

A Computational Model for First Language Acquisition

Paula Buttery

Natural Language and Information Processing Group
Computer Laboratory, University of Cambridge
JJ Thomson Avenue, Cambridge CB3 0FD
paula.buttery@cl.cam.ac.uk

Abstract

A model of lexical acquisition has been developed in order to learn both the semantic and syntactic category of lexemes associated with a generalised categorial grammar. Computational simulation shows that acquisition is possible when some prior language knowledge is assumed. The effects on acquisition have been explored when aspects of this prior knowledge are varied.

1 Motivation and Background

The study of child language acquisition is prolific in terms of cognitive and socio-linguistic experiments. These experiments are designed to give an insight into the way in which children go about acquiring language. Detailed theorems are proposed using the results of these experiments as evidence.

Computational simulations can be a useful tool in the study of language acquisition. If a model of the learning procedure can be built then the theorems suggested by experimenters can be investigated. The significance of such an investigation could be to add weight to the validity of a theorem or to highlight problem areas where the theorem will come unstuck. The model may also be able to add something to discussion in the event that there is dispute over which theorem is the most relevant for a particular phenomenon.

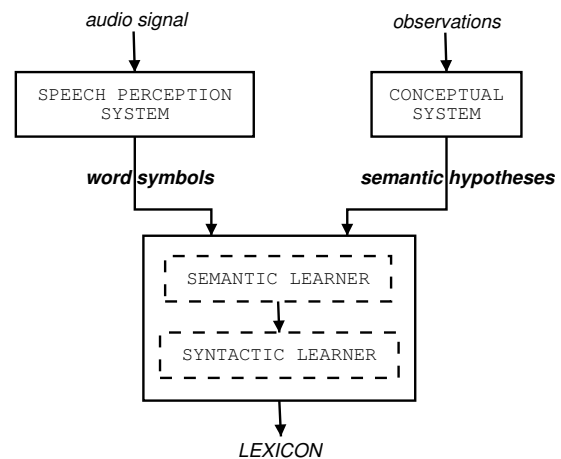
One benefit of using a computer model of a learner rather than the real thing is that the learning simulation can be stopped at any point and the internal state of the learner can be exposed. This of course is not possible with a real child.

Until the problem of language acquisition is solved a model of the problem can not be perfect. The next version of the model can be im-

proved on the basis of shortfalls of the previous one. The current model will be built on a set of assumptions. Any predictions made by the model are only as relevant as the original assumptions it makes. Bearing this in mind, the purpose of this work is to create an initial model of language acquisition and show that this model can indeed learn. This original model based on strong assumptions can then be made more complex (the assumptions can be relaxed) if the model promises to be useful.

2 Starting Assumptions

For the basis of this work Siskind's cognitive model (Siskind, 1996) is assumed. This is a simple model representing the interaction between a speech perception system and a conceptual system.



On hearing an utterance the purpose of the speech perception system is to break up the acoustic signal and pass a series of word symbols to its output. At the same time the conceptual system is responsible for producing semantic hy-

potheses for the utterance¹.

The conceptual system need not only rely on visual observation to produce semantic hypotheses. This means that a blind child will also be able construct hypotheses without witnessing the details of the scene. The number of hypotheses that a blind child produces may be somewhat larger than that of a seeing child but, as will be shown, this won't matter too much as long as the correct hypothesis is present.

A possible problem with the conceptual system is that it could produce an infinite number of semantic hypotheses for a given utterance. Siskind avoids this problem by stating that the learner will only entertain likely semantic hypotheses. However, he does not specify the distinction between likely and unlikely. Pinker, on the other hand, (Pinker, 1994) suggests that semantic hypotheses will be constrained for two reasons. First, they are constrained by the semantic structures that constitute mental representations of a word's meaning. This is referred to as the Universal Lexical Semantics which is analogous to Chomsky's Universal Grammar (Jackendoff, 1990). Secondly, hypotheses may be constrained by the way the child's lexicon is constructed. It seems that children are fairly unwilling to admit true synonyms to their lexicon and consequentially a child would rather not hypothesize an existing word's meaning for a new word (Clark, 1987). Even when these constraints are taken into account there may still be several plausible semantic hypotheses for a given utterance. The conceptual system is therefore expected to produce a set of semantic hypotheses.

