COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2012) 1–14

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2011.1226.x

Exploring the Robustness of Cross-Situational Learning Under Zipfian Distributions

Paul Vogt

Tilburg Centre for Cognition and Communication, Tilburg University

Received 26 January 2011; received in revised form 30 June 2011; accepted 4 July 2011

Abstract

Cross-situational learning has recently gained attention as a plausible candidate for the mechanism that underlies the learning of word-meaning mappings. In a recent study, Blythe and colleagues have studied how many trials are theoretically required to learn a human-sized lexicon using cross-situational learning. They show that the level of referential uncertainty exposed to learners could be relatively large. However, one of the assumptions they made in designing their mathematical model is questionable. Although they rightfully assumed that words are distributed according to Zipf's law, they applied a uniform distribution of meanings. In this article, Zipf's law is also applied to the distribution of meanings, and it is shown that under this condition, cross-situational learning can only be plausible when referential uncertainty is sufficiently small. It is concluded that cross-situational learning is a plausible learning mechanism but needs to be guided by heuristics that aid word learners with reducing referential uncertainty.

Keywords: Word learning; Cross-situational learning; Lexicon learning time; Referential uncertainty; Zipf's law

1. Introduction

The ability to acquire a relatively large set of arbitrary, although conventionalized, mappings between word-forms and references is a key aspect of cognition that has been considered uniquely human. Despite many studies, no consensus has been reached on what is the exact nature of human word learning (see, e.g., Bloom, 2000; Clark, 2003; Hall & Waxman, 2004, for overviews). *Cross-situational learning* (e.g., Pinker, 1984; Siskind, 1996) has been hypothesized as a potential candidate to explain this ability.

Correspondence should be sent to Paul Vogt, Tilburg Centre for Cognition and Communication, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: p.a.vogt@uvt.nl

Cross-situational learning is a straightforward mechanism that theoretically does not require any heuristics for learning. It works based on multiple exposures of a word in varying situations, where a word's reference is taken as the one that occurs in most of these situations. There is growing evidence that children (and adults) can and do use cross-situational learning (Akhtar & Montague, 1999; Houston-Price, Plunkett, & Harris, 2005; Smith & Yu, 2008). One problem with cross-situational learning is that it fails to explain the *fast-mapping phenomenon* (i.e., the observation that many words are acquired after only one exposure, Carey & Bartlett, 1978). However, the referents of many—if not most—words are acquired through slow mapping (Carey, 1978; Smith & Yu, 2008). Another problem with cross-situational learning is that it has often been considered to be too slow of a mechanism to explain the ability to acquire a human-sized lexicon before adulthood.

Estimates show that humans acquire roughly 60,000 words within the first 18 years (Anglin, 1993; Bloom, 2000). Blythe et al. (2010) have recently explored mathematically the time required to learn such large vocabularies based on cross-situational learning. They have found that, even under high levels of referential uncertainty, learning times are such that humans could easily acquire 60,000 words within the first 18 years of life. The model they derived, however, is based on a large number of simplifying assumptions, of a through which the results of Blythe et al. can be more optimistic than they actually are.

One of the assumptions made by Blythe et al. (henceforth BSS) is that the meaning (or referent) space is distributed uniformly, whereas they assume words are distributed following Zipf's law (Zipf, 1949). Consequently, the situations in which a word is heard can contain distracting referents of both high-frequency words and low-frequency words, all with equal probability. So under this assumption the word ''papa,'' for instance, could be heard in situations of equal likelihood that contain—in addition to the actual father—another man or a male lion as a distractor meaning. Clearly, the latter situation would in reality occur far less frequently than the former. BSS agree that this is an unrealistic assumption, but they mention that they do not know what the actual distribution is, and expect the speed of cross-situational learning to slow down. In this report, it is assumed (and argued) that a Zipf distribution of referents is more likely, and what happens to the learning times of (large) vocabularies under various levels of referential uncertainty is shown.

The purpose of this report is not to present a highly realistic model of cross-situational learning; such models already exist (e.g., Siskind, 1996; Yu, Ballard, & Aslin, 2005; Frank, Goodman, & Tenenbaum, 2008; Fazly, Alishahi, & Stevenson, 2010). Also, no exhaustive review is provided on the evidence from psychology on cross-situational learning (for such a review consult, for instance, Smith & Yu, 2008). Instead, the objective is to illustrate some theoretical limitations of *pure cross-situational learning* (cf. BSS) using straightforward computer simulations. Before presenting the computer model and the results, the BSS study is briefly reviewed.

