To appear in: The Handbook of Brain Theory and Neurla Networks, second edition, (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, 2002

# Language Acquisition

**Brian MacWhinney**

Language is a uniquely human achievement. All of the major social achievements of human culture -- architecture, literature, law, science, art, and even warfare -- rely on the use of language. Although there have been attempts to teach language to primates, the ability to learn language is a tightly copyrighted mark of the human species. This view of language led Chomsky (1965) to voice this assessment:

> It is, for the present, impossible to formulate an assumption about initial, innate structure rich enough to account for the fact that grammatical knowledge is attained on the basis of the evidence available to the learner. Consequently, the empiricist effort to show how the assumptions about a language acquisition device can be reduced to a conceptual minimum is quite misplaced. The real problem is that of developing a hypothesis about initial structure that is sufficiently rich to account for acquisition of language, yet not so rich as to be inconsistent with the known diversity of language.

To address this challenge, neural network researchers have explored a wide variety of network architectures and linguistic problems. This work has shown how children can learn language without relying on specifically linguistic, innate initial structure. However, several problems must be addressed before we can say that neural networks have answered Chomsky's challenge.

## An Example

Let us consider, as an example of this type of research, the neural network developed by MacWhinney, Leinbach, Taraban, and McDonald (1989). This model was designed to explain how German children learn to select one of the six different forms of the German definite article. In English, we have a single word "the" that serves as the definite article. In German, the article can take the form *der, die, das, des, dem,* and *den,* as indicated in Table 1. The choice of a particular form of the article depends on three additional features of the noun: its gender (masculine, feminine, or neuter), its number (singular or plural), and its role within the sentence (subject, possessor, direct object, prepositional object, or indirect object). There are 16 cells in the paradigm for the four cases and four genders (masculine, feminine, neuter, plural), but there are only six forms of the article. This means that a given form of the article, such as *der* can be used for either masculine-nominative-singular or feminine-genitive-singular, and so on.

| | **Masc.** | **Fem.** | **Neut.** | **Plural** |
|---|---|---|---|---|
| **Nom.** | der | die | das | die |
| **Gen.** | des | der | des | der |
| **Dat.** | dem | der | dem | der |
| **Acc.** | den | die | das | die |

*Table 1.* The German Definite Article Paradigm

To make matters worse, assignment of nouns to gender categories in German is quite nonintuitive. For example, the word for “fork” is feminine, the word for “spoon” is masculine, and the word for “knife” is neuter. Acquiring this system of arbitrary gender assignments is particularly difficult for adult second language learners. Mark Twain expressed his consternation at this aspect of German in a treatise entitled “The aweful German language” in which he accuses the language of unfairness and capriciousness in its treatment of young girls as neuter, the sun as feminine, and the moon as masculine. Along a similar vein, Maratsos and Chalkley (1980) argued that, because neither semantic nor phonological cues can predict which article accompanies a given noun in German, children could not learn the language by relying on simple surface cues.

These relations are so complex that a careful linguistic description of the system occupies well over 200 pages. However, MacWhinney et al. show that it is possible to construct a connectionist network that learns this system from the available cues. The model uses a simple feed-forward architecture. The input is structured into two pools of units. The first pool has 143 phonological units and 5 token meaning units. The second pool has 17 case cues from syntactic structure and 11 phonological cues from endings on the noun. These two input pools feeds into two separate pools of collector units that then feed together into a second level of hidden units. The output is a set of six nodes for the six possible forms of the German article. It is important to remember that each of these 6 articles must serve several functions to fill up the 16 cells of the declensional paradigm.

