been used successfully in the past to study the processing of multiword sequences (Arnon & Snider, 2010; Jolsvai et al., 2013). In this task, participants see multiword sequences on the screen and are asked to decide – as quickly and accurately as possible – if the sequence is a possible one in English. Fillers consisted of impossible sequences like '*full the out*' or '*I as said*'. Similar to a lexical decision task, participants are asked to press one key if the sequence is possible, and another if it is not. Each participant saw all of the experimental items (total = 92) intermixed with 92 impossible fillers to yield an equal number of *yes* and *no* responses over the course of the experiment. Order of presentation was randomized for each participant. The task took about 15 min to complete.

Results and discussion

Accuracy was high overall (mean of 97%) for both the early-acquired (mean 98%) and late-acquired items (97%), as is expected in lexical decision tasks. We excluded responses under 200 ms or more than 2.5 standard deviations from the mean of each condition. This resulted in the loss of 6% of the data. Incorrect responses were also excluded from the analysis.

We use mixed-effect regression models to analyse the results. All models had the maximal random effects structure justified by the design (cf. Barr, Levy, Scheepers, & Tily, 2013). The frequencies of the unigrams, bigrams, and trigrams were entered as control variables into all analyses in order to measure the effect of AoA while controlling for frequency. We ran a principal component analysis to reduce the collinearity between all the unigram, bigram and trigram frequency measures, which were collinear. This led to three components (instead of the six frequency measures), and ensured that collinearity in all reported models was small (all variance inflation factors [vif's] were under 2). We added the plausibility ratings to all analyses since they differed between the two conditions. We also controlled for the average lexical AoA of the words in the two trigrams, since that differed between the two conditions.

Reaction times

As predicted, reaction times were faster for early-acquired items compared to later ones (early: 685 ms (SD = 68), late: 731 ms (SD = 74)). A mixed-effects linear regression model was used to predict logged reaction times. We included *type* (early vs. late), log(*plausibility*) (logged to reduce skewness), *number-of-letters*, *average-lexical*-AoA (the averaged lexical AoA of the three words in the trigram), and the three PCA frequency components as fixed effects. We had *subject* and *item-pair* as random effects, as well as a by-subject random slope for *type*, and a by-*item* slope for *type* (to ensure the effects hold beyond items and subjects).

As expected, early items were decided on faster than later ones ($\beta = -.04$ [SE = .01], $p < .01$; model comparison chi-square = 6.35, $p < .05$, see Table 3). The effect was significant controlling for syntactic completeness, all frequency measures, lexical AoA, and plausibility. *Plausibility* did not predict reaction times ($\beta = -.08$, SE = .05, $p > 0.2$, chi-square = 2.3, $p = 0.1$), and neither did *lexical AoA*, even though it differed between the conditions ($\beta = -.01$, SE = .009, $p > .2$, chi-square = 0.97). Unsurprisingly, items with more letters were responded to more slowly ($\beta = .01$, SE = .004, $p < .001$, chi-square = 14.14). Two of

Table 3
Mixed-effect regression with AoA as a binary measure (early vs. late) for Experiment 1. Significance obtained using the lmerTest function in R.

Fixed effects	Coef.	SE	*T*-value	*P*-value
Intercept	**6.52**	**.11**	**57.8**	**<.001**
AoA-Early	**−.04**	**.01**	**2.78**	**<.05**
Plausibility	−.08	.05	−1.44	>0.1
Lexical-AoA	−.01	.009	−1.08	>0.2
Num-Let	**.01**	**.004**	**3.85**	**<.001**
pca1	**.03**	**.008**	**3.7**	**<.01**
pca2	**−.04**	**.008**	**−4.65**	**<0.01**
pca3	−.007	.008	−.88	>0.3

Variables in bold were significant ($p < .05$).

the three aggregate frequency measures from the principal component analysis were significant. The first principal component – which was most highly correlated with the third word frequency – led to slower reaction times (pca1: $\beta = .03$, SE = .008, $p < .01$, chi-square = 4.5) while the second principal component – most highly correlated with the first bigram frequency – led to faster reaction times (pca2: $\beta = -.04$, SE = .008, $p < .001$, chi-square = 18.2). The effect of the third component was not significant (pca3: $\beta = -0.007$, SE = .008, $p > .7$, chi-square = 0.07). These frequency effects should be interpreted with caution: Since the purpose of the study was to control for frequency effects, rather than investigate them the two conditions were matched on all frequencies, and the items were not selected to be from a wide frequency range. Importantly, the effect of multiword AoA persisted after controlling for all adult frequencies.

If speakers' ability to estimate AoA extends to multiword sequences, then the subjective rating – collected from a different sample - should be predictive of reaction times in our study. We ran an additional analysis to see how well the subjective AoA ratings predicted reaction times. We used the exact same model (in terms of fixed and random effects), but replaced the binary variable of *type* (early vs. late) with the log(*subjective rating*) for each trigram (logged to reduce skewness). The random slope between *type* and *item* was also removed because items were no longer treated as pairs.

Interestingly, the *subjective ratings* were highly predictive of reaction times. Items estimated as learned later were responded to more slowly than earlier ones, after controlling for lexical AoA, syntactic completeness, all frequency measures and plausibility ($\beta = .01$, SE = .02, $p < .001$, chi-square = 43.00, see Table 4). As in the previous model, *plausibility* ($\beta = -.04$, SE = .03, $p > .2$, chi-square = 1.38) was not significant. Unlike in the previous model, the effect of *lexical AoA* in this model was significant, though it went in an unexpected direction: items with a higher average lexical AoA resulted in shorter reaction times ($\beta = -.02$, SE = .006, $p < .01$, chi-square = 20.1). This unexpected pattern – which was not found when the binary classification was used – may be a spurious effect driven by the high correlation between average lexical AoA and the subjective ratings ($r = .54$, $p < .01$), indeed, when we remove the subjective ratings from the model, lexical AoA is no longer significant ($\beta = -.003$, SE = .005, $p > .5$). Unsurprisingly, items with more letters were responded to more slowly ($\beta = .01$, SE = .003, $p < .001$, chi-square = 25.3, $p < .001$). The same two principal components measures were significant in this analysis (pca1: $\beta = .02$, SE = .008, $p > .01$, chi-square = 9.1; pca2: $\beta = -.04$, SE = .008, $p < .001$, chi-square = 29.1; pca3: $\beta = 0.004$, SE = .008, $p > .9$, chi-square = .002; pca4: $\beta = 0.001$, SE = .007, $p > .8$, chi-square = .06).