2. Learning times for cross-situational learning

Cross-situational learning can, in brief, be explained by the following example: Consider learning a novel word, say ''wakabu,'' from a new language in a situation with

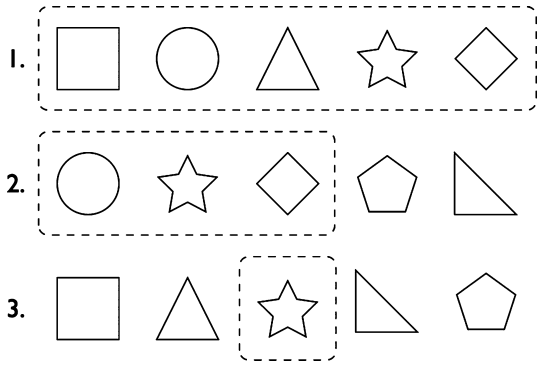

Fig. 1. An example of cross-situational learning in three trials. When a novel word, say ''wakabu,'' is expressed in the first situation (top row), where there is a square, a circle, a triangle, a star, and a diamond, and also suppose the learner considers shapes as a possible meaning, then there are five possible meanings of ''wakabu'': square, circle, triangle, star, and diamond. In the second situation there is a circle, a star, a diamond, a pentagon, and a rectangular triangle; the set of hypothetical meanings is reduced to contain only circle, triangle, and diamond. In the final situation, only the meaning star remains.

two objects as displayed in trial 1 (top row) of Fig. 1. Suppose you know that this word only relates to a shape. The *referential uncertainty* in this case is five, that is, ''wakabu'' refers to one of the following five different referents: square, circle, triangle, star, and diamond. Now suppose you hear the same word in the situation as depicted in trial 2 of Fig. 1 where there is neither the square nor the triangle. Here three possible referents remain: circle, star, and diamond. When in a third trial the circle and diamond no longer feature, you can infer that the word refers to the only remaining candidate star.

Suppose there would have been more objects with distinct features in each situation. The referential uncertainty and consequently the learning time as well would have increased. BSS have derived an analytical approximation to estimate the learning times for cross-situational learning under different levels of referential uncertainty. They have assumed that in each trial, a learner would be exposed to a word in conjunction to a *target meaning* (or referent) and C different *incidental meanings*. The C incidental meanings are randomly selected, with a uniform distribution, from a set of M different ''incidental meanings that *might* be inferred alongside the true target meaning'' (Blythe et al., 2010). Hence, C is a measure of referential uncertainty.

BSS have estimated learning times for three different strategies of using cross-situational learning: a minimal cross-situational strategy (Minimal XSL), an approximate cross-situational strategy (Approximate XSL), and a pure cross-situational strategy (Pure XSL, or XSL for short). In all strategies, the learner forms a hypothesis on the first exposure and sticks to that until it is proved to be incorrect, at which point he/she selects a new hypothesis. For Minimal XSL, this is a meaning randomly selected from the situation at that moment. In Pure XSL, a new hypothesis is selected based on the complete history of previous situations. Approximate XSL is somewhere in between using any

other selection strategy for selecting the new hypothesis from the current situation. BSS have shown that the time (t_{appr}) to learn a lexicon using the latter strategy is between the time it takes to learn a lexicon using Pure XSL (t_{xsl}) and that of Minimal XSL (t_{min}), that is,

$$t_{\texttt{xsl}} \leq t_{\texttt{appr}} \leq t_{\texttt{min}} \tag{1}$$

As Pure XSL is the most efficient strategy they have proposed, the remainder of this article will focus on this strategy.

Assuming that words are presented to the learner with a uniform distribution, BSS have estimated the time taken to learn a lexicon of size W with a probability $P_W(t)$ sufficiently close to unity, that is, $P_W(t) \geq 1-\epsilon$ with ϵ a small parameter. Using findings from a previous study (Smith, Smith, Blythe, & Vogt, 2006), BSS have shown that, provided $t_{\texttt{xsl}}$ is sufficiently large and ϵ sufficiently small,

$$t_{\texttt{xsl}} \approx W \frac{1}{1-\frac{C}{M}} \ln\left(\frac{MW}{2\epsilon}\right) \tag{2}$$

This function behaves such that the learning time initially rises slowly for an increasing C/M ratio, and only when this ratio becomes relatively large, the learning time explodes. Note that as $C/M \to 1$, $t \to \infty$.

