Journal of Child Language (2018), **45**, 1073–1090
doi:10.1017/S0305000918000089

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Self-reported parental vocabulary input frequency for young children

Dorthe BLESES[1], Werner VACH[2], and Philip S. DALE[3]*[†]

[1]TrygFonden's Centre for Child Research and School of Communication and Culture, Aarhus University, Denmark, [2]University Hospital Basel, Switzerland, and [3]Department of Speech & Hearing Sciences, University of New Mexico, USA
*Corresponding author. Department of Speech & Hearing Sciences, University of New Mexico, 1700 Lomas Blvd NE, Suite 1300, Albuquerque, NM 87131. E-mail: dalep@unm.edu

(Received 3 March 2017; revised 5 December 2017; accepted 14 March 2018;
first published online 15 April 2018)

Abstract

Vocabulary input frequency influences age of acquisition, and is also an essential control for investigating the influence of other factors. We propose a new method of frequency estimation, self-report. 918 Danish-speaking parents of 12–36-month-old children estimated their frequency of use of 725 words. Self-report was substantially correlated with both language sample based frequencies (0.67) and frequencies of a large written corpus of Danish (0.58). Correlations within vocabulary categories between frequency and age of acquisition, restricted to words occurring in the language samples, were comparable for the two estimates. Overall, self-report based frequency estimates appear to have a promising degree of validity, which reflects their greatest strength, independence of the situation.

Keywords: vocabulary; input; parent

The acquisition of vocabulary is one of the major accomplishments of childhood. As early as age 2;6, the median vocabulary size is about 570 words for children acquiring American English (Fenson, Marchman, Thal, Dale, Reznick, & Bates, 2007), but substantial cross-linguistic variation has been documented (Bleses *et al.*, 2008); estimates of vocabulary at school entry often exceed 10,000 words (Carey, 1978). Beyond size, there is much diversity of words with respect to phonological, grammatical, syntactic, and pragmatic features, which adds to the challenge for the child, and for acquisition theories as well.

Differences in age of acquisition across words can therefore serve as a natural laboratory for studying the influence of various factors, including both the nature of linguistic input and features inherent in the words. The validity of such an approach

[†]We are grateful to the parents who provided the information about their children for this study. We also thank Rune Jørgenson of Southern Denmark University, who set up the online system for data collection, Ivan Iachine for data preparation, and Laila Kjaerbaek for assistance with the language samples. Access to datafiles and the web-administered questionnaire is available via the Open Science Framework at <http://osf.io/bjgct>.

© Cambridge University Press 2018

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

is also supported by some commonalities in sequence of acquisition which have been found across children within a language, and even cross-linguistically. For example, nominals make up the great majority of early vocabulary; predicates (verbs and adjectives) begin to emerge somewhat later; and closed-class words such as articles and prepositions are last to emerge (Bates *et al.*, 1994). Within nominals, as Brown (1958) first noted and Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976) were to elaborate, basic-level words are learned before subordinates and superordinates.

The acquisition of each word is a complex, multi-faceted process which takes place over time, not a single dichotomous change of state (Hoff, 2014, ch. 5; Marchman & Dale, 2018). A variety of criteria have been used to operationalize acquisition, including comprehension, production, generalization to new exemplars, and providing an oral definition. For the purposes of the present paper, ACQUISITION will be used to represent spontaneous (i.e., not elicited) production of a word with both pronunciation and meaning related to, though not necessarily identical with, the adult form. In addition to the face validity of this criterion, it has the advantage of being more straightforward to assess than comprehension, or mastery of the precise meaning.

Within the set of potentially influential factors, input frequency plays a primary role. An extensive discussion of the role of frequency in vocabulary and other aspects of language development is provided in Ambridge, Kidd, Rowland, and Theakston (2015), and the following commentaries. Most theoretical accounts assume, either explicitly or implicitly, that the more often a word is heard, the earlier it will be learned. Undoubtedly some exposures are more productive than others, such as exposure during joint attention or semantic contingency (Tomasello, 2003), utterance-final position (Naigles & Hoff-Ginsberg, 1998), and degree of morphological variation (Naigles & Hoff-Ginsberg, 1998). Nevertheless, it is unlikely that these other factors would completely obscure the role of frequency.

Input frequency has a second important role to play in the study of age of acquisition of vocabulary items. It is an essential control in the study of other factors, including phonological, syntactic, semantic, and pragmatic features. For example, if single-syllable words are acquired on average earlier than two-syllable words, or verbs before adjectives, this might reflect differences in the relative frequency of these categories rather than phonological or syntactic properties. Here, the size and diversity of vocabulary is an asset for research, allowing some statistical disentangling of frequency and other properties.

Finally, input frequency for words is an interesting and important topic in its own right. To what extent is the frequency of individual words in child-directed speech affected by the gender and birth order of the child, and socioeconomic factors such as maternal education? Frequency differences may play a causal role in child variance, which has been demonstrated to be correlated with those socioeconomic factors, e.g., faster vocabulary growth for girls than boys, for first born than later born, and for children whose mothers have more education, though all effect sizes are modest (Fenson *et al.*, 2007). Most research on gender differences, for example, has been motivated by interest in the source of the well-documented and consistent, though small, differences favoring girls in early language development (Eriksson *et al.*, 2011). However, it has largely concentrated on holistic and pragmatic aspects such as total quantity and conversational topic, rather than vocabulary composition and frequency and other structural aspects. For example, parents are more likely to acquire and play with action-oriented toys with boy, whereas symbolic play is more common with girls.

