Modelling Language Acquisition: Grammar from the Lexicon?

Steve R. Howell (showell@hypatia.psychology.mcmaster.ca)
Department of Psychology, McMaster University, 1280 Main Street West, Hamilton, ON Canada

Suzanna Becker (becker@mcmaster.ca)
Department of Psychology, McMaster University, 1280 Main Street West, Hamilton, ON Canada

Abstract

A neural network model of language acquisition is introduced, based on and motivated by current research in psychology and linguistics. It includes both semantic-feature representations of words and localist linguistic representations of words. The network learns to associate the semantic features of words to their linguistic labels, as well as to predict the next word in the corpus. This is interpreted to model both the acquisition of a lexicon, and the beginnings of syntax or grammar (word order). The relationship of lexical learning to grammar learning is examined, and similarities to the human data found. The results may provide support for the 'Grammar from the Lexicon', or 'emergent grammar' position.

Introduction

How do children acquire language? More generally, how does any abstract language learner acquire language? When we attempt to model language processing via computer simulation, what should we be attempting to model, mature adult performance, or the developmental schedule of a child? What can such a model hope to tell us about the process of language acquisition in human infants?

These are some of the questions motivating our effort to model language processing. Much evidence exists as to the usefulness of the connectionist modelling enterprise for the understanding of human language in general. However, as we seek to model more fully the actual processing, and even production, of language, in a behavioural fashion, we consider it very important to take a developmental approach to human language processing. That is, a complete model of language processing should first become a model of language acquisition. Evidence suggests that a model of language acquisition in children should provide the foundation necessary to scale up to a model of more mature language processing, as we shall see.

Developmental Language Acquisition

In considering a developmental model of language, one important aspect is the limits of the enterprise. That is, where does language acquisition start, and where does it end? Language is a very complex cognitive activity, and our connectionist modelling techniques still maturing. We do not want to include any more than absolutely necessary in a model of language if we are to be successful. Thus, it is important to be explicit about our assumptions, in terms of pre-linguistic mental representations, or of what we can exclude from our model or include only as inputs.

We assume here that modelling any of the low-level acoustic properties of language is unnecessary for our purposes. While issues such as phonemic segmentation are important for language, those auditory tasks are arguably well-learned by the time of vocabulary acquisition. Further, modelling to the level of acoustics is too computationally demanding to include in a model of language acquisition at present.

If we consider the start of vocabulary acquisition to be at the age of the child's first word, typically 8-12 months, then we can ask the following question. What cognitive capacities does the child have prior to that point? What does language have to build upon? Some suggest that there is a considerable amount.

Lakoff and colleagues (Lakoff, 1986; Lakoff & Johnson, 1999) suggest that the child has reached an adequate level of concept formation prior to the development of language. Few would argue, we believe, that pre-linguistic children must have some kind of internal representation of the world, some understanding that a cat is fuzzy and can be patted, even if they don't know the words cat, or pat, or fuzzy. Lakoff argues that children's sensorimotor experience is continually building up these pre-linguistic concepts, concepts that are very specific and concrete, and that these concepts enable the child to function in their limited world.

With all of this cognitive machinery already well established, the language learning problem has happily become much simpler. If a child already has a concept for things like 'cat', then when it begins to learn the word for cat, it is really only attaching a linguistic label to a category of sensorimotor experience that it has previously built up. The learning of words is thus reduced to the learning of labels for things. The attributes of those things and the relationships between them are all predetermined (at least at this stage) by the child's environmental experience. Of course, nouns fit into this viewpoint with greater ease than do verbs; it is harder to point to a verb than a noun.

This is the traditional view in developmental psycholinguistics according to Gillette et al. (Gillette, Gleitman, Gleitman, & Lederer, 1999). As they point out however, this view has limits. Specifically, they show evidence that only some words can be derived solely via extralinguistic context.