In short, the speech perception system and conceptual system are acting as black boxes. We do not know the exact mechanism they use to produce their output. What is important is that there is some process that allows streams of words symbols and semantic hypotheses to be

¹In this paper the word symbols which are produced by the speech perception system will be written in italics. The semantic symbols that are produced by the conceptual system to make up the semantic hypotheses of an utterance will be written in bold. So, a child hearing the utterance "Kitty eats food" would theoretically produce the word symbols *Kitty*, *eats* and *food* from their speech perception system. If the child was simultaneously observing the cat eating something they would hopefully also produce the semantic expression **eat(Kitty, food)** from their conceptual system.

created for each utterance. These data streams are the starting point from which the modeled learner can attempt to acquire a lexicon.

2.1 Generating Input for the Learner

Since real speech perception systems and conceptual systems are not available, their output must be simulated to provide input for the learner.

The purpose of the speech perception system is to segment the audio signal. This is not a straight forward task (Brent, 1999). Speech doesn't contain any reliable markers analogous to the blank spaces between words in text. For adults segmentation is an easier (although not fool-proof) task. The corpus used for this work has already been transcribed. The sets of words symbols required can be simply created from the textual representation of each utterance.

Simulating the output of the conceptual system is more difficult. The approach used here was to parse utterances using an existing grammar and extract the semantic representations produced. The set of semantic hypotheses for a particular utterance could then be selected from these. This method gives control over how much noise the learner is exposed to. At one extreme the set of semantic hypotheses could contain the one single correct hypothesis and at the other extreme many incorrect hypotheses.

3 The Corpus and Grammar

The Sach's Corpus of the Childes database (MacWhinney, 1995) has been used for this work. The corpus contains a selection of interactions between a child and her parents from the age of one year one month to five years one month. A unification based generalised categorial grammar has been created by Villavicencio (Villavicencio, 2002) to describe this corpus. The grammar uses the rules of application, composition and generalised weak permutation.

- Forward Application
 $X/Y Y \rightarrow X$
- Backward Application
 $Y X \setminus Y \rightarrow X$
- Forward Composition
 $X/Y Y/Z \rightarrow X/Z$
- Backward Composition
 $X \setminus Y Y \setminus Z \rightarrow X \setminus Z$

- Generalised Weak Permutation
 $((X | Y_1) \dots | Y_n) \rightarrow ((X | Y_n) \dots | Y_1)$
 where $|$ is a variable over \setminus and $/$.

The corpus was preprocessed so that the child’s sentences were also excluded. Only the parents’ sentences are given as input to the system. Also all phonological annotations and grammatical structures not covered by Villavicencio’s grammar, including interjections and elliptical material, have been removed.

4 The Learning Algorithms

The learning algorithm consists of two parts: A semantic learner and a syntactic learner. Each utterance in turn is first processed for the acquisition of semantic information and subsequently for the acquisition of syntactic information. The utterance is then discarded and not referred to again. It is assumed that children behave in a similar manner, acquiring all the information they can from an utterance upon the moment of hearing it. It is unlikely that children have the inclination to store utterances for processing later. It would also be impractical since the child would have to remember all the situational information associated with the utterances.

4.1 Semantic Learner

The semantic learner receives as input a set of word symbols from the speech perception system and a set of semantic hypotheses from the conceptual system. At the simplest extreme the set of semantic hypotheses will contain only the one correct meaning. In such a case the learner’s task is to find a mapping between the word symbols and parts of this semantic expression. The semantic learner attempts to produce this mapping using the theories of cross-situational learning and covering constraints and is based on Siskind’s work on Cross Situational Techniques (Siskind, 1996).

4.1.1 Cross Situational Learning

Cross situational learning has been suggested as a method of learning for hundreds of years but more recently by Pinker (Pinker, 1989) among others. The theory speculates that lexical acquisition may be achieved by finding the common factors across all observed uses of a word. Hearing a word in enough contexts should therefore allow the learner to rule out all incorrect hypotheses and converge on a unique meaning.