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

Despite the value of input frequency information, it is not easily obtained. The 'gold standard' for frequency is the analysis of parent–child language samples. Because of the size of the vocabulary used with young children, most of which can occur only rarely, large language samples are needed. Even more challenging is the role of the context, or situation. Language samples are most often recorded in conditions of parent–child play or book reading. These contexts make some words more common, and others less frequent or entirely absent. The less frequent and absent words might occur frequently in eating, dressing, or bedtime contexts. Simply adding more samples will not necessarily address this problem unless other contexts are sampled.

In part due to the challenges just mentioned, previous research on vocabulary input has generally examined either the total number of words addressed to children, or the lexical diversity of the input (Zauche, Thul, Darcy Mahoney, & Stapel-Wax, 2016). Both of these measures are correlated with vocabulary and other aspects of early language development. Pan, Rowe, Singer, and Snow (2005) found that lexical diversity of the input was a stronger predictor of vocabulary growth than total input. However, this research does not address the question of frequency as a predictor of acquisition for individual words. There is a smaller body of research which has examined the frequency of individual words, but these studies have typically focused on a smaller set of words, e.g., Blackwell's (2005) study of the relation between frequency of individual adjectives in maternal speech, along with other properties of the adjectives, and their age of acquisition.

In the first study to empirically relate input frequency to age of acquisition for a substantial proportion of the vocabulary, Goodman, Dale, and Li (2008) used all language samples for English in the Child Language Data Exchange System (CHILDES; MacWhinney, 2000) which included children between 0;7 and 7;5 (M = 3;0), for a total of about 3.8 million word tokens. Age of acquisition was based on the norming data for the MacArthur-Bates Communicative Development Inventories (Fenson *et al.*, 2007). The correlation between input frequency and age of acquisition based on all words was positive; that is, more frequent words were acquired later, not earlier. However, this was due to the well-established findings about vocabulary categories mentioned above, e.g., although nominals are frequent in language to children, individual nominals are not used often by parents though they are learned early, whereas closed-class items are used very frequently by parents, but are acquired later than nominals. Thus the primary analysis was within word categories and, in every case, there was a significant negative correlation between frequency and age, ranging from –0.24 to –0.55. The more frequent words in each category were produced earlier than less frequent words in that category. This finding was replicated and extended to six additional languages by Braginsky, Yurovsky, Marchman, and Frank (2016).

Braginsky *et al.* (2016) used language corpora in CHILDES for their additional languages in those few cases for which there exists a substantial body of samples. For the great majority of languages of the world there is little or no such evidence available. This is particularly unfortunate because in this area, as for so many, cross-linguistic comparisons are often the most illuminating.

An alternative method of estimating input frequency is much needed. In the present study, we explored the value of a new method: asking parents how frequently they use individual words via a web-based questionnaire. Parents have proven themselves highly valid in reporting the production of individual words by their children (Fenson *et al.*, 2007; Law & Roy, 2008; for Danish specifically, Bleses *et al.*, 2008); can they do the same for themselves? If so, this would be a highly cost-effective method for obtaining

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

frequency information. Dale, Tosto, Hayiou-Thomas, and Plomin (2015) utilized a parent questionnaire to obtain information about input, but it was largely focused on overall quantity of talking with the child, rather than specific words; in contrast, Tan and Schafer (2005) obtained reports from parents about their ostensive naming of 24 familiar objects. In both cases, the measures appeared to have some validity in the sense of being correlated with aspects of child language development, which was encouraging for the development of the present measure.

The present study

Our broad goal in this research is to evaluate parents as reporters of input frequency. In the present study, we examined parental self-reports of frequency of vocabulary use to young children in Danish, a language for which there are only very limited language samples available. The aim is to characterize the input to children in general, not to provide a measure for individual parents to be compared with their own child's development, a goal which would require very rich language samples, in both quantity and breadth of contexts, such as the work of Roy, Frank, DeCamp, Miller, and Roy (2015). The present strategy has an advantage noted by Goodman *et al.* (2008): there might be a correlation between input frequency and age of acquisition, but input frequency might vary so much across dyads that no overall correlation results. If children's age of acquisition is correlated with input frequency from a different group of parents, then the effect is quite general.

To evaluate the validity of parental self-reports, we correlate the results with measures based on parent–child language samples and also with frequency information from a large sample of written Danish. In addition to comparing these self-reports with the available language sample information and the adult corpus, we asked whether the findings overall were similar to those found for English and other languages with respect to variability in frequency of use and correlation with age of acquisition. Specifically, we addressed the following research questions:

1. Does self-reported input frequency have reasonable face validity with respect to identifying words with high/low frequency, and with respect to age trends?
2. Does self-reported input frequency correlate with estimates derived from language samples and from written language corpora?
3. Does self-reported input frequency replicate the within-category input age of acquisition correlations found in English?
4. Do discrepancies from the correlations in #3 suggest the role of other factors?
5. Do self-reported input frequencies vary with the demographic variables of a child's gender and the presence of siblings, and maternal education?

Method

Participants

The participants in this study were native Danish-speaking parents of children between 12 and 36 months, all residing in Denmark. A postcard was mailed to 6,000 randomly selected parents of children in the month prior to ages 12, 18, 24, 30, and 36 months; addresses were provided by the Danish Central Office of Civil Registration. It included a link to a web-based, Danish adaptation of the MacArthur-Bates Communicative Development Inventory: Words & Sentences (CDI:WS), which includes a vocabulary checklist of 725 items (Bleses

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

et al., 2008). 1213 parents responded to the form, but not all responses were complete. 918 responded to at least 625 items (no more than 100 omitted), and this was judged sufficiently complete for analysis. Parents often responded prior to the target age; that is, data were obtained in the age periods 11–12, 17–18, 23–24, 29–30, and 35–36 months. A mother's report of their education was coded using the categories established by Statistics Denmark: (1) primary/elementary school only; (2) high school (general, business, or technical); (3) vocational post-secondary education (e.g., retail, health assistant, security guard); (4) short higher education (2–3 years, e.g., financial management); (5) medium higher education (3–4 years, e.g., teacher); (6) higher education (min. 5 years, e.g., lawyer); (0) missing or not classifiable. There was considerable diversity in the sample, but overall parents were well educated, with a median value of 5 compared to a census median of 4. Parents with lower education were particularly under-represented; only 4% had only a high school education or less, compared to the census proportion of 18%.