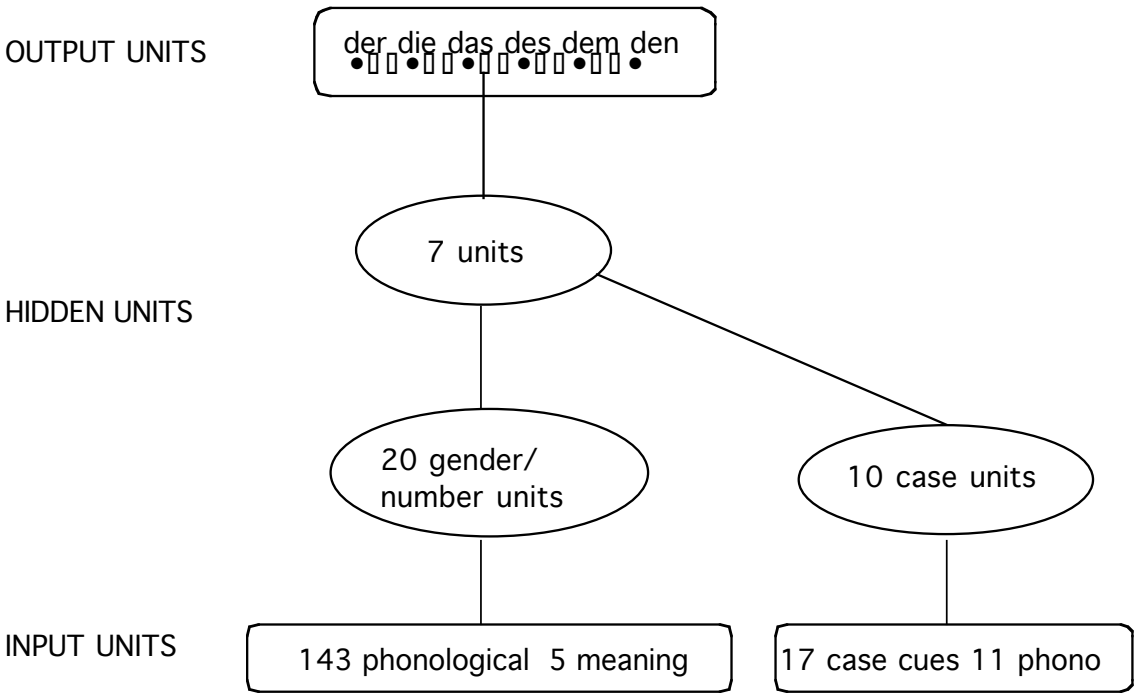


Figure 1: A network model of the acquisition of German declensional marking

The network was trained using the back-propagation algorithm. After 40 epochs of training on a set of 102 real German nouns, the network was able to choose the correct article 98% of the time. This meant that it not only succeeded in getting the gender of the noun right, but also figured out how to use the case cues to correctly select one of the 16 cells of the paradigm. To test the network's generalization abilities, we presented it with old nouns in new case roles. In these tests, the network chose the correct article on 92 percent of trials. This type of cross-paradigm generalization provides clear evidence that the network went far beyond rote memorization during the training phase. In fact, the network quickly succeeded in learning the whole of the basic formal paradigm for the marking of German case, number, and gender on the noun.

In addition, the simulation was able to generalize its internalized knowledge to solve the problem that had so perplexed Mark Twain -- guessing at the gender of entirely novel nouns. We presented the network with 48 new high frequency German nouns in a variety of sentence contexts. On this completely novel set, the simulation chose the correct article from the six possibilities on 61 percent of trials, versus 17 percent expected by chance. Thus, the system's learning mechanism, together with its representation of the noun's phonological and semantic properties and the context, produced a good guess about what article would accompany a given noun, even when the noun was entirely unfamiliar. In a subsidiary simulation, we showed that, when the model only has to guess the gender of the noun, and not its position in the paradigm, it achieves over 70% accuracy on new nouns. This is a level that comes close to that achieved by native speakers.

The network's learning paralleled children's learning in a number of ways. Like real German-speaking children, the network tended to overuse the articles that accompany feminine nouns. The reason for this is that the feminine forms of the article have a high frequency, because they are used both for feminines and for plurals of all genders. The simulation also showed the same type of overgeneralization patterns that are often interpreted as reflecting rule use when they occur in children's language. For example, although the noun "Kleid" (clothing) is neuter, the simulation used the initial "kl" sound of the noun to conclude that it was masculine. Because of this, it invariably chose the article that would accompany the noun if it were masculine. Interestingly, the same article-noun combinations that are the most difficult for children were also the most difficult for the network.

Demonstrations of this type illustrate how children can acquire linguistic knowledge without relying on stipulated, hard-wired constraints of the type envisioned by Chomsky. Similar demonstrations have been produced in a wide variety of areas including: the English past tense, Dutch word stress, universal metrical features, German participle acquisition, German plurals, Italian articles, Spanish articles, English derivation for reversives, lexical learning from perceptual input, deictic reference, personal pronouns, polysemic patterns in word meaning, vowel harmony, historical change, early auditory processing, the phonological loop, early phonological output processes, ambiguity resolution, relative clause processing, word class learning, speech errors, bilingualism, and the vocabulary spurt.

