Psychology of Language and Communication 2001, Vol. 5. No. 1

PIOTR KOCHAŃSKI
Polish Academy of Sciences, Warsaw
ANDRZEJ NOWAK
University of Warsaw, Warsaw School for Advanced Social Psychology
PETER CULICOVER
Ohio State University
WOJCIECH BORKOWSKI
University of Warsaw

ON COMPUTATIONAL SIMULATIONS OF THE ACQUISITION OF NATURAL LANGUAGE SYNTAX

In this article we present various distributional approaches to language acquisition. These are based on statistical parsing of sentence corpora and model certain aspects of the process of language learning. Our aim is to give a non-specialist in this field a general idea of when such methods are useful, and what their advantages and limitations are. We show how distributional models have been applied to fundamental linguistic questions about the innate properties of human language. We also discuss possible directions of development of distributional methods, so they might be able to account for difficult linguistic processes such as the acquisition of syntax.

Introduction

Since 1965 – the year when Chomsky's *Aspects of the theory of syntax* was published – linguistic theory has experienced an intriguing and often heated debate about first language acquisition. The most fundamental questions are: how do humans learn language? why is language acquisition so rapid and accurate? Chomsky's answer to these questions was that there must be an innate faculty of the mind with which all humans are endowed that is dedicated to the acquisition of language. It is sufficient for a learner to be exposed to and to interact with a language or languages in the surrounding environment.

On the other hand many linguists and psychologists have claimed that language should be treated on an equal footing with other human cognitive activities, in the sense that there need not be any special capacity for language beyond what is required for other more general aspects of cognition. A question that has been raised along with those above has been: can language acquisition be accounted for in terms of human capacities that are not specific to language, e.g. those that enable us to swim, sing, dance, play a sport or a musical instrument, and so on?

---

[1] Send requests for reprints to the first author: Piotr Kochański, Center for Theoretical Physics of the Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warszawa, Poland.

Chomsky's answer to this question all along has been that no such general theory of learning has even begun to approach a satisfactory account of the linguistic capacity of human beings. Chomsky and others (Baker 1979; Pinker 1994) have presented strong arguments in support of Chomsky's position. The first argument is known as "poverty of the stimulus". Children acquire language so well in a relatively short time through being exposed to linguistic input, which is often of a poor quality and seems to be insufficient to explain how humans know so much about their language. The poverty of the stimulus argument is related to the fact that there appears to be little or no negative evidence in the input to children. Adults speak mostly grammatically correct sentences. Those that contain grammatical errors are not marked as such, and children are not systematically corrected for their own grammatical errors.

The alternative approach has concentrated on showing that the language acquisition mechanism can be subsumed under a general account of human cognition. A typical strategy involves creation of a model to explain certain specific aspects of language acquisition, and testing the model in behavioral experiments or in computer simulations. Serious computational modeling of language acquisition has been carried out only recently because of the emergence of relatively cheap access to powerful computers.

Among the computational approaches represented in the literature are the following: machine learning, connectionist networks, and various statistical methods of retrieving linguistic structure from the input sentences corpora, e.g. hidden Markov models, cluster analysis, discriminant analysis, multidimensional scaling. Following Redington and Chater (1998) we refer to all of these approaches as 'distributional' approaches.

How do distributional approaches cope with the task of revealing the nature of learning language by children? Do they lead to linguistically and psychologically interesting results and conclusions or do they provide a misleading view of how language acquisition takes place? In the present paper we try to partially answer these questions. In order to accomplish this goal we review some key concepts and illustrate them with the appropriate practical applications. Rather than providing a comprehensive review of the literature, we concentrate on those specific proposals that appear to us to be the most realistic as accounts of how children actually acquire language.

At the beginning we note that distributional approaches have been in disfavor for many years. In the article "Finite State Languages" (Chomsky and Miller, 1958) the authors show that any finite state grammar is unable to produce all of the sentences of a natural language. Moreover, a model that generates sentences using the probability of a word based on its position with respect to other words cannot explain any interesting aspect of language, in particular, linguistic structure. The "poverty of stimulus" and "no negative evidence" arguments render distributional approaches still more unattractive, since speakers appear to have knowledge that has a completely null distribution in the learner's experience.

Other arguments against distributional approaches were discussed at length by Pinker (1984) and in Redington and Chater (1998); among them we cite two. Pinker claims that there are properties of the language that must be deduced. As an example he gives the complex noun phrase constraint, i.e. that it is impossible to extract a constituent of a complex NP (such as one that contains a relative clause). He argues that evidence that this is the case does not appear to be deducible on distributional grounds. A second problematic aspect of distributional methods is the possibility of spurious correlation. Given the three sentences: "John eats meat", "John eats slowly" and "Meat is good" the conclusion drawn

might be that "slowly is good" is a proper English sentence. The words *meat* and *slowly* are used here in the same distributional context in two sentences; as a result, it is possible to conclude on distributional grounds that they must have related meanings, or, at least, that it should be possible to use them interchangeably. Crucially, most distributional approaches are based on a principle of this type: if two words are used in similar contexts, they must be in some sense similar.

This criticism was partially addressed by linguists working with computer simulations (see Redington and Chater (1998) for a thorough discussion), for instance, the "no negative evidence" argument is doubtful. It is not true that in order to learn something we need negative evidence. For instance, we do not need negative evidence to learn addition, or to learn that the sun rises in the East (Redington and Chater, 1998).

While it is interesting and useful to consider 'in principle' arguments for or against various classes of approaches, we wish to concentrate here on what kinds of distributional information is actually present in natural language, whether various computational techniques are capable of extracting this information, and whether the success of such computational techniques would be sufficient to account for the acquisition of natural language. We find, in fact, that certain techniques are entirely suitable as engineering approaches to extracting information from corpora, but cannot be seriously entertained as components of an account of human language acquisition.