Measures and procedure

Self-reported vocabulary input frequency

For all 725 words on the Danish CDI, parents responded to the question "How often do you say this word to your child?" by choosing among the alternatives: every day (3 points) / 2–4 times per week (2 pts) / less often (1 pt). The mean of this score was used for analysis. A translation of the directions for parents is provided in the 'Appendix'. Access to the web-administered questionnaire is available via the Open Science Framework at <http://osf.io/bjgct>.

Vocabulary input frequency from language samples

Language samples are available in the CHILDES database for two children learning Danish, who are described in Plunkett (1986). These are longitudinal studies with monthly recordings from 8 to 36 months. The corpora include 94 language samples, 41 for Anne and 53 for Jens. The fortnightly visits were approximately 1.5 hrs in length; they included Piagetian sensorimotor tests, as well as a final free play episode (no indication of the length of this episode is given). The authors state "…parent and child were encouraged to engage in a variety of social situations. An attempt was made to establish some regularity in the kind of situations observed across visits (feeding time, solving a problem together, story-telling). However, importance was attached to collecting naturalistic data and so coercion was avoided." (Plunkett, 1986, p. 65). There was a total of 31,242 maternal utterances which include 128,419 word tokens, comprising 5,730 word types. This input frequency measure was obtained for 664 of the 725 words, excluding animal sounds, CDI items which were multiword forms (e.g., *ved siden af* = 'next to') and family-specific proper names (e.g., name of babysitter). These exclusions were similar to those of Goodman *et al.* (2008). For analysis, the input frequency was indexed by $\log_{10}$ (total frequency in samples).

Vocabulary frequency from Korpus 2000

The Society for Danish Language and Literature has compiled several large corpora of written language, with the primary goal of supporting the development of dictionaries (Asmussen, 2006). The corpora are also publicly available online at <http://www.dsl.dk/korpus2000>. Korpus 2000 contains approximately 28 million words from texts published 1998–2002. The texts are primarily drawn from newspaper articles (approximately 2/3 of the corpus), along with novels and magazine articles.

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

Age of acquisition for individual vocabulary items

Drawing on the normative dataset (N = 6112) for the Danish CDI (Bleses *et al.*, 2008), an age of acquisition for each word (Age50) was defined as the age at which 50% of the children are reported to produce the word. This criterion was based on fitting a logistic growth curve for each word, which enabled determining an estimate by extrapolation even when fewer than 50% of the 36-month-old children were reported to produce the word. However, seven words with reported production by less than 25% of the children at 36 months were omitted.

Results

Which are high and low input frequency words, and which words show increasing or decreasing age trends? (Research question 1)

Table 1 includes the 11–13 words (ties for input frequency necessitated this variability) which had the highest and lowest input frequency, aggregating across all child ages, and also the words which showed the greatest increase or decrease in reported input frequency from 12 to 36 months. The words with highest input frequency were primarily social interaction words ('yes', 'no', 'hi', 'night-night', 'thank you'), people's names, and words associated with ingestion ('breakfast', 'food', 'to drink') or sleep routines ('bed', 'to sleep'). Words which decreased from 12 to 36 months were primarily social routines and animal sounds. Words which became more frequent over this age range were quite diverse, including some words for clothes ('underpants', 'boots'), hygiene items ('toilet paper', 'soap'), play items/location ('bicycle', 'playground'), and a few verbs ('to draw', 'to hurry', 'to run'). Overall, these trends are consistent with studies of vocabulary composition in children (Bleses *et al.*, 2008; Graham, San Juan, & Vukatana, 2015; Hoff, 2014).

Does self-reported frequency in Danish correlate with estimates derived from language samples and from written corpora? (Research question #2)

When self-reported input frequency was correlated with (log) frequency in the Plunkett corpus, the resulting Spearman *rho* correlation = 0.14 ($p < .001$, two-tailed). This extremely weak overall relation was primarily due to the fact that so many of the 664 words did not occur at all (i.e., frequency = zero) in the Plunkett corpus. Because the Plunkett corpus was quite small for the purpose of estimating frequency, and included only a very narrow range of contexts, it was decided to focus on the 147 words on the Danish CDI which did occur at least once in the language samples. A qualitative analysis of words not occurring in the Plunkett corpus despite relatively high self-reported frequency confirmed the role of context. Of the ten words with highest input score which were completely absent in the Plunkett corpus, eight were associated with eating or bedtime routines ('night-night', 'sleep', 'eat', 'breakfast', 'drink', 'toothbrush', 'dinner', 'pyjamas'). The exceptions were 'pants' and 'play'. For the restricted subset of words, the correlation was moderately strong ($rho = 0.67$, $p < .01$). A scatterplot of the relationship is shown in Figure 1. Figure 1 also identifies a few words for which there is a large discrepancy between the two sources. Of the five words which were frequent in the language samples but not in parental self-reports, four ('clown', 'bad', 'sick', and 'roof') are likely to be words which occurred in book-reading or story-telling with the children. Most of the words that

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

Table 1. Words with Highest and Lowest Average Score, and Greatest Increase and Decrease from 8–36 Months, Based on Parental Self-Report