For a trivial example consider the utterances “Santa laughs” and “Santa eats pies”. They would have word symbol sets and semantic expressions as follows:

$$\{santa, laughs\} \mapsto \mathbf{laugh}(santa)$$

$$\{santa, eats, pies\} \mapsto \mathbf{eat}(santa, pies)$$

From these two utterances it is possible to ascertain that the meaning associated with the word symbol *santa* must be **santa** since it is the only semantic element that is common to both utterances.

4.1.2 Covering Constraints

The idea of covering constraints is essentially the reverse of cross-situational learning. The idea requires that the semantic expression representing a complete utterance is built up only from the semantic expressions relating to words within that utterance, i.e. it doesn’t contain any external semantic information. Given that this is the case, consider the situation where the semantic mapping for all but one of the word symbols is known. The semantic expression associated with the final word symbol is necessarily what is left over when the all the known semantic expressions are removed from the expression representing the entire utterance.

Consider the example “Grinch hates Xmas”. If the following is already known:

$$\{grinch, hates, Xmas\} \mapsto \mathbf{hate}(grinch, xmas)$$

$$grinch \mapsto \mathbf{grinch}$$

$$hate \mapsto \mathbf{hate}(x, y)$$

Then the necessary conclusion is:

$$xmas \mapsto \mathbf{xmas}$$

4.1.3 Constraining Hypotheses with Partial Knowledge

Cross situational learning and covering constraints are most useful if the correct semantic expression is known. In the situation where there are several semantic hypotheses, the learner tries to reduce the number before applying the techniques.

The hypotheses are constrained by removing all those that are impossible given what has already been learnt. To show how this works, imagine the learner has heard the utterance

“Mice like cheese” and hypothesized the following semantic expressions:

like(**mice**, **cheese**) (1)

madeOf(**moon**, **cheese**) (2)

madeOf(**moon**, **cake**) (3)

If it has already established that *cheese* maps to **cheese** then 3 can be ruled out as a possible meaning since it doesn't contain the necessary semantic expression. Hypothesis 2, however, can not be ruled out. The learning algorithm attempts to learn from all remaining hypotheses. If all semantic hypotheses are ruled out then the learner assumes that one of the words in the utterance has multiple senses.

4.1.4 Dealing with Noise

The learner is able to recover from using incorrect semantic hypotheses. This is achieved by associating a confidence factor with each mapping. Whenever the mapping is confirmed the confidence factor is increased. Periodically mappings are removed if their confidence factor is below a threshold level.

4.2 The Fundamental Assumption

The syntactic learner relies on output from the semantic learner. This idea is consistent with the theorem of semantic bootstrapping (Pinker, 1987) where the child uses its semantic knowledge of some words to help it to begin acquiring syntax.

A fundamental assumption is employed to link the two learners; *The semantic arity of a word is the same as its number of syntactic arguments.*

For example, if it is known that *likes* maps to **like(x, y)**, then the fundamental assumption says that its syntactic category will be in one of the following forms: $a \setminus b \setminus c$, $a / b \setminus c$, $a \setminus b / c$, $a / b / c$ or more concisely $a \mid b \mid c$ (where a, b and c may be basic or complex syntactic categories themselves).

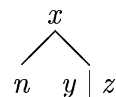
4.3 Syntax Learner

The syntax learner attempts to create valid parse trees for the current utterance using the categorial grammar rules. Finding a valid parse can then help to resolve unknowns in the syntactic category types. The algorithm proceeds as follows:

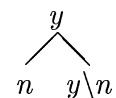
On receiving an utterance each word is looked up. If it already has a syntactic category associated with it that category is retrieved otherwise skeleton categories are assigned to the word by means of the fundamental assumption. Note that skeleton categories can only be assigned if a semantic mapping has been learnt. Utterances which contain some words that are yet to have a semantic mapping are discarded at this point.

When syntactic information is available for every word in the utterance, binary tree structures are created that have as many leaf nodes as words in the utterance². The leaves of these trees are labeled with the known or skeleton syntax categories. It is sufficient to consider only binary trees since the rules of the categorial grammar involve binary operations. The permutation rule is incorporated by allowing any ordering of the functor-category pairs. All of these pairs are tried when attempting to apply one of the categorial grammar rules.