The importance of the in-depth understanding of the real task of computer simulation has been already underlined in Brent (1997). Following Marr (1982), Brent makes the distinction between so called computational theories and algorithmic theories. According to Marr, computational theories answer the question "what is computed and why" – those are what/why theories. The algorithmic level is concerned with the particular representation of input or output and covers the details of how the computations are done. As an example Marr gives the mechanical cash register. The what/why theory gives abstract and formal definition of addition (i.e. it tells us that this is the function that takes two numbers and maps them to another number in a certain way). This definition is free of the details how the addition is actually done by the register, whether the numbers are represented in the Arabic or Roman system, or if we use decimal or binary base. Depending on the representation that is used, details of the addition algorithm would be different, which is also given by the "how" (the implementation) theory. Note that all implementation problems are immaterial from the point of view of the what/why theory. For instance, the fact that the mechanical register is able to add numbers that are smaller than the register's (finite) range does not invalidate the definition of addition.

Let us consider how the above distinction applies to the formulation of a theory of language acquisition. A typical simulation model tries to test the following hypothesis:

H: Strategy S partially explains children's ability to perform task T.

In the spirit of Marr's approach we can ask how children perform task T, investigating whether or not a given model resembles children's behavior. Alternatively, we can concentrate on the what/why level theory to find out if a given strategy S really facilitates task T.

Working with hypotheses H, we cannot concentrate on proving that in order to perform task T we have to use strategy S, that is, that S is necessary to do T. Typically we must check simply if strategy S is sufficient to perform task T. In other words, we try to prove

that the strategy in question is effective (Brent, 1998). Let us imagine that someone proposed the theory that drinking water leads to understanding language. No doubt it is necessary to drink water to learn language (this is an empirical fact) but we understand that drinking water alone would not help to learn language (it is not a sufficient strategy).

Similarly, we may use a computer simulation to check if a given strategy is effective. In fact, as long as we stick to the what/why level theories at the conceptual level, the computer simulation does not differ from a behavioral experiment. Both are done for the same reason, to test whether a particular approach is sufficient for a particular task.

As noted earlier, it is generally held that theories of language that are based on the assumption that the position of a word in the sentences is given by the probability determined using co-occurrence information are linguistically uninformative. This claim is obviously true if it is interpreted as the claim that children perform complicated calculations simulated on the computer using a particular algorithm, or that they solve some particular arithmetic problem, both of which produce a measure of the probability of the word being in some particular place in the sentence. It is of course tremendously difficult to determine how a child's mind realizes a particular learning strategy for word meanings. A computer program is not the appropriate tool to answer such a question – it is difficult to use the computer to verify the "how" level theory (although we do not claim that it is impossible). Most researchers using computer modeling do not claim that they are investigating the "how" level theory. Typically, one constructs a what/why theory and tries to determine if strategy S in hypotheses H is effective for task T. In this case, the general criticism of distributional approaches does not apply.

To elaborate this point, let us consider a particular example. No reasonable linguist would claim that in order to place a word in the sentence a child calculates mutual entropy using logarithms of base 2 (not 3 or 4). Similarly, nobody would suggest that in order to bring out the similarity of the distribution of the words *cat* and *dog* a child calculates the distance between them using given, e.g. Euclidan metrics to the exclusion of all other possibilities. But there is nothing wrong in claiming that children, on the basis of a set of sentences, are able to infer that two words used several times similarly might be always used similarly, that is, might be in the same category. If we hear that both cats and dogs eat, sleep, can be teased, are smaller than our dad, have fur and so on, it is reasonable to assume both dog and cat are similar creatures; it is reasonable for a learner to make the same assumption.

Crucially, from the point of view of the what/why theory, it is irrelevant what measure we use to calculate the distance between the words *dog* and *cat*. The important thing is that similar use usually reflects the similarity of words. Of course, we would not want to claim that *cat* and *dog* are two names for the same thing, but we can safely put them in the same category and we can use them similarly. This is even clearer in the case of verbs like *run* and *go*.

The important thing about a particular computer implementation of the model is that it has to be robust. If a model gives reasonable results for some particular choice of model parameters and fails completely for another, but a fairly close choice, this means that either the model or the simulation is erroneous.

The criticism most often used against the simulation approach is that the program might work well, not because of the well-chosen what/why strategy but because of some detail that is incidental to the implementation of the theory. But, if we vary model parameters, or

add some noise to the model, we may convince ourselves that the proof of our hypothesis is not an effect of blind luck, but an indication that the chosen theory models the particular behavior well.

Robustness is a very important feature of any simulation in social sciences, not only in linguistics. This is sometimes forgotten, since the first science where simulations were used was physics. In a physical model we don't need robustness, since we know the parameters of the model with great precision: the mass or charge of the electron are well known and it is not necessary to check if our theory gives physically valid results if we vary the electron charge by 10 percent. In the social sciences, on the other hand, we do not have an underlying theory that would enable us to make categorical claims about this or that feature of the model we use. This applies undoubtedly to linguistics as well.

Another common criticism of computer simulations is that they are not realistic. Conditions simulated by the program are far from those in the real world. The point is that the same argument might be used against behavioral experiments. Conditions in the laboratory are different from those in the natural child's environment. As we have seen already, the difference between what/why level theory and the experiment is conceptually unimportant. Any attempt to understand the world is limited by our experimental abilities; this is true in any science.

Obviously we cannot always divide a theory into the pure what/why part and the implementation; they are both dependent and can impose constraints on each other (Brent, 1998).

The great advantage of distributional methods is that various assumptions about language acquisition may be tested with the help of computer simulation. Let us consider again the "poverty of the stimulus" argument. We can record sentences spoken to children, transcribe them, and then build the model, which would test if, for instance, it is possible to extract semantic or syntactic information using certain statistical methods. Regardless of the results of such procedure we are able to learn something, to find out to what extent this argument is justified. Similarly, we can treat any other assumption made about language acquisition; possibly computer simulation is not a fully satisfactory method, but it is objective, repeatable and, especially when it gives a positive result, may lead to strong evidence supporting the model. The problem arises when we do not obtain a positive result, i.e. we arrive at the result that something cannot be done by our method. This might not mean that our hypothesis is wrong, since possibly we simply used the wrong model or a wrong implementation.

