

# Statistical Models of Language Learning and Use

Mark Johnson and Stefan Riezler

DRAFT of 7th March 2001

## 1 Introduction

This paper summarizes our recent work in developing statistical models of language which are compatible with the kinds of linguistic structures posited by current linguistic theories. Unlike most work in statistical computational linguistics, we are interested in models which are capable of capturing the kinds of non-local context-sensitive dependencies captured by modern theories of syntax. In a series of papers we have developed tools for estimating or “learning” such models from data (Johnson et al., 1999; Johnson and Riezler, 2000; Riezler et al., 2000) and this paper provides a high-level overview of both the general approach and the methods we developed; however, the reader should refer to the original papers for full details and experimental results.

The goal of this work is to understand how human language can be learnt and used in comprehension and generation. These investigations are couched in a statistical framework because we think that much of language learning and use involves weighing weak, unreliable cues. Humans seem exquisitely sensitive to frequency information. In psycholinguistics, lexical frequency effects are often stronger than purely non-lexical syntactic phenomena, and researchers in the field of language acquisition note that children can detect prosodic cues that are often (but not always) correlated with syntactic structure. In language comprehension, there are prosodic and lexical cues which are weakly correlated with syntactic structure or semantic interpretation, and it seems reasonable to expect human language users to exploit these regularities. Turning to theoretical results on learning, it seems that statistical learners may be more powerful than non-statistical learners. For example, while Gold’s famous results showed that neither finite state nor context free languages can be learnt from positive examples alone (Gold, 1967), it turns out that *probabilistic* context free languages can be learnt from positive examples alone (Horning, 1969). Informally, a class of languages may be statistically learnable even though its categorical counterpart is not because the statistical learning framework makes stronger assumptions about the training data (i.e., it is distributed according to some probabilistic grammar from the class) and accepts a weaker criterion for successful learning (convergence in probability).

Our statistical approach requires that the space of possible human languages be described in terms of a finite dimensional parameter space. In this respect

our approach is similar to the “Principles and Parameters” approach (Chomsky, 1981), except that our parameters are real-valued rather than binary-valued. Clearly, the requirement that all human languages can be described in such terms implies that there are innate constraints on possible human languages. In principle our statistical learning methods can be applied to highly constrained models with a small number of parameters (which would presumably have to be very abstract) as well as to very general models with a very large number of parameters. In this sense our approach is agnostic concerning the nature of the innate constraints. As we explain below, we believe that statistical approaches to language learning provide a general theoretical framework in which diverse models of learning can be investigated empirically, which we expect will be far more interesting than philosophical speculation.

Our adoption of a statistical approach does not mean that we claim that humans are actually performing numerical calculations during language learning or language use. Indeed, in our experiments we are often comparing vastly different probabilities that vary by several orders of magnitude, so high precision numerical calculation might not be necessary. It may turn out that one can obtain similar results by classifying events along a finite ordinal scale, say, as “likely”, “unlikely” or “impossible”. Even if this can be done, however, we claim that statistics still plays a foundational role in the study of language learning and use, since statistics provides the basic mathematical theory of inference from uncertain data. As such, it provides the theory of optimal learners and optimal comprehenders (optimal in an information-theoretic sense) which serve as idealizations of, and upper bounds to, human performance. If an optimal statistical learner fails to learn a language given certain kinds of inputs (say, phonological forms alone) under certain assumptions about universal grammar, then we can be fairly certain that human beings either have access to richer data or have stronger biases that restrict the class of possible grammars.

Even though our emphasis in this paper is scientific, we would like to point out that this research has important technological applications. Unlike many of the models used in current statistical computational linguistics, the kinds of grammars we are working with, Lexical Functional Grammars, provide detailed syntactic and semantic descriptions of the sentences of the language. This kind of detailed description is essential if computers are to perform tasks that require a deeper and more detailed “understanding” of natural language than the current superficial models permit.

An immediate goal of this research is to find a way of defining probability distributions over linguistically realistic structures in a way that permits us to define language learning and language comprehension as statistical problems, and the rest of this paper concentrates on these questions. The next section describes the linguistic theory, Lexical-Functional Grammar, which defines the linguistic structures used in this research, and the following section explains how we define a probability distribution over these structures. Section 4 describes how one can learn the parameters that define probability distributions over these structures in principle, and points out some of the practical problems that make straight-forward ways of estimating these distributions infeasible. This leads us

to the “pseudo-likelihood” estimation methods described in section 5, which also raise interesting questions concerning the nature of the data available to the child and modularity of language learning and processing.