Highest average		Lowest average		Greatest decrease		Greatest increase	
Word	Score	Word	Score	Word	Change	Word	Change
No	2.99	Cinema	1.02	Peekaboo	−1.29	Underpants	1.35
Yes	2.98	'children's egg' (chocolate egg with toy inside)	1.02	Patty cake	−1.14	To draw	0.97
Mommy	2.98	Maelkesnitte (chocolate treat)	1.02	Porridge	−0.82	Boots	0.86
Hi	2.98	Backyard	1.02	So big!	−0.80	Toilet paper	0.79
Child's name	2.97	Sled	1.03	Stroller	−0.68	Soap	0.78
Night-night	2.97	Peanut butter	1.03	To clap	−0.66	Bicycle	0.74
Food	2.98	Swimming pool	1.04	Yum yum	−0.59	Playground	0.72
To sleep	2.96	Crib	1.06	Woof woof	−0.54	To hurry	0.72
Daddy	2.96	Camping	1.06	moo	−0.46	Pillow	0.71
Thank you	2.96	Father's brother	1.07	Meow	−0.44	Story	0.70
Bed	2.95	Snowman	1.07	Quack quack	−0.40	To run	0.84
Breakfast	2.95	Turkey	1.07	Baa baa	−0.34		
To drink	2.95						

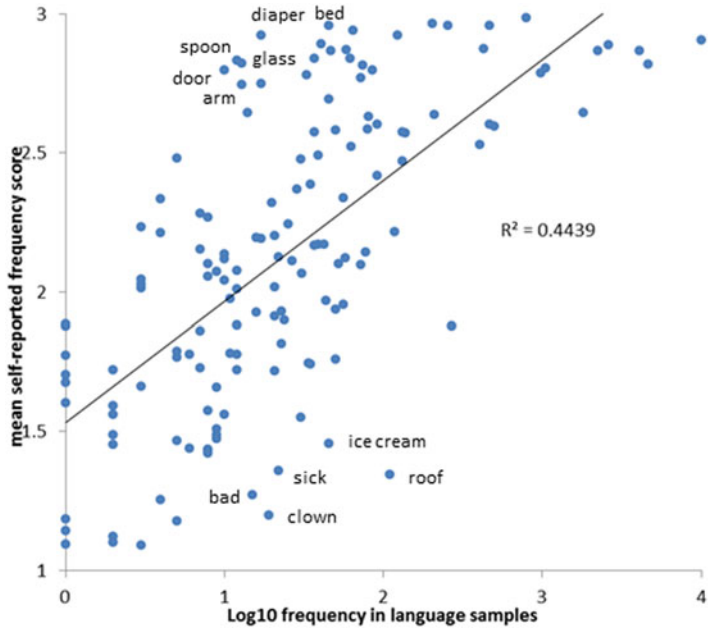


Figure 1. Scatterplot of mean self-reported input frequency with $\log_{10}$(frequency) in Plunkett (1986) language samples for 147 words which are present in the latter.

are reported to be used frequently by parents, but not found in the language samples, are likely to be associated with routines which did not occur in the language sample contexts, such as eating ('spoon', 'glass'), personal care ('diaper'), and sleeping ('bed').

A second source of validity coefficients is based on frequency in Korpus 2000. Not all the 664 words occur in Korpus 2000, again because some of the CDI vocabulary entries are homographs or multiword phrases. But for the 519 words which do occur, the correlation between frequency and mean reported frequency is *rho* = 0.58 ($p < .001$). Interestingly, the two observational measures of frequency are only weakly to moderately correlated (*rho* = 0.21, $p < .001$) for all 519 words in Korpus 2000 (*rho* = 0.51, $p < .001$ if the correlation is restricted to words which do appear in the Plunkett corpus). Overall, it appears that reported frequency is more closely correlated with frequency information from Korpus 2000 (0.58) than with frequency information from the language samples (0.14 if words which do not occur in the language samples are scored as zero), likely due to the greater reliability of the former given its much larger sample size.

Does self-reported frequency in Danish correlate with age of acquisition, within categories? (Research question #3)

The relation between age of acquisition and input frequency was examined first at a between-category level. For each of the 22 categories of the Danish CDI, a mean input frequency and mean age of acquisition was determined. A scatterplot of the results is shown in Figure 2. It illustrates the positive correlation (*rho* = 0.57, $p < .01$) between input frequency and age of acquisition, contrary to a naive prediction that

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

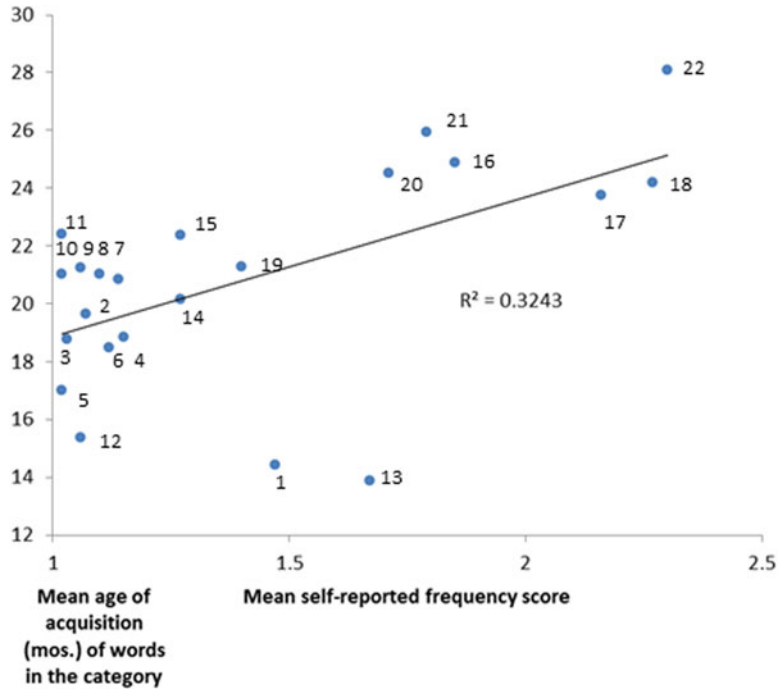


Figure 2. Scatterplot of mean self-reported frequency with mean age of acquisition for words in the 22 subcategories of the Danish CDI. 1 = Sound effects and animal sounds; 2 = animal names; 3 = vehicles; 4 = toys; 5 = food and drink; 6 = clothing; 7 = body parts; 8 = small household items; 9 = furniture and rooms; 10 = outside things; 11 = places to go; 12 = people; 13 = games and routines; 14 = action words; 15 = descriptive words; 16 = words about time; 17 = pronouns; 18 = question words; 19 = prepositions and locations; 20 = quantifiers; 21 = helping verbs; 22 = connecting words.