It is well known that there is an overwhelming preponderance of nouns in children's early speech, not only in English but in most languages, while adults, of course, have a much more equal balance. Several explanations have been offered for this distinction. The discontinuity hypothesis holds that the cognitive capacities of children are fundamentally different from adults. Thus, at some point after the start of development of language children's cognitive capacity for language changes. Gentner describes the noun learning advantage as due to the conceptual complexity of the ways in which the two classes, noun and verb, describe the world (Cited in Gillette et al, 1999). That is, nouns describe object concepts, while verbs describe relations between objects. The latter would obviously be the more complicated task, since it depends on the success of the former. As Gillette et al point out, by this interpretation learning words is not just a matter of associating labels to concepts. Significant conceptual learning must occur as well. If true, this interpretation would argue against the conceptualization of language-age children as relatively conceptually stable, and would also invalidate one of the assumptions of our modelling approach.

Fortunately, Gillette et al. offer a different interpretation, the continuity hypothesis, which assumes that children are conceptually equipped to understand at least those concepts that underlie the words that adults typically use with them, both nouns and verbs. However, they argue that it is still possible to account for children's initial restriction to noun learning, using instead the different informational requirements of words that are necessary to uniquely identify them from extralinguistic context. They refer to their hypothesis as an information-based account, and describe several experiments that support this account.

Most importantly Gillette et al. provide strong evidence that learnability is not primarily based on lexical class. That is, it is not whether a word is a noun or a verb that determines if it can be learned solely from observation. Rather, they demonstrate that the real distinction is based upon the word's imageability or concreteness.

It is obvious that the very first words must be learned solely by the child attempting to discover contingencies between sound categories and aspects of the world, over many different exemplars. Gillette et al. demonstrate that the very first words used by mothers to their children are the most straightforwardly observable ones, and that as a group, the nouns are in fact more observable than the verbs. Furthermore, the imageability of a word is more important than the lexical class. The most observable verbs are learned before the less observable initial nouns, accounting for the few rare early verbs in children's vocabularies.

So, imageability or concreteness is the most important aspect of the early words, nouns and verbs alike, and it determines the order in which they tend to be learned by children. This result argues against the discontinuity hypothesis, and supports Lakoff's early concepts and the borders that we have drawn for our language modelling enterprise. However, what of the less imageable words? How are they learned?

Gillette et al. also find evidence for the successive importance of noun co-occurrence information and then argument structure. That is, for later learning of the less imageable words (mostly verbs), observing which previously known nouns co-occur in a sentence with the yet unknown word label helps greatly to uniquely identify the concept. Thus rather than imageability determining exactly which object we are talking about over multiple experiences, for many verbs the nouns involved act to identify it. Thus if the noun 'ball' is paired with a yet unknown word, the concept 'throwing' may be activated for many learners, allowing them to infer that the unknown word means 'to throw' (Gillette et al, 1999). Argument structure is yet a further step to verb inference. Gillette et al. show that the number and position of nouns in the speech stream reliably cues which verb concept the unknown word could be.

At this point in the child's language learning we have moved beyond initial lexical learning and are in the realm of syntax. The first words (mainly nouns) have been learned without reference to other words, their sheer imageability enabling them to be inferred from the adult to child speech stream and the extralinguistic evidence. The next step involves the use of these concrete nouns to help infer the less imageable verb meanings in the speech stream, and from there the child is no longer learning words solely from the extralinguistic context. The lexical structure of utterances now assists the child as well. For example, the first few verbs learned, when experienced in adult speech and involving a novel object, will cue the inference of the new noun label and, depending on the particular verb, even the type of noun involved. The circular, bootstrapping process of language learning is on its way (for further evidence concerning verbs and nouns respectively, see Goldberg, 1999; Smith, 1999). Before long new words will no longer require explicit extralinguistic context at all. The school-age child will begin reading and acquiring new words solely by lexical constraints, allowing them to exhibit the incredible word acquisition rates that have been reported (e.g. Bates & Goodman, 1999).

Of course, once the child's lexicon has reached a certain level of complexity, perhaps 300 words (Bates and Goodman, 1999) the multi-word stage begins, and grammar acquisition begins to be a consideration as well as just lexical acquisition.