What, then, is the conclusion? First of all, we should decide what kind of theory we are constructing. Usually, if we going to use a computer simulation (i.e. distributional methods) we have in mind the what/why level theory. Secondly, we have to remember that we must check the robustness of our model and its particular implementation. We should also be aware of the value of the experimental results. Behavioral experiments often give precious clues concerning both theory and its implementation (the "how" level theory).

One last observation is that the majority of studies in the field of distributional approaches to language concern the English language, the features of which make it relatively easy for statistical analyses of word order distribution. English has an impoverished case system and uses constituent order to denote grammatical function. This is not the case in many languages. Consequently, it is possible to develop specific statistical methods attending to the distribution of words that simulate acquisition of English but that fail when applied to other languages. In a sense, this is again a question of robustness. In

building a simulation model we should always ask whether it would work for more then one language. Only if it is adaptable to a range of different ways of expressing grammatical relations could we realistically claim that our model explains some element of language acquisition independently of the particular language under consideration.

Although criticism towards application of distributional theories is often justified we hope that we have succeeded in arguing for their potential usefulness with respect to questions that are appropriate for computer simulation.

In the next section we describe the range of application of distributional methods. The third and last section of this paper is devoted to a discussion of statistical methods leading to the description of learning words and morphology.

What distributional approaches can and cannot prove

Now we concentrate on the possible questions that distributional methods are meant to answer. It is not surprising that any given method has its limitations. A particular model might deal well with certain problems and at the same time be insufficient to explain other phenomena.

A nice analogy concerning this problem was noted by Redington and Chater (1998). They compare language acquisition models with computer approaches to vision. In the analysis of visual processes they distinguish low-level and high-level information perceived by humans. Segmentation of the image into parts is governed by low-level factors as texture, color and so on. Perception of those factors is usually an easy task for a computer program. On the other hand, the task of image segmentation is also affected by high-level factors, for instance, identity of the observed object. Recall the experiment in which people were presented with a picture of a Dalmatian dog on the spotted background. When observers where informed about the content of the stimuli the recognition of the object was much faster. Inputting the recognition of the high-level is rather difficult, and it is not clear if it is fully possible.

Thinking of language acquisition in a similar way, we may speculate that in learning the meaning of words we acquire low-level information, whereas learning syntax is a matter of high-level structure. As we shall see, this claim conforms to the majority of distributional approaches. It is fairly easy to build a computer program, which, when exposed to sentence corpora, is able to extract semantic information, i.e., to gather words into groups where all of the words have similar meaning. At the same time it seems that familiar distributional approaches (at least in their most common versions) are unable to learn syntax.

We categorize the tasks that have been treated with using distributional approaches into the following main groups:

1. Finding semantic structure (word classes).
2. Finding elements of syntactic structure (e.g. phrase structure).
3. Identification of meaning of words, or, conversely, treatment of meaning as a perceived part of the world and using it as an input for further computer analysis.

Beyond these main research problems there are numerous important sub problems. Before going into detail we take note of the character of the input used in language acquisition simulations. Typically computational linguists use so-called ‘word corpora’ – large files containing sentences. The oldest and very popular corpus is the Brown corpus, created in the 1960s and 1970s. It was created as a representative sample of American English at that time and so it contains legal texts, scientific texts, fiction, and many others. Another widely used

corpus is the Penn Treebank, which contains sentences taken from the *Wall Street Journal*. Moreover, this corpus is tagged: words are categorized into syntactic categories. There are also corpora of sentences from languages other than English, bilingual corpora, and so on.

Most of those linguistic data sets are made up of written sentences. This is a crucial point and must be kept in mind, especially if we are interested in language acquisition, because children are exposed only to spoken language, not written language. There does exist a corpus of spoken language consisting of sentences that are transcribed from recorded natural interactions between adults and children. This is the CHILDES database (MacWhinney and Leinbach, 1991).

Apart from natural language corpora linguists also use sentences created artificially, usually with the help of some artificial simplified syntax. Such corpora have their advantages for testing a particular model or presenting some of its features, since if a carefully constructed corpus presents a fundamental difficulty for a given approach, corpora of naturally occurring sentences will be even more problematic. We believe that no conclusion concerning the human capacity to learn language may be drawn using distinctly artificial sentences. Some approaches work well for artificial sentences but fail completely when confronted with real data. One reason for this is that in creating artificial corpora it is typical to take a group of tags (A, B, C) and call them 'nouns', to take another group of tags (X,Y,Z) and call them 'verbs'. Then a number of rules are introduced to generate sentences. These rules randomly choose a verb or a noun and use them properly. A sentence generated using this method contains only syntactic information, but there is no semantic information present (A and B are treated as nouns and are not differentiated further). This is just the opposite of natural language corpora, where semantic information overwhelms syntax, in the sense that the most robust distributional regularities are determined by the semantic properties of the words, not their syntactic properties.

The situation just described is relevant to some extent in the case of written language corpora. As we have observed, such sentences are not what children hear when they are learning language. The most valuable corpora from the perspective of language acquisition are those of transcribed speech, such as the CHILDES database, mentioned above. The spoken language and written language are in some important aspects very different. Written language is invariably grammatical, almost entirely declarative, and almost all sentences are complete. In contrast, spoken language corpora contain many phrases that are not grammatically correct sentences, and there is an abundance of questions, exclamations, and imperatives.

Another rarely mentioned but important problem is that most corpora have built in hidden assumptions. We have mentioned the first one already: children listen to spoken language, not to text taken from a newspaper, etc. (obviously the CHILDES database takes care of this problem). A second hidden assumption is reflected in the fact that in almost all corpora the words are separated by spaces. In natural spoken language there are no distinctly audible boundaries between words. There are cues such as prosody that may help learners (and hearers in general) distinguish words in some languages, but there are many languages in which the lexical stress is not in fixed position, e.g., English and Russian.

As a result, even before we try to investigate problem 1 above (i.e. finding word classes), we have to solve a problem of segmentation. We need to understand the way that children segment strings of sounds into complete words. This is not trivial, since, as we have already noted, in real speech there are no audible gaps between words.

