A communicative framework for early word learning

Benjamin C. Morris[1] & Daniel Yurovsky[1,2]

[1] University of Chicago

[1] Carnegie Mellon University

Author Note

Correspondence concerning this article should be addressed to Benjamin C. Morris, Department of Psychology, University of Chicago, 5848 S University Ave, Chicago, IL 60637. E-mail: yurovsky@uchicago.edu

Abstract

Enter abstract here. Each new line herein must be indented, like this line.

Keywords: keywords

Word count: X

A communicative framework for early word learning

Introduction

Word learning as a statistical inference problem.

From Quine on. (Quine, 1960)

three kinds of uncertainty – over statistical time and in the moment

constraints, pragmatics, etc deal with uncertainty in the moment

uncertainty over consistent meanings – priors of some kind to deal with this tenenbaum & xu (Tenenbaum, 1999,@xu2007)

statistical co-occurrence structure deals with unceraินty reduction over time (Siskind, 1996,@yu2008,@blythe2010,@blythe2016)

these two scales are linked (Frank, Goodman, & Tenenbaum, 2009)

linking priors and in the moment scales (Frank & Goodman, 2012,@frank2014)

All of the arguments in these domains are about the relative difficulty of these different kinds of problems (Trueswell, Medina, Hafri, & Gleitman, 2013,@smith2014,@yurovsky2014,@yurovsky2015)

but all of this stuff is still about speakers talking to no one! (Tomasello, 2000, @tomasello2001)

Indeed, it looks like it matters whether speech is to children - structural reasons (Aslin, Woodward, LaMendola, & Bever, 1996,) - evidence from weisleder, hoff, etc. (Weisleder & Fernald, 2013) - argument from ruthee about structure of contra evidence from Akhtar (Akhtar, Jipson, & Callanan, 2001,@akhtar2005,foushee2016)

In contrast, pedagogical inference – shafto, bonawitz, etc. (Bonawitz et al., 2011,@shafto2012) - evidence for some of this kind of stuff from follow-in labeling. tomasello, baldwin, yu - but this is probably not what parents are doing most of the time (although c.f. tamis-lemonda) (Tamis-LeMonda, Kuchirko, Luo, Escobar, & Bornstein, 2017) - old arguments from newport, etc. (Newport, Gleitman, & Gleitman, 1977)

An intermediate position: Speakers goal is to communicate - Grice (1969)

reference games and transmission of language - Kirby, Tamariz, Cornish, and Smith (2015) - Gibson et al. (2017) - Baddeley and Attewell (2009)

Critically, reference games and information theory (in general) assume that speaker and receiver share the same code

But what if only one person knows the code? In this case, in order to communicate successfully, speakers need to take into account the listener's knowledge of the language - evidence for some speaker design - brown-schmidt and tanenhaus (Brown-Schmidt, Gunlogson, & Tanenhaus, 2008)

In this case, ambiguity will be controlled in part by the speaker's communicative goals, and scale with the listener.

We show that without any explicit pedagogical goal, can get speaker design in reference games that leads to better learning

A spectrum of models from pedagogical to adversarial. Figure?

A model of learning and production

Brief explanation of the general reference game framework

Experiments 1 and 2

speakers adapt to beliefs about points and also speaker knowledge

Method

Participants.

Material.

Procedure.

Data analysis.

Results

Discussion

Experiments 3 and 4

this leads to better learning, but not as good as ostension (obviously)

A model of teaching

Experiment 5

teaching!

Consequences for Learning

In the model and experiments above, we asked whether the pressure to communicate successfully with a linguistically-naive partner would lead to pedagogically supportive input. These results confirmed its' sufficiency: As long as linguistic communication is less costly than deictic gesture, speakers should be motivated to teach in order to reduce future communicative costs. Further, the strength of this motivation is modulated by predictable factors (speaker's linguistic knowledge, listener's linguistic knowledge, relative cost of speech and gesture, learning rate, etc.), and the strength of this modulation is well predicted by a rational model of planning under uncertainty about listner's vocabulary.

In this final section, we take up the consequences of communicatively-motivated teaching for the listener. To do this, we adapt a framework used by Blythe et al. (2010) and colleagues to estimate the learning times for an idealized child learning language under a variety of models of both the child and their parent. We come to these estimates by simulating exposure to successive communicative events, and measuring the probability that successful learning happens after each event. The question of how different models of the parent impact the learner can then be formalized as a question of how much more quickly learning happens in the context of one model than another.

We consider three parent models:

1. *Teacher* - under this model, we take the parents' goal to be maximizing the child's linguistic development. Each communicative event in this model consists of an ostensive labelling event (Note: this model is equivalent to a *Communicator* that ignores communicative cost).