Table 4
Mixed-effect regression with subjective AoA ratings for Experiment 1. Significance estimates were obtained using the lmerTest function in R.

Fixed effects	Coef.	SE	*T*-value	*P*-value
Intercept	**6.26**	**.08**	**76.08**	**<.001**
Subjective-AoA	**.01**	**.08**	**7.61**	**<.001**
Plausibility	−.03	.03	−.89	>0.3
Lexical-AoA	**−.02**	**.02**	**−4.64**	**<0.01**
Num-Let	**.01**	**.002**	**6.9**	**<.001**
pca1	**.03**	**.009**	**3.4**	**<.01**
pca2	**−.04**	**.009**	**−4.79**	**<0.001**
pca3	−.003	.009	−.41	>0.6

Variables in bold were significant ($p < .05$).

In sum, participants were faster to respond to early-acquired trigrams compared to later-acquired ones, after controlling for adult usage patterns, plausibility and lexical AoA. Moreover, the estimated age at which a trigram was acquired was a significant predictor of reaction times, as is the case for individual words. These findings provide the first demonstration of AoA effects for units larger than single words.

To make sure these findings are not limited to a specific set of items, we conduct a second experiment using a different set of items extracted in the same way. This second study will also address a potential shortcoming of the first: despite the great care taken in constructing and selecting the items, the early- and late-acquired conditions in the first study did differ in lexical AoA. While all the words in the phrases were acquired early (before the age of six), later-acquired phrases contained words that were acquired on average a year-and-a-half later than those of the early-acquired phrases (early-phrase: average lexical AoA of three years and 8-months vs. later-phrases: average lexical AoA of five years and 5-months). Since our goal is to demonstrate an effect of phrase AoA that goes beyond the documented word AoA, we need to make sure that this difference is not driving our effect. Finally, to further ensure that the effect is not driven by frequency differences between the variants in adult usage, we decided to impose an even more stringent frequency criterion in the second study: the early- and late- variants had to have similar word (unigram), bigram, and trigram frequencies in adult speech within a window of ±10% and not 20% as in the first study.

Experiment 2

Methods

Participants

Seventy undergraduate students from Cornell University participated in the study in exchange for course credit (mean age: 19.7, range 18–22; 46 females and 24 males). All participants were native English speakers, did not have any language or learning disabilities, and reported normal or corrected-to-normal vision. We collected data from the same number of participants as in Experiment 1.

Materials

Corpus-based item extraction. We used the same procedure used in Experiment 1 to extract an additional set of item pairs. We used the same child-directed corpus as in the previous study. We extracted three-word sequences that appeared over 10 times per million in the corpus and then matched each of the frequent trigrams with another trigram that differed only in one word and satisfied the following constraints: First, the two variants had similar word (unigram), bigram, and trigram frequencies in adult speech (using the same combined Fisher and Switchboard corpus used in the previous study). We decreased the window to ±10% (from ±20% in Experiment 1) to further ensure our effect is not driven by differences in adult usage patterns between the two variants. Second, the other, late-acquired trigram did not appear in the speech produced by any child in the aggregated corpus, and occurred rarely in the speech directed to children (average of less than one occurrence [0.95] in the whole corpus). Third, all of the single words in the two variants were early acquired (based on Kuperman et al., 2012), and fourth, both variants were complete intonational phrases (and not sentence fragments) and had to be judged as complete syntactic constituents by an independent research assistant.

Applying these criteria to our early-acquired candidates resulted in 33 item pairs: each pair consisted of an early-acquired and late-acquired variant (see Table 5 for examples of early and late variants, and Appendix B for the full item list). The early and late items did not differ in adult unigram frequency (word$_1$: $t(64) = -0.002$, $p > .9$, word$_2$: $t(64) = 0.003$, $p > .9$, word$_3$: $t(64) = .003$, $p > .9$), bigram frequency (bigram$_1$: $t(64) = .29$, $p > .7$, bigram$_2$: $t(64) = .25$, $p > .8$), and trigram frequency ($t(64) = -.05$, $p > .9$). See Table 6 for the frequency properties of the items. As intended, the items here were better controlled than in Experiment 1. The early and late items did not differ in the number of letters (early: 12.66, late: 13.3, $t(64) = -1.4$, $p > .1$), and more importantly, the early and late items did not differ in average lexical AoA (early: 4.03, late: 4.08, $t(64) = -0.32$, $p > .7$).

As in the Experiment 1, to make sure that the frequency difference found between our item pairs reflects a real difference in the language used with children and adults (and is not merely the result of comparing two different corpora), we applied our item selection process to two different sets of spoken adult corpora (Switchboard vs. Fisher), using the same 10% frequency window used in Experiment 2. Using these two large corpora (larger than the ones we used for extracting the experimental items), we only found **21** such pairs (compared with 980 when comparing child and adult speech). Of these, only **3** complied with the additional criteria used in our paper that all trigrams had to be syntactic constituents and form one prosodic unit.

To ensure that burstiness (Katz, 1996) did not bias our material selection, we applied the exact same analyses as in Experiment 1, collecting mean frequencies and standard deviations for all trigrams from 100 random contiguous chunks. We then compared counts across the Early and Late conditions to ensure that neither the standard deviation ($t = -0.5682$, df = 71.207, p-value = 0.5717) nor the mean ($t = 0.0223$, df = 77.971, p-value = 0.9823) of the groups differed. A significant difference in standard deviation would indicate that one of the conditions was more "bursty" than the other – such a difference was not found suggesting that our items were well-matched in terms of adult frequencies.

Plausibility ratings. We used the same procedure as in Experiment 1 to collect plausibility ratings for each trigram using AMT. Thirty-four native English speakers (19 females and 15 males, screened in the same way as in the previous rating study) were asked to rate the plausibility of the items on a scale from 1 to 7 (1: highly implausible – 7: highly plausible). In addition to the 66 experimental items, participants also rated 40 implausible filler sequences. While all the experimental items were judged as more plausible than the implausible fillers (experimental: 5.6, fillers: 4.3, $t(124) = 8.5$, $p < .0001$), the early acquired items were more plausible than the late acquired ones (early: 6.0, late: 5.24, $t(64) = 4.19$, $p < .001$). The plausibility rating of each item was therefore controlled for in the statistical analyses reported below (see Table 6).