BSS derived their model such that they could easily use different distributions with which words are presented to the learner. As words tend to be distributed according to Zipf's law (Zipf, 1949), which states that the frequency of the kth most common word is proportional to $1/k$, BSS changed the uniform with a Zipfian distribution, yielding the following estimate:

$$t_{\texttt{xsl}} \approx \frac{W\mu}{1-\frac{C}{M}} \mathcal{W}_0\left(-\frac{MW}{2\ln(1-\epsilon)}\right), \tag{3}$$

where $\mathcal{W}_0(x)$ is the principal branch of Lambert's W function (Corless, Gonnet, Hare, Jeffrey, & Knuth, 1996). As this function largely behaves as a logarithm, Eq. (3) only differs from Eq. (2) by the factor μ, provided ϵ is small. BSS found that $\mu = 11.579\ldots$, so that the learning time for a Zipf distribution is not exorbitantly higher than for a uniform distribution.

It is important to note that, although the words are now selected following a Zipfian distribution, the C incidental meanings are still sampled uniformly. As already argued in the introduction, this seems to be an unrealistic assumption. However, what is a realistic distribution of meanings or referents? Assuming the frequency of meaning is usage-based, a logical candidate is the Zipfian distribution. (A more elaborated argument follows in the discussion.)

The remainder of this article explores what happens if the incidental meanings occurring in the different episodes are selected following Zipf's law. Rather than deriving a mathematical formula to estimate learning times, Monte Carlo simulations are used. The main reason for using Monte Carlo simulations is because attempts at solving this problem analytically

have been unsuccessful so far: In calculating the probability that a whole lexicon is learned at some time t, one has to average the probabilities of all possible sequences of situations leading up to time t. For uniform distributions as in Blythe et al. (2010) and Smith et al. (2006), the probabilities of these sequences are all the same, so it suffices to calculate the number of possible sequences and multiply this number with the probability of a sequence. However, assuming a Zipfian distribution of referents, every possible sequence yields a different probability, and so, one cannot simply multiply, and when t becomes large, the number of possible sequences becomes immensely large.

3. Model and methods

The computer model used in this study is basically the same as the model used by BSS. In this model, a learner has to learn a lexicon of W different word-meaning mappings. In each learning episode, one target meaning m_T and C distinct incidental meanings are selected to construct the situation or *context* $\mathscr{C}$. Each of these meanings is sampled without replacement from a set of $M \leq W$ meanings following the Zipfian distribution $f \propto \frac{1}{k}$, where k is the rank of the meaning, and meanings are ranked from most frequent to least frequent (Zipf, 1949). For convenience, these meanings are ranked from $k = 1$ to $k = M$, rather than sampling the meanings from W different meanings. After the context is thus constructed, the learner is presented with the target's word and the context.

For each word w_T, the learner maintains a list H_T of hypothetical meanings, which on the first exposure of that word equals the context $\mathscr{C}$ of that episode. Then, in each following exposure of word w_T, H_T is updated by removing all meanings $m_j \in H_T$ that are not in the situation, that is, all $m_j \notin \mathscr{C}$. The mapping for word w_T is learned when the size of H_T equals 1. When all word-meaning mappings are thus learned, the simulation stops and the number of episodes is stored.

This model is tested for various values of W, M, and C, and for each condition, a Monte Carlo simulation is carried out by repeating the simulation 2,000 times with different random seeds. The number of episodes t^* required to learn the lexicon is measured and the time t within which 99% of all repetitions have terminated is reported. Hence, there is a 99% confidence that a lexicon under the specified conditions can be learned within t episodes (this is equal to setting the parameter $\epsilon = 0.01$ as BSS did). Fig. 2 summarizes the basic algorithm of the Monte Carlo simulations.