Grammar From the Lexicon

Bates and Goodman (1999) examine the highly linked development of grammar and the lexicon. They provide evidence for the emergence of grammar directly from the lexicon itself. Specifically, they show the lack of evidence for any dissociation of lexical and grammatical processes (drawn from studies of early and late talkers, focal brain lesions, and developmental deficits), along with the very tight developmental ties between the two. For example, lexical status at twenty months (during children's vocabulary burst) is the single best predictor of grammatical status at 28 months (during children's grammar burst), with a correlation coefficient of between .70 and .84. This is in fact as good a statistical relationship as that between separate measures of grammar! This is good evidence that grammar does emerge, at least partially, from the very growth of the lexicon itself.

This finding, as well as those of Gillette et al, is important to the development of our model of language acquisition, as if grammar development is emergent from lexical development, then we want to be sure that we do not model them as two separate modules or components. Rather, a central tenet of our model is to use a single processor architecture to learn both lexicon and grammar. Furthermore, lexical development should precede grammatical, and grammatical development should not take off until sufficient lexical development has occurred. Our model should exhibit the same sort of acquisition (and production, eventually) behaviour as a child.

A Dynamical Systems Approach

Elman (1995) suggests viewing the process of initial lexical and grammatical development as a dynamical system, or attractor model, which can be learned through a process of predicting the input. Roughly speaking, this viewpoint is as follows. A language learner's semantic representations are very limited at first, much like a flat three-dimensional landscape. Then as the learner develops stable categories and concepts, the landscape gradually develops depressions or basins, each basin corresponding to a word or concept, and each experience of that concept deepening the basin, until eventually the landscape is full of deep and wide basins of attraction. These are "attractors" since, while any partial or confused activation of a semantic representation will tend to indicate a place on the landscape not in one of these basins, the slope of the 'terrain' is such that the representation will tend to be drawn down into one basin or another, and the larger basins will be more likely to capture the activation. They "attract" the activation.

Furthermore, this attractor representation is hierarchical. General or superordinate concepts might have very large basins, containing within them smaller basins corresponding to more specific but semantically related terms.

Obviously this landscape representation only applies to the lexicon. How does grammar enter into the picture? Well, if the lexicon is viewed as basins in this representation landscape, or state-space, then grammar is contained in the transitions that occur between these states. That is, a true dynamical system consists not only of these representations in state space, but also relationships that influence movement from one representation to another. Further details can be found in Elman (1995), but for our present purposes it is sufficient to realize that this dynamical systems approach provides a possible mechanism for the implementation of the word-inference processes described earlier (Gillette et al. (1999). Certainly a recurrent net like the one we will describe in our model is capable of exhibiting the behaviour of a dynamical system, with the hidden unit representations corresponding to the state-space vectors and the operation of the network providing the transitions between them based on the values stored in its weights. It can also be argued that the cortex operates in this fashion (Elman, 1995; Sulis, 2001, personal communication), and thus that the same explanation can be offered for human language processing.

The 'Complete' Early Language Acquirer

Let us assume, then, that the child (or model) starts with pre-existing pre-linguistic concepts of the world, upon which linguistic labels will be learned by direct instruction as well as simple exposure. This pre-existing conceptual structure implies either a pre-existing mental representation (semantic landscape) or one that is quickly built up as words are matched to concepts.

In our model, we assume that the child begins syntax or grammar learning at the same time as it begins learning vocabulary. However, since there is little evidence that grammar is directly instructed (Bates & Goodman, 1999), unlike noun acquisition (Smith, 1999), and since grammar is inherently more complex, grammar learning does not really succeed until after the most primal of the lexical attractors have been firmly set and the lexical and syntactical bootstrapping has begun. In essence, grammar exposure begins at the same time as lexical learning, but grammar learning doesn't effectively take place until the lexical representations are solidified.

Thus we would expect to see exactly that behaviour that is seen in real children; lexical development proceeds at an ever accelerating pace, then when the lexical foundation is firm enough (the 'noise' or uncertainty in the language environment is reduced enough) the mental machinery can focus on syntactic relationships, and grammatical learning should accelerate. Our model should exhibit exactly this behaviour if it is capturing the essence of human language acquisition.