For each node of the tree the learner looks at the information in the adjacent nodes and applies a rule of the categorial grammar if possible. Consider a node x where there is a basic syntactic category n at the left child and a category with one argument $y \mid z$ at the right node (with x , y and z being unknown variables):

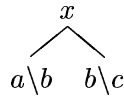


The learner notices that the backward application rule can be applied here and updates the nodes accordingly:

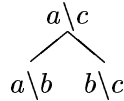


Now consider the following tree where a, b and c are syntax categories and x is unknown:

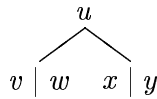
²All possible binary tree shapes for the utterance length are considered in turn. Unfortunately there are an exponentially increasing number of such trees for a utterance of length n given by Catalan(n). Also some words have more than one syntax category, in these cases it is necessary to consider trees for all syntactic permutations of the utterance. However, if all binary trees are tried for all labellings this becomes computationally intractable. This problem can be solved by using heuristics or dynamic programming (a chart) to share subanalyses and cut down the computation to polynomial time.



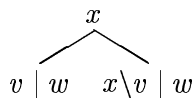
The learner would apply the backward composition rule to give:



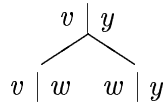
In some situations, usually when there are many unknowns, more than one rule can be applied. In the next example u , v , w , x and y are variables which may be basic or complex syntactic categories.



Applying an application rule would give something like:



While applying a composition rule could give something like:



In these situations, when more than one rule applies, there are two options of how to proceed. The first option is to try all possibilities. The second is to apply heuristics to make the decision on how to apply the rules. Such a heuristic might be to take an argument from the right when there is a choice since this is most common in English. In the case of short utterances there is no reason not to try all possibilities. Heuristics are applied only when the sentence is long.

Unparseable trees are discovered when none of the grammar rules can be applied to a node and its children. These trees are discarded. The remaining valid trees can provide evidence for the syntax categories of words at their leaf nodes.

4.3.1 Learning the Basic Syntax Categories

Without making further information available to the learner the syntax categories at the leaf nodes will have unknown variables in them. The learner needs to know something about the basic categories of the categorial grammar. This is achieved in two ways:

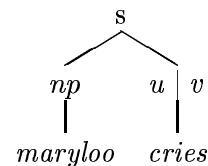
1. The learner is constrained to only allow trees where the top node is either s or $s \backslash np$ (in the case of imperatives)³.
2. The learner is given the syntactic category of some group of words.

In terms of a real learner this second point is equivalent to having an innate ability to recognise groups of entities linked by some common theme and labeling all the entities in that group with the same mental tag. This can't be too far off what children must actually do. For instance, it seems probable that children are innately aware of the concept of an object (Piaget, 1954) and might therefore label books, tables and chairs with the same object tag.

For a very simple example consider the utterances "Marylou cries" and "The grinch cries". Suppose that the learner is predisposed to recognise and label the groups that are captured by the atomic categories of the categorial grammar. Then a learner could know that *marylou* is of semantic category **marylou** and syntactic category np , *grinch* is of semantic category **grinch** and syntactic category n .

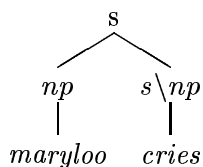
In order to find a skeleton category for the remaining words we look at their semantic category and then apply the fundamental assumption: *the* has semantic category **the(p)** and therefore skeleton syntactic category $x \backslash y$, *cries* has semantic category **cry(q)** and therefore skeleton syntactic category $u \backslash v$.

An initial tree for the first utterance would be:

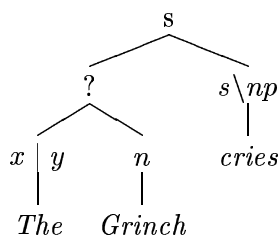


³In fact $s \backslash np$ should be mapped to $s[\text{imp}]$ by a lexical rule applied to verbs so the heads of imperatives are constrained to appropriate verb forms: Close the window! vs. *Closes the window!

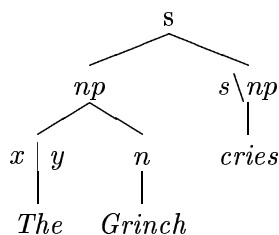
Crawling this tree the learner would discover a possible mapping between *cries* and syntactic category $s \setminus np$.