One further problem concerns the development of inflectional morphology. In English this is a relatively minor problem. Besides the added past tense ending /-*ed*/, and the third person singular present tense /-*s*/, overt inflectional morphology is almost absent from English. In contrast, in languages like Russian and Polish, morphology is a key part of the grammar – syntactic relations are encoded by inflectional morphology affixed to verbs, nouns, adjectives, and so on. Descriptive linguistics provides evidence of many such devices in the languages of the world, including suffixes, prefixes, infixes, ablauts and umlauts, tonal morphemes and truncations.

Segmentation and morphology. Speech recognition[2]

A. Segmentation

The way in which children solve the word segmentation problem remains unclear. There are various hypotheses which suggest possible sources of information that children could in principle use to recognize complete words out of the stream of spoken sounds.

The simplest hypothesis, but rather doubtful, is that children learn words in isolation and later they identify them in fluent speech (Suomi, 1993). There is no evidence in the CHILDES database to suggest that children are systematically exposed to tutorial demonstrations of individual words and their meanings. Rather, the evidence suggests they listen to speakers uttering expressions in which the words to be learned are embedded in a variety of longer expressions. In fact, even if children were truly exposed to separate words in context, it is not clear how they could figure out that they were hearing single words, e.g, that *butter* is not made of two words "but" (which a child could heard in another occasion) and "ter".

Other authors have proposed various solutions to this problem; they have claimed that there are subtle acoustic and phonetic markers of a word's beginning and end (Lehiste, 1971). There is some evidence that intonation, stress, melody and rhythm can help in identification of words (Slobin, 1973; Peters 1985).

Another possibility is that prosodic clues like stress or vowel lengthening might enable the learner to mark the word boundaries (see Cutler and Norris, 1988; Gleitman, Gleitman, Landau and Wanner, 1988). Saffran, Aslin and Newport (1996) found that parents often pronounce topic words in a sentence in such a way that they are highlighted.

We turn next to a discussion of distributional approaches.

The Metrical Segmentation Strategy (MSS) is a simple strategy for segmentation proposed by Cutler (1993) and Cutler and Norris (1988). MSS assumes that prosodic marking correlates with the beginning of a new word. In English, for instance, strong syllables appear at the beginning of a word; in other languages, prosodic cues may be different (Cutler and Butterfield, 1992; Otake, Hatano, Cutler and Mehler, 1993). There is even experimental evidence that infants are sensitive to prosodic clues, strong and weak vowels etc. (Jusczyk, Cutler and Redanz, 1993). A distributional analysis of the stress pattern of English agrees with the assumptions of MSS (Cutler and Carter, 1987); the problem is that such regularities haven't been found in other languages.

The most widespread approach was proposed by Wolff (1975, 1977, 1988). Wolff started with a dictionary of atomic symbols (letters or phonemes). Then, using appropriate word

---

[2] What follows is based in large part on Redington and Chater (1998) and Brent (1997).

corpora, he calculated the co-occurrence pattern of those symbols and added the most frequent combinations back to the dictionary as new symbols. This procedure was applied iteratively, leading to the grouping of longer symbols. This method has various limitations; first of all, it works best for small corpora (thousands of words). In addition, it predicts that shorter words (*as*, *of*, etc.) are learned earlier, which is demonstrably not the case. On the other, hand it has been proposed that children do know short words very early, although they do not use them (Gerken, Landau and Remez, 1990).

A new and very interesting approach was presented by Brent and Cartwright (1997). Their Distributional Regularity (DR) strategy makes use of the following preferences:

a. The number of novel words should be minimized
b. The sum of the lengths of the novel words should be maximized
c. The product of the relative frequencies of all the words should be also maximized

The implementation of an algorithm based on the above postulates performs quite well. For instance, their algorithm, at a particular stage, manages to recover the proper segmentation from an input string "apet adog": i.e., *a pet a dog*. Apart from this strategy these authors propose a phonotactic constraints strategy, which further limits the number of possible words that can be built out of phonemes. The axioms obeyed during segmentation are:

a. Vowel constraint: every word must contain a vowel
b. Boundary cluster constraint: each language has its own finite set of consonant clusters that can occur at the beginning of the word. As an example we can take cluster "gd" which can never occur as a word beginning. Clearly, each language has a different set of such clusters; in Polish "gd" may appear at the beginning of a word.
c. Internal cluster constraint: each language specifies a finite set of consonant sequences that can appear between two vowels within a word.

B. Morphology

Here the problem is somewhat similar to the segmentation problem. We concentrate on identification of various morphological elements that might carry both semantic and syntactic information. In English the former is much more important, but we need to keep in mind other languages, where morphology encodes syntactic relations. In English we would like to extract "semantic" (i.e. derivational) affixes (e.g. /-*ly*/, /-*ness*/, /-*tion*/) as well as morphosyntactic affixes, such as /-*ed*/ or /-*ing*/.

The distributional methods that have been used to deal with the segmentation problem have been used almost unchanged for morphology. Wolff's (1975, 1977, 1988) model can be used in the form presented above. Brent (1993) developed an approach based on the minimum description length (MDL) principle (Rissanen, 1989). MDL tells how to compare various arrangements of the elements of data (here the elements are phonemes and arrangements are words or morphemes). The "goodness" of a particular arrangement is given by the length of description necessary to describe a given arrangement. This description is twofold: first, of postulated structure (the postulated morphemes) and, second, of data in terms of that structure. MDL gives a trade-off between the two parts of an arrangement, preferring the description that is minimal, i.e. we want to find the optimal arrangement of letters (phonemes) that describes the initial word corpora (the data). Such a description captures regularities in the data and uses them to encode them effectively. This is in a sense similar to well-known file compression utilities. It is worth noting that Brent's algorithm works very well, at least on the written (*The Wall*

Street Journal) English text. About 85%-90% of identified morphemes were linguistically relevant.

There are numerous other approaches that differ from those already described but only in details. We return to this subject when we discuss a class of connectionist models.

C. Speech recognition. Markov Models

Speech recognition, although not directly related to language acquisition, is also worth mentioning. Working on this topic, scientists have developed a very powerful simulation method that has also been used to model language acquisition, namely Markov Models. First, we introduce an abstract definition of Markov models and then we illustrate it with an example of its application in linguistics. Our derivation follows Manning and Schütze (1999).