## 2 Lexical-functional grammar

This research differs from most work in statistical computational linguistics in that it is compatible with and builds on the results of modern linguistic theory. While our approach is compatible with virtually all existing theories of grammar (including transformational grammar and minimalist grammars), we have adopted the framework and structures of Lexical-Functional Grammar (LFG) in our research. LFG has several properties that make it especially well-suited for research involving linguistically-oriented probabilistic grammars. The formal definition of LFGs and the structures they generate is clear and precise (Kaplan, 1995), and LFG provides simple, clean descriptions of a wide range of typologically diverse linguistic phenomena (Bresnan, 1982). Recent linguistic work integrating Optimality-theory with LFG helps to identify features that are likely to be useful in the probabilistic models discussed in the next section (Bresnan, 1998; Johnson, 1998). Equally important, the LFG framework outlines all of the components of grammar and there are existing, reasonably comprehensive, LFGs for languages such as English and German. There is also a substantial amount of existing computational research on LFG, including on efficient parsing with large grammars (Maxwell III and Kaplan, 1993), which we exploit in our research.

An LFG structure of a sentence consists of a small number of distinct components, such as the phonological structure, the syntactic structure, the semantic interpretation, etc. Unlike linguistic theories such as HPSG (Pollard and Sag, 1987; Pollard and Sag, 1994), these components are heterogeneous, permitting the “natural” data-structures for each component to be used (e.g., syntactic structures are trees, semantic interpretations are represented by terms from the  $\lambda$ -calculus, etc.).

To keep things simple in this paper, however, we will only use a subset of these components and simplify them where appropriate. For example, we take the phonological component of a sentence to be just a string of words, and ignore prosody and other phonological details. Similarly, we take the semantic interpretation of a sentence to be its predicate-argument structure (roughly, “who did what to whom”), and ignore mood, tense, etc. We make extensive use of two components in this paper. The constituent or *c-structure* of a sentence shows the temporal arrangement of words, phrases and clauses organized as a tree structure. The functional or *f-structure* of a sentence is an attribute-value structure that shows the grammatical function relationships between the phrases and clauses of a sentence, abstracting away from details of linear order. The particular grammatical function relationship involved (e.g., subject, object, etc.) is represented by the attribute name, and f-structures also encode the argument-adjunct distinction. Although it probably deserves to be a

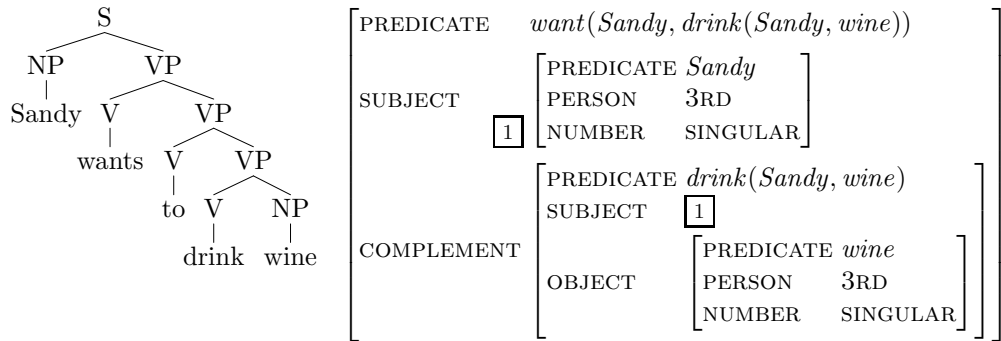


Figure 1: The c-structure and the f-structure for the English sentence *Sandy wants to drink wine*.

component in its own right, for simplicity we follow early work in LFG that encodes the predicate-argument structure of a phrase or sentence as the value of the PREDICATE attribute in an f-structure. Figure 1 depicts the c-structure and f-structure of the English sentence *Sandy wants to drink wine*.