Subjective AoA ratings. As in Experiment 1, we collected subjective AoA ratings for all trigrams from 32 native English speakers (19 females and 13 males). We followed the exact same procedures and instructions used in Experiment 1. Participants rated all sixty-six experimental items (33 early-acquired and 33 late-acquired) as well as seventy single words taken from the Kuperman et al. norms (again, to ensure that our sample provides similar word AoA estimates). On each trial, participants saw a trigram or word on the screen and were asked to estimate the age (in years)

Table 5
Examples of matched early- and later- acquired trigrams and their plausibility and frequency measures for Experiment 2.

Early-trigram	Early child-directed-freq	Early-plausibility	Early-adult-freq	Late-trigram	Late-child-directed-freq	Late-plausibility	Late-adult-freq
a good girl	203	6.47	10	a good dad	0	6.52	9
take them off	84	6.36	27	take time off	0	6.47	28
you push it	77	5.8	3	you mail it	1	5.72	3
can eat it	60	5.75	6	can change it	1	5.47	8

Table 6
Adult frequency properties in the two conditions (per million words) for items in Experiment 2.

Condition	Word1	Word2	Word3	Bigram1	Bigram2	Trigram
Early	10,375	11,929	3798	467	36	0.55
Late	10,380	11,919	3793	416	30	0.56

when they first understood the item (even if they did not use it at the time).

All participants completed the task. The results corroborated our corpus-based classification: our early-acquired items were rated as learned earlier than our later-acquired ones: the early items were acquired only 5 days on average before the later ones (early: 4;03, late: 4;08, $t(64) = -0.32$, $p > .7$). Importantly, the correlation between the lexical AoA in our participant sample and that in the large-scale lexical AoA study (Kuperman et al., 2012) was very high ($r = .96$). The correlation between the current ratings and the ones collected for the same words in Experiment 1 was also very high ($r = .95$), further confirming the validity of the sample and the reliability of the subjective rating method.

Procedure

The procedure was identical to that of Experiment 1.

Results

Accuracy was high overall (mean of 97%) for both early-acquired (98%) and late-acquired items (95%), as is expected in lexical decision tasks. We excluded responses under 200 ms or more than 2.5 standard deviations from the mean of each condition. This resulted in the loss of 7% of the data. Incorrect responses were also excluded from the analysis.

We use the same mixed-effect regression models as in Experiment 1 to analyse the results. All models had the maximal random effects structure justified by the design (cf. Barr et al., 2013). The frequencies of the unigrams, bigrams, and trigrams were entered as control variables into all analyses in order to measure the effect of AoA while controlling for frequency. We ran a principal component analysis to reduce the collinearity between all the unigram, bigram and trigram frequency measures, which were collinear. This led to four components (instead of the six frequency measures), and ensured that collinearity in all reported models was small (all variance inflation factors [vif's] were under 2). We added the plausibility ratings to all analyses since they differed between the two conditions. We also controlled for the lexical AoA of the words in the two trigrams.

Reaction times

As predicted, reaction times were faster for early-acquired items compared to later ones (early: 720 ms (SD = 50), late: 771 ms (SD = 70)). A mixed-effects linear regression model was used to predict logged reaction times. We included *type* (early vs. late), log(*plausibility*) (logged to reduce skewness), *number-of-letters*, *average-lexical*-AoA (the averaged lexical AoA of the three words in the trigram), and the four PCA frequency components as fixed effects. We had *subject* and *item-pair* as random effects, as well as a by-subject random slope for *type*, and a by-*item* slope for *type* (to ensure the effects hold beyond item pairs – in each pair there was an early and a late variant - and subjects).

As expected, and as found in Experiment 1, early items were decided on faster than later ones ($\beta = -.04$ [SE = .01], $p < .05$; model comparison chi-square = 5.03, $p < .05$, See Table 7). The effect was significant controlling for all frequency measures, lexical AoA, and plausibility. *Plausibility* did not predict reaction times ($\beta = -.04$, SE = .05, $p > 0.4$, chi-square = 0.81) and neither did *lexical AoA*, which was better matched between the conditions ($\beta = .001$, SE = .01, $p > .9$, chi-square = 0.03). Unsurprisingly, items with more letters were responded to more slowly, $\beta = .02$, SE = .005, $p < .001$, chi-square = 16.33). None of the four aggregate frequency measures from the principal component analysis were significant (pca1: $\beta = .009$, SE = .008, $p > .3$, chi-square = 1.24; pca2: $\beta = -.002$, SE = .008, $p > .9$, chi-square = .15; pca3: $\beta = -0.01$, SE = .009, $p > .2$, chi-square = 1.44; pca4: $\beta = 0.005$, SE = .008, $p > .5$, chi-square = 0.53*)*. Because the two conditions were matched on all frequencies, and the items were selected to be from a small frequency range (smaller than that of Experiment 1), it not surprising that the frequency measures were not predictive of reaction times.

As in Experiment 1, we wanted to see if the subjective ratings (collected from a different sample) would predict reaction times. We ran an additional analysis to see how well the subjective AoA ratings predicted reaction times. We used the exact same model (in terms of fixed and random effects), but replaced the binary variable of *type* (early vs. late) with the log(*subjective rating*) for each trigram (logged to reduce skewness). The random slope between *type* and *item* was also removed because items were no longer treated as pairs.

Table 7
Mixed-effect regression with AoA as a binary measure (early vs. late) for Experiment 2. Significance obtained using the lmerTest function in R.

Fixed effects	Coef.	SE	*T*-value	*P*-value
Intercept	**6.36**	**.12**	**51.49**	**<.001**
AoA-Late	**.04**	**.01**	**2.23**	**<.05**
Plausibility	−.04	.05	−0.84	>0.4
Lexical-AoA	.001	.01	0.09	>.9
Num-Let	**.02**	**.005**	**4.04**	**<.001**
pca1	.009	.008	1.04	>3
pca2	.002	.008	0.02	>.9
pca3	−.01	.009	−1.09	>0.2
pca4	.005	.009	.63	>.5

Variables in bold were significant ($p < .05$).

Table 8
Mixed-effect regression with subjective AoA ratings for Experiment 2. Significance estimates were obtained using the lmerTest function in R.

Fixed effects	Coef.	SE	*T*-value	*P*-value
Intercept	**6.26**	**.08**	**75.3**	**<.001**
Subjective-AoA	**.08**	**.02**	**4.22**	**<.001**
Plausibility	−.07	.02	−2.67	<0.01
Lexical-AoA	−.004	.01	0.048	>0.9
Num-Let	**.02**	**.003**	**7.44**	**<.001**
pca1	.01	.008	1.35	>.1
pca2	−.006	.008	−0.68	>0.4
pca3	−.008	.009	−.98	>0.3
pca4	.004	.009	0.42	>0.6

Variables in bold were significant ($p < .05$).

