Proc. INTERSPEECH, Brighton UK, 6-10 September 2009

Modelling Vocabulary Growth from Birth to Young Adulthood

Roger K. Moore[1], *Louis ten Bosch*[2]

[1] Department of Computer Science, University of Sheffield, UK
[2] Centre for Language and Speech Technology, Radboud University Nijmegen, Netherlands
r.k.moore@dcs.shef.ac.uk, l.tenbosch@let.ru.nl

Abstract

There has been considerable debate over the existence of the 'vocabulary spurt' phenomenon - an apparent acceleration in word learning that is commonly said to occur in children around the age of 18 months. This paper presents an investigation into modelling the phenomenon using data from almost 1800 children. The results indicate that the acquisition of a receptive/productive lexicon can be quite adequately modelled as a *single* growth function with an ecologically well founded and cognitively plausible interpretation. Hence it is concluded that there is little evidence for the vocabulary spurt phenomenon as a separable aspect of language acquisition.

Index Terms: child language acquisition, lexical development, vocabulary size

1. Introduction

In recent times there has been considerable controversy and debate over the existence of the phenomenon known as 'the vocabulary spurt' - an apparent acceleration in word learning exhibited by very young children [1][2][3][4][5][6]. The spurt (or 'naming explosion') is commonly said to occur around the age of 18 months, when infants speed up from acquiring one or two new words a week [7] to acquiring up to nine words a day [8][9]. Although many different processes are implicated in early language learning [10], the vocabulary spurt continues to be of special interest to the research community concerned with language acquisition.

The vocabulary spurt appears to be more apparent in some children than in others, and a wide range of different reasons have been suggested to explain the effect. For example, MacWhinney [11] proffers three broad explanations: (i) the development of control over articulation, (ii) the role of syntactic patterns in the learning of new words, and (iii) the underlying growth of cognitive capacity. On the other hand, Gopnik & Meltzoff [12] propose that the effect is related to the ability to categorise objects, whereas Nazzi & Bertonchini [7] and McCune [13] suggest that the spurt results from a shift from an 'associationist' to a 'referential' lexical acquisition mechanism.

To our knowledge, only a few authors have attempted to analyse the developmental data using statistical methods. Bates & Carnevale [5] and van Geert [14] have modelled vocabulary growth as a dynamic system using logistic growth functions. Likewise, Ganger & Brent [15] fitted logistic curves to developmental data from 38 children, and they found that only one in five could be said to exhibit the phenomenon at all. More recently, McMurray [16] has argued that the supposed 'vocabulary explosion' could be adequately accounted for by the distributional effects of words with varying ease of acquisition. McMurray claims that a minority of words are either relatively easy to acquire and thus learnt quickly, or relatively hard to acquire and thus learnt slowly. The rate of learning thus appears to accelerate as the larger number of words with average learning difficulty is acquired.

Clearly the jury is still out as to the nature and extent of the vocabulary spurt phenomenon. Many perceptual, computational, social and neural constraints affect what a child can learn and at what time it can be learnt [17], and isolating a particular driving function appears to be difficult to achieve. Nevertheless, this paper presents the results of a new attempt to explain the vocabulary spurt phenomenon. The authors have applied similar principles to those used by Ganger & Brent [15] to a much larger data set (based on almost 1800 children), and the modelling has also been extended from infants/toddlers to young adults. These new results suggest that there is little evidence to support the existence of the vocabulary spurt phenomenon.

2. Data

2.1. Vocabulary size in infants & toddlers

The data used in this study were derived from the MacArthur Communicative Development Inventories [18] which are made available online at http://www.sci.sdsu.edu/lexical. Data were selected for the American English language (data is also available for Spanish) and, rather than submit multiple queries to the online website, the entire corpus was downloaded to facilitate easier analysis and modelling.

The downloaded dataset consisted of lexical development norms for a total of 1,789 children organised into month-by-month norms for the comprehension and production of 384 words from 8 to 16 months, and the production of 652 words from 16 to 30 months. For each word within each specified age group, the data indicates the proportion of children who had been reported as having understood or produced it. For example, 90% of 8-month-old infants were reported to have *understood* the word "mommy" whereas only 21% of 8-month-old infants were reported to have *said* the word "mommy". Likewise, around one third (38%) of 16-month-old toddlers had been observed to produce the word "airplane", but nearly all (97%) 30-month-old toddlers had been observed to produce the word.