Let us imagine that we have a sequence of random variables $X = (X1, ..., XT)$, and each of them can take some value out of set $S=\{s1, ... , sN\}$. We call sequence X a Markov process (or chain) if

$$P(Xt+1 = sk \mid X1, ... Xt) = P(Xt+1 = sk \mid Xt).\ P(X = a \mid Y = b)$$

is a conditional probability, i.e. the probability that X has a value a if Y=b. Putting it simply: the value of random variable X at time t depends only on the value of random variable at time t-1. There are obviously extensions, where X at time t depends on X at time t-1 and t-2 and so on up to t-K (where K is finite). In linguistics the Markov model with some K is often called the K-gram model, where K is the length of the "history horizon".

If the Markov process has a property that $P(Xt+1 = sk \mid Xt) = P(X2=sk \mid X1)$ then we call this Markov process 'stationary.'

Now, let us suppose that sequence $X = (X1, ... XT)$ is a sequence of phonemes creating a word X or a sequence of words creating a sentence or a sequence of sounds produced by a speaker. We can ask what is the probability of observing a given phoneme/word/sound sequence: $P(X1, X2, ... XT)$. If we can teach our system how such a probability function should look, then, if we observe a new sequence (even one never seen before) we are able to say if it is correct. We know (with a given probability) if a sequence of phonemes might be a word, if a sequence of words creates a correct sentence, or if a string of sounds might be translated as a part of the word or sentence.

The central problem is how to get (an efficient!) probability function P. We need a "training set" of words, sentences sounds, and now Markov process theory comes into play. Its properties give us the formula for P

$$P(X1, ... XT) = P(X1)P(X2 \mid X1)P(X3 \mid X2)...P(XT \mid XT\text{-}1)$$

We see that it is enough to know the conditional probabilities of pairs of words in order to calculate P. Conditional probabilities of pairs are easily calculated if we have a "training" corpus. Even if the corpus does not contain a given word/sentence/sound string we are able to predict if it is correct. In case of speech recognition we use P to map sound strings to written language.

The model described above is known as a Visible Markov Model (VMM), because we calculate P by "seeing" the sequence of Xt. VMM appears to be not general enough to give reasonable function P in more complicated cases.

In contrast, in a Hidden Markov Model we do not assume that we know the exact path leading from word to word; we know only the probability of one of those paths. HMM models are obviously much more flexible. Let consider the K-gram model (i.e. a Markov Model with a "history horizon" of length K). It turns out that a VMM model with fixed K performs rather poorly. We can use instead a HMM model where double conditional probability functions are given by

$$P(Xt \mid Xt\text{-}1, Xt\text{-}2) = a1\ P1(Xt) + a2\ P2(Xt \mid Xt\text{-}1) + a3\ P3(Xt \mid Xt\text{-}1, Xt\text{-}2)$$

We do not know the path from, for instance Xt-1 to Xt; the possible path is hidden in the parameters of the model a1, a2 and a3, which must be adjusted. Another great advantage of HMM models is that there exists an efficient method of training them, known as the Expectation/Maximization algorithm (Manning and Scheutze (1999) and Jelinek (1997)).

The bulk of applications of HMM lies within the domain of speech recognition, but Harris (1955), and Saffran, Newport and Aslin (1996) and Newport (1996) suggest that both infants and adults may use transitional probabilities to identify words. Transitional probability is nothing other than the conditional probability of two words or phonemes being adjacent in a sequence.

Distinguishing word classes

The problem of acquisition of word classes is a central subject that has been treated using distributional methods. This is because there is a wide agreement that, even though language is an innate human ability, the specific vocabulary must be learned after the child's birth. Many questions arise regarding what strategies are used by children in learning words, as well as what information they use to learn the vocabulary of a particular language.

In fact, there are two related problems: discovering that there are different classes of words and discovering which word is a member of which word category. As far as the first problem is concerned, it is often proposed that the child has innate pre-existing word categories (Pinker, 1984). We believe that children can discover the existence of various categories on the basis of observation of the world (i.e. the ability to do that is just one more cognitive competence). Children observe that in the world there are objects that have names (we call them nouns). In addition, we can manipulate those objects. Different ways of using and operating also have names – we call them verbs. Although this explanation seems plausible, even if it is correct it does not address the second question, that of determining which word belongs to which category. (It is of course true that there are nouns and verbs that do not have direct correlates in the world.)

This second problem is more interesting and no unique solution exists. All approaches have two main components. First, each assumes a method of specifying the similarity of words, i.e. it is necessary to say words should land in the same category. Secondly, each has a way of visualizing the result, so that once the corpus is recognized as having words that are similar, the words can be grouped and the groups can be listed.

In order to find the similarity between words using distributional methods, researchers have used various sentence corpora and have tried to calculate the difference between words on the basis of co-occurrence statistics. One counts how many times two words

happen to be adjacent; in some approaches more distant neighbors of a word are taken into account. Then one obtains, for a given word, the co-occurrence pattern of this word with respect to all other words in the corpus. In practice researchers concentrate only on those words that are most often used in the corpus. Then one compares those co-occurrence patterns and calculates the difference between words. The co-occurrence patterns of word A and B are represented as vectors Ai and Bi. For instance A = (2,5,3,...) means that word A follows the first word in the corpus 2 times, the second word 5 times and so on. In fact, for every word two such patterns are necessary: one for words that follow a given word and one for those that precede it in the corpus. Next, one can calculate the distance between words as a function of those patterns. For instance,

d2AB=?(Ai-Bi)2

is a Euclidean distance between two patterns (? means the summation over i from 1 to the number of words used). This approach has been applied in various versions in many studies; see e.g. Brill, Magerman, Marcus, and Santorini (1990), Finch and Chater (1991, 1992, 1993), Grünwald (1996), Hughes and Atwell (1994), Kiss (1973), Marcus (1993), Rosenfeld, Huang and Schneider (1969), Schütze (1993).

Another way of finding the similarity of words is to use a connectionist network. This approach was introduced by Elman (1990). Predictions of this model have a further range than just the identification of word categories (Elman, 1993, 1998). Elman's approach uses artificial neural networks. Before we continue, let us recall a few of the most important features of neural networks.