Similar to Experiment 1, the *subjective ratings* were highly predictive of reaction times: items estimated as learned later were responded to more slowly than earlier ones ($\beta = .08$, SE = .02, $p < .001$, chi-square = 17.56, See Table 8), controlling for all frequency measures, lexical AoA and plausibility. Unlike the previous analysis, *Plausibility* was a significant predictor, with more plausible items being responded to faster ($\beta = -.07$, SE = .02, $p < .01$, chi-square = 7.15). This difference may be impacted by the higher correlation between plausibility and the subjective ratings ($r = -.34$, $p < .01$). Importantly, as in the previous model, *lexical AoA* was not significant ($\beta = -.005$, SE = .01, $p > .9$, chi-square = 0.09), suggesting that the unexpected pattern found when using subjective ratings in Experiment 1 was a spurious one. Items with more letters were responded to more slowly, $\beta = .02$, SE = .004, $p < .001$, chi-square = 55.1, $p < .001$). None of the four pca frequency measures were in this model as well (pca1: $\beta = .02$, SE = .008, $p > .1$, chi-square = 2.06; pca2: $\beta = -.006$, SE = .007, $p > .4$, chi-square = .42; pca3: $\beta = -0.008$, SE = .009, $p > .3$, chi-square = 1.04; pca4: $\beta = 0.004$, SE = .009, $p > .6$, chi-square = .26)

In sum, participants were faster to respond to early-acquired trigrams compared to later-acquired ones, after controlling for adult usage patterns, plausibility and lexical AoA. Moreover, the estimated age at which a trigram was acquired was a significant predictor of reaction times, as is the case for individual words. These findings replicate and strengthen the results of Experiment 1: they show that speakers are sensitive to multiword AoA even after matching the items on lexical AoA and applying a more stringent frequency criterion for matching the variants on adult usage patterns.

Discussion

The research on lexical AoA has demonstrated that early-acquired words show a processing advantage in adults compared to words that are acquired later. In this study, we extend these findings to show that the effect is not limited to words, but is also found for multiword sequences. We used a phrasal decision task to compare processing times between early- and late-acquired trigrams that differed only in one word and were matched on all adult frequencies, as well as word AoA (e.g., *for the <u>baby</u>* vs. *for the <u>men</u>*). The results of two studies – using two different sets of items - show that trigrams that were learned earlier – as estimated using both child-directed corpus frequencies and subjective ratings – were responded to faster compared to later acquired trigrams. The effect was significant both when using the corpus-based classification (early vs. late) and when using the subjective AoA ratings gathered from a different set of speakers. The effect cannot be attributed to usage patterns in adult language since it was found when controlling for all adult frequencies as well as plausibility: adults responded to early-acquired trigrams faster than later-acquired ones even though both the trigrams and the individual words were equally frequent in adult language (and after controlling for all frequencies in the analyses). These effects were found using two different sets of items, suggesting they are not limited to a particular set of phrases.

The combined results of the rating studies and the phrasal decision tasks show that (a) speakers are able to estimate the relative order of acquisition of multiword sequences, and (b) that these subjective estimates predict processing times, as they do for individual words. Speakers were faster to respond to phrases that were estimated as learned earlier (by a different set of participants). Both measures (the corpus-based ones and the subjective ratings) capture the *relative order of acquisition* of different sequences and provide an indication of what early building blocks for language look like. The findings indicate that, similar to words, multiword sequences that were learned earlier showed a processing advantage, after controlling for many properties in adult language use.

As in the case of lexical AoA effects, it is hard to prove a causal relation between order of acquisition and the processing advantage seen in adults. It is possible that early-acquired items were learned earlier because they are easier on some other dimension of meaning or form. Neither the current study, nor the large literature on lexical AoA effects can provide a definitive answer to this challenge: while studies can (and do) control for many of the linguistic properties of the items, it is theoretically possible that there are additional factors that were not accounted for and that drive the effect. One way of addressing this challenge is by using artificial language learning to study AoA effects: such settings provide full control of both the linguistic properties and the learning settings of the different items. Two studies have used such a design to show AoA effects (Izura et al., 2011; Catling et al., 2014): when participants were taught nonce words for novel objects (e.g., Greeble shapes), early-learned items showed processing advantages compared to later-learned ones. Since all other factors were kept equal (frequency, meaning, learning setting), such findings provide convincing evidence for the claim that order-of-acquisition has an independent and real role in generating the well-documented AoA effects.

Two additional factors are worth considering in more depth. The relation between AoA and two frequency measures that capture experience throughout the life span – cumulative frequency and frequency trajectory – has been debated in the lexical AoA literature. Cumulative frequency refers to the overall experience with a word throughout life: early-acquired words have a higher cumulative frequency compared to later-acquired ones by virtue of being known for more years (Lewis, Gerhand, & Ellis, 2001). Frequency trajectory refers to the change in experience during the lifespan: early-acquired words tend to have a high-to-low trajectory; they are encountered a lot early in life and less in adulthood (Zevin & Seidenberg, 2002, 2004). Both factors have been argued to be the real force behind AoA effect and both were not controlled for in the current study: could they be driving our effects? There is quite a lot of evidence in the lexical AoA literature against cumulative frequency being the underlying factor creating AoA effects. AoA effects persist after controlling for cumulative frequency (Ghyselinck et al., 2004; Perez, 2007;Moore & Valentine, 1998); AoA effects are found in lab conditions where both cumulative frequency and AoA are fully controlled (Stewart & Ellis, 2008); and AoA effects are not larger in older adults compared to younger ones, as predicted by the cumulative frequency hypothesis (Morrison et al., 2002, but see Catling et al., 2013). Moreover, cumulative frequency effects are rarely found for lexical decision tasks like the one we used: such effects (when found) seem to be limited to tasks where there is a non-arbitrary mapping between spelling and sound (like reading aloud tasks). In general, there seems to be a difference in the magnitude and stability of AoA effects between tasks that rely on more non-arbitrary spelling-sound mappings, like reading aloud tasks, and ones that draw on more arbitrary spelling-meaning mapping like the lexical decision task we used where the relation between an object and its label is mostly arbitrary (Bonin, Barry, Méot, & Chalard, 2004; Bonin, Méot, Mermillod, Ferrand, & Barry, 2009; Maermillod et al., 2012). Importantly, the hypothesis is even less applicable to our study since all of our items had very low frequency in the adult corpus (around one per million for both the early and late items): any cumulative difference between them would be very small since they are all low frequency in adult usage. In sum, cumulative frequency does not seem like a plausible explanation for our effects.