Method

Our experiment consists of training our model of language acquisition many times from different initial conditions, and analyzing the performance results for their fit to the human data and improvements over the control models.

The Model

The model of language acquisition discussed herein (see Figure 1) takes as input uniquely identified words (localist input representations), and learns how those words can be used in sentences. This is not a novel undertaking (see Elman, 1990, 1993; Howell & Becker, 2000). However, what is new to this model is the addition of a second set of inputs, semantic-feature inputs. By 'semantic', however we actually mean pre-linguistic semantics or meaning (e.g. sensorimotor features). Thus, instead of abstractly manipulating locally-distributed word representations, a process that has been characterized by McClelland as "learning a language by listening to the radio" (Elman, 1990), our model attempts to ground the word representations in reality by associating them with a set of these semantic features for each word.

Furthermore, the network is not performing only the prediction task that is argued (Elman, 1990) to lead to an internalization of basic aspects of grammar, specifically word-order relationships. Instead, it is also learning, simultaneously, to memorize its linguistic inputs, memorize its semantic inputs, and associate the two together, such that either one alone will elicit the other.

Why construct a neural network model in this way? First, using a simple recurrent architecture and prediction task retains the successful grammar learning capabilities that have been shown so well by Elman and colleagues. Second, adding a semantic layer will eventually allow for the use of phonemic input representations and the binding of those phonemes into words (through semantic constancy across each individual word) although the network discussed in this paper does not deal with phonemic inputs, only whole-word inputs. Third, the inclusion of the semantic input layer and a semantic output layer means that semantic features can be read off for any given linguistic input, indicating whether the network has learned the "meaning" of the word.

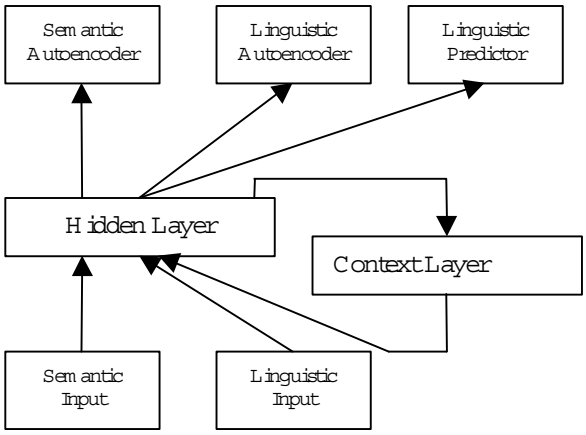

Figure 1: Modified SRN architecture, including standard SRN hidden layer and context layer, standard linguistic prediction layer, and novel semantic autoencoder and linguistic autoencoder.

Finally, the inclusion of both linguistic autoencoding (word learning) and linguistic prediction (grammar learning) allows us to explore the dynamics of the model, and determine if the learning behaviour of the model maps to the human developmental data. That is, does the word learning have to reach a critical mass before the grammar learning proceeds? Does a jump in lexical competence lead to a linked jump in grammatical competence? If so, then perhaps the model can provide evidence for the view that grammar emerges from the lexicon (Bates and Goodman, 1999).

Model Details

There are two input layers and three output layers. The semantic output layer is paired with the semantic input layer. Both are 68 nodes in size, since the semantic feature dimensions taken from Hinton & Shallice (1991) have 68 dimensions.

The linguistic input and the linguistic outputs are of size 29, since the vocabulary has 29 words. Both linguistic outputs are tied to the same set of linguistic inputs, but where the linguistic autoencoder's training signal is the present input, the linguistic predictor's training signal is the input at the next time step.

Both the hidden and the context layer are of size 75, and the hidden-to-context transfer function is a one-to-one copy with no hysteresis (see Howell & Becker, 2000). The hidden-to-context connection is not

trainable, but the context-to-hidden feedback connection is trained exactly as is either of the input-to-hidden connections.

Training Environment

The network is trained on a corpus of text derived from a small (390 word) subset of Elman's original corpus of two and three word sentences with a 29 word vocabulary (Elman, 1990).