words heard more often should be learned earlier, but consistent with Goodman *et al.* (2008). It largely reflects the fact that closed-class words (categories 16–22) are very frequent, but late in appearance due to their syntactic and semantic complexity and somewhat distinct phonological properties (Bates *et al.*, 1994; Höhle, 2015).

Following Goodman *et al.* (2008), we then examined the correlation between input and age of acquisition within vocabulary subcategories. For all 22 subcategories of the Danish CDI there was a consistent and substantial negative correlation as expected. The correlations ranged from –0.58 to –0.96, with a median correlation of –0.83 (all *p*s < .05).

The 22 vocabulary subcategories were then aggregated into five broad categories, following the logic of Bates *et al.* (1994) and Goodman *et al.* (2008). The categories have conceptual coherence and somewhat distinct developmental trajectories as proportions of vocabulary. In addition, the larger sample size of words in each category provided more robust estimates of correlation and enabled a more detailed analysis and understanding of this relationship. Table 2 summarizes the composition and mean self-reported input frequency of the five categories, that is, means of the frequencies for individual words in each category, along with mean age of acquisition from Danish normative data. Some words do not fit into the five categories, and therefore the sum of the second column (638) is less than 725.

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

Table 2. Composition and Mean Parental Self-Reported Input Frequency and Age of Acquisition (Age50) of Broad Vocabulary Subcategories

Vocabulary subcategory	Number of words	Input frequency *M (SD)*	Age of acquisition *M (SD)*
Common nouns	332	1.84 (0.56)	26.6 (4.9)
Animal names			
Vehicles			
Toys			
Food and drink			
Clothing			
Body parts			
Small household items			
Furniture and rooms			
Outside things			
Places to go			
People	35	1.64 (0.51)	29.3 (5.4)
People			
Verbs	92	2.19 (0.44)	27.0 (3.2)
Action words			
Adjectives	59	2.01 (0.40)	28.7 (3.7)
Descriptive words			
Closed class	120	2.45 (0.34)	30.9 (4.6)
Words about time			
Pronouns			
Question words			
Prepositions & locations			
Quantifiers			
Helping verbs			
Connecting words			

Notes. The CDI categories of "Sound effects and animal sounds" and "Games & Routines" are not included, as they do not fit into any of these categories. This is consistent with the classification of Goodman, Dale, and Li (2008).

Common nouns and people had the lowest self-reported frequency, verbs and adjectives were intermediate, and closed-class words were the most frequent, consistent with the result for English reported in Goodman *et al.*

Table 3 summarizes correlational information within these five categories. As shown in the second and third columns of Table 3, the correlations were substantially higher based on Danish parental report than the results of Goodman *et al.* (2008) for English. Note that very similar methods were used for estimating age of acquisition in the two studies. Although a very substantial body of language samples from CHILDES were used to estimate input frequencies by Goodman *et al.*, their validity was limited by the restricted range of contexts sampled, as discussed earlier. That limitation results in substantial differences in validity across words. Parental self-report is less affected by context, and thus the present measure is likely to include less measurement error. The fourth column of Table 3, based on input as estimated from Danish language samples, provides correlations which are generally comparable to those from English (Goodman *et al.*, 2008), though somewhat reduced, possibly reflecting reduced reliability from a smaller corpus. The fifth column of Table 3 is based on just the 147 CDI words which occur in the Danish language samples. The correlations are very similar to those based on parental self-report. The sixth and final column of

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

Table 3. Correlations (Rho) within Vocabulary Subcategories of Age of Acquisition with Selected Input Measures

Vocabulary subcategory	Danish: parental report	English: CHILDES input frequency	Danish: Plunkett input frequency (all 664 words)	Danish: only CDI words occurring in Plunkett samples	Danish: Korpus 2000 frequency
Common nouns	−0.73	−0.55	−0.42	−0.69	−0.45
People	−0.85	−0.52	−0.37	−0.78	−0.24
Verbs	−0.58	−0.22	−0.23	−0.70	−0.11
Adjectives	−0.75	−0.28	−0.31	−0.72	−0.19
Closed class	−0.85	−0.24	−0.16	−0.78	−0.41

Notes. The correlations for English are from Goodman, Dale, and Li (2008).

Table 3 is based on Korpus 2000. The correlations are comparable to those for the Danish language samples, and somewhat less than found for English language samples.

Do discrepancies from the input-age of acquisition correlations suggest the role of other factors? (Research question #4)

Although the correlations reported in the previous section are almost all large, in Cohen's (1988) rubric, other factors are likely to play an additional role in determining the age of acquisition of individual words. Figure 3 is a 'close-up' scatterplot for verbs, which identifies words which appear to be easier to learn than would be expected on the basis of frequency (as reported by parents) alone, and words which appear to be more difficult to learn. The former category is primarily composed of words for simple, easily observable actions which young children can perform. In contrast, the latter category includes mental state words ('think', 'listen'), contrast / change of state words ('stop', 'fix'), non-motion words ('sit', 'stay'), and a causative, interpersonal word ('feed'). These trends, suggesting a substantial influence of semantic factors, are highly consistent with other research literature (e.g., Graham *et al.*, 2015) on factors which influence the difficulty of learning specific verbs.

