Cognitive Psychology 98 (2017) 1–21

Contents lists available at ScienceDirect

Cognitive Psychology

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

Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth

Gary Jones [a,*], Caroline F. Rowland [b,c]

[a] Nottingham Trent University, UK
[b] Max Planck Institute for Psycholinguistics, Netherlands
[c] University of Liverpool, UK

ARTICLE INFO

Article history:
Accepted 27 July 2017
Available online 10 August 2017

Keywords:
Input quantity
Lexical diversity
Vocabulary acquisition
CLASSIC
Language acquisition

ABSTRACT

Children who hear large amounts of diverse speech learn language more quickly than children who do not. However, high correlations between the amount and the diversity of the input in speech samples makes it difficult to isolate the influence of each. We overcame this problem by controlling the input to a computational model so that amount of exposure to linguistic input (quantity) and the quality of that input (lexical diversity) were independently manipulated. Sublexical, lexical, and multi-word knowledge were charted across development (Study 1), showing that while input quantity may be important early in learning, lexical diversity is ultimately more crucial, a prediction confirmed against children's data (Study 2). The model trained on a lexically diverse input also performed better on nonword repetition and sentence recall tests (Study 3) and was quicker to learn new words over time (Study 4). A language input that is rich in lexical diversity outperforms equivalent richness in quantity for learned sublexical and lexical knowledge, for well-established language tests, and for acquiring words that have never been encountered before.

1. Introduction

The onset and speed of vocabulary acquisition differs enormously from child to child, in all languages. A significant proportion of this variation is explained by children's different experiences of language in their environment. Children who hear a large amount of lexically diverse child-directed speech learn language more quickly and have larger vocabularies than children who do not (for a review, see e.g. Hoff, 2006). However, it is almost impossible to directly compare the effect of input quantity with the effect of lexical diversity in natural speech samples because the two are highly correlated. In this paper, we use computational modeling to investigate how the quantity of input and the lexical diversity of that input impact on a model's learning throughout development. We show how the lexical diversity of the language input is more important than its quantity for learning sublexical and lexical knowledge, for performing language tests that are highly predictive of the children's language ability, and for vocabulary learning more generally.

* Corresponding author at: Department of Psychology, Nottingham Trent University, Chaucer Building, Nottingham NG1 4FQ, UK.
E-mail address: gary.jones@ntu.ac.uk (G. Jones).

http://dx.doi.org/10.1016/j.cogpsych.2017.07.002
0010-0285/

Preschool children show large individual differences both in the age of onset, and in the speed, of vocabulary learning (Fenson et al., 2007). These differences are strongly predicted by environmental factors, particularly the quantity and quality of the linguistic input children receive. In terms of quantity, several studies have established a direct strong or medium correlation between measures of the sheer quantity of linguistic input children hear and the size of children's vocabularies (see e.g. Bornstein, Haynes, & Painter, 1998; Bornstein & Tamis-LeMonda, 1995; Cartmill et al., 2013; Hart & Risley, 1992; Hart & Risley, 1995; Hoff & Naigles, 2002; Hurtado, Marchman, & Fernald, 2008; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991; Huttenlocher, Waterfall, Vasilyeva, Vevea, & Hedges, 2010; Rowe, 2012). For example, Huttenlocher et al. (1991) showed that 20% of the variance in children's vocabulary growth between 16 and 24 months of age (N = 22) was predicted by variations in the amount of speech produced by the children's mothers. Analyses of Hart and Risley's (1992) and Hart and Risley's (1995) American data show that some children (particularly those defined as of high socio-economic status) were exposed to up to 153,000 more words per week than others, which had a significant effect on these children's vocabulary growth. Pearson, Fernandez, Lewedeg, and Oller (1997) even showed that the relative vocabulary size of bilingual English-Spanish speaking one-to-two year olds was predicted by the relative amount of input they received in each language, which suggests a direct correlation between the likelihood of hearing a word in a language and the ability to learn it (for similar results, see also Hoff et al., 2012).