The model is to a large extent the same as that of BSS. So it also adopts among others these five simplifying assumptions: (a) a word is always exposed in a situation where its meaning is present, (b) each word and situation are perceived without errors, (c) all situations have the same number of distinct incidental meanings, (d) all meanings are given, so there is no need for categorization, and (e) the lexicon is unambiguous. The model only differs from that of BSS in that the learner does not guess-and-test in case there are still other candidate meanings. Although the guess-and-test method is slightly faster than the current model, the difference can be ignored for large t.

```
repeat X times with a different random seed
  learning-time = 0
  let H_i = ∅ for all i = 1, ..., W
  while (number-words-learned< W)
  do
    select target m_T and add to the context C
    select C incidental meanings to add to C
    if H_T == ∅
      then H_T = C
    express-word-w_T
    learning-time++
    if H_T > 1
      remove-meanings-from-H_T-not-in-C
      if H_T-size == 1
        then number-words-learned++
  done
  store learning-time
print learning-time t
```

Fig. 2. The basic procedure of the Monte Carlo simulation.

4. Results

Fig. 3 shows how the learning time t' relates to the ratio C/M for different values of M for learning a vocabulary of size $W = 100$. The learning time t' is the ratio between actually measured learning time t and the time t_0 required to learn the lexicon, would there be no incidental meaning ($C = 0$), that is, the learning time for a fast-mapper (Blythe et al., 2010):

$$t' = \frac{t}{t_0}. \tag{4}$$

For each $M = \{10,25,50\}$, the context size C was varied such that $C/M = \{0,0.2,0.4,0.5,0.8\}$. For the condition where $M = 50$, no results are reported, because the simulation was abandoned due to the extremely long run-time this simulation required. The

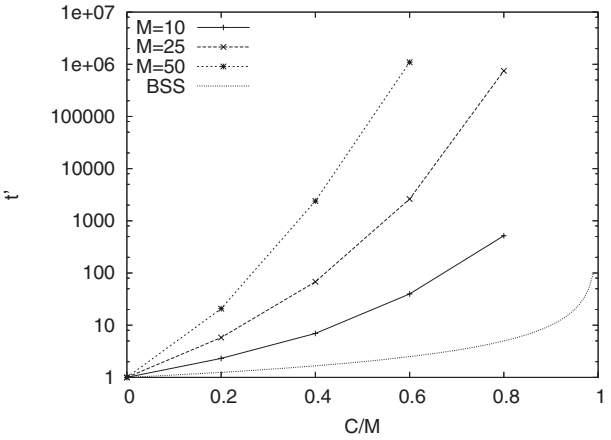

Fig. 3. Learning times for cross-situational learning as a function of C/M for learning a lexicon of size $W = 100$, and different numbers of incidental meanings $M = \{10,25,50\}$. The bottom line shows the results from BSS with $W = 100$ and $M = 50$ (cf. Eq. 3).

learning time for the fast learner t_0 was taken from the simulations for the case where $C = 0$, which, for all conditions was $t_0 = 3{,}903$ episodes. The ratio t' for all conditions with $C/M = 0$ was thus 1. The bottom line in the figure displays the results following the BSS model (Eq. 3) for $W = 100$ and $M = 50$.

It is clear from Fig. 3 that in all cases, the learning time t' quickly diverges from the results of BSS and shows a super-exponential growth. So learning times rapidly run out of hand, and for $M = 50$ and $C/M = 0.6$, for instance, the actual learning time t was measured to be $t \approx 4.2 \times 10^9$ episodes.

BSS have evaluated their results using a study by Hart and Risley (2003), who have estimated that parents speak between 600 and 2,100 per hour to their children. Assuming that a learner requires 18 years to learn the 100 words with $M = 50$ and $C/M = 0.6$, he or she would need to hear ~746,000 words *per day*. This is between 362 and 1,267 times more than Hart and Risley's estimation.

Extrapolating from Hart and Risley (2003), BSS have shown that in their model, a large lexicon of 60,000 words can be learned within the estimated number of words humans are exposed to in their first 18–19 years, even when the situations are as uncertain to allow the context size be $C = 90$, that is, $C/M = 0.9$.