In the simplest case a neural network is composed of two layers: the input layer and the output layer. In each layer are placed "neurons", i.e. objects (physical or mathematical) that can be in two or more states (a neuron can be active, hence it has a value +1 or its voltage can be 5V, or it can be inactive: it has a value 0 or voltage 0V in some electronic realization of the neural network). Neurons in input and output layers might be connected; the strength of this connection tells if an activated input neuron can activate (or deactivate) an output neuron. Before using a neural network we "teach" it by assigning values to connections between neurons.

A simple example is a neural network that recognizes animals. Each input neuron tells us if the presented animal has or does not have a given feature: fur, tail, horns, is small, big, etc. The output neurons tell us what animal is at the input. For instance, let us imagine that at the input neurons "small", "furry", "tail", "eats meat" are active; then a properly trained neural network would give us on the output layer that the animal is either a cat or a dog, not a cow.

Elman used a more complicated network, which also had a "hidden" layer (between the input and output layers). This network was taught as follows: as an input and output it has words, and the activating connection was created if two words appeared in the sentence one after the other. Then the output value, which appeared in the output layer, was "back propagated" to the input, so the network could deal with arbitrarily long sentences. The next step was to translate the structure of the connections in the network into distances between words; these distances were used to calculate word similarity measures.

Numerous other authors have used related methods: Scholtes (1991a, 1991b), Finch, S.P., Chater, N., & Redington, M. (1995).

As we mentioned in the previous section, we may also use Hidden Markov Models to search for word categories.

Up till now, we have concentrated on the problem of finding distances between words. Now we need a method to translate distances between words into groups of words. This is a standard task and there are well-developed statistical methods of doing this, namely, clustering algorithms. Clustering can be hierarchical and non-hierarchical. In hierarchical clustering, words are first put into one large cluster, then they are divided into smaller subgroups, and so on. In this way we obtain a hierarchy of clusters. In the non-hierarchical method, words are simply put into several clusters, the number of which must be initially specified. For detailed discussion, see Manning and Schütze (1999).

Those methods work surprisingly well, at least for English language sentences (Finch and Chater, 1991, 1992, 1994; Finch, Cater and Redington, 1995). English provides ideal material for such a method, because it is a fixed word order language, and word order in such a language carries all of the information about the semantic and syntactic structure of the language. It appears that these distributional methods lead to the semantic categorization of words, rather than syntactic. This fact is important to remember, because even with a semantically determined grouping of words we are still far from understanding language acquisition, given that syntactic knowledge must also be acquired.

Let us consider whether it is possible to extract some information about the syntactic structure in spite of the strong semantic bias. One method that has been applied often is generalization or 'bootstrapping' (Grimshaw, 1981; Pinker, 1984; Schlesinger, 1981, 1988), and more specifically, 'semantic bootstrapping'. Typically, we start with a small number of words that we consider to be basic, in the sense that children learn them early, and also that they clearly belong to the basic syntactic categories (verbs, nouns, etc.). It is assumed that children already have this knowledge. During the learning process the system uses this prior knowledge to develop the basic syntactic categories. Another method tries to connect various semantic clusters (for example, those that contain nouns) into one syntactically meaningful cluster. To some extent we can treat this approach as a model of one aspect of syntax acquisition.

The bootstrapping method might be justified by one of Elman's results (Elman, 1993), where he has shown that for an artificial neural network it is necessary to start teaching with simple sentences, otherwise the network will not succeed in finding proper categories. Bootstrapping also solves one other difficulty. By using distributional methods we tacitly assume that children analyze sentences spoken to them. But how can they analyze new utterances without already knowing something about the language, and in particular the grammar? Bootstrapping's answer is that some partial knowledge about language is acquired from extra-linguistic sources, such as the observation of parents while they are doing something (feeding a child, playing) and talking about their activity.

Finding syntactic structure

Finally, we have come to the most important and difficult problem for distributional methods: finding syntactic structure. In the previous chapter we suggested that we might try to group words into syntactic categories. This would be an achievement, but even then we are still far from discovering the syntactic structure.

The goal of computational approaches to language acquisition is to find phrase structure in the sentence corpora. In most research this goal has been sharpened to focus on the discovery of the basic English sentence structure Subject-Verb-Object. There is a hope in this research that properly used semantic information might give sufficient clues concerning superficial phrase structure. The problem is that we have to use distributional methods in such a way that the results are not biased by purely semantic information. Moreover, it is important that the approach be neutral across different languages if it is to be relevant to the acquisition of language by humans.

As we have already noted, the solution lies in the use of generalization or bootstrapping. Both methods lead to the discovery of syntactic categories of words. Bootstrapping assumes that a child understands that there are nouns or verbs, etc. and is able to put all words into one of those groups. Generalization does not assume any prior knowledge, but simply tries to group together categories that contain the same parts of speech. Redington and Chater (1998), using methods developed in work by Finch and Chater (1991,1992), obtained interesting results; they were able to extract noun phrases, verb phrases and prepositional phrases from sentences . This method is far from ideal since it captures not only the phrase structure of the sentence, but also proto-sentences, which are irrelevant from the point of view of syntax. Moreover, it is difficult to define a good measure of success of their method.

A very interesting idea concerning the extraction of phrase structure was presented by Yuret (1998). He combines the well-known distributional method, which evaluates the likelihood that two words are adjacent using an entropy-like measure, with a number of additional, non-statistical, postulates. Yuret's postulates are concerned with the nature of possible relationships among words in a string.

In order to find phrases Yuret joins all the words in the sentences with links. The strength of each link is given by the entropy-like distance between words. If the value of this measure is large it means that the words are often adjacent, and such a link is called 'strong.' But not all links are allowed. The forbidden links are those that violate the following rules:

a. Cycles are forbidden, i.e. if we have a sentence with three words (w1, w2, w3) and there is a link between w1 and w2, w2 and w3, w1 and w3 then we have a cycle and the weakest link must be removed.
b. Crossings of links are forbidden. If we have a four word sentence (w1, w2, w3, w4) where w1 and w3 are linked and w2 and w4 are linked, then we have a link crossing and the weaker of them must be removed.