Input to the semantic input layer was derived from the above corpus by converting each word in the corpus to the word's semantic featural representation, using a set of features derived from Hinton and Shallice (1991). This feature set includes only the sensory features and excludes the semantic-association ones found in the original. This resulted in a binary distributed representation for the semantic layer. It is important to note that on language tasks a binary distributed representation would often be expected to learn faster than a localist representation, as it provides more information to the network.

The network's weights were randomly initialized, and training proceeded as usual for Simple Recurrent Networks, using the backpropagation algorithm (Rumelhart, Hinton, and Williams, 1986).

Training proceeded until reasonable levels of accuracy were achieved. Trial runs of up to 1500 epochs indicated that the net asymptoted near 500 epochs, so training did not in any case proceed beyond 500 epochs.

Error measures and accuracy measures were logged at each input presentation, but averaged over the 390 patterns to one value per epoch of training.

Results & Discussion

The first finding from the various runs of the network is that the net does in fact learn. There had been some concern that the demands of three different tasks sharing a single hidden layer might cause significant or even catastrophic interference in the learning tasks. On the contrary, with a hidden layer size only slightly larger than the largest input layer (75 compared to 68 for the semantic input layer) the net learned all three tasks.

Furthermore, the tasks were learned in the expected order. That is, judging from the error curves the binary distributed semantic representations were learned most quickly (since they provide more information for the network to learn on, i.e. more bits turned on) followed by the localist linguistic autoencoding and then the localist linguistic prediction. Prediction, of course, is a more difficult task than autoencoding or 'memorization', just as verb learning is a more difficult task than noun learning.

For the present purposes, our analysis is limited to the lexical-grammatical relationship (and further semantic results are not reported). Specifically, over 24 simulation runs the mean peak lexical accuracy was 96.6 percent correct, while the mean peak grammatical accuracy was 37.33 percent correct (See Figure 2).

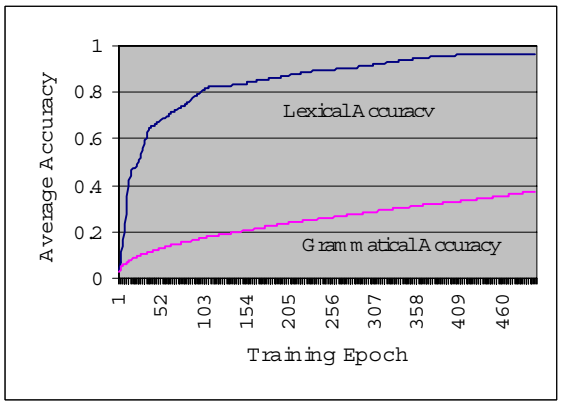

Figure 2: Average Accuracy Curves Over 24 Runs

Comparisons with 'control' or partial networks lacking the semantic or lexical autoencoder task also indicate that each task is learned faster and more accurately in the experimental network than in the control networks. Only the grammatical results are reported here, however.

For control network 1, which included only the linguistic prediction task (i.e. an original Elman net) the peak prediction accuracy was lowest, with a mean of 18.5 percent correct, and significantly different from the experimental network via t-test ($n = 10, p < 0.0001$).

For control network 2, which excluded only the semantic layers, the peak prediction accuracy, achieved at epoch 500, was significantly lower than the experimental network ($n=10, m = 28.4, p < 0.0001$).

For control network 3, which excluded only the linguistic autoencoder, the peak prediction accuracy was still lower than the experimental network ($m = 37.1$) but the difference did not reach significance ($n=10, p = 0.137$).

Thus, training all three tasks through a single hidden layer apparently creates synergies that allow each to proceed faster than it would alone.

Most interesting, however, was the relationship between the lexical and grammatical accuracy curves for the experimental network. We expect that if our model is catching important elements of the human language learning experience, then it should exhibit lexicon-then-grammar behavior. Certainly, as discussed above, the speed of learning (rate of error decline) exhibits this relationship, but that is only to be expected by the difficulty of the tasks. A better question is

whether the network exhibits the lexical-to-grammatical performance correlations that Bates and Goodman (1999) discuss. That is, does the lexical performance at time T correlate well with the grammatical performance at some later point?