Fig. 4 shows the actual learning times t of applying the current model to learning $W = 60{,}000$ words and $M = 100$ incidental meanings for various values of C/M. Here, the Monte Carlo simulations have not consisted of 2,000 repetitions of each simulation, but only up to 600 replications (for $C/M = \{0.01,0.05,0.10,0.15\}$) and 100 replications (for $C/M = 0.20$). Larger context sizes have not been processed for limitations of computational processor time. Again, the lower line displays results of the BSS model.[1]

The figure shows that when $C/M = 0.1$, the learning time $t \approx 10^8$, which in 18 years would amount to 15,221 words per day. When $C/M = 0.2$, however, the learning time

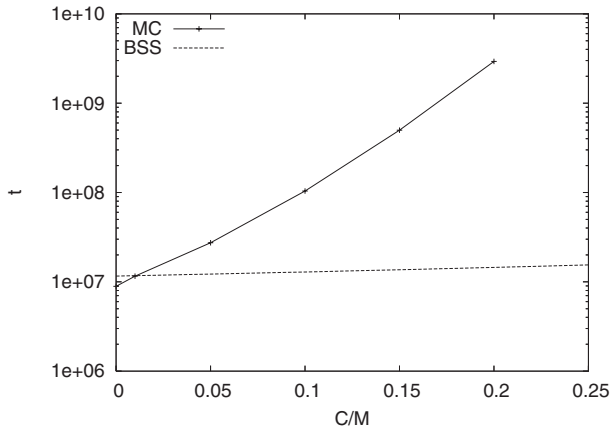

Fig. 4. Learning times for cross-situational learning as a function of C/M for learning a lexicon of size $W =$ 60,000 and $M = 100$ incidental meanings.

$t \approx 2.9 \times 10^9$, which comes to about 441,400 words per day in 18 years. Assuming the upper bound of the number of words that parents are estimated to speak to their children (Hart & Risley, 2003), when hearing 2,100 words for 7.2 h per day, learning 60,000 words in 18 years is feasible with XSL, if $C/M = 0.1$. However, children would require 210 h of input per day if $C/M = 0.2$. Clearly, the former situation is plausible, and the latter is not.

5. Discussion

The results clearly show that when the occurrence of meanings follow a Zipfian distribution, cross-situational learning of large vocabularies is only possible within a reasonable amount of time when there is little referential uncertainty. It has been estimated that humans acquire approximately 60,000 words in their first 18 years (Anglin, 1993; Bloom, 2000). Moreover, estimations are that parents speak about 600–2,100 words per hour (Hart & Risley, 2003). Extrapolating from the above figures, BSS have estimated the upper bound of words heard in 18 years to be slightly more than 10^8. Although the learner from the BSS model could have a referential uncertainty of around $C = 90$ for $M = 100$, the current results show that this is only possible when the uncertainty is not much higher than $C = 10$. Even when XSL mainly applies to early word learning, as suggested by one reviewer, under these conditions, it cannot deal with high amounts of referential uncertainty as the results from Fig. 3 show.

To understand why the learning times found here are so large, it is instructive to look at a situation that can serve as a lower bound of the learning problem. To learn the entire lexicon, the lowest ranked word-meaning mapping also needs to be learned and this tends to take longest for two reasons: First, this mapping is not exposed frequently; second, this mapping requires more exposures than higher ranked mappings before the competing incidental

meanings no longer occur in the context. When the context size is small, this is not a major problem. However, when it is large, it is very likely that the most frequent mapping almost always occurs in the context. In fact, the probability $P(m_i \in \mathscr{C})$ that a meaning m_i occurs in a context $\mathscr{C}$ of size C equals

$$P(m_i \in \mathscr{C}) > 1 - (1 - p_i)^C, \tag{5}$$

where p_i is the probability with which the meaning m_i occurs. So the probability $P(m_i \in \mathscr{C})$ approaches 1 for large context sizes C and does so faster for higher ranked meanings.

Although we cannot analytically estimate the learning time of a whole lexicon assuming the Zipfian distribution of meanings, we can estimate the time it takes before an incidental meaning m_i is disambiguated from a target meaning m_T (see the Appendix). Fig. 5 shows the estimation of the number of episodes required to disambiguate the least frequent target meaning from the most frequent incidental meaning. It is clear that the number of trials required explodes rapidly and exceeds the BSS learning times substantially. The difference with the results from the Monte Carlo simulation is that in the simulation all meanings need to be learned and each needs to be disambiguated from each incidental meaning.