The relation between frequency trajectory and AoA is more complex: rather than viewing the two as contradictory, a recent proposal sees them as complementary (Maermillod et al., 2012). Frequency trajectory provides a richer, two-dimensional measure of AoA that encodes both the order of acquisition and the amount of exposure to items during learning and offers a way to make more precise predictions on how (and why) age-related effects occur in learning. This proposal is backed up by a series of computational simulations that assess the effects of AoA and frequency trajectory on learning arbitrary mappings (as those found in picture naming or lexical decision tasks) and non-arbitrary ones (like those found in reading aloud tasks). The results show effects of both AoA (early vs. late) and frequency trajectory in tasks with more arbitrary spelling-sound mappings but not with less arbitrary sound-form ones, a pattern that is consistent with other findings (e.g., Zevin & Seidenberg, 2004, vs. Perrez, 2007, but see Monaghan, Shillcock, Christiansen, & Kirby, 2014 for findings that earlier-acquired words have less arbitrary sound-mappings more generally). Interestingly, our study provides possible additional evidence for the utility of frequency trajectory in studying AoA effects: because we used actual corpus measures in constructing our items we know their frequency in both child-directed and adult speech. Our early-acquired items did indeed have the high-to-low frequency trajectory that is expected to result in a processing advantage. The fact that RTs were affected both by our corpus-based classification and by the subjective ratings (which are a measure of order of acquisition) is consistent with viewing the two as complementary measures of age-effects on learning.

The challenge of isolating order-of-acquisition from the other linguistic characteristics with which it is correlated or associated becomes even harder for multiword sequences. The relative paucity of research on the properties and processing of multiword sequences (compared to words) means that there are no established norms on the linguistic properties known to affect lexical processing for multiword sequences (e.g., imageability, neighbourhood density). Even more challenging is the fact that it is not clear how to operationalize such features for larger sequences (e.g., how does one measure the imageabiltiy of a sequence?). Consequently, this study should be seen as a first step in establishing AoA in multiword phrases: While there may indeed be additional differences between the early and late trigrams that were not taken into account, we controlled for many of the prominent factors affecting word processing (e.g., frequency, word AoA, plausibility).

This is the first study, to our knowledge, to uncover AoA effects for units larger than single words. The existence of such effects has psycholinguistic and developmental implications. From a developmental perspective, finding that multiword sequences show AoA effects provides strong support for their role as building blocks for language learning. Because both frequency and lexical AoA were controlled for, the effects in our study could not come about unless learners (at some point) had treated the sequence as one unit. Speakers' sensitivity to multiword AoA requires that they (a) remain sensitive to the AoA of the sequence in addition to (and independently from) that of its parts, and (b) keep track of both frequency and AoA for multiword sequences. Consequently, finding AoA effects for multiword sequences challenges the commonly held view that children first learn words and then use these basic lexical units to develop more complex and structured representations. Instead, our results suggest that children are sensitive to distributional information computed at multiple granularities (between sounds, words, and sequences of words), and draw on units of varying sizes in the process of learning (Christiansen & Chater, 2016b; Tomasello, 2003). Our findings further support the claim that children are sensitive to input frequencies (see Christiansen & Chater, 2016a; Diessel, 2007; Lieven, 2010, for reviews), and provide novel evidence for the usage-based prediction that multiword units serve as building blocks for language learning (Abbot-Smith & Tomasello, 2006). This prediction is hard to test by looking at child language because of the difficulty involved in identifying the units children learn from, especially given production-comprehension asymmetries. Looking at adult processing to find traces of early units – as in the current study – offers a novel way of examining children's building blocks, and provides additional evidence for their role in learning.

From a psycholinguistic perspective, the findings highlight the parallels between words and larger sequences. They blur the long-held lexicon-grammar distinction – multiword sequences show a key signature of lexical storage – and challenge the notion that words and larger patterns are processed by qualitatively different systems (Pinker, 1999; Pinker & Ullman, 2002). Instead, our findings are better accommodated by a single-system view of language (Bybee, 1998; Christiansen & Chater, 2016a,b; Croft, 2001; Elman, 2009; Langacker, 1987; McClelland et al., 2010; Wray, 2002) where all linguistic experience is processed by similar cognitive mechanisms. Our results corroborate and extend previous findings on the role of multiword units in online processing (e.g., Arnon & Snider, 2010; Jolsvai et al., 2013; Tremblay & Tucker, 2011) and underscore the importance of incorporating larger units into production and comprehension models (McCauley & Christiansen, 2013).

The findings also have methodological implications. They position multiword AoA as an additional factor that needs to be taken into account and controlled for in psycholinguistic studies. At the same time, the current studies highlight the link between child-directed frequencies and subjective AoA ratings and in doing so, offer an additional way to estimate AoA that is of relevance also for the lexical AoA literature. Despite the well-studied relation between input frequencies and language learning (e.g., Diessel, 2007; Ambridge et al., 2015), studies of lexical AoA have not explored the relationship between the frequency of a word in child-directed speech and its' assessed AoA. Instead, AoA has been estimated using subjective ratings (validated using child norming data). However, as in the case of multiword phrases, words with early AoA may be ones that appear often in child-directed speech. To explore this possibility, we compared the child-directed frequencies of words that were rated as early- and later-acquired based on the Kuperman et al. (2012) norms. We treated all words that had an AoA of under three as early-acquired words (there were forty such words) and matched this set with an equally sized set of words that were rated as learned after the age of five and had comparable frequency in adult language (adult frequency for early set: 435 per million, adult frequency for later set: 434 per million). At the group level, the early-acquired words indeed appeared more frequently in our child-directed corpus than the later-acquired ones (early: 363 per million, late: 79 per million, $t(78) = 3.83$, $p < .001$). Moreover, a regression analysis revealed that early-acquired words were more frequent in child-directed speech, *after* controlling for the relation between adult frequency and child-directed frequency[2]. While this analysis is preliminary, it suggests that child-directed corpus frequencies are correlated with estimated lexical AoA and can serve as a proxy for AoA.