By analogy to the methods cited in Bates and Goodman (1999), a point on the lexical accuracy curves that could be considered the 'lexical burst' was identified (approx. Epoch 108). Then, since there was no explicit 'grammar burst' within our time window a set of correlations was calculated to the grammatical performance at various time lags from the lexical burst (see Figure 3). The results indicate that the highest correlation, approximately .80, is from the lexical burst to grammatical performance 75 epochs later.

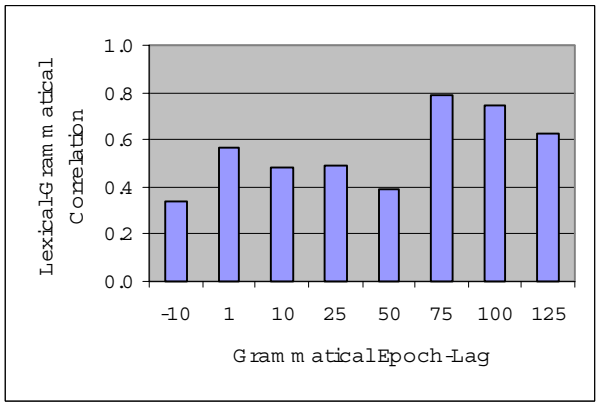

Figure 3: Lexical-Grammatical Correlations (n = 24)

This is similar to Bates & Goodman's cited correlation between lexical status at 20 months and grammatical status at 28 months in children. At first, the similarity may seem limited, since our model uses only 29 words, not the 300-plus that is suggested to be the critical mass required for grammar learning. Also, our sentences use only the 29 words from the model's vocabulary, and no unfamiliar words, and word learning is being represented by average accuracy curves. Further, grammatical status is being measured by accuracy of prediction rather than Mean Length of Utterance (MLU).

However, we believe these results are promising, and that further study is warranted. We have already begun to run simulations that use larger vocabularies, and that provide analogues of MLU measurements for grammatical status, in order to elucidate further the model's relationship to human performance.

Acknowledgments

Thanks to George Lakoff, whose writings and personal conversations inspired some of this work, and to Damian Jankowicz, whose comments were most helpful throughout. This work has been supported by a Post-graduate Fellowship from the National Sciences and Engineering Research Council of Canada (NSERC) to SRH, and a research grant from NSERC to SB.

References

Bates, E. and Goodman, J. C. (1999). On the emergence of grammar from the lexicon. In MacWhinney, B. (Ed.) (1999). The Emergence of Language. New Jersey: Lawrence Erlbaum Associates.

Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14, 179-211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48. 71-99.

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. van Gelder (Eds.), Mind as Motion: Explorations in the Dynamics of Cognition. Cambridge, MA: MIT Press.

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

Goldberg, A. E. (1999). The emergence of the semantics of argument structure constructions. In MacWhinney, B. (Ed.) (1999). The Emergence of Language. New Jersey: Lawrence Erlbaum Associates.

Hinton, G. E. & Shallice, T. (1991). Lesioning a connectionist network: Investigations of acquired dyslexia, Psychological Review, 98, 74-75.

Howell, S. R. & Becker, S. (2000). Modelling language acquisition at multiple temporal scales. Proceedings of the Cognitive Science Society, 2000, 1031.

Lakoff, G. and Johnson, M. (1999). Philosophy in the flesh: The embodied mind and its challenge to western thought. New York: Basic Books.

Lakoff, G. (1987). Women, fire and dangerous things: What categories reveal about the mind. Chicago and London: University of Chicago Press.

Rumelhart, D E., Hinton, G. E. & Williams, R. J. (1986) Learning internal representations by error propagation. In J. L. McClelland, D. E. Rumelhart and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations (pp. 318-362). Cambridge, MA: The MIT press.

Smith, L. B. (1999). Children's noun learning: How general learning processes make specialized learning mechanisms. In MacWhinney, B. (Ed.) (1999). The Emergence of Language. New Jersey: Lawrence Erlbaum Associates.