2. *Communicator* - under this model, we take the parents' goal to be maximizing communicative success while minimizing communicative cost. This is the model we explored in the previous section.

3. *Indifferent* - under this model, the parent produces a linguistic label in each communicative event regardless of the child's vocabulary state. (Note: this model is equivalent to a *Communicator* who ignores communicative cost).

SOME STUFF ABOUT CROSS SITUATIONAL LEARNING

One important point to note is that we are modeling the learning of a single word rather than the entirety of a multi-word lexicon (as in Blythe et al., 2010). Although learning times for each word could be independent, an important feature of many models of word learning is that they are not (Frank et al., 2009; Yu, 2008; Yurovsky et al., 2014; although c.f. McMurray, 2007). Indeed, positive synergies across words are predicted by the majority of models and the impact of these synergies can be quite large under some assumptions about the frequency with which different words are encountered (Reisenauer, Smith, & Blythe, 2013). We assume independence primarily for pragmatic reasons here–it makes the simulations significantly more tractable (although it is what our experimental participants appear to assume about learners). Nonetheless, it is an important issue for future consideration. Of course, synergies that support learning under a cross-situational scheme must also support learning from communcators and teachers (Markman & Wachtel, 1988, @frank2009, @yurovsky2013). Thus, the ordering across conditions should remain unchanged. However, the magnitude of the difference sacross teacher conditions could potentially increase or decrease.

Method

Teaching. Because the teaching model is indifferent to communicative cost, it engages in ostensive an ostensive labeling (pointing + speaking) on each communicative event. Consequently, learning on each trial occurs with a probability that depends entirely on the learner's learning rate ($P_k = p$). Because we do not allow forgetting, the probability that a learner has failed to successfully learn after n trials is equal to the probability that they have failed to learn on each of n successive independent trials (The probabiliy of zero successess on n trials of a Binomial random variable with parameter p). The probability of learning after n trials is thus:

$$P_k(n) = 1 - (1 - p)^n$$

The expected probability of learning after n trials was thus defined analytically and required no simulation. For comparison to the other models, we computed P_k for values of p that ranged from .1 to 1 in increments of .1.

Communication. To test learner under the communication model, we implemented the same model described in the paper above. However, because our interest was in understanding the relationship between parameter values and learning outcomes rather than inferring the parameters that best describe people's behavior, we made a few simplifying assumptions to allow many runs of the model to complete in a more practical amount of time. First, in the full model above, speakers begin by inferring their own learning parameters (P_s) from their observations of their own learning, and subsequently use their maximum likelihood estimate as a standin for their listener's learning parameter (P_l). Because this estimate will converge to the true value in expectation, we omit these steps and simply stipulate that the speaker correctly estimates the listener's learning parameter.

Second, unless the speaker knows apriori how many times they will need to refer to a particular referent, the planning process is an infinite recursion. However, each future step in the plan is less impactful than the previous step (because of exponential discounting), this infinite process is in practice well approximated by a relatively small number of recursive steps. In our explorations we found that predictions made from models which planned over 3 future events were indistinguishable from models that planned over four or more, so we simulated 3 steps of recursion[1].

Hypothesis Testing. The literature on cross-situational learning is rich with a variety of models that could broadly be considered to be "hypothesis testers." In an eliminative hypothesis testing model, the learner begins with all possible mappings between words and objects and prunes potential mappings when they are inconsistent with the data according to some principe. A maximal version of this model relies on the principle that

[1] It is an intersting empirical question to determine how the level of depth to which that people plan in this and similar games (see e.g. bounded rationality in Simon, 1991, resource-ratinoality in @griffiths2015). This future work is outside the scope of the current project.

every time a word is heard its referent must be present, and thus prunes any word-object mappings that do not appear on the current trial. This model converges when only one hypothesis remains and is provably the fastest learner when its assumed principle is a correct assumption (Smith, Smith, & Blythe, 2011).

A positive hypothesis tester begins with no hypotheses, and on each trial stores one ore more hypotheses that are consistent with the data, or alternatively strengthens one or more hypotheses that it has already stored that are consistent with the new data. A number of such models have appeared in the literature, with different assumptions about (1) how many hypotheses a learner can store, (2) existing hypotheses are strengthened, (3) how existing hypotheses are pruned, and (4) when the model converges (Siskind, 1996; Smith et al., 2011; Stevens, Gleitman, Trueswell, & Yang, 2017; Trueswell et al., 2013; Yu & Smith, 2012).