The ability to identify units of learning in adult processing may also be relevant for the study of differences between first- (L1) and second-language (L2) learning. It has recently been proposed that some of the difference between children's and adults' language learning can be related to adults relying less on multiword units as building blocks for language learning (Arnon, 2010; Arnon & Ramscar, 2012; Ellis, Simpson-Vlach, & Maynard, 2008; Wray, 1999). This proposal assumes (a) that multiword units do serve as building blocks for native speakers, and (b) that L2 learners draw on them less (or in a different manner) when acquiring a second language. Both assumptions are difficult to test because of the challenges involved in identifying early building blocks. Finding AoA effects for multiword sequences provides strong support for the first assumption. Recent simulations by McCauley and Christiansen (in press) provide evidence for the second assumption. Employing a computational model—the chunk-based learner (CBL; McCauley & Christiansen, 2011, 2014)—that learns to process language by chunking together words, they compared the "chunkedness" of utterances produced by adult and child native speakers of English and German with the productions of native Italian speakers, learning English or German as their L2. The results indicated that the productions of adult and child native speakers were easier to recreate using multiword chunks compared to the language produced by the L2 learners. Although these results are preliminary, they highlight the importance of multiword building blocks in L1 language use as suggested by our AoA results, and the possible different use of such units in L2 learning.

In sum, the current study sharply undermines a long-held assumption in the study of language that treats words as ontologically distinct from larger sequences. Instead, we argue that multiword units, like words, serve as early building blocks that leave traces in adult language. The study revealed a novel effect of multiword AoA: speakers are sensitive to the order of acquisition of multiword sequences. This finding highlights the role of multiword units as important building blocks in language learning and use, and calls for their incorporation into models of language learning and processing.

Acknowledgments

We would like to thank Alex B. Fine, Ram Frost and Victor Kuperman for helpful discussion, as well as Doron Ariav, Klejda Bejleri, Jess Flynn, Scott Goldberg, Jordan Limperis, Limor Raviv, Samantha Reig, and Sven Wang for assistance with data collection and item preparation. This work was partially supported by BSF Grant number 2011107 awarded to IA and MHC.

[2] Words in the early set had a higher child-directed frequency ($\beta = .01$, SE = .03, $p < .001$), after controlling for the effect of adult-frequency on child-directed frequency, which was also significant (more frequent words in adult speech were also more frequent in child-directed speech, $\beta = .07$, SE = .01, $p < .001$).

Appendix A

Full list of items from Experiment 1 with plausibility ratings (scale from 1 to 7), frequencies (per million words based on child-directed speech), subjective AoA ratings (in years), and adult trigram frequency (per million words based on Fisher corpus)

Early trigram	Early plausibility	ET-child-freq	ET-age-rating	Late trigram	LT-plausibility	LT-childfreq	LT-age rating	Adultfreq
a good boy	6.36	58.2	3.43	a good mother	6.44	0	4.31	0.52
a good girl	6.47	44.2	3.56	a good dad	6.52	0	5.06	0.39
all the pieces	5.61	15.0	4.65	all the rooms	5.33	0.4	4.87	0.25
are you done	6.58	26.3	4.18	are you real	5.02	0.2	5	0.47
at school today	6.63	15.4	4.78	at two today	5.63	0	6.71	0.04
can eat it	5.75	13.0	3.46	can change it	5.47	0.2	5.62	0.7
can you push	5.63	13.3	4.25	as you push	4.83	0	6.62	0.04
don’t throw it	6.22	20.2	3.56	don’t fear it	4.94	0	5.06	0.47
done with this	6.13	13.0	5.15	through with this	5.63	0.4	4.53	0.08
for the baby	6.44	19.6	4	for the men	5.69	0.2	4.5	0.68
gonna fix it	5.91	13.7	4.65	gonna treat it	4.19	0	6.25	0.04
have a bite	6.41	17.6	3.84	have a wheel	2.47	0.6	5.67	0.08
in the bed	6.36	28.7	3.53	in the hands	4.69	0.2	6.46	2.08
in the train	5.27	12.8	4.56	in the fourth	4.47	0	4.34	0.76
make a car	3.16	12.4	6.06	maybe a car	4.19	0.2	6.71	0.08
on the box	5.69	16.1	3.81	on the eye	4.08	0.2	5.09	0.21
on the paper	5.3	62.6	3.87	on the third	4.44	0.8	6.56	0.87
on the potty	5.77	15.9	2.75	on the beds	4.75	0.6	4.03	0.08
on the slide	5.16	14.8	4.25	not the slide	4.11	0.2	4.75	0.04
on this page	6.13	25.7	4.56	on this salad	5.14	0	5.56	0.04
on your shirt	6.27	20.7	3.78	for your shirt	4.16	0.2	4.5	0.04
play with something	5.94	13.5	3.28	play with too	2.86	0.2	3.71	0.04
play with those	5.69	15.7	3.37	play with kids	6	0	4.87	0.12
read some books	6.63	11.3	4.34	read more books	6.55	0.4	6.87	0.16
show me where	6.47	18.3	3.65	show me things	5.52	0	6.06	0.12
sit up here	6.02	14.8	3.12	anybody up here	5.94	0	3.84	0.04
smell the flowers	6.47	10.9	3.81	fix the flowers	4.22	0	4.12	0.04
take them off	6.36	18.3	3.87	take time off	6.47	0	4.71	1.13
that’s a car	6.02	18.5	3.43	that’s a phone	5.97	0.4	3.62	0.08
the red one	5.97	41.8	3.09	the earlier one	5.25	0	9.18	0.04
use this one	6.52	12.4	4.15	made this one	5.61	0.2	3.62	0.08
wanna do that	5.83	28.3	4.18	couldn’t do that	6.25	0.2	5.93	4.37
wanna sit down	6.13	18.5	3.12	couldn’t sit down	6.16	0	7.12	0.16
want some help	6.61	10.9	4.12	want good help	4.66	0	6.9	0.04
you can’t play	6.33	25.5	4.21	you can’t give	4.63	1.5	5.03	1.22
yougonna help	5.63	15.7	4.21	yougonna change	4.11	0.2	4.43	0.04
you push it	5.8	16.8	3.81	you mail it	5.72	0.2	4.71	0.12
you threw it	5.91	14.1	3.75	you fed it	4.5	0	7.51	0.04
youwanna talk	6.22	13.0	4.9	you couldn’t talk	5.25	0	6.12	1.73
you’regonna fall	6.13	14.6	3.4	you’regonna quit	5.91	0	6.25	0.04

Appendix B

Full list of items from Experiment 2 with plausibility ratings (scale from 1 to 7), frequencies (per million words based on child-directed speech), subjective AoA ratings (in years), and adult trigram frequency (per million words based on Fisher corpus)