BSS have further shown that the learning times under a Zipfian distribution of words (not incidental meanings) differs from the learning times under a uniform distribution by a factor $\mu = \sum_{i=1}^{W} 1/i$ that depends on the lexicon size W (provided the confidence factor $1-\epsilon$ is sufficiently close to 1). So, in their analysis, the difference between the two distributions is independent of C and M. When not only the lexicon follows Zipf's law but also the occurrence of incidental meanings in the contexts, the difference in learning times depends clearly on, at least, M, but possibly also on C, given the differences in the super-exponential learning curves shown in Fig. 3.

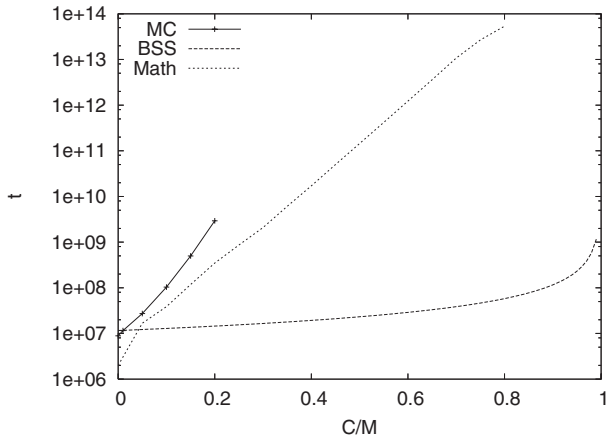

Fig. 5. The dotted line shows the number of trials required to disambiguate the lowest frequency target meaning from the highest frequency incidental meaning. The upper line shows the results from the Monte Carlo simulations and the bottom line those from BSS. All results assumed lexicon size $W = 60{,}000$ and number of incidental meanings $M = 100$.

The results from this model are in line with the experimental results presented in Kachergis, Yu and Shiffrin (2009), who have shown that high-frequency words are easier learned than low-frequency words under XSL. Moreover, they have shown that contextual diversity (i.e., referential uncertainty) is an important factor, similar to the current study. In a computational study that replicated Kachergis et al.'s experiment, Fazly, Ahmadi-Fakhr, Alishahi, and Stevenson (2010) have shown that some of their results can be explained based on a combination of XSL and the principle of contrast (Clark, 1993). Their results indicate that context familiarity (i.e., already having seen a meaning in a previous context) and age of exposure (due to which learners tend to have already learned more frequency words) are important factors in learning. So humans seem to use additional heuristics, such as the principle of contrast. The current study provides a theoretical explanation why such heuristics are necessary to explain word learning by XSL.

The findings make clear that the time required to learn a large lexicon using Pure XSL when the meanings occur in situations following a Zipfian distribution is much larger than the analysis of BSS reveals. The question remains as to whether a Zipfian distribution of meanings is realistic or not. This question is hard to answer, mostly because it remains unclear as to what exactly constitutes a meaning, and we cannot measure meaning occurrence in the brain. However, various studies indicate that a Zipfian distribution is a likely candidate. When meanings correspond one-to-one with the words used in language, that is, they are usage-based, Zipf's law applies consequently (Zipf, 1949). When meanings correspond to the categories or concepts we form in our brains, and assuming that the meaning space is organized following a hierarchical taxonomy as proposed by Rosch (1978), then there is formal suggestive evidence that the meanings occur following Zipf's law (Manin, 2008). This even holds when meanings are categorized in a hierarchical taxonomy from a uniform distribution of referents using the principle of least effort (Vogt, 2004). Last, but not least, Zipf's law is the most universal distribution of natural phenomena, and it appears to be an inevitable consequence of a general class of stochastic systems (Corominas-Murtra & Solé, 2010). Further research is required as to the exact nature of meanings and their occurrence distribution. A promising way for doing this would be to investigate whether the occurrence of meanings (or referents or concepts) apply to the class of stochastic systems used by Corominas-Murtra and Solé (2010).

6. Conclusion

In this article, the theoretical limitations of pure cross-situational learning are studied under the assumption that both words and meanings occur following Zipf's law. The results show that large vocabularies can only be learned in reasonable time when referential uncertainty is very low, that is, with small context sizes. This finding is in conflict with the findings presented by Blythe et al. (2010), who have shown that pure cross-situational learning is highly robust for increasing referential uncertainty. The reason is that Blythe et al. have assumed Zipf's law only to apply to the occurrence of words, but they have assumed a uniform distribution of the meaning space.