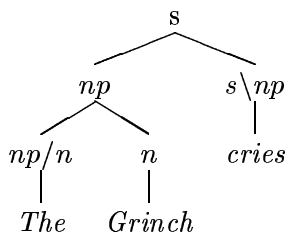
Now possible initial trees are drawn for the next utterance. One is shown below. Any mappings already learnt are used.



Crawling this tree the syntax learner would first discover the following:



and then:



As with the semantic learner confidence values are attached to the syntax mappings. With periodic disposal of those mappings with low confidence values, mislearnt mappings are not propagated.

5 Results and Evaluation

The learner has been run on the first 2000 utterances of the corpus. The input to the system is a collection of utterances paired with sets of

semantic representations. These semantic representations are the hypothesized meanings a child may assume for an utterance. The correct interpretation may or may not be included amongst the semantic representations and the number of representations per utterance may be varied.

Initially the learner was run with each utterance being paired with only one correct semantic expression. In terms of a real learner this would mean that a child knows exactly what is meant whenever she hears an utterance. The semantics learner achieved recall of 68% with precision 98% under these conditions

The syntactic learner is dependent upon the recall of the semantic learner in that it can only proceed when a skeleton syntactic category is available for every word in the utterance; when this is the case the success rate is high with precision 82%.

5.1 Increasing the Number of Hypotheses

The learner was run with increasing numbers of semantic hypotheses per utterance. The extra hypotheses were chosen randomly and the correct semantic expression was always present in the set. Hypotheses sets of sizes 2, 3, 5, 10 and 20 were used. Recall remained fairly constant regardless of the number of hypotheses. The precision also remained very high moving only as low as 93% for the set with 20 hypotheses.

The recall values had small variance for the following reason: As soon as the meanings of the most frequently occurring words have been found many of the incorrect hypotheses can be ruled out (using the method of constraining hypotheses explained above). Once this starts to happen the problem rapidly reduces itself to that of just having the one correct semantic hypothesis.

The reduction in precision was caused when utterances containing a particular word had been repeatedly paired with hypotheses containing the same incorrect semantic expression. This scenario was fairly likely given this particular corpus since many of the utterances contain the same type of information. Some semantic expressions are much more likely than others when choosing hypotheses at random from all those generated by the corpus. For instance, the nouns *baby* and *coffee* appear very frequently. It

is possible for an infrequently occurring word to get paired with semantic expressions for one of the very frequent words every time it appears. This could lead to it being learnt incorrectly.

Children appear to have a mechanism for dealing with this problem. As discussed earlier, in real life children don't like to admit synonyms to their lexicon (Clark, 1987). This behaviour could be built into the learner by not allowing two words symbols to map to the same semantic expression. To avoid the situation where an incorrect mapping is acquired first, thus precluding the correct mapping, it would be necessary to implement this using the confidence factors. Perhaps mappings to a particular semantic expression could be prohibited once one mapping exists which has a confidence value higher than some threshold value.

5.2 Introducing Noise

Finally, the learner was run with some utterances being completely mismatched with semantic hypotheses (i.e. the correct hypothesis was not present amongst the set). This is analogous to the case where the child was not able to understand the meaning of the utterance from its observations. The results were found to be highly dependent on the utterances that were chosen to be mismatched. If many utterances were chosen that contained infrequently occurring words then the recall would plummet. There is a clear reason for this result. The distribution of words in the corpus is Zipfian. Most words appear very infrequently (over 250 words appear just once and more than 125 appear twice). In the original experiment (where only the correct hypothesis was paired with the meaning) 36% of words could be learnt with only one exposure. This capability is useless if a word that appears only once in the corpus is paired with an incorrect hypothesis. In such a situation the word will never be learnt.

5.3 The Case for Semantic Bootstrapping

This learner uses the semantics properties of words as a clue to which syntactic category they belong. This is referred to as semantic bootstrapping (Pinker, 1987). The theory suggests that semantics are needed in order to begin the syntax learning process.