Apart from the large number of child subjects involved, the advantage of the MacArthur corpus over other data collection methodologies is that the use of fixed vocabulary checklists provides a much more reliable estimate of the words that a child knows. The disadvantage is that the responses necessarily saturate as the size of each child's individual vocabulary approaches the number of items on each checklist.

2.2. Vocabulary size in older children and adults

Estimates of the vocabulary size in both childhood and adulthood vary quite considerably. However, according to Bates & Carnevale [5] the vocabulary of an average English speaker is about 14,000 at the age of six years [19][20], rising to about 40,000 at the age of 40 [1]. This agrees with Nation & Waring's [21] observation that a five year old beginning school will have a vocabulary consisting of around 4000 to 5000 word families, which equates to 12,000-15,000 words. On the other hand, Goulden *et al* [22] suggest that a university graduate would have a vocabulary of around 20,000 word families, which probably equates to around 60,000 words by the age of 18 [23].

Based on these data, in this study we have assumed that an average five year-old child would have a vocabulary of around 14,000 words, and that an average 16 year-old would have a vocabulary of around 40,000 words.

3. Methodology

As argued by previous authors, the logistic function (or 'growth curve') is an appropriate model for a process of word acquisition [5][14][15]. As with many other biological processes, growth is initially slow, rises to a maximum, and then slows again as the system becomes saturated – a classic 'S-shaped' curve. Proponents of the vocabulary spurt phenomenon would argue that a maximum rate of vocabulary growth is observed around the age of 18 months. The issue in this study was thus to determine growth curves based on single or multiple growth functions, and to observe the goodness of fit to the developmental data.

In order to analyse the patterns of behaviour embedded within the lexical development norm dataset, the data was subjected to different thresholds with respect to the proportion of children who had been observed as having acquired particular words. For good coverage of the whole dataset, these thresholds were arbitrarily chosen as 20%, 50% and 80%. This generated three subsets of the overall corpus: those words that were used by at least 20% of the children (i.e. the faster learners), 50% of the children (i.e. the average learners) and 80% of the children (i.e. including the slower learners).

As mentioned above, the MacArthur corpus incorporates data on the receptive and productive vocabularies of infants (8-16 months) and the productive vocabulary of toddlers (16-30 months). Hence six sets of data were available to be modelled, and this was achieved by obtaining least squares fits with various mathematically defined growth functions.

The first function was the standard 'logistic' curve of the form:

$$v_t = (V.v_0.e^{rt}) / (V + v_0(e^{rt} - 1)) \qquad (1)$$

where v_t is the estimate of the vocabulary size at time t, V is the eventual vocabulary size, v_0 is the initial vocabulary size and r is the rate of growth.

The second function was the 'Gompertz' curve (a function often used to model biological growth, e.g. tissue development):

$$v_t = V.e^{a.exp(r.t)} \qquad (2)$$

where v_t is the estimate of the vocabulary size at time t, V is the eventual vocabulary size, a is a delay factor and r is the growth rate.

For the MacArthur corpus, V was 384 for the infant data and 652 for the toddler data. For the adult data, V was set to be 40,000.

4. Results

4.1. Baseline conditions using single growth functions

4.1.1. Infant and toddler data

For the data extracted from the MacArthur corpus, both the standard logistic and the Gompertz growth curves fitted the data very well with very little difference to choose between them. Figure 1 illustrates the fits achieved using the logistic function for the *receptive* vocabulary of infants between the ages of eight and 16 months. The three curves indicate the results for the 20%, 50% and 80% samples from the corpus (as explained above). Continuous lines represent the best mathematical fit to the individual data points.