**Challenges**

Researchers have contested the logic underlying these demonstrations. Some of the problems that have been raised relate only to minor implementational features of the earliest models (MacWhinney & Leinbach, 1991), but others are more fundamental. The five most fundamental challenges are:

1. Dual route. Neural networks provide a good account of associative processes, but fail to account for the learning of regular rules.
2. Lexical learning. Neural networks have problems learning large numbers of words.
3. Syntax. Neural networks have problems dealing with the compositional aspects of complex syntax.

4. Neuronal reality. Some neural network architectures make inappropriate assumptions regarding neural processing.
5. Embodiment. Neural networks have not been able to model the ways in which the mind is linked to the body.

Let us look at responses to each of these five challenges in greater detail.

**Dual route**

Pinker (1999) has been a key proponent of the application of a dual-route model to language acquisition. He contends that irregular forms, such as *fell, went, feet,* or *broken*, are processed through an associative memory grounded on neural networks, but that regular forms, such as *jumped, wanted, cats,* and *dropped*, are produced by rule. Pinker views his defense of the psychological reality of linguistic rules as a part of a general defense of the linguistic theory of generative grammar.

The fact that irregulars are processed differently from regulars does not prove the existence of symbolic rules. Neural networks have no problem representing both regular and irregular patterns in a single network. For example, the network developed by Kawamoto (1994) encodes regular forms as nodes in competition with less regular nodes. In that homogeneous recurrent network architecture, regular and irregular forms display quite different temporal activation patterns, in accord with empirical observations. Kawamoto's model also accords with the fact that even the most regular patterns display phonological conditioning and patterns of gradience (Bybee, 1995) of the type modeled by neural networks.

**Lexical learning**

Neural networks have problems learning large numbers of words. Typically, neural network architectures have been used primarily as methods for extracting and classifying patterns. Word learning is differs from classification in two ways. First, the association between a word's meaning and its sound is almost entirely arbitrary. There is nothing in the specific sounds of *table* that depicts the shape or purpose of a table. This means that the learning of words cannot rely on the methods for pattern detection that are so important in neural network research. Second, the number of words that a speaker must learn is extremely large. If we look just at word stems, adult English speakers control between 10,000 and 50,000 words. Many of these words have multiple meanings and many can be further combined into compounds and rote phrases. Thus, the effective lexicon of an adult English speaker is from 20,000 to 80,000 words.

Self-organizing maps (SOMs) (Farkas & Li, 2001) offer a promising framework for dealing with the encoding of lexical items. Precision of encoding can be obtained by increasing the dimensionality of the coding space and then recompressing the additional dimensions. Conflicts between related words that are close on the lexical map can trigger a process of focused learning that concentrates specifically on words that are being confused. Catastrophic interference can be avoided by adding new nodes without forgetting older patterns (Hamker, 2001).

Because SOMs provide a relatively local encoding of words, they then allow us to address four additional problems that stem from problems of representing words in neural networks.

1. **U-shaped learning**. Children often produce a form like *went* correctly for several weeks or months and then shift to occasionally saying *go-ed*. Later, they move back to saying *went* consistently. This pattern, known as u-shaped learning requires an ability to learn some forms first by rote. Back propagation networks are good at producing overgeneralizations like *go-ed* but weak at producing and holding on to a rote form like *went* (Plunkett & Marchman, 1993). By default, SOMs place an emphasis on early rote learning and are slower to generalize out the regular patterns.
2. **Homophony.** Because most neural network models do not have discrete representations for lexical items, they have problems distinguishing homophonous forms. Consider what happens to the three homophones of the word *ring* in English. We can say *the maid wrung out the clothes*, *the soldiers ringed the city*, or *the choirboy rang the bell*. These three different words all have the same sound /rIN/ in the present, but each takes a different form in the past. In SOMs, these three words have clearly different representations in semantic space.
3. **Compounds.** Without discrete representations for lexical items, neural networks have problems with compound words. The fact that the past tense of *undergo* is *underwent* depends on the fact that *undergo* is a variant of the stem *go*. When the compound itself is high enough in frequency, the network can learn to treat it as an irregular. Networks have problems learning the past tense of low frequency irregular compounds. However, if the network can detect the present of "go" inside "undergo," it can solve this problem.