Early trigram	Early plausibility	ET-child-freq	ET-age-rating	Late trigram	LT-plausibility	LT-child-freq	LT-age-rating	Adult-freq
a big truck	6.29	9.4	3.3	a big key	4.97	0.19	4.1	0.45
a good book	6.29	10.6	5.2	a good team	6.26	0.00	7.2	1.95
a good girl	6.14	44.5	4.8	a good mother	6.26	0.00	4.6	0.55
all the pieces	5.67	13.0	4.7	all the parties	4.65	0.00	6.9	0.28
are you drawing	6.15	11.1	3.6	are you proud	6.38	0.19	6.3	0.05
are you eating	6.53	35.3	3.8	are you black	5.38	0.00	6.2	0.35
don't see it	6.21	24.7	4.0	don't see that	5.26	0.57	4.4	11.15
down the slide	5.82	24.5	3.5	down the bend	5.00	0.00	7.8	0.05
for the baby	6.02	19.2	3.6	for the teacher	6.18	0.38	5.2	0.80
get the book	6.47	9.8	4.3	get the number	5.74	0.38	5.9	0.53
going to bed	6.59	14.3	3.0	going to pass	5.41	0.19	7.0	1.03
had a farm	5.59	15.7	3.9	had a boat	4.88	0.19	5.0	0.90
have a cookie	6.56	10.9	2.9	have a costume	4.65	0.57	4.4	0.15
have a snack	6.62	9.4	3.2	have a photo	4.32	0.00	5.3	0.20
i got one	6.09	10.0	3.6	i got out	6.03	0.94	4.3	12.78
if you're happy	6.18	24.0	4.7	if you're running	5.26	0.00	5.6	0.60
in a box	6.35	17.5	3.4	in a magazine	6.21	0.19	5.7	1.93
in my hair	6.06	10.4	3.8	in my view	5.97	0.00	7.6	0.60
in my pocket	6.44	12.3	3.7	in my forties	6.26	0.00	9.5	2.15
in the bag	6.26	76.6	4.7	in the magazines	5.88	0.00	5.9	0.55
in the bathroom	6.56	14.3	3.2	in the streets	6.26	0.19	4.8	4.33
in the bed	6.26	39.8	3.4	in the heart	4.94	0.19	6.2	2.30
in the hat	4.59	18.5	4.1	in the ears	4.47	0.00	4.3	0.23
in the pool	6.29	11.5	3.9	in the papers	5.21	0.00	6.3	3.93
in the potty	5.44	10.9	2.9	in the boot	4.41	0.19	6.1	0.15
in the sandbox	6.03	9.8	3.6	in the shadows	5.91	0.00	6.0	0.10
in the store	6.41	11.7	4.5	in the restaurant	6.44	0.38	5.5	5.58
in the sun	5.56	9.8	4.3	in the bible	6.35	0.00	5.8	4.55
in the trash	6.41	15.8	3.8	in the hills	5.06	0.19	5.3	1.60
in the zoo	5.91	13.6	4.0	in the beaches	3.24	0.00	5.7	0.10
on the book	4.74	13.4	4.1	on the cat	4.29	0.00	3.9	0.98
on the tree	5.24	11.7	3.9	on the windows	4.62	0.38	4.5	0.45
on this page	6.38	23.8	5.5	on this hill	5.47	0.19	4.9	0.05
pick you up	6.21	26.6	4.0	set you up	5.44	0.38	7.7	0.98
see the book	5.76	10.0	2.9	see the number	5.15	0.00	4.1	0.20
sit up here	5.82	16.4	3.1	run up here	5.50	0.00	4.1	0.05
some more juice	5.94	13.8	3.1	some more seats	5.18	0.00	5.8	0.05
that was nice	6.79	11.9	3.7	that was real	6.03	0.19	5.8	4.80
that's a baby	5.85	24.3	3.2	thats a plane	6.35	1.13	3.6	0.05
to the baby	5.12	11.5	3.2	to the girls	4.56	0.57	4.7	0.50
to the bath	4.74	10.0	3.1	to the babysitter	5.47	0.38	4.9	0.15
under the chair	6.15	10.6	3.2	under the sea	5.88	1.51	4.8	0.05
wipe it off	6.62	11.9	3.6	throws it off	4.32	0.00	6.4	0.10
with the toys	5.38	12.8	3.4	with the boxes	5.03	0.00	4.6	0.10
you can write	6.32	9.8	4.6	you can win	6.41	0.00	4.9	1.70
you pull it	5.38	14.3	4.3	you pass it	4.50	0.19	5.4	0.68

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage based account of syntactic acquisition. *The Linguistic Review, 23*, 275–290.

Ambridge, B., Kidd, E., Rowland, C., & Theakston, A. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*, 239–273.

Arnon, I. (2010). *Starting Big: The role of multiword phrases in language learning and use* PhD dissertation : . Stanford University.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth – Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development, 7*(2), 107–129.

Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech, 56*(3), 349–371.

Arnon, I., & Cohen Priva, U. (2014). Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon, 9*, 377–400.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*(3), 292–305.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67–82.

Bannard, C. (2006). *Acquiring phrasal lexicons from corpora* : . University of Edinburgh.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences, 106*(41), 17284–17289.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science, 19*(3), 241–248.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *Linguistic Review, 23*(3), 291–320.

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science, 33*(5), 752–793.

Bonin, P., Barry, C., Méot, A., & Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: A never ending story? *Journal of Memory and Language, 50*, 456–476.

Bonin, P., Méot, A., Mermillod, M., Ferrand, L., & Barry, C. (2009). The effects of age of acquisition and frequency trajectory on object naming. *Quarterly Journal of Experimental Psychology, 62*, 1–9.

Borensztajn, G., Zuidema, W., & Bod, R. (2009). Children's grammars grow more abstract with age–evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science, 1*(1), 175–188.

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition, 13*(7–8), 992–1011.

Bybee, J. (1998). The emergent lexicon. In *The 34th Chicago linguistic society* (pp. 421–435).

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of don't in English. *Linguistics, 37*(4), 575–596.

Buttery, P., & Korhonen, A. (2005). Large scale analysis of verb subcategorization differences between child directed speech and adult speech. Paper presented at the *Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*. Saarbrucken, Germany.

Catling, J., Dent, K., Preece, E., & Johnston, R. (2014). Age-of-acquisition effects in novel picture naming: A laboratory analogue. *Quarterly Journal of Experimental Psychology, 66*(9), 1756–1763.