All remaining links should join words, which create phrases.

Those postulates are linguistically motivated, at least in the case of English. Yuret's method gives very good results (although there is no easy measure of quality – Yuret evaluated whether or not the links obtained reveal phrase structure by looking at the number of sentences and phrases found by his algorithm that are syntactically meaningful in the presented examples.) The problem with this approach is that, along with the distributional analysis, the algorithm incorporates some additional linguistic knowledge which is specific to English. Clearly, if we put into the algorithm the full grammar of the language we surely can get it back. In the current case we have introduced a bit of information in the form of postulates a and b. The question is, whether or not this information is not too much for this method to be still interesting.

Meaning: do we need it for language acquisition?

A central tenet of modern theorizing about natural language syntax is that the generalizations about syntactic structure can be formulated purely in terms of syntactic primitives, without reference to meaning. Of course, the main function of language, however, is to encode thought and communicate meaning. Even if meaning is not necessary for the description of language competence in adults, it may be a necessary component of the environment for language acquisition (cf. Culicover 1999). In this approach it is assumed that linguistic utterances are not presented alone, rather they are being accompanied by some perception of the context in which they are produced. The perceptual component may be represented in models of learning as a meaning of the perceived scene expressed in any convention that can encode meaning (e.g. Jackendoff 1990, Kintsch 1974). The task of the learner is to find correspondences between the spoken language and the meaning of the observed scene. In this approach we are interested not only in whether the learner can figure out the rules of syntax, but also how meaning is extracted on the basis of spoken language, and how a grammatical sentence is produced that expresses a desired meaning.

One task for a learner is to find out the correspondence between spoken words and concepts corresponding to perceived objects. This task is not easy, since in child directed speech words usually occur in multiple word utterances paired with multiple possible perceptual interpretations. There are not enough examples of single words accompanying perceived objects to figure out the meaning of the words by comparing a single word with a single perceived object. This task may be solved, however, by finding covariance between specific heard words and perceived objects (for an algorithm see Siskind 1997). As Siskind points out, the task becomes easier as the learner knows more words as a result of a bootstrapping mechanism. The meaning of words already known eliminates some possibilities of the meaning of the novel word and suggests some other meanings.

Kirby (1998, 1999) has demonstrated how syntax can emerge through evolution in a population of learners that try to communicate meaning. In his model each learner tries to communicate meaning by a sequence of (initially random) sounds. Each sequence is assumed to correspond uniquely to the meaning with which it was paired. The initial phase of the population can be described as a proto-language, with no syntax. The sequence of unanalyzed sounds JOHNLIKESMARY acts as a single word that refers to the meaning "John likes Mary". The critical step in the acquisition of syntax occurs when a part of the spoken sequence is substituted by a variable. This occurs when a difference in sounds of two sentences is accompanied by a single difference in meaning. For example when sentences: JOHNLIKESMARY and PAULLIKESMARY are accompanied by meaning "John likes Mary" and "Paul likes Mary", the learning system will represent it as X LIKESMARY, where X may take a value of either John or Paul. As a result of this rule the system will create a category containing both John and Paul and X will represent this category. With further generalizations this rule may be used to express the fact that anyone likes Mary. Another rule states that if two categories are related in the same way to other categories, the two will be merged. Introduction of every new rule of syntax results in the speaker's ability to express meaning in a more general way.

Kirby (1998, 1999) assumes a replicator dynamics. Each of the correspondence rules may be transmitted with equal probability during a single interaction, but since the more general rules will be heard much more often than the idiosyncratic ones, they have a much

higher probability of being transmitted, and thus with time they start to dominate in the population. In fact, the probability of transmission of a given rule in a population is directly proportional to its generality. As a result, grammatical rules take the form of a sequence X,Y, Z, where the variables can range over syntactic categories in a language. In Kirby's approach language acts as a self-organizing adaptive system, in which a learner can make a generalization that the previous population has not made. Since generalization has better chances of survival than an idiosyncratic rule, the whole system moves toward syntactic language, and away from idiosyncratic proto-language.

Most of the problems with the distributional approach to acquisition of grammar might be solved if one assumed the existence of a bootstrapping mechanism. For children, the meaning may be necessary for learning syntax only in the early stages of the acquisition of syntax, where it provides a bootstrapping device. When a child has learned the basic syntactic categories and the basic rules of grammar, distributional information may suffice for categorizing new words on the basis of similarities to ones that are already known. The solution to the problem of the role of distributional analysis may be that the distributional cues are used in conjunction with the cues coming from perception and concerning meaning. The role of the distributional cues as compared with perceptual cues would likely grow with the knowledge of the language.

Conclusions

The classical dilemma of the theory of language acquisition is whether the syntax in its detailed organization is innate or learned. In the view of the above analysis the proper question may be of a more gradual nature: how much of prior (innate) mechanisms do we have to assume before syntax may be learned, and to what degree may syntactic rules be learned on the basis of exposure to language. It is clear, that a lot must be assumed about the linguistic faculty before it can perform all of the distributional analysis, perform generalizations, figure out the rules of syntax, and find out correspondence rules. On the other hand there is clearly a lot of information in purely distributional analysis, and the learner may be able to learn a lot on the basis of distribution alone. However, we suspect that distribution in itself does not provide sufficient information to learn the syntax of a natural language, especially when the syntactic richness and complexity of natural languages is taken fully into account. On the other hand, the task may be feasible if the system has a bootstrapping mechanism. We suggest that the meaning, which is accessed by the learner through perception of the context of spoken sentences, may provide the basis for such a mechanism.

References

Brent, R.B. (1997). *Computational approaches to language acquisition*. Elsevier Science Publisher/ MIT Press.

Brent, M.R., & Cartwright, T.A. (1997). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 63, 121-170.