Finally, Bayesian models have been proposed that leverage some of the strengths of both of these different kinds of model, both increasing their confidence in hypotheses consisten with the data on a given learning event and decreasing their confidence in hypotheses inconsistent with the event (Frank et al., 2009).

Because of its more natural alignment with the learning models we use Teaching and Communication simulations, we implemented a positive hypothesis testing model[2]. In this model, learners begin with no hypotheses and add new ones to their store as they encounter data. Upon first encountering a word and a set of objects, the model encodes up to h hypothesized word-object pairs each with probability p. On subsequent trials, the model checks whether any of the existing hypotheses are consistent with the current data, and prunes any that are not. If no current hypotheses are consistent, it adds up to h new

[2] Our choice to focus on hypothesis testing rather than other learning frameworks is purely a pragmatic choice–the learning parameter p in this models maps cleanly onto the learnin parameter in our other models. We encourage other researchers to adapt the code we have provided to estimate the long-term learning for other models.

hypotheses each with probability p. The model has converged when it has pruned all but the one correct hypothesis for the meaning of a word. This model is most similar to the Propose but Verify model proposed in Trueswell et al. (2013), with the exception that it allows for multiple hypotheses. Because of the data generating process, storing prior disconfirmed hypotheses (as in Stevens et al., 2017), or incrementing hypotheses consistent with some but not all of the data (as in Yu & Smith, 2012) has no impact on learner and so we do not implement it here. We note also that, as described in Yu and Smith (2012), hypothesis testing models can mimic the behavior of associative learning models given the right parameter settings (Townsend, 1990).

In contrast to the Teaching and Communication simulations, the behavior of the Hypothesis Testing model depends on which particular non-target objects are present on each naming event. We thus began each simulation by generating a copus of 100 naming events, on each sampling the correct target as well as (C-1) competitors from a total set of M objects. We then simulated a hypothesis tester learning over this set of events as described above, and recorded the first trial on which the learner converged (having only the single correct hypothesized mapping between the target word and target object). We repeated this process 1000 times for each simulated combination of $M = (16, 32, 64, 128)$ total objects, $C = (1, 2, 4, 8)$ objects per trial, $h = (1, 2, 3, 4)$ concurrent hypotheses, as the learning rate p varied from .1 to 1 in increments of .1.

General Discussion

Conclusion

Acknowledgement

The authors are grateful to XX and YY for their thoughtful feedback on this manuscript. This research was supported by a James S MacDonnel Foundation Scholars Award to DY.

References

Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental Science*, *8* (2), 199–209.

Akhtar, N., Jipson, J., & Callanan, M. A. (2001). Learning words through overhearing. *Child Development*, *72* (2), 416–430.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, 117–134.

Baddeley, R., & Attewell, D. (2009). The relationship between language and the environment: Information theory shows why we have only three lightness terms. *Psychological Science*.

Blythe, R. A., Smith, A. D., & Smith, K. (2016). Word learning under infinite uncertainty. *Cognition*, *151*, 18–27.

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

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120* (3), 322–330.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107* (3), 1122–1134.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998–998.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Frank, M. C., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *3*, 201619666–201619666.

Grice, H. P. (1969). Utterer's meaning and intention. *The Philosophical Review*, *78*, 147–177.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

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

Newport, E. L., Gleitman, H., & Gleitman, L. R. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. A. Ferguson (Ed.), *Talking to children language input and interaction* (pp. 109–149). Cambridge University Press.

Quine, W. V. O. (1960). *Word and object. Cambridge, Mass.* Cambridge, Mass.: MIT Press.

Reisenauer, R., Smith, K., & Blythe, R. A. (2013). Stochastic dynamics of lexicon learning in an uncertain and nonuniform world. *Physical Review Letters*, *110*(25), 258701.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization Science*, *2*(1), 125–134.

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

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical wordreferent learning. *Trends in Cognitive Sciences*, *18*(5), 251–258.

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive Science*, *41*, 638–676.

Tamis-LeMonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., & Bornstein, M. H. (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental Science*.

Tenenbaum, J. B. (1999). *A bayesian framework for concept learning* (PhD thesis). Massachusetts Institute of Technology.

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, *10*, 401–413.

Tomasello, M. (2001). Could we please lose the mapping metaphor, please? *Behavioral and Brain Sciences*, *24*(6), 1119–1120.

Townsend, J. T. (1990). Serial vs. Parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychological Science*, *1*(1), 46–54.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156.

Weisleder, A., & Fernald, A. (2013). Talking to children matters early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, *4*(1), 32–62.

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychological Review*, *119*, 21–39.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53–62.

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review*, *21*, 1–22.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*, 891–921.