Modelling Language Acquisition: Lexical Grounding Through Perceptual Features

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

Abstract

A neural network model of language acquisition is introduced, motivated by current research in psychology and linguistics. It uses both extra-linguistic perceptual features and symbolic representations of words. The network learns to auto-associate these inputs 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). Furthermore, the inclusion of the extra-linguistic perceptual features is argued to be a form of direct developmental grounding in embodied concepts, which will allow the later learning of more abstract concepts to be grounded indirectly in meaning through relations to the first words. Through this bootstrapping process, future versions of the network may be scalable to large vocabularies, and may bridge the gap between high-dimensional and embodied theories of meaning.

Introduction

In attempting to model the actual processing and production of language, in a behavioural fashion, we consider it very important to take a developmental approach. 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. Furthermore, taking the developmental approach may offer solutions to the problem of symbol grounding, and provide a bridge between high-dimensional and embodied theories of meaning.

Developmental Language Acquisition

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. Children must have some understanding that a dog barks, is furry, and can be played with, even if they don't know the words 'dog', or 'barks', or 'furry'. Lakoff argues that children's sensorimotor experience is continually building up these embodied, pre-linguistic concepts, concepts that are very specific and concrete, and that these concepts enable the child to function in their particular limited world.

If we assume that this conceptual machinery is already well established by the time of the first words, the language learning problem becomes much simpler. If a child already has an embodied concept for things like 'dog', then when it begins to learn the word for dog, 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, Glietman, & 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. Gillette et al. offer a new interpretation of this difference, using 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. Thus the early words may be profitably considered different from the later words in language acquisition; they act as a foundation for the rest. 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. They are grounded directly in the child's embodied reality. The next step involves the use of these concrete nouns to help infer the less imageable verb meanings in the speech stream (still well embodied), 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, and grammar learning begins to emerge. 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 most 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 learner is acquiring new words without reference to extra-linguistic context, we are dealing with abstract symbols again. Or are we? The new words that are acquired through listening to speech or reading, without perceptual referents, are defined by their relations to other words in the context. To the extent that those other words, those symbols, have been directly grounded in meaning through associating with embodied concepts, then the new word becomes *indirectly* grounded through its relations to the grounded words.

The initial, imageable words that were directly grounded in embodied concepts serve as a foundation for the later words that will not be. Meaning can propagate up through the lexicon. But how exactly might this work? Neural network modelling might shed some light on the process, as we shall see. However, it is a complex proposition, and must be approached in progressive stages of investigation. The first stage, presented in the following neural network simulations, deals with the earliest aspects of language acquisition, initial lexicon and grammar learning, and with accompanying direct grounding in perceptual features. Extension of these first simulations to later, indirect grounding is also discussed.

Method

The model of language acquisition discussed herein (see Figure 1) takes as input arbitrary symbols for 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 localist word representations (arbitrary symbols), 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.

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. 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.

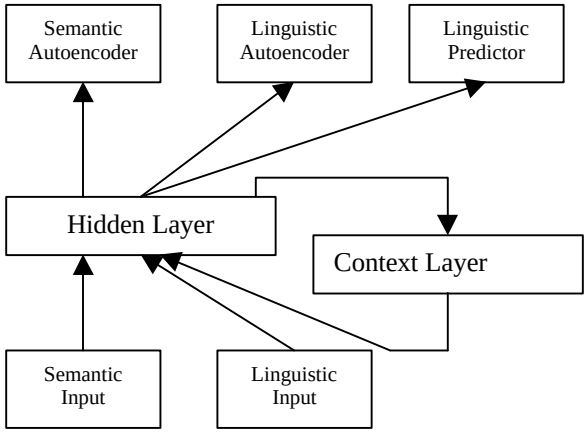

Figure 1: Modified SRN architecture, including standard SRN hidden layer and context layer, standard linguistic prediction layer, and novel semantic autoencoder and linguistic autoencoder.

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 demonstrated by Elman and colleagues. Second, adding a semantic layer will eventually allow for the use of phonemic input representations. The constancy of the semantic input (an analogue to focused attention to an object) across the successively presented phonemes will serve to bind the phonemes together into a word. The network discussed in this paper does not deal with phonemic inputs, however, 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, or whether it is still treating the word only as an arbitrary symbol.

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. This aspect of the model is reported in Howell and Becker (2001), and will not be considered in detail here.

Model Details

There are two input layers and three output layers. The semantic output layer is auto-encoding 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 via back-propagation exactly as are both 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.

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 near-asymptotic accuracy was achieved, found empirically to be at 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 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. Future work will address more explicitly the implications of hidden layer size for this type of network.

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) 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.

Complete lexical-grammatical analysis is presented in Howell and Becker (2001). For the present purposes, our analysis is limited to the semantic-linguistic relationships. Specifically, does the inclusion of extra-linguistic semantic features help or hinder the lexical and grammatical learning?

The experimental network, over 24 simulation runs, reached a mean peak lexical accuracy of 96.6 percent correct, while the mean peak prediction (grammatical) accuracy was 37.33 percent correct. Comparisons with ‘control’ or partial networks lacking the semantic or lexical autoencoder task indicate that each task is learned faster and more accurately in the experimental network than in the control networks.