Do self-reported input frequencies vary with gender, presence of siblings, and maternal education? (Research question #5)

Two child characteristics were examined as possible correlates of relative use of individual words by parents. In both cases, we began by correlating the rank order of frequency of use of words between the relevant groups, as it is relative frequency that is most relevant for most basic research on vocabulary growth. We followed this by examining some of the absolute differences qualitatively, to determine if there are patterns which merit further investigation in future research. The first characteristic was the gender of the child. The relative order of frequency of word production to males and females was correlated ($rho = 0.98$), indicating a very consistent overall order. The 15 words which showed the largest difference in each direction are listed

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

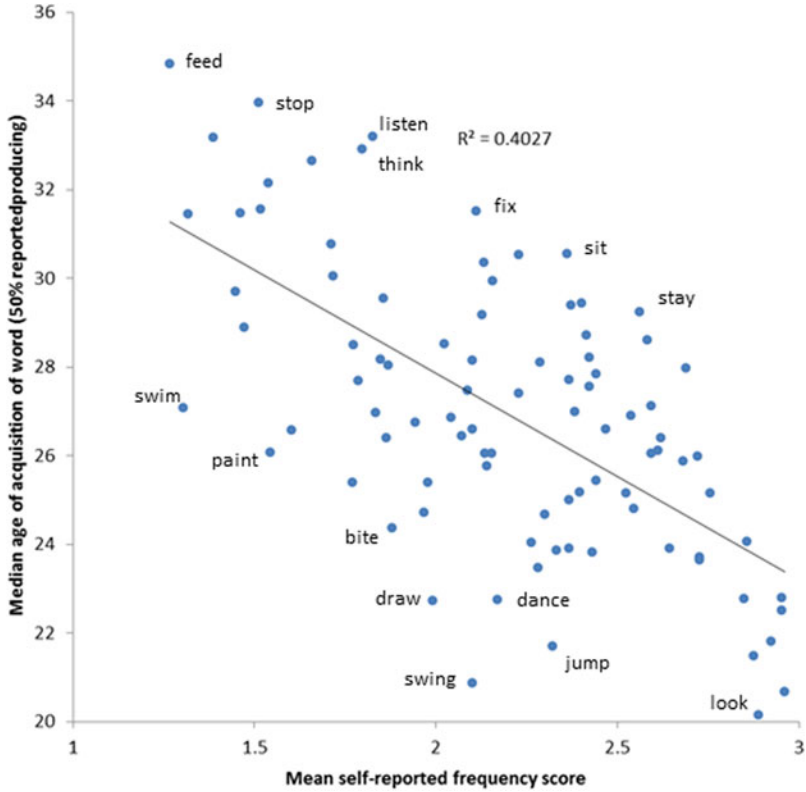


Figure 3. Scatterplot of mean self-reported input frequency with age of acquisition for verbs, with identification of words which are notably early or late in acquisition compared with input frequency.

in Table 4. For greater use to boys, the words were primarily vehicles and tools, along with gender-relevant 'boy' and *tissemand* (child word for 'penis'). For girls, the words were primarily clothing, doll, and child related ('doll', 'baby', 'stroller'), and gender-relevant 'girl' and *tissekone* (child word for 'vagina').

A second aspect of the child which was examined as a possible influence on word use was the presence of siblings in the home. The relative order of frequency of word production to only children and to children with siblings was correlated (*rho* = 0.98), again indicating a very consistent overall order. The words in each of these categories which showed the largest difference are listed in Table 5. For greater use to children with siblings, the words were primarily pronouns and family words, social interaction words ('please', 'to share'), and a word likely to be relevant for an older sibling ('school'). For greater use to children without siblings, the words were primarily animal noises and animal names.

In the present study, the potential influence of maternal education was considered a methodological question. Given that our goal was to characterize input generally, does the composition of the informant (mother) sample influence the results, or are the findings consistent across variation in maternal education? The maternal education categories defined earlier were combined into a dichotomous variable which

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

Table 4. Words with Greatest Difference in Average Score of Reported Use to Boys and to Girls (15 Each, in order of Size of Difference, of a Total of 69 Words More Frequent to Boys, and a Total of 48 Words More Frequent to Girls). Words without Reaching a Significant Difference Are Excluded.

Words used more frequently to boys	Words used more frequently to girls
penis	doll
firetruck	dress
truck	tights
vroom	vagina
tractor	brush
hammer	sister
motorcycle	stroller
train	hers
bus	girl
police	draw
fireman	baby
airplane	button
boat	puzzle
helicopter	swing
boy	teddybear

contrasts mothers with relatively low education (categories 1–4; 49.5% of the sample) with mothers with relatively high education (categories 5–6; 48.1%; 2.5% were missing or unclassifiable). As shown in Table 6, mothers with a high education responded with higher reported frequency overall, and this held for all five broad vocabulary categories. The relative ordering of the five categories was also identical for the two groups. Most uses of input frequency are less concerned with absolute levels of use than with the relative frequency of words and word categories. For this reason, the correlation of frequency across the 664 words was computed to determine if the two groups of mothers were ordering the words in a similar way; it yielded $rho = 0.99$ ($p < .001$). Thus the two groups yield virtually identical information on the relative frequency of words. However, it should be noted that the present sample was considerably under-representative of mothers with very low education (high school or less). Access to the data files for the present paper is available via the Open Science Framework at <http://osf.io/bjgct>.