The quality, as well as the quantity, of the language input is important. Input quality can be operationalized in many different ways. Some studies have tested the complexity of input utterances, using for example, a Mean Length of Utterance (MLU) measure (Bornstein et al., 1998; Hoff & Naigles, 2002) or constituent and clausal complexity (Huttenlocher et al., 2010). Others have focused on communicative devices thought to increase the likelihood of the child interpreting a word correctly, such as the degree to which the referent of the word is easily identifiable or not (referential uncertainty; Cartmill et al., 2013) or the number of utterances spoken during periods of joint attention (Hoff & Naigles, 2002). Still others focus on features of the conversations that parents hold with their children, and measure, for example, how effectively parents engage the child's attention during episodes of joint engagement or how often they model language during routines and rituals (e.g. Hirsh-Pasek et al., 2015). However, the metric that is used most often, and which produces the most consistent effects, is lexical diversity, operationalized as the number of different word types produced by the caregiver during a set timeframe (e.g. Bornstein et al., 1998; Demir-Vegter, Aarts, & Kurvers, 2014; Hoff & Naigles, 2002; Huttenlocher et al., 2010; Pan, Rowe, Singer, & Snow, 2005; Rowe, 2012; see also Hsu, Hadley, & Rispoli, 2015, for results for verb lexical diversity). For example, Bornstein et al. (1998) found that both maternal lexical diversity and maternal MLU were significant predictors of child vocabulary at 18 months in a structural equation model. Similarly, Hoff and Naigles (2002) reported that the number of different word types produced by the mothers of 63 two-year olds was a strong predictor of the number of different words their children produced ten weeks later (though in this study, MLU was a stronger predictor). An input that models a high proportion of rare words is particularly effective, especially at older ages (e.g. Beals, 1997; Rowe, 2012; Weizman & Snow, 2001). Thus, there is good evidence for a role for both the quantity of the input and its lexical diversity in determining the rate at which children develop vocabulary.

However, almost all research to date has tested input quantity and lexical diversity using speech samples where both are free to vary. Very few studies, if any, directly compare the independent effects of input quantity and lexical diversity on child language, and virtually none test how these relationships change with development. This is problematic, since any complete theoretical account of vocabulary acquisition has to explain not only how different properties of the input are utilized during learning, but also how changes in the amount and structure of this learned knowledge affect subsequent learning.

Studying the effect of the input developmentally is important since there is tentative evidence for differential age-related effects. Rowe (2012) recently assessed age-related changes in the contribution of input quantity and quality to vocabulary growth, looking at the effects of input at 18, 30, and 42 months on children's vocabulary 12 months later. Quantity (number of word tokens produced in a 90 min interaction) was a strong predictor of vocabulary at 30 months but lexical diversity (total number of different word types) and sophistication (total number of rare words) were the strongest predictors by the time the children reached 42 months. By 54 months, the amount of decontextualized talk in the input (explanations, talk about pretense, narratives) was the strongest predictor of vocabulary. This progression makes sense when seen in the light of the acquisition tasks facing children at different ages. In the earliest stages of learning, when the child knows few words, hearing any word multiple times will lead to learning. Thus, frequent exposure to a small number of words will be effective. However, later on, when many high frequency words have already been learnt, exposure to a larger variety of words becomes more important. Even later, decontextualized language introduces children to a different set of (infrequent) words (e.g. explicit explanation of word meanings, words about the future or past), which is why it has its strongest effect later in learning.

In this paper, we directly compare the effects of input quantity and lexical diversity on vocabulary acquisition across development. We do this using a computational model, because directly comparing these effects is very difficult to do within the context of natural language settings. Parents who are verbose also use a wider variety of different words (e.g., Hart & Risley, 1992; Smolak & Weinraub, 1983; Weizman & Snow, 2001), so, while maternal type and token data might both predict child vocabulary growth, the correlation between them is usually so high as to render meaningless any statistical differences between the two. For example, Hoff and Naigles (2002) were unable to enter both number of word tokens (input quantity) and number of word types (lexical diversity) into the same regression model because high correlations between these measures ($r = 0.89$) made their effects indistinguishable statistically. The majority of research outlined above is therefore difficult