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 also significantly lower than the experimental network ($m = 28.4$, $n=10$, $p <0.0001$).

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

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

Most relevant to our present argument, however, is the difference between the experimental network and control network 2. When the semantic input layer and output layer are removed, the performance of the network over the time frame drops significantly. That is, grammatical prediction is less accurate (28.4% vs. 37.33%).

Thus the semantic learning, which occurs first, can be viewed as building up embodied conceptual representations, or at least sensory representations, since our current extra-linguistic representations are mostly perceptual. With these learned, the lexical learning and grammatical learning are accelerated, and the arbitrary word representations become grounded in reality, or at least in perceptual features.

Is this a high-dimensional or an embodied model of meaning and language? We would argue that it is both, or at least has the potential to become both. Landauer and colleagues (e.g. Landauer, Laham & Foltz, 1998) provide perhaps the best example of a high-dimensional model of meaning, learning 'meaning' solely from word to word relations (although see also Burgess & Lund, 2000, for a different method, HAL, using a moving window over the text). Landauer’s Latent Semantic Analysis (LSA) technique takes a large corpus of text, such as an encyclopedia, and creates a matrix of co-occurrence statistics for words in relation to the paragraphs in which they occur. This yields a very high-dimensional vector representation for the word, which is then reduced in dimensionality through singular-value decomposition until a smaller 'meaning' vector is obtained for the word, usually about 300 elements long. These meaning vectors have been used by Landauer et. al. to demonstrate performance at the human level in such tasks as multiple choice vocabulary or domain knowledge tests and emulation of expert exam grading. These methods (LSA & HAL) have the advantage of realistically-sized vocabularies, the ability to handle large corpora, and near-human performance. What they lack, however, is any incorporation of syntax, since the words are treated as unordered collections (a 'bag of words'). More importantly, LSA or HAL 'meaning' vectors lack any grounding in reality. Experiments by Glenberg and Robertson (2000) have shown the LSA method to do poorly at the kinds of reasoning in novel situations that human semantics makes trivial, thanks largely to the embodied ‘meaning’ of human semantics.

As mentioned above, we believe that our method shows the potential to bridge these two forms of meaning. As Burgess & Lund (2000) discuss, their HAL method using their smallest text window produces similar results in word meaning clustering to an Elman SRN. In addition, they state that the SRN is a little more sensitive to grammatical nuances. Since our SRN architecture is becoming more complex in order to capture more aspects of grammar, we would expect to retain the advantage in grammatical relations. Further, Burgess & Lund point out that the two methods have in common the fact that words are represented in a high dimensional distributed meaning space; in our SRN, it is the hidden layer representation. So, our approach can be viewed as high-dimensional. In fact, once we have switched to using phonemic input representations, as the present architecture makes possible, we may be able to approach the effective vocabulary sizes of LSA. In addition, however, our feature vectors can be viewed as embodied, at least to the degree that the feature lists we use are empirically derived, such as the feature norms of McRae & colleagues (e.g. McRae, de Sa, & Seidenberg, 1997) to which we will be switching in future. This gives our network the important advantage of direct, embodied, grounding of meaning.

So, our network learns to associate arbitrary symbols with meaningful embodied features, and combines the high-dimensional and embodied approaches to meaning. That may be interesting in itself, but the real

value of our approach should be more evident as this architecture is scaled up. After the initial words are trained along with their feature vectors, generating an initial lexicon of directly grounded words, we intend that later words will receive less and less in the way of feature vectors; they will bootstrap themselves into meaning based on relations to earlier words. This will map to the way humans begin to experience words with less and less frequent perceptual context, until they are inferring word meanings solely from textual context. In our network, this corresponds to later words being introduced into the training corpus without feature representations. To the degree that our approach is successful, new words should begin to demonstrate at semantic output the same sort of features that similar words would.

Thus if the network's initial lexicon includes 'dog', and the sensorimotor features for dog, then when the new word 'wolf' is introduced, along with new corpus text that discusses wolves, the similarity of wolf to dog should become evident. After all, dogs and wolves are perceptually similar, may occur in many of the same kinds of real world situations and hence sentences (chasing things, eating things, running, howling, etc.). Thus we expect that over time the symbol 'wolf' will come to produce much the same perceptual output as 'dog' does, without being explicitly trained to do so.

Results from the simulations reported here are suggestive of this. We are already working on more realistically sized vocabularies and corpora to test this theory more rigorously.

Acknowledgments

Thanks to George Lakoff, whose writings and personal conversations inspired some of this work. This work has been supported by a Post-graduate Fellowship from the National Sciences and Engineering Research Council of Canada (NSERC) to SRH, a research grant from NSERC to SB, and an Ontario Graduate Scholarship to DJ.

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.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In Dietrich & Markham (Eds.) Cognitive Dynamics: Conceptual change in humans and machines.

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.

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

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language, 43,* 379-401.

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.*

Howell, S. R. & Becker, S. (In press). Modelling language acquisition: Grammar from the Lexicon? *Proceedings of the Cognitive Science Society, 2001.*

Landauer, T. K., Laham, D., & Foltz, P. W., (1998). Learning human-like knowledge by Singular Value Decomposition: A progress report. In M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems 10,*(pp. 45-51). Cambridge: MIT Press.

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.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126, 99-130.

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.