Discussion

This paper is a report of an initial exploration of an innovative approach to obtaining input frequency estimates for vocabulary. Given the small size of the datasets for both parental self-report and language samples, all conclusions must be considered tentative. Nevertheless, several aspects of the results are encouraging for further work on this method. As shown in Table 1 (Research question 1), the words identified as most

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

Table 5. Words with Greatest Difference in Average Score of Reported Use to Children with and without Siblings; in order of Size of Difference, of a Total of 222 Words More Frequent to Children with Siblings, and a Total of 35 Words More Frequent to Children without Siblings. Words not Reaching a Significant Difference Are Excluded

Words used more frequently to children with siblings	Words used more frequently to children without siblings
sister	woof woof
brother	vroom
school	moo
your	peekaboo
you (singular)	quack quack
to hurry	baa baa
please	choo choo
you (plural)	elephant
to share	dog
underpants	grandpa
his	cow
push	truck
to make noise	
hers	
him	

and least frequent in input, and those which show the greatest increase and decrease between 8 and 36 months, are highly plausible based on previous literature on vocabulary composition and development in this period (Bates *et al.*, 1994; Hoff, 2014). The obtained correlation of 0.67 between input frequency and frequency in the Plunkett corpora for individual words (Research question 2) is substantial, and probably an underestimate given the limitations of the corpora as a comparison

Table 6. Mean Self-Reported Frequency of Words for High-Education and Low-Education Mothers

Vocabulary subcategory	Low-education mothers *M (SD)*	High-education mothers *M (SD)*	Effect size *d*
Common nouns	1.80 (.23)	1.85 (.23)	.22
People	1.60 (.25)	1.66 (.25)	.23
Verbs	2.13 (.36)	2.24 (.34)	.32
Adjectives	1.95 (.39)	2.05 (.39)	.27
Closed class	2.36 (.45)	2.53 (.36)	.44
Total	1.89 (.26)	1.97 (.25)	.34

Notes. All differences between the two groups are significant.

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

measure. It is reassuring that there was also a strong correlation with frequencies from an adult written language corpus. The correlations in Table 3 between age of acquisition and input frequency, computed within categories (Research question 3) are substantial, and comparable to those for frequency information from the Plunkett corpora when analysis is based on words occurring in those corpora. Verbs which are acquired notably earlier or later than would be predicted by input frequency based on parental self-report and age of acquisition (Figure 3) are consistent with previous research which has identified, e.g., the difficulty of mental state words (Ridgeway, Waters, & Kuczaj, 1985) and words which describe contrast and change of state (Bloom, 2000; Tomasello & Merriman, 1995).

Research question 5 addressed two questions which have not to our knowledge been previously addressed, the extent to which the input frequency of individual words differs based on the gender of the child, and on whether the child has siblings. As noted earlier, the words showing the greatest gender differences (Table 4) are consistent with gender-related stereotypes for children. As correlations, they may have a diverse range of explanations. They may reflect parental stereotypes which lead to gender differences in speaking to children; conversely, they may reflect child-to-parent effects in that these differences are the result of parental estimates, valid or invalid (biased), of their child's interests or language comprehension. Further research, especially in designs with experimental manipulations, will be needed to explore this issue. Birth-order differences in vocabulary input have also not been previously studied. The differences shown in Table 5 may reflect a greater involvement of parents as play partners with a first child than with later children, and as social mediators with later children.

The final portion of Research question 5 concerned the consistency of self-reported frequency information across diversity in maternal education. At least with respect to the relative frequency of production of individual words, there was very high consistency between the higher-education and lower-education mothers.

Limitations and future directions

Several aspects of this initial study limit the generalizability of the results. Some of these concern the amount and nature of the language sample measures used as potential validators of parental self-report. The total amount of child language sample data is very small by comparison with that for English. And as is typical for most language sample corpora, the range of contexts is limited. The latter limitation is shown vividly by the fact that only 164 of the 664 target words from the Danish CDI occurred at all in the language samples. As discussed above under Research question 2, context played a large role in determining which words did not occur, or other discrepancies between the two sources. It is our hope that similar research with parental self-report of input frequency will be conducted for English and other languages with more extensive language samples, which will enable more powerful estimates of validity.

The correlations between self-reported input frequency and age of acquisition reported in Table 2 (column 2) are substantially higher than found for English (column 3). We believe that the primary reason for this is the greater validity of determining input frequency with this new approach compared to using available language samples. It is also possible that our approach is prone to bias in the sense that parental knowledge of their child's production of a word may affect their

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

judgment of input frequency, inflating the correlation spuriously. To evaluate this possibility, it would be necessary to collect CDI measures of vocabulary acquisition and both observed and self-reported input frequency estimates from the same parents. It might be possible to determine size of this bias, and adjust it for research which uses self-reported input frequency. However, in our opinion this is unlikely to be a substantial influence on our own results. Even if parents were influenced by knowledge of their child's vocabulary, the size of the vocabulary and hence the identity of emerging words would still differ substantially from child to child, due to the well-documented large variability in vocabulary size at each age. The correlations reported here are not comparisons with parental report with the vocabulary of their own child, but with that of an average child of that age. Nevertheless, it is possible that parental reports are affected by social desirability and knowledge of vocabulary development trends, and this merits further investigation.

It is likely that the directions for parents could be improved. For example, it might be helpful to parents to remind them to consider the full range of interactive situations with their child, e.g., mealtime, bedtime, bathing. The response wording "2–4 times a week" may also have been ambiguous with respect to whether it referred to 2–instances, or 2–4 days.

The present volunteer sample of parents providing self-reports of input was skewed, with an under-representation of parents with low education. Parents with low education are especially difficult to recruit, and it is also possible that they are less able to provide accurate self-reports. These considerations may limit the usefulness of the method for some research questions, particularly ones which are focused on determinants of parental input, although further research would be highly valuable.