This learner cannot currently learn anything

about a word's syntax until it knows about its semantics. This doesn't imply that it is impossible to learn syntax before semantics. Consider the case where the syntactic category of every word but one is known. The category of the new word will be easy to infer by looking at the parse tree. This would be an easy extension to the current learner.

The syntactic category of a word can tell us some useful things about its semantics (its arity) but not everything; knowing the syntax category of a word does not help us decide which predicate name to use for the semantic representation. Likewise, the semantic representation tells us nothing about how the word behaves structurally with others in sentence but can tell us something about the number of syntactic arguments it takes.

Gleitman asserts that there are some verbs for which the semantics can not possibly be learnt without resorting to their sub-categorization frame. This is what she calls syntactic bootstrapping ((Gleitman, 1990) and (Fisher *et al.*, 1994)). Pinker on the other hand says this information is interesting "like a puzzle" and therefore potentially useful to clever children/adults but is not essential.

If Gleitman is correct and it is impossible to ascertain the meaning of some words without first resolving their syntactic category, then it will be essential to provide feedback within the model from the syntactic learner back to the semantic learner.

What is important, however, is that this learner can not begin to learn syntax in the first instance without first learning some semantics and using it to infer some initial syntactic categories. Hence the learner presented here tends to support the semantic bootstrapping theory given the assumptions made.

5.4 The Validity of the Fundamental Assumption

It is not always the case that the number of syntactic arguments for a word is the same as the semantic arity. For example consider the following sentences:

1. "Marylou tries to be happy"
2. "Marylou seems to be happy"

tries and *seems* in these sentences have two syntactic arguments but their semantic arities are not both two. Sentence 1 would have a semantic representation something like **try(marylou, happy(marylou))** where *try*'s semantic arity is two. Sentence 2, however, needs a semantic representation something like **seem(happy(marylou))** where *seem*'s semantic arity is only one.

To improve the model to deal with these types of words the fundamental assumption needs to be relaxed.

6 Conclusions and Future Work

The learner presented here can acquire a lexicon from real data. The model appears to be a useful starting point for investigations into acquisition. The model's initial assumptions have been relaxed in this work to show that learning is possible in the presence of multiple semantic hypotheses and in the presence of noise. The model also gives support to the theory of semantic bootstrapping.

The learner has obvious scope for further improvement. For instance, the fundamental assumption could be relaxed and a mechanism for not allowing exact synonyms could be introduced.

The potential of the model for studying other acquisition theories is great. One interesting experiment would be to look at the known lexicon at any given point in the corpus and compare this lexicon with the utterances the child was producing at this stage. This study may be able to highlight the reasons why the child was making particular errors. Also something may be added to the discussion of language production vs. language comprehension by looking at the difference between the size of lexicon the child uses and the lexicon the model has acquired.

References

M Brent. 1999. Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Science*, 3:294–301.

E Clark. 1987. The principle of copntrasr: A constraint on language acquisition. In B MacWhinney, editor, *Mechanisms of language aquisition*. Erlbaum, Hillsdale, NJ.

C Fisher *et al.* 1994. When it is better to receive than to give: syntactic and conceptual con-

straints on vocabulary growth. *Lingua*, 92(1-4):333–375, April.

L Gleitman. 1990. The structural sources of verb meaning. *Language Acquisition*, 1:3–55.

R Jackendoff. 1990. *Semantic Structures*. MIT Press, Cambridge, MA.

B MacWhinney, 1995. *The CHILDES project: Tools for analysing talk*. Lawrence Erlbaum Associates, Hillsdale, NJ, second edition.

J Piaget. 1954. *The construction of reality in the child*. Basic Books, New York.

S Pinker. 1987. The bootstrapping problem in language acquisition. In B MacWhinney, editor, *Mechanisms of language aquisition*. Erlbaum, Hillsdale, NJ.

S Pinker. 1989. *Learnability and Cognition*. MIT Press, Cambridge, MA.

S Pinker. 1994. How could a child use syntax to learn verb semantics. *Lingua*, 92(1-4):377–410, April.

J Siskind. 1996. A computational study of cross situational tecniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91, Nov/Oct.

A Villivicencio. 2002. *The acquisition of a unification-based generalised categorial grammar*. Ph.D. thesis, University of Cambridge. Thesis published as Technical Report UCAM-CL-TR-533.