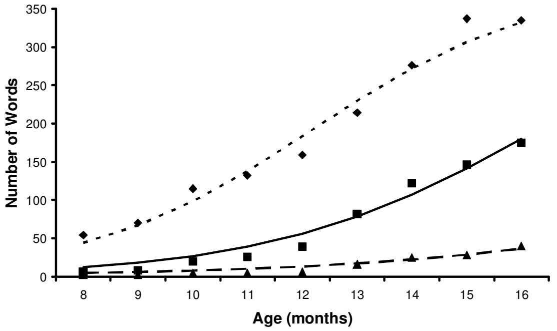

Figure 1: *Receptive vocabulary growth curves for 20% (dotted line), 50% (solid line) and 80% (dashed line) of infants aged 8 to 16 months.*

Figure 2 illustrates the fits achieved using the logistic function for the *productive* vocabulary of infants between the ages of eight and 16 months. As in Figure 1, the three curves indicate the results for the 20%, 50% and 80% samples from the corpus, and continuous lines represent the best fit to the individual data points.

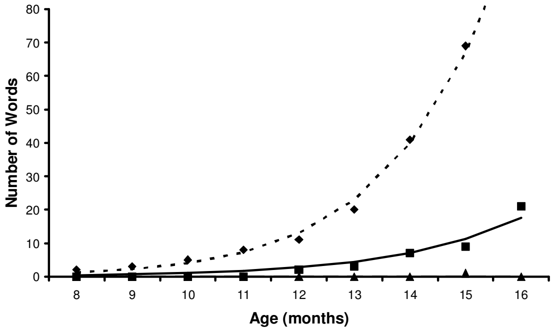

Figure 2: *Productive vocabulary growth curves for 20% (dotted line), 50% (solid line) and 80% (dashed line) of infants aged 8 to 16 months.*

Figure 3 illustrates the fits achieved using the logistic function for the *productive* vocabulary of toddlers between the ages of 17 and 30 months.

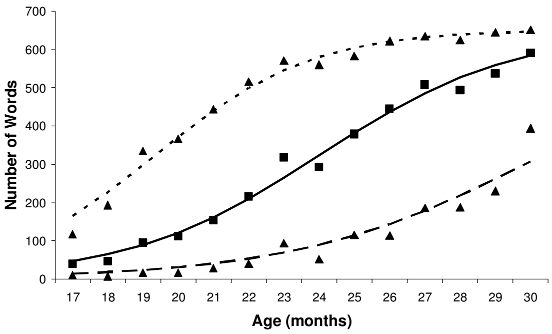

Figure 3: *Productive vocabulary growth curves for 20% (dotted line), 50% (solid line) and 80% (dashed line) of toddlers aged 17 to 30 months.*

Both Figure 2 and Figure 3 refer to the growth in *productive* vocabulary. Hence it is possible to combine the early and late stage results to cover the period from eight to 30 months. Although this is not strictly correct (as the vocabulary sizes are different in each case), it is not unreasonable given that the saturation effect only applies in the later stages for the early condition. The combined result shown in Figure 4 employs exactly the same logistic functions as used in Figure 3, and the goodness of fit confirms the validity of combining the data.

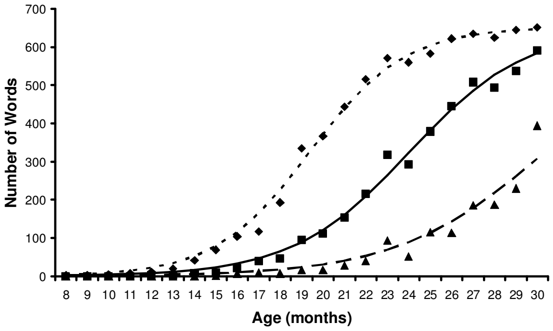

Figure 3: *Productive vocabulary growth curves for 20% (dotted line), 50% (solid line) and 80% (dashed line) of children aged 8 to 30 months.*

The results thus far clearly demonstrate the saturation effect that arises from the finite word lists employed in the Dale & Fenson study [18]. They also illustrate the different learning rates exhibited by different groups of children. For example, it can be seen from Figures 2 and 3 that at two years of age, 50% of children will have a productive vocabulary of around 300 words, whereas the faster learners will have already achieved that around 19 months, and slower learners will take until they reach around 30 months. The results thus confirm the observation that infants reach an average vocabulary of 300 words by the age of 24 months [24].