Brill, E., Magerman, D., Marcus, M., & Santorini, B. (1990). Deducing linguistic structure from the statistics of large corpora. *DARPA Speech and Natural Language Workshop*. Hidden Valley, PA: Morgan Kaufmann.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Culicover, P. W. (1999). *Syntactic nuts*. Oxford: Oxford University Press.
Cutler, A. (1993). Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research*, 22, 109-131.
Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation-evidence from juncture misperception. *Journal of Memory and Language*, 31, 218-236.
Cutler, A., & Carter, D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133-142.
Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113-121.
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., & McClelland, J.L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27, 143-165.
Finch, S.P., & Chater, N. (1991). A hybrid approach to the automatic learning of linguistic categories. *Artificial Intelligence and Simulated Behaviour Quarterly*, 78, 16-24.
Finch, S.P., & Chater, N. (1992). Bootstrapping syntactic categories. *Proceedings of the 14th annual conference of the Cognitive Science Society of America* (pp. 820-825). Bloomington, IN: Cognitive Science Society.
Finch, S.P., & Chater, N. (1993). Learning syntactic categories: A statistical approach. In M. Oaksford & G.D.A. Brown (Eds), *Neurodynamics and psychology*. London: Academic Press.
Finch, S.P., & Chater, N. (1994). Distributional bootstrapping: From word class to proto-sentence. In A. Ram & K. Eiselt (Eds), *Proceedings of the 16th annual meeting of the Cognitive Science Society* (pp. 301-306). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Finch, S.P., Chater, N., & Redington, M. (1995). Acquiring syntactic information from distributional statistics. In J. Levy, D. Bairaktaris, J.A. Bullinaria, & P. Cairns (Eds), *Connectionist models of memory and language* (pp. 229-242). London: UCL Press.
Gerken, L.A., Landau, B., & Remez, R. (1990). Function morphemes in young children's speech perception and production. *Developmental Psychology*, 26, 204-216.
Gleitman, L.R., Gleitman, H., Landau, B., & Wanner, E. (1988). Where learning begins: Initial representations for language learning. In F.J. Newmeyer (Ed.), *Linguistics: The Cambridge surv*ey, Vol. 3 (pp. 150-193). Cambridge, UK: Cambridge University Press.
Grimshaw, J. (1981). Form, function, and the language acquisition device. In C.L. Baker & J.McCarthy (Eds), *The logical problem of language acquisition*. Cambridge, MA: MIT Press.
Grünwald, P. (1996). A minimum description length approach to grammar inference. In S.Wermter, E. Riloff, & G. Scheler (Eds), *Symbolic, connectionist, and statistical approachesto learning for natural language processing*. Springer Lecture Notes in Artificial Intelligence1040 (pp. 203-216). Berlin, Germany: Springer-Verlag.
Hughes, J., & Atwell, E. (1994). the automated evaluation of inferred word classifications. InT. Cohn (Ed.), *Proceedings of the European conference on Artificial Intelligence (ECAI)* (pp. 535-539). Chichester, UK: John Wiley.

Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT Press.
Jusczyk, P.W., Cutler, A., & Redanz, N.J. (1993). Infants' preference for the predominantstress patterns of English words. *Child Development*, 64, 657-687.
Kiss, G.R. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation*, 7, 1-41.
Lehiste, I. (1971). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, 51, 2018-2024.
MacWhinney, B and Leinbach, J. (1991). The Child Language Data Exchange System. *Journal of Child Language*, 22, 271-296.
Manning, C. D. and Schütze, H (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
Marcus, G.F. (1993). Negative evidence in language acquisition. *Cognition*, 46, 53-85.
Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, 32, 258-278.
Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.
Peters, A.M. (1985). Language segmentation: Operating principles for the perception andanalysis of language. In D.I. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol.2. Theoretical issues* (pp. 1029-1067). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Radford, A. (1988). *Transformational grammar* (2nd edn.). Cambridge, UK: Cambridge University Press.
Redington, M. & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: A distributional perspective. *Language and Cognitive Process*, 13 (2/3), 129-191.
Rissanen, J. (1989). *Stochastic complexity and statistical enquiry*. Singapore: World Scientific Publishers.
Rosenfeld, A., Huang, H.K., & Schneider, V.B. (1969). An application of cluster detection to text and picture processing. *IEEE Transactions on Information theory*, 15, 672-681.
Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical cues in language acquisition: Word segmentation by infants. In G.W. Cottrell (Ed.), *Proceedings of the 18th annual conference of the Cognitive Science Society* (pp. 376-380). Mahwah, NJ: Lawrence Erlbaum Associates Inc.
Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35, 606-621.
Searle, John, R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.
Schlesinger, I.M. (1981). Semantic assimilation in the acquisition of relational categories. In W. Deutsch (Ed.), *The child's construction of language*. New York: Academic Press.
Schlesinger, I.M. (1988). The origin of relational categories. In Y. Levy, I.M. Schlesinger, & M.D.S. Braine (Eds), *Categories and processes in language acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Scholtes, J.C. (1991a). Kohonen's self-organising map applied towards natural language processing. *Proceedings of the CUNY conference on Human Sentence Processing*.
Scholtes, J.C. (1991b). Using extended feature maps in a language acquisition model. *Proceedings of the 2nd Australian conference on Neural Networks*.

Schütze, H. (1993). Word space. In S.J. Hanson, J.D. Cowan, & C.L. Giles (Eds), *Advances in neural information processing systems, vol. 5*. San Mateo, CA: Morgan Kaufmann.

Slobin, D.I. (1973). Cognitive prerequisites for the development of grammar. In C.A. Ferguson & D.I. Slobin (Eds), *Studies of language development*. New York: Holt, Rinehart & Winston.

Suomi, K. (1993). An outline of a developmental model of adult phonological organization and behavior. *Journal of Phonetics*, 21, 29-60.

Wolff, J.G. (1975). An algorithm for the segmentation of an artificial language analogue. *British Journal of Psychology*, 66, 79-90.

Wolff, J.G. (1977). The discovery of segmentation in natural language. *British Journal of Psychology*, 68, 97-106.

Wolff, J.G. (1988). Learning syntax and meanings through optimization and distributional analysis. In Y. Levy, I.M. Schlesinger, & M.D.S. Braine (Eds), *Categories and processes in language acquisition* (pp. 179-215). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Yuret, D. (1998). *Discovery of linguistic relations using attraction*. MIT PhD Thesis. Available on-line: xxx.lanl.gov/cmp-lg/9805009.