The results do not disqualify cross-situational learning as a plausible mechanism for learning word-meaning mappings. Instead, the results indicate the requirement of additional heuristics to reduce referential uncertainty, such as, for example, joint attention (Tomasselo & Todd, 1983), the whole object bias (Macna- mara, 1982), mutual exclusivity (Markman & Wachtel, 1988), the principle of contrast (Clark, 1993), or syntactic cues (Gillette, Gleitman, Gleitman, & Lederer, 1999). Blythe et al. (2010) also acknowledge that additional heuristics are required, but the results presented here suggest that these need to be far stronger than those proposed by them. Various computational models of cross-situational learning have already implemented some of these heuristics (e.g., Smith, 2005; Yu & Ballard, 2007; Frank et al., 2008; Alishahi & Fazly, 2010; Fazly, Ahmadi-Fakhr, et al., 2010), and they tend to benefit learning greatly. The present study provides a theoretical explanation why these heuristics have to be present: If the meaning space is distributed following Zipf's law, pure cross-situational learning cannot explain human lexicon acquisition within reasonable time.

To conclude, cross-situational learning can be the fundamental learning mechanism, provided it is guided by various additional heuristics to downsize referential uncertainty substantially. The lack of these heuristics in other species could be the reason why only humans have the ability to acquire large vocabularies. The evolution of such heuristics among humans, then, could have been one of the major adaptations in the evolution of language.

Acknowledgments

Paul Vogt is funded by a VIDI grant provided by the Netherlands Organization for Scientific Research (NWO). The writing of this article would not have been possible without the excellent comments from Tony Belpaeme, Richard Blythe, Emiel Krahmer, Doug Mastin, Andrew Smith, Kenny Smith, and two anonymous reviewers.

Note

1. The curve from BSS for low values of C/M exceeds the values from the simulations. This is because the formula from BSS does not hold for low values of C/M (Blythe, Smith, & Smith, 2010).

References

Akhtar, N., & Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. *First Language*, *19*, 347–358.

Alishahi, A., & Fazly, A. (2010). Integrating syntactic knowledge into a model of cross-situational word learning. In R. Camtrabone & S. Ohlsson (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2452–2457). Austin, TX: Cognitive Science Society.

Anglin, J. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, *58*, 1–166.

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

Blythe, R., Smith, K., & Smith, A. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*(4), 620–642.

Carey, S. (1978). The child as word-learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 265–293). Cambridge, MA: MIT Press.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Clark, E. V. (1993). *The lexicon in acquisition*. Cambridge, England: Cambridge University Press.

Clark, E. V. (2003). *First language acquisition*. Cambridge, England: Cambridge University Press.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., & Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics*, *5*, 329–359.

Corominas-Murtra, B., & Solé, R. (2010). Universality of Zipf's law. *Physical Review E*, *82*(1), 011102.

Fazly, A., Ahmadi-Fakhr, F., Alishahi, A., & Stevenson, S. (2010). Cross-situational learning of low frequency words: The role of context familiarity and age of exposure. In R. Camtrabone & S. Ohlsson (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2615–2620). Austin, TX: Cognitive Science Society.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2008). A Bayesian frame-work for cross-situational word-learning. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems 20* (pp. 457–464). Cambridge, MA: MIT Press.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–176.

Hall, D. G., & Waxman, S. R. (2004). *Weaving a lexicon*. Cambridge, MA: The MIT Press.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, *27*, 4–9.

Houston-Price, C., Plunkett, K., & Harris, P. (2005). ''Word-learning wizardry'' at 1; 6. *Journal of Child Language*, *32*, 175–190.

Kachergis, G., Yu, C., & Shiffrin, R. (2009). Frequency and contextual diversity effects in cross-situational word learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (Vol. 31, pp. 2220–2225). Austin, TX: Cognitive Science Society.

Macnamara, J. (1982). *Names for things: A study of human learning*. Cambridge, MA: MIT Press.

Manin, D. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science*, *32*(7), 1075–1098.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, *20*, 121–157.