4.1.2. Child and adult data

One set of results for the child and adult data are illustrated in Figure 4. The diagram shows plots of the logistic and Gompertz curves that best fit the early learning productive data for the 50% of children case. It can be clearly seen that the logistic rises far too quickly, suggesting that an average five year-old child would have a 40,000 word vocabulary. On the other hand, the Gompertz growth curve provides a much more reasonable model. Note that the Figure also includes a linear growth curve for comparison (corresponding to a growth rate of 200 words per month).

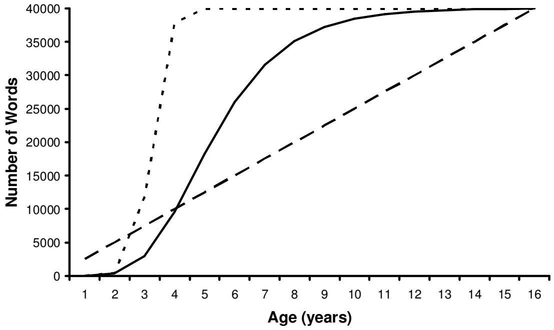

Figure 4: *Vocabulary growth curves from birth to young adulthood. The dotted line signifies a logistic function, and the solid line signifies a Gompertz function. The dashed line represents purely linear growth.*

Figure 5 illustrates the excellent fit of both receptive and productive Gompertz curves with the early-stage learning data.

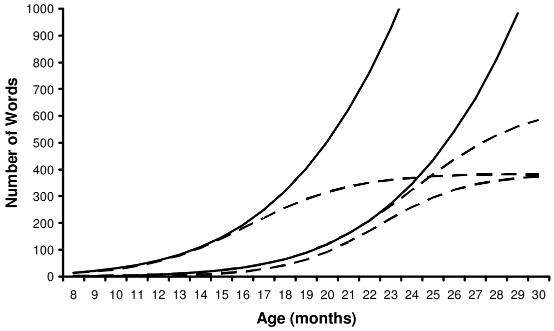

Figure 5: *Vocabulary growth curves showing the high degree of fit between the late-stage receptive and productive Gompertz functions (solid lines) with the early-stage logistic models (dashed lines).*

The overall picture to emerge from the fitting of single growth curves to the development data (for 50% of children) is as follows:

- the receptive vocabulary increases by around 40 words per month at the age of 17 months;
- the productive vocabulary increases by around 45 words per month at 23 months;
- the productive vocabulary increases by around 60 words per month at 25 months;
- the productive vocabulary increases at a maximum rate of around 700 words per month at five years of age.

Overall these models suggest that from birth to early adulthood the rate of growth of the acquired vocabulary increases steadily with a peak acquisition rate at about five years of age and with no evidence for an earlier spurt.

4.2. Evidence for multiple growth functions

Closer inspection of the MacArthur data reveals some interesting micro-structure. Figure 6 illustrates the rate of change of *receptive* vocabulary for the infants. Although the overall data fits well with a single growth function (Figure 2), it is clear that there is more than one peak in Figure 6. The peak around 13-14 months must be discounted due to the finite size of the vocabulary used, and the evidence for a minor spurt at around 10 months for all children is weakened by the suspiciously coincident dips between 11 and 12 months.

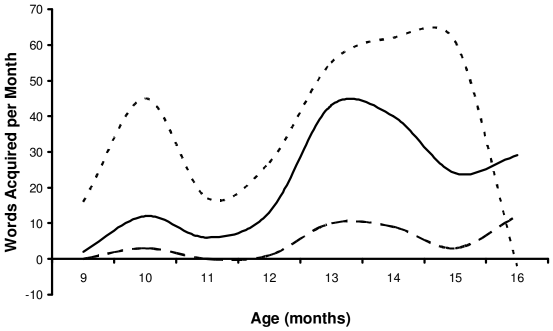

Figure 6: *Rate of change of the receptive vocabulary for20% (dotted line), 50% (solid line) and 80% (dashed line) of infants.*

By contrast, the rate of change of the *productive* vocabulary for infants (not shown) shows no subsidiary peaks, and that for toddlers (Figure 7) shows multiple peaks that appear to be more due to random variation than to any patterned behaviour (indeed the coincident dips at 20, 24 and 28 months strongly suggest hidden artefacts in the data collection process).