4. **Derivational status.** Neural networks have problems utilizing information regarding the derivational status of lexical items. In English, the past tense forms of denominal verbs always receive the regular past tense suffix. For example, the word *ring* can be used as a verb in *the groom ringed her finger*, but we would never say *the groom rung her finger*. Without an ability to know that a word derives from a noun, neural networks cannot encode this pattern. German provides even clearer examples of the importance of derivational status. All German nouns that derive from verbs are masculine. For example, the noun *der Schlag* ('blow'; 'cream') derives from the verb *schlagen* ('to hit'). However, there is no motivated way of indicating this in the model. In general, the model includes no independent way of representing morphological relationships between words. Thus, no distinction is made between true phonological cues such as final /e/ or initial /kn/ and derivational markers for the diminutive, such as *-chen* or *-ett*. This leads to some very obvious confusions. For example, masculines such as *der Nacken* ('neck') and *der Hafen* ('harbor') end in phonological /en/, whereas neuters such as *das Wissen* ('knowledge') and *das Lernen* ('learning') end in the derivational suffix *-en*. Confusion of these two suffixes leads to inability to correctly predict gender for new nouns ending in /en/. Without having a way of representing the fact that derivational morphemes have an independent lexical status, neural networks cannot process these patterns.

These four difficulties reflect a single core problem. By working with neural networks that flexibly encode lexical items, we can begin to address these additional features of word structure.

**Syntax**

Elman (1990) has provided demonstrations of the ability of neural networks to process complex syntactic structures. His model uses recurrent connections to update the network's memory after it listens to each word. The network's task is to predict the next word. This framework views language comprehension as a highly constructive process in which the major goal is trying to predict what will come next. Psycholinguists recognize the importance of prediction, but they view the major task of language processing as the construction of mental models. It is not clear how understanding prediction will help us understand the construction of mental models, although the two processes are certainly related.

An alternative to the predictive framework relies on the older neural network mechanisms of spreading activation and competition. For example MacDonald, Seidenberg, and Perlmutter (1994) have presented a model of ambiguity resolution in sentence processing that is grounded on competition between lexical items. Models of this type, do an excellent job of modeling the temporal properties of sentence processing. Such models assume that the problem of lexical learning in neural networks has been solved. They then proceed to use localist representations to control interactive activation during sentence processing. Until we have indeed solved the problem of lexical learning, this is a very effective way of advancing the research agenda.

Another approach that makes similar assumptions uses a linguistic framework known as Construction Grammar. This framework emphasizes the role of individual lexical items in early grammatical learning (Tomasello, 2000). Early on, children learn to use

simple frames such as *my + X* or *his + X* to indicate possession. As development progresses, these frames are merged into general constructions, such as the possessive construction. In effect, each construction emerges from a lexical gang. Sentence processing then relies on the child's ability to combine constructions online. When two alternative constructions compete, errors appear. An example would be **say me that story*, instead of *tell me that story*. In this error, the child has treated say as a member of the group of verbs that forms the dative construction. In the classic theory of generative grammar, recovery from this error is supposed to trigger a learnability problem, since such errors are seldom overtly corrected and, when they are, children tend to ignore the feedback. Neural network implementations of Construction Grammar address this problem by emphasizing the direct competition between *say* and *tell* during production. The child can rely on positive data to strengthen the verb *tell* and its link to the dative construction, thereby eliminating this error without corrective feedback. In this way, models that implement competition provide solutions to the logical problem of language acquisition.

These various approaches to syntactic learning must eventually find a way of dealing with the compositional nature of syntax (Valiant, 1994). A noun phrase such as "my big dog and his ball" can be further decomposed into two segments conjoined by the "and". Each of the segments is further composed of a head noun and its modifiers. Our ability to recursively combine words into larger phrases stands as a major challenge to connectionist modeling. One likely solution would use predicate constructions to activate arguments that are then combined in a short-term memory buffer during sentence planning and interpretation. To build a model of this type, we need to develop a clearer

mechanistic link between constructions as lexical items and constructions as controllers of the on-the-fly process of syntactic combination.

**Neuronal realism**

Some researchers have criticized neural network models in the area of language acquisition for a failure to properly represent basic facts about the brain. To the degree that the back propagation algorithm relies on reciprocal connections between units, this criticism is well-founded. However, work in this area has begun to rely on models such as self-organizing feature maps, adaptive resonance, and Hebbian learning that have closer mappings to the features of neural organization. In fact, Elman (1999) has shown how the imposition of biologically realistic assumptions, such as the brain's preference for short connections, can lead to more effective language learning. Thus, this particular challenge to neural network theory may end up being more of a searchlight than a barrier.