One of the reasons for adopting an attribute-value representation of f-structure in LFG is that such structures can describe the multiple functional roles that a single constituent can play in a single sentence. For example, in *Sandy wants to drink wine* the NP *Sandy* functions both as the subject of both the verb *wants* and the verb *drink* (cf., *Sandy wants Sam to drink wine*). This is indicated by a re-entrancy in the f-structure, depicted by the shared index “1” in Figure 1.

Similar re-entrancies are used to indicate the functional roles played in relative clauses and WH-questions, where a functional dependency may span an unbounded distance in the constituent structure. For example, in the question *Which bottle did Sandy want Sam to open?*, the WH-phrase *which bottle* functions as the object of the verb *open* even though the two elements are discontinuous in c-structure terms. As explained in section 3, lexical dependencies between governor-govenee pairs play an important role in our probabilistic model, and their explicit representation in LFG’s f-structure makes the construction of our probabilistic model much easier. (As far as we can tell, because other current stochastic parsing models compute such lexical dependencies directly from c-structure, they tend to ignore most of the more complicated cases of dependencies).

F-structures also make explicit other important linguistic information. For example, the f-structure in Figure 1 encodes person and number features on noun phrases (important for subject-verb and pronoun agreement); although not shown here, f-structures also encode verb tense and other semantically important information. Notice that the f-structure makes explicit dependencies that may be non-local or only indirectly marked in the c-structure, and represents these in a relatively language-independent way. This gives LFG the power to provide simple descriptions of phenomena such as crossed serial dependencies, which cannot be described using weaker, purely tree-based approaches such as context-

free grammars (Bresnan et al., 1982; Shieber, 1985; Kaplan and Zaenen, 1995).

To be well-formed, an LFG linguistic structure must satisfy certain constraints. Some of these constraints ensure the consistency of the component structures; e.g., that the words of the phonological structure are (by and large) the same as the words in the syntactic structure. In a conventional (non-stochastic) LFG for a particular language, other language-specific constraints are also imposed, which force the structure to be grammatical in that language. For example, a conventional LFG for English might impose the constraint that a direct object NPs follow their governing verbs or prepositions, yielding familiar English SVO word order, whereas a corresponding LFG for German would impose different word order constraints that permit SOV word order.

An account of language acquisition should explain how the properties that differentiate the language being learnt from other possible human languages are acquired. Since one of the goals of this research is to determine the extent to which language learning can be viewed as a statistical parameter estimation problem, the restrictions or constraints imposed on possible linguistic structures should be universal, i.e., satisfied by all possible human languages. Thus the set of candidate linguistic structures (which we call  $\Omega$  below) should include all structures possible in any human language. (This desideratum also applies to Optimality-theory versions of LFG). Unfortunately, such “universal grammars” are not yet available: indeed, there are still major conceptual issues to be resolved before such a universal grammar (for any linguistic theory) can be constructed. For example, what is the nature of the lexicon in a universal grammar? Since a universal grammar should by definition be able to generate any linguistic structure from any possible human language, and since these linguistic structures include arbitrary lexical details (e.g., phonological forms and semantic interpretations), it would seem that a universal grammar must be capable of generating an infinite number of (possible) lexical items. While this is not incompatible with a finite universal grammar (since the lexicon could itself be specified by a generative process), the nature of the universal constraints on possible lexical entries seems largely unexplored.

Because of the lack of any reasonable candidate for a universal grammar, our computational experiments to date have utilized grammars for specific languages such as English (Johnson et al., 1999; Johnson and Riezler, 2000) and German (Riezler et al., 2000). Thus the statistical models developed in these experiments in effect learn how likely each grammatical linguistic structure of the particular language are: thus these models are capable of interpreting and disambiguating phonological forms in comprehension, but do not directly model language learning per se.

### 3 Probability distributions over linguistic structures

This section explains how we define a probability distribution over a set of possible linguistic structures  $\Omega$ . In a model of language learning,  $\Omega$  should be the set of all structures that could appear in any human language, but for a model of parsing of a single language one might take  $\Omega$  to be the set of grammatical structures of that language. Note that in either of these cases,  $\Omega$  is a countably infinite set, even if it is highly constrained by innate, language-specific constraints.