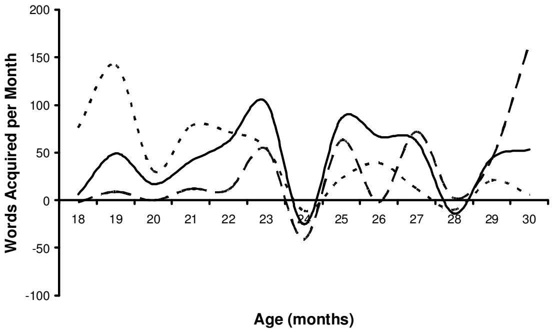

Figure 7: *Rate of change of the productive vocabulary for20% (dotted line), 50% (solid line) and 80% (dashed line) of toddlers.*

5. Final observations and conclusion

The growth functions can be explained ecologically and cognitively: the curves are growth simulations in which the three relevant parameters at each time instant are (i) the capacity for growth, (ii) the level of current knowledge, and (iii) the amount of information presented thus far.

Overall, the results reported in this paper illustrate that data relating to the acquisition of a receptive/productive lexicon can be quite adequately modelled as a *single* growth function. Although there is clearly variation in the rate at which words are acquired, the evidence for a significant 'spurt' in word acquisition appears to be rather slim. It is concluded that the Gompertz function appears to offer a very satisfactory ecologically-motivated model of lexical growth from birth to young adulthood and, in general, it is not necessary to assume that children undergo a vocabulary spurt during language acquisition.

6. Acknowledgement

This research was funded in part by the European Commission, under contract number FP6-034362, in the ACORNS project (www.acorns-project.org), and the Dutch Organisation for Scientific Research (NWO).

7. References

[1] McCarthy, D. (1954). Language development in children. In L. Carmichael (Ed.), *Manual of Child Psychology* (2nd ed., pp. 492-630). New York: John Wiley & Sons.

[2] Bloom, L. (1973). *One Word at a Time: The Use of Single-Word Utterances Before Syntax*. The Hague: Mouton.

[3] Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, 38(149).

[4] Dromi, E. (1987). *Early Lexical Development*. Cambridge: Cambridge University Press.

[5] Bates, E., & Carnevale, G. (1993). New directions in research on language development. *Dev. Review*, 13, 436-470.

[6] Bloom, P. (1993). *The Transition from Infancy to Language: Acquiring the Power of Expression*. Cambridge, MA: Cambridge University Press.

[7] Barrett, M. (1995). Early lexical development. In P. Fletcher & B. McWhinney (Eds.), *The Handbook of Child Language* (pp. 211-241). Cambridge MA: Blackwell.

[8] Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of Child Language*, 17(1), 171-183.

[9] Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: two modes of word acquisition? *Developmental Science*, 6(2), 136-142.

[10] Kuhl, P. K. (2000). A new view of language acquisition. *Proc. National Academy of Science*, 97, 11850-11857.

[11] MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, 49, 199-227.

[12] Gopnik, A., & Meltzoff, A. N. (1987). The development of categorization in the second year and its relation to other cognitive and linguistic developments. *Child Development*, 58, 1523-1531.

[13] McCune, L. (2008). *How Children Learn to Learn Language*. New York: Oxford University Press.

[14] van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, 98, 3-53.

[15] Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, 40(4), 621-632.

[16] McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, 317(5838), 631.

[17] Kuhl, P. K. (2004). Early language acquisition: cracking the speech. *Nature Reviews: Neuroscience*, 5, 831-843.

[18] Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, 28, 125-127.

[19] Templin, M. C. (1957). *Certain Language Skills in Children: Their Development and Interrelationships*. Minneapolis, MN: University of Minnesota Press.

[20] Carey, S. (1982). Semantic development: The state of the art. In E. Wanner & L.Gleitman (Eds.), *Language Acquisition: The State of the Art*. New York: Cambridge University Press.

[21] Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, Pedagogy* (pp. 6-19). New York: Cambridge University Press.

[22] Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, 11, 341-363.

[23] Aitchinson, J. (1994). *Words in the Mind: An Introduction to the Mental Lexicon* (2nd ed.). Oxford: Blackwell.

[24] Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59(5, Serial No. 242), 1-189.