Pinker, S. (1984). *Learnability and cognition*. Cambridge, MA: MIT Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 189–206). Hillsdale, NJ: Lawrence Erlbaum Association.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Smith, A. D. M. (2005). Mutual exclusivity: Communicative success despite conceptual divergence. In M. Tallerman (Ed.), *Language origins: Perspectives on evolution* (pp. 372–388). Oxford, England: Oxford University Press.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Smith, K., Smith, A., Blythe, R., & Vogt, P. (2006). Cross-situational learning: a mathematical approach. In P. Vogt, Y. Sugita, E. Tuci, & C. Nehaniv (Eds.), *Symbol grounding and beyond* (pp. 31–44). Berlin: Springer.

Tomasselo, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First Language*, *4*, 197–212.
Vogt, P. (2004). Minimum cost and the emergence of the Zipf-Mandelbrot law. In J. Pollack, M. Bedau, P. Husbands, T. Ikegami, & R. A. Watson (Eds.), *Artificial life IX proceedings of the ninth international conference on the simulation and synthesis of living systems* (pp. 214–219). Cambridge, MA: The MIT Press.
Yu, C., & Ballard, D. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13–15), 2149–2165.
Yu, C., Ballard, D., & Aslin, R. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science: A Multidisciplinary Journal*, *29*(6), 961–1005.
Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Appendix

Fig. 5 shows the time required to disambiguate the least frequent target meaning from the most frequent incidental meaning. Here, the estimation of this time is derived analytically. Assume all meanings are ranked based on the occurrence frequency with the meaning of rank 1 being most frequent. Following Zipf's law, each meaning m_k occurs with a probability of

$$p_k = \frac{1}{k\mu}, \tag{6}$$

where $\mu = \sum_{i=1}^{N} 1/i$ is the normalization factor. As in the computer model, for the target meanings $N = W$, and for the incidental meanings $N = M$. We assume that words have a one-to-one correspondence to their target meanings.

To disambiguate target meaning m_T from incidental meaning m_i, the latter should not occur in the situation $\mathscr{C}$. This happens with probability $P(m_i \notin \mathscr{C})$. For simplicity, it is assumed, contrary to the computer model, that the meanings are sampled with replacement. Then,

$$P(m_i \notin \mathscr{C}) = (1 - p_i)^C, \tag{7}$$

where, C is the context size. This gives us the relation that

$$P(m_i \in \mathscr{C}) = 1 - (1 - p_i)^C. \tag{8}$$

So the probability that the highest ranking incidental meaning is in the context approaches 1 when referential uncertainty is high, that is, $P(m_i \in \mathscr{C}) \to 1$ for large C.

It is now possible to estimate the time t_T it takes to disambiguate target meaning m_T from incidental meaning m_i with a probability larger than $1-\epsilon$. The first step is to calculate the probability $P(m_i \notin \mathscr{C}; t_i)$ that after t_i trials of selecting a context of incidental meanings, the meaning m_i does not occur in the context at least once:

$$P(m_i \notin \mathscr{C}; t_i) = 1 - (P(m_i \in \mathscr{C}))^{t_i} \geq 1 - \epsilon. \tag{9}$$

After some rearrangement and substitutions, the number of trials t_i at which meaning m_i did not occur in the context alongside the target at least once with a probability larger than or equal to $1-\epsilon$ is

$$t_i \geq \frac{\ln \epsilon}{\ln P(m_i \in \mathscr{C})} = \frac{\ln \epsilon}{\ln(1 - (1 - p_i)^C)}. \tag{10}$$

So the target meaning m_T has to occur at least t_i times to be confident with a probability equal to $1-\epsilon$ that meaning m_i did not occur as an incidental meaning at least once, thus disambiguating m_T from m_i.

Now how many trials t_T are required to know with a probability larger than $1-\epsilon$ that the target m_T was drawn at least t_i times? The probability that after $t_T \geq t_i$ trials, meaning m_T was drawn at least t_i times is

$$P_T(t_T, t_i) = \sum_{n=t_i}^{t_T} \binom{t_T}{t_i} p_T^{t_i} (1 - p_T)^{t_T - t_i} = I_{p_T}(t_i, t_T - t_i - 1), \tag{11}$$

where $I_x(a,b)$ is the *regularized beta function*. This function is hard to solve analytically for b, but knowing p_T and t_i, it is possible to find the smallest t_T for which $P_T(t_T,t_i) \geq 1-\epsilon$ numerically. This was done to generate Fig. 5 for target $m_W (W = 60{,}000)$ and the highest ranked incidental meaning m_1 $(M = 100)$.