The probability distribution over  $\Omega$  is defined in terms of a finite vector of *features*  $\mathbf{f} = (f_1, \dots, f_m)$ , where each  $f_j$  is a function mapping a linguistic structure  $x \in \Omega$  to a real number  $f_j(x)$ . (The term ‘feature’ is used both in statistics and linguistics; we follow the standard usage in statistics here, and use the term ‘attribute’ to refer to components of attribute-value structures or node labels). While the mathematics impose few constraints on what the features can be, we generally take  $f_j(x)$  to be the number of times that a given construction appears in the linguistic structure  $x \in \Omega$ , which means that  $f_j(x)$  is a non-negative integer.

The features can be *lexicalized*, i.e., they can make reference to a specific words or word classes, but they need not be. For an example of a non-lexicalized feature, let  $f_1(x)$  be the number of times that a direct object immediately precedes its governing verb in  $x$ ; this is presumably almost always zero for sentences of English, but is often non-zero for a head-final language like German. For an example of a lexicalized feature, let  $f_2(x)$  is the number of times the verb *eat* appears with a direct object in the structure  $x$ ; if this is close to number of times *eat* appears in  $x$  then presumably *eat* is a primarily transitive verb. The selection of features is presumably an empirical linguistic issue (just as the selection of constraints in an Optimality Theory grammar or of parameters in a Principles and Parameters model are empirical issues).

In our experiments with probabilistic LFGs we use a wide variety of features (see Johnson et al. (1999) for a more detailed description). Inspired by probabilistic context-free grammars, we introduced a feature  $f_A$  for each category  $A$  that can label a c-structure node, and define  $f_A(x)$  to be the number of times a node labelled  $A$  appears in the c-structure of  $x$ . Additionally, the probabilistic LFGs evaluated below used the following kinds of features, whose selection was guided by the principles proposed by Hobbs and Bear (1995). Adjunct and argument features indicate adjunct and argument attachment respectively, and permit the model to capture a general argument attachment preference. In addition, there are specialized adjunct and argument features corresponding to each grammatical function used in LFG (e.g., SUBJECT, OBJECT, COMPLEMENT, ADJUNCT, etc.) There are features indicating both high and low attachment (determined by the complexity of the phrase being attached to). Another feature indicates non-right-branching nonterminal nodes. There is a feature for non-parallel coordinate structures (where parallelism is measured

in constituent structure terms). Each f-structure attribute-atomic value pair which appears in any feature structure is also used as a feature. We also use a number of features identifying syntactic structures that seem particularly important in the particular corpora we used in our experiments, such as a feature identifying NPs that are dates (it seems that date interpretations of NPs are preferred if they are available).

Ideally we would like to include lexical features directly in our experiments to capture the dependencies between governors and the heads of the phrases that they govern, but we did not have enough training data to estimate these directly in our experiments. However, probabilistic models of such dependencies can be constructed by other means, and we can include information from such “auxiliary” models in our model as follows (Johnson and Riezler, 2000; Riezler et al., 2000). Suppose we have an auxiliary model  $R$  which assigns a positive numerical preference score  $R(x)$  to each  $x \in \Omega$ . ( $R$  might define a probability distribution over  $\Omega$ , but need not). Then we define a new feature  $f_R(x) = \log R(x)$ , and treat it otherwise just as another feature in our model. In effect, the preference information from the auxiliary model  $R$  is treated as another source of information that will be taken into account in the model we construct. This provides a general mechanism whereby a range of complex preferences (possibly including innate ones) can be included in a statistical model, which generalizes the “reference distribution” approach described in Jelinek (1997).

We now explain how the probability of a particular linguistic structure  $x$  is defined in terms of its feature values  $\mathbf{f}(x) = (f_1(x), \dots, f_m(x))$ . While there are many ways in which this can be done, we use the class of *log-linear models* in our research (Abney, 1997). We justify our choice of log-linear models after we have explained how they are defined.

Given a set of linguistic structures  $\Omega$  and a feature vector  $(f_1, \dots, f_m)$ , a log-linear model is defined by a *parameter vector*  $\theta = (\theta_1, \dots, \theta_m)$ , where each  $\theta_j$  is a real number. Informally,  $\theta_j$  is the “weight” assigned to the corresponding feature  $f_j$ . If  $\theta_j$  is positive then higher values of  $f_j(x)$  increase the probability of  $x$ , and if  $\theta_j$  is negative then higher values of  $f_j(x)$  decrease the probability of  $x$  (assuming that the values of  $f_{j'}(x), j' \neq j$  stay the same).