Based on this initial study, parental self-report appears to have considerable potential, and merits further research. Its greatest advantage is that it is situation-independent, unlike language sampling. Furthermore, it is highly cost-effective. The greatest limitation at present is the extent to which self-report is biased by knowledge of the child's vocabulary; this limitation may be addressable with new kinds of data and analysis.

References

Ambridge, B., Kidd, E., Rowland, C. P., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*, 239–73.

Asmussen, J. (2006). Towards a methodology for corpus-based studies of linguistic change: contrastive observations and their possible diachronic interpretations in the Korpus 2000 and Korpus 90 General Corpora of Danish. In A. Wilson, D. Archer, & P Rayson (Eds.), *Corpus linguistics around the world* (pp. 33–48). Amsterdam: Brill Academic Publishers.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S. … Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*, 85–123.

Blackwell, A. A. (2005). Acquiring the English adjective lexicon: relationships with input properties and adjectival semantic typology. *Journal of Child Language*, *32*, 535–62.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. & Basbøll, H. (2008). The Danish Communicative Development Inventories: validity and main developmental trends. *Journal of Child Language 35*, 651–69.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2016). From *uh-oh* to *tomorrow*: predicting age of acquisition for early words across languages. In A. Papafragou, D. Grodner,

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

D. Mirman, & J. C. Trueswell (Eds.), *Proceeding of the 38th Annual Conference of the Cognitive Sciences, Society* (pp. 1691–6). Austin, TX: Cognitive Science Society.

Brown, R. (1958). How shall a thing be called? *Psychological Review*, *65*, 14–21.

Carey, S. (1978). The child as word learner. In J. Bresnan, G. Miller, & M. Halle (Eds.), *Linguistic theory and psychological reality* (pp. 264–93). Cambridge, MA: MIT Press.

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

Dale, P. S., Tosto, M. G., Hayiou-Thomas, M. E., & Plomin, R. (2015). Why does parental language input style predict child language development? A twin study of gene-environment correlation. *Journal of Communication Disorders*, *57*, 106–17.

Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Perez Pereira, M. … Gallego, C. (2011). Differences between girls and boys and emerging language skills: evidence from 10 language communities. *British Journal of Developmental Psychology*, *30*, 326–43.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: user's guide and technical manual*, 2nd ed. Baltimore. MD: Paul H. Brookes.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*, 515–31.

Graham, S. A., San Juan, V., & Vukatana, E. (2015). The acquisition of words. In E. L. Bavin & L. R. Naigles (Eds.), *The Cambridge handbook of child language*, 2nd ed. (pp. 369–87). Cambridge University Press.

Hoff, E. (2014). *Language development*, 5th ed. Belmont, CA: Wadsworth.

Höhle, B. (2015). Crosslinguistic perspectives on segmentation and categorization in early language acquisition. In E. L. Bavin & L. R. Naigles (Eds.), *The Cambridge handbook of child language*, 2nd ed. (pp. 159–82). Cambridge University Press.

Law, J., & Roy, P. (2008). Parental report of infant language skills: a review of the development and application of the communicative development inventories. *Child and Adolescent Mental Health*, *13*, 198–206.

MacWhinney, B. (2000). *The CHILDES project: the database*, Vol. *2*. Mahwah, NJ: Lawrence Erlbaum.

Marchman, V. A., & Dale, P. S. (2018). Assessing receptive and expressive vocabulary in child language. In A. M. B. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics and the neurobiology of language: a practical guide* (pp. 40–67). Hoboken, NJ: Wiley.

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.

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, *76*, 763–82.

Plunkett, K. (1986). Learning strategies in two Danish children's language development. *Scandinavian Journal of Psychology*, *27*, 64–73.

Ridgeway, D., Waters, E., & Kuczaj, S. A. (1985). Acquisition of emotion-descriptive language: receptive and productive vocabulary norms for ages 18 months to 6 years. *Developmental Psychology*, *21*, 901–8.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*, 12663–8.

Tan, S. H., & Schafer, G. (2005). Toddlers' novel word learning: effects of phonological representation, vocabulary size, and parents' ostensive behavior. *First Language*, *25*, 131–55.

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

Tomasello, M., & Merriman, W. E. (1995). *Beyond names for things: young children's acquisition of verbs*. Hillsdale, NJ: Erlbaum.

Zauche, L. H., Thul, T. A., Darcy Mahoney, A. E., & Stapel-Wax, J. L. (2016). Influence of language nutrition on children's language and cognitive development: an integrated review. *Early Childhood Research Quarterly*, *36*, 318–33.

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press

Appendix

Instructions (online) for Parents

(TITLE) Help us become more familiar with which words small Danish children hear the most often by completing this questionnaire

We need more knowledge about the linguistic environment that Danish children experience in their daily lives to understand children's language acquisition better. We are now investigating which words Danish children hear at different ages and how often they hear them.

We are investigating 725 words that have been included in studies of Danish children's early language development. You can read more about the study here [LINK TO CDI-PROJECT PAGE].

For each of the 725 words, please check whether you say the words 'every day', '2–4 times a week' or 'less often'. (go directly to the survey here.) Remember that words which you, for example, read to your children, sing or use in other activities should be marked as well.

It is important that you do not spend too long considering each word, but trust your immediate judgment into how often you use the words. You can expect to take between 20 and 30 minutes to complete the form. The entire form should be filled in at once.

We also ask you to answer a few questions about the child's gender, age, and number of siblings, and your education and position.

The questionnaire survey has been reported to the Danish Data Protection Agency. The results of the study will be treated confidentially, and all participation is anonymous.

Go to completion of the form here [ICON]

Cite this article: Bleses D, Vach W, Dale PS (2018). Self-reported parental vocabulary input frequency for young children. *Journal of Child Language* **45**, 1073–1090. https://doi.org/10.1017/S0305000918000089

https://doi.org/10.1017/S0305000918000089 Published online by Cambridge University Press