Neural networks must also achieve a closer match to what we are now learning about functional neural circuitry. We know that auditory cortex, Broca's area, temporal word storage, and frontal attentional areas are all involved in various ways in language processing. However, we have not yet figured out exactly how these separate brain structures map onto separate aspects of lexical and syntactic processing.

**Embodiment**

The final challenge to neural network modeling comes from researchers who have begun to explore the ways in which the mind is grounded on the body. This relatively new line of research emphasizes the importance of findings that mental imagery makes

use of the reactivation of perceptual systems to recreate physically grounded images. A convergence of work in neuroscience, psychology, and cognitive linguistics points to the view of language use not as disembodied symbol processing, but as indirectly grounded on basic mechanisms for perception and action which themselves operate on the human body. Neural network models have just begun to deal with this new challenge. One approach emphasizes the ways in which distal learning processes can train action patterns such as speech production on the basis of their perceptual products (Plaut & Kello, 1999). Another approach, adopted by the NTL (Neural Theory of Language) group (Bailey, Feldman, Narayanan, & Lakoff, 1997) relies on the higher-order formalism of Petri nets to represent the control structure of body motions such as *pull* or *stumble*. The architecture then includes transparent methods of linking the higher-level representation to a neural network implementation.

One trend that will facilitate this work, as well as all modeling of language acquisition is the increasing availability of transcript and multimedia data from children interacting with their caretakers. The Child Language Data Exchange System (CHILDES) at http://childes.psy.cmu.edu now provides thousands of hours of transcripts of child language data, much of it linked to audio and some to video. In the context of the broader TalkBank Project at http://talkbank.org, this data is being recoded in XML format and linked to a variety of computational tools for analyzing gestural, phonological, morphological and syntactic structure. This growing database provides increasingly rich targets for neural network modeling.

## Conclusion

Neural networks have addressed many aspects of Chomsky's challenge. They have been used to develop useful models of virtually all aspects of language learning and processing. However, further challenges lie ahead. Of these, the most pressing is the need to develop methods for simulating the learning of a realistically sized lexicon of several thousand words. If this problem can be solved, it will have further positive consequences for models of syntactic development that emphasize the importance of item-based learning. It is likely that a good solution to this problem will need to rely on an improved understanding of the ways in which the brain stores and processes lexical items. An even greater challenge will be developing models that express the ways in which language processing is grounded on embodied cognition. Together, these challenges guarantee vitality in this area for years to come.

## References


* Bailey, D., Feldman, J., Narayanan, S., & Lakoff, G. (1997). Modeling embodied lexical development. *Proceedings of the 19th Meeting of the Cognitive Science Society*, 18-22.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*, 425-455.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

* Elman, J. (1990). Finding structure in time. *Cognitive Science, 14*, 179-212.

Elman, J. L. (1999). The emergence of language: A conspiracy theory. In B. MacWhinney (Ed.), *The emergence of language* (pp. 1-28). Mahwah, NJ: Lawrence Erlbaum Associates.

Farkas, I., & Li, P. (2001). Modeling the development of lexicon with a growing self-organizing map. *NIPS*.

Hamker, F. H. (2001). Life-long learning Cell Structures -- continuously learning without catastrophic interference. *Neural Networks, 14*, 551-573.

Kawamoto, A. (1994). One system or two to handle regulars and exceptions: How time-course of processing can inform this debate. In S. D. Lima & R. L. Corrigan & G. K. Iverson (Eds.), *The reality of linguistic rules* (pp. 389-416). Amsterdam: John Benjamins.

* MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676-703.

* MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 29*, 121-157.

* MacWhinney, B. J., Leinbach, J., Taraban, R., & McDonald, J. L. (1989). Language learning: Cues or rules? *Journal of Memory and Language, 28*, 255-277.

Maratsos, M., & Chalkley, M. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Children's language: Volume 2* (pp. 127-214). New York: Gardner.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech conrephension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381-416). Mahwah, NJ: Lawrence Erlbaum Associates.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building. *Cognition, 49*, 21-69.

* Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences, 4*, 156-163.

Valiant, L. (1994). *Circuits of the mind*. Oxford: Oxford University Press.