Mathematically, the probability  $P_\theta(X = x)$  of  $x$  given the parameter vector  $\theta$  is defined as follows. We define the *weight*  $V_\theta(x)$  of  $x$  as the exponential of a linear combination of the feature values of  $x$ , weighted according to the parameter vector. (Thus the logarithm of  $V_\theta(x)$  is a linear combination of the feature values, hence the name log-linear model).

$$V_\theta(x) = \exp\left(\sum_{j=1}^m \theta_j f_j(x)\right)$$

A probability distribution over linguistic structures  $\Omega$  must satisfy the normalization constraint that the sum of probability of the structures in  $\Omega$  is 1, i.e.,  $\sum_{x \in \Omega} P_\theta(X = x) = 1$ . We cannot set  $P_\theta(X = x) = V_\theta(x)$  because in general  $V_\theta$  does not satisfy the normalization constraint. However, we can make  $P_\theta(X = x)$

proportional to  $V_\theta(x)$  by dividing the latter by a normalization factor known as the *partition function*  $Z_\theta$  (the name comes from statistical physics, which was the first major application of log-linear models).

$$Z_\theta = \sum_{x \in \Omega} V_\theta(x) \quad (1)$$

$$P_\theta(X = x) = \frac{V_\theta(x)}{Z_\theta} \quad (2)$$

To be honest, the class of log-linear models is just one of large number of classes of probabilistic models, and we do not know if human language is best described by such a model. Never the less, given the lack of even weak evidence for other model classes, the flexibility of log-linear models seem to make them a good class to investigate first. Unlike probabilistic context free grammars and related models, log-linear models permit essentially arbitrary dependencies between features (include context-sensitive dependencies), which makes them ideal for defining probability distributions over linguistically realistic structures (Abney, 1997). Additionally, there are information-theoretic reasons for preferring log-linear models over other model classes. The class of log-linear models is also the class of *maximum entropy* models; roughly speaking, these are the models which contain the minimum additional information over and above the information contained in the training data (see Jelinek (1997) for a textbook introduction). Virtually all of the well-known probabilistic models of language are subclasses of the class of log-linear models (e.g., probabilistic context-free grammars, hidden Markov models, etc.). Finally, even though one might suspect that the restriction to *linear* combinations of the feature values is unduely restrictive, because no restrictions are placed on the features themselves, we can define a feature which is a nonlinear combination of other features, so the class of log-linear models is much less restrictive than it may first seem.

Because the parameter vector  $\theta$  determines the distribution of linguistic structures in the language, estimating  $\theta$  corresponds to learning the language: we discuss this problem in the next and following sections. Suppose we have the parameter vector for a language—how might it be used in comprehension and production? To keep things simple, ignore contextual and other factors, and assume that linguistic structures are in fact distributed according to  $P_\theta$ . Comprehension is the problem of identifying the intended semantic interpretation  $s$  given a phonological form  $w$  (here taken to be a word string), while production is the inverse:  $s$  is given and we need to identify the corresponding  $w$ . Given these assumptions, it is possible to show that a language user minimizes their expected error rate by selecting the most probable linguistic structure whose phonological form is  $w$  in comprehension, and by selecting the most probable linguistic structure whose semantic interpretation is  $s$  in production.

We can give a more formal account of the parsing problem as follows (the account of the generation problem is similiar). Let  $S$  be the function mapping a linguistic structure to its semantic interpretation, and (abusing notation



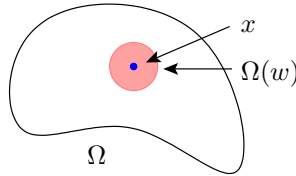


Figure 2: Parsing the phonological form  $w$  only involves optimization over the set  $\Omega(w)$  of linguistic structures with phonological form  $w$ .

somewhat) let  $\Omega(w)$  be the set of linguistic structures with phonological form  $w$ . Then the semantic interpretation  $s'$  with minimum expected error in the parsing problem is:

$$s' = \operatorname{argmax}_s P_\theta(S = s | W = w) \quad (3)$$

$$= \operatorname{argmax}_s \sum_{\substack{x \in \Omega(w) \\ \text{s.t. } S(x)=s}} V_\theta(x) \quad (4)$$