Catling, J., South, F., & Dent, K. (2013). The effect of age of acquisition on older individuals with and without cognitive impairments. *The Quarterly Journal of Experimental Psychology, 66*, 1963–1973.

Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language, 89*, 244–254.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Christiansen, M. H., & Chater, N. (2016a). *Creating language: Integrating evolution, acquisition, and processing*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (2016b). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral & Brain Sciences, 39*, e62. http://dx.doi.org/10.1017/S0140525X1500031X.

Cieri, C., Miller, D., & Walker, K. (2004). The Fisher Corpus: A resource for the next generations of speech-to-text. In *Proceedings of the language resources and evaluation conference*. .

Clark, E. V., & Hecht, B. F. (1983). Comprehension, production, and language acquisition. *Annual Review of Psychology, 34*, 325–349.

Croft, B. (2001). *Radical construction grammar*. Oxford: Oxford University Press.

Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax*. Oxford: Oxford University Press.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology, 25*(2), 108–127.

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 24*(2), 515–523.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375–396.

Elman, J. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science, 33*, 547–582.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using mechanical turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass, 5*(8), 509–524.

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologia, 115*, 43–67.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92* (pp. 517–520).

Goldberg, A. (2006). *Constructions at work*. Oxford: Oxford University Press.

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.

Grimm, A., Muller, A., Hamann, C., & Ruigendijk, E. (Eds.). (2011). *Production-Comprehension Asymmetries in Child Language*. Berlin: De Gruyter.

Izura, C., Pérez, M. A., Agallou, E., Wright, V. C., Marín, J., Stadthagen-González, H., & Ellis, A. W. (2011). Age/order of acquisition effects and the cumulative learning of foreign words: A word training study. *Journal of Memory and Language, 64*(1), 32–58.

Janssen, N., & Barber, H. (2012). Phrase frequency effects in language production. *PLoS One, 7*. http://dx.doi.org/10.1371/journal.pone.0033202.

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition, 13*(7–8), 789–845.

Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (2013). Meaning overrides frequency in idiomatic and compositional multiword chunks. In N. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 692–697). Austin, TX: Cognitive Science Society.

Juhasz, B. J. (2005). Age-of-Acquisition effects in word and picture identification. *Psychological Bulletin, 131*(5), 684–712.

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition, 13*(7–8), 846–863. http://dx.doi.org/10.1080/13506280544000075.

Katz, S. M. (1996). Distribution of content words and phrases in text and language modeling. *Natural Language Engineering, 2*, 15–59.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978–990. http://dx.doi.org/10.3758/s13428-012-0210-4.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* : . Stanford University Press.

Lewis, M. B., Gerhand, S., & Ellis, H. D. (2001). Re-evaluating the age-of-acquisition effect: Are they simply cumulative frequency effects? *Cognition, 78*, 189–205.

Lieven, E. (2010). Input and first language acquisition: Evaluating the role of frequency. *Lingua, 120*(11), 2546–2556.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language, 30*(2), 333–370.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20*(3), 481–507.

Lieven, E., & Tomasello, M. (2008). Children's first language acquisition from a usage-based perspective. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 168–196). New York and London: Routledge.

MacWhinney, B. (2000). *The CHILDES project* (3rd edn). Hillsdale, NJ: Erlbaum.

Maermillod, M., Bonin, P., Meot, A., Ferrand, L., & Paindavoine, M. (2012). Computational evidence that frequency trajectory theory does not oppose but emerges from Age-of-Acquisition theory. *Cognitive Science, 36*, 1499–1531.

McCauley, S. M. & Christiansen, M. H. (in press). *Computational investigations of multiword chunks in language learning*. Special issue on Multiword Units in Language, Topics in Cognitive Science, xx-xx.

McCauley, S. M., & Christiansen, M. H. (2013). Towards a unified account of comprehension and production in language development. *Behavioral & Brain Sciences, 36*, 38–39.

McCauley, S. M., & Christiansen, M. H. (2014). Acquiring formulaic language: A computational model. *Mental Lexicon, 9*, 419–436.

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 1619–1624). Austin, TX: Cognitive Science Society.

McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science, 2*(4), 751–770.

McClelland, J. L., Botvinick, M., Noelle, D., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to understanding cognition. *Trends in Cognitive Sciences, 14*, 348–356.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*, 20130299.

Moore, V., & Valentine, T. (1998). Naming faces: The effect of AoA on speed and accuracy of naming famous faces. *The Quarterly Journal of Psychology, 51*, 458–513.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology Section A, 50*(3), 528–559.

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(1), 116–133.

Morrison, C. M., Hirsh, K. W., Chappell, T., & Ellis, A. W. (2002). Age and age of acquisition: An evaluation of the cumulative frequency hypothesis. *European Journal of Cognitive Psychology, 14*(4), 435–459.

Perez, M. A. (2007). Age of acquisition persists as the main factor in picture naming when cumulative frequency and frequency trajectory are controlled. *Quarterly Journal of Experimental Psychology, 60*, 32–42.

Pierrehumbert, J. B. (2012). Burstiness of verbs and derived nouns. In *Shall we play the Festschrift game?* (pp 99–115). Berlin: Springer.

Pinker, S. (1991). Rules of language. *Science, 253*(5019), 530–535.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6*(11), 456–463.

Reali, F., & Christiansen, M. H. (2007). Word chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology, 60*(2), 161–170.

Shi, R., Werker, J., & Cutler, A. (2006). Recognition and representation of function words in English-learning infants. *Infancy, 10*, 187–198.

Stewart, N., & Ellis, A. (2008). Order of acquisition in learning perceptual categories: a laboratory analogy of the age-of-acquisition effect? *Psychonomic Bulletin & Review, 15*, 70–74.

Soderstrom, M. (2007). Beyond baby talk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review, 27*, 501–532.

Stadthagen-Gonzalez, H., Bowers, J. S., & Damian, M. F. (2004). Age-of-acquisition effects in visual word recognition: Evidence from expert vocabularies. *Cognition, 93*, B11–B26.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods, 38*, 598–605.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*(2), 569–613.

Tremblay, A., & Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon, 6*(2), 302–324.

Wray, A. (1999). Survey article. *Language Teaching, 32*, 213–231.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language, 47*, 1–29.

Zevin, J. D., & Seidenberg, M. S. (2004). Age-of-acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory and Cognition, 32*, 31–38.