In equation (4) we see that stochastic parsing only involves an optimization of  $V_\theta$  over the set  $\Omega(w)$ , and does not involve computation of the partition function  $Z_\theta$ . The domain of optimization is depicted graphically in Figure 2. The “off line parsability constraint” of LFG ensures that  $\Omega(w)$  (the set of possible parses of  $w$ ) is always finite (Kaplan and Bresnan, 1982; Pereira and Warren, 1983) even though the set of possible linguistic structures  $\Omega$  is not,<sup>1</sup> so the optimization (3) can be carried out by enumerating the members of  $\Omega(w)$ . Since the functions  $V$  and  $S$  are effectively computable, this implies that  $s'$  is effectively computable as well. Moreover, standard parsing algorithms for non-probabilistic grammars can be used to enumerate  $\Omega(w)$  (see Maxwell III and Kaplan (1993) for a discussion of LFG parsing). Thus for hand-crafted grammars which produce only a moderate number of detailed parses per sentence all of the technology needed to find the parse with the minimum expected error already exists.

## 4 Learning grammars

The previous section described how we define a log-linear probability distribution over linguistic structures  $\Omega$ . We now turn to the problem of determining the parameter vector  $\theta$  from some observational data  $D$ . At this abstract level, this is just a statistical estimation problem and a wealth of statistical theory applies

<sup>1</sup>The off-line parsability constraint only implies that  $\Omega(w)$  is finite if the lexicon itself is finite, which may not be true if  $\Omega$  is taken to be the set of structures generated by a universal grammar.

to it. The difficulty of this problem depends on the set of possible structures  $\Omega$ , the features  $\mathbf{f}$  which describe each structure, and the data  $D$  from which the estimate is to be made. There is no guarantee that the parameter vector can be estimated at all, and even if it can be, it may be the case that an unreasonably large amount of data required to estimate it accurately, no matter what estimation procedure is used. In general, there is a trade-off between model *bias* (i.e., constraints on possible languages) and the *variance* in estimates of the model’s parameters from data: changing the model class to lower the bias raises the variance in the estimated models (Geman, Bienenstock, and Doursat, 1992). This argument is familiar in the study of language acquisition, where it is standardly used to argue for restrictive innate constraints on possible languages. We would like to suggest that a statistical perspective enriches the debate in several ways. First, “soft” probabilistic biases can lower variance, so it may not be necessary to posit “hard” categorical innate constraints to ensure accurate learning. Second, statistics provides quantitative tools for investigating bias/variance trade-offs, which may add an additional empirical dimension to the study of universal grammar and language acquisition.

In our experiments we use *maximum likelihood* estimators (but see the discussion of regularization in section 5). A maximum likelihood estimator selects a parameter vector  $\theta$  which makes the data  $D$  as likely as possible, i.e., it ignores the prior term and maximizes the log-likelihood  $L_D(\theta) = \log P(D|\theta)$  of the training data in  $\theta$ . Under very general conditions, maximum likelihood estimation is *unbiased* (the expected value of parameter estimate is its true value), *consistent* (as the size of the data grows, the estimated parameters converge on the true value) and *asymptotically efficient* (there is no other estimation procedure whose parameter estimates have uniformly lower variance). Further, given the independence assumptions below the maximum likelihood estimator for a log-linear model selects the closest model to the training data distribution in terms of Kullback-Leibler divergence (an information-theoretic measure of the distance between two distributions).

More formally, suppose that  $D$  consists of a sequence of *fully observed* parses  $D = (x_1, \dots, x_n), x_i \in \Omega$ . (“Fully observed” means that the learner has access to the complete linguistic structures; we consider the problem of learning from phonological forms alone below). We make the standard statistical assumptions that each observation  $x_i$  is independent of the other observations  $x_{i'}, i' \neq i$ , and that each  $x_i$  is identically distributed according to  $P_\theta$  for some unknown  $\theta$  (these assumptions are undoubtedly incorrect, but we hope that they are approximately true). Given these assumptions, the likelihood  $L_D$  of the data  $D$  and the corresponding maximum likelihood estimate  $\hat{\theta}$  of  $\theta$  are:

$$L_D(\theta) = \prod_{i=1}^n P_\theta(X = x_i) \tag{5}$$

$$\hat{\theta} = \operatorname{argmax}_\theta L_D(\theta)$$

Figure 3 graphically depicts this maximum likelihood estimation. Informally,



Figure 3: Maximum likelihood estimation from fully observed (parsed) data

maximum likelihood estimation adjusts  $\theta$  to make the weight  $V_\theta(x_i)$  of each training datum as large as possible relative to the partition function  $Z_\theta$  (the sum of the weights of all linguistic structures  $\Omega$ ).

It is straight forward to show that  $L_D$  has a unique maximum value, and at this maximum the expected value  $E_{\hat{\theta}}(f_j)$  of each feature under the distribution  $P_{\hat{\theta}}$  is equal to its expected value under the “empirical distribution” of the training data  $D$ , i.e.,:

$$E_{\hat{\theta}}(f_j) = \frac{1}{n} \sum_{i=1}^n f_j(x_i), \quad j = 1, \dots, m.$$

Thus maximum likelihood estimation selects a parameter vector  $\hat{\theta}$  so that the expected value of each feature under the estimated distribution  $P_{\hat{\theta}}$  is the same as the average value of that feature in the training data, which intuitively seems to be a reasonable thing to do.

Now we turn to the case where the training data is *partially hidden* and consists of phonological forms alone, i.e.,  $D' = (w_1, \dots, w_n)$ , where each  $w_i$  is a phonological form (here taken to be a string of words). In this situation the training data does not uniquely identify the linguistic structure corresponding to each phonological form  $w_i$ ; all we know is that it lies somewhere inside the set  $\Omega(w_i) = \{x : W(x) = w_i\}$  of linguistic structures whose phonological forms are  $w_i$ . Making the same independence assumptions as before, the likelihood  $L'_{D'}$  of the data  $D'$  is now a product of the marginal probability of each  $w_i$ , where the marginal probability of  $w$  is the sum of the probability of each  $x \in \Omega(w)$ .

$$\begin{aligned} P_\theta(W = w) &= \sum_{x \in \Omega(w)} P_\theta(X = x) \\ L'_{D'}(\theta) &= \prod_{i=1}^n P_\theta(W = w_i) \\ \hat{\theta} &= \operatorname{argmax}_\theta L'_{D'}(\theta) \end{aligned} \tag{6}$$

Figure 4 graphically depicts the quantity being maximized during estimation from phonological forms alone. Notice that the maximum likelihood estimator selects the  $\theta$  that places maximum weight on the  $\Omega(w_i)$  as compared to the whole of  $\Omega$ .

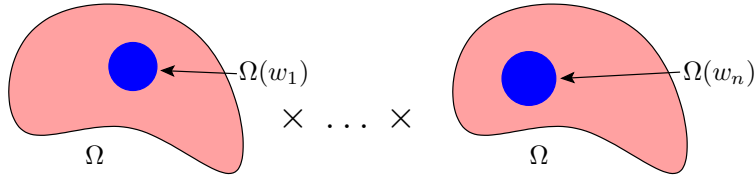


Figure 4: Maximum likelihood estimation from partially visible (phonological form) data

There is a standard technique known as the *Expectation-Maximization (EM)* algorithm which reduces the optimization required in maximum likelihood estimation from partially hidden data to a series of optimizations of the kind involved in maximum likelihood estimation from fully visible data (Dempster, Laird, and Rubin, 1977). The technique requires an initial guess  $\theta^{(0)}$  of the parameter vector as well as the partially observed data  $D'$ , and it produces a sequence of estimates  $\theta^{(1)}, \theta^{(2)}, \dots$ . This sequence has the property that each additional estimate typically increases, and provably does not decrease, the likelihood of the partially observed data, i.e.,  $L'_{D'}(\theta^{(k+1)}) \geq L'_{D'}(\theta^{(k)})$ . Informally, the technique involves treating each partial observation  $w_i$  as a set of fully observed data consisting of each  $x \in \Omega(w_i)$ , with each full observation  $x$  weighted according to  $P_{\theta^{(k)}}(x)$ , where  $\theta^{(k)}$  is the estimate of  $\theta$  at the  $k$ th iteration. Thus EM “pays most attention to” the  $x \in \Omega(w_i)$  that its current estimate of  $\theta$  assigns the highest probability to.

Unlike the fully visible case, there is no guarantee that the likelihood function for partially hidden data has only a single local maximum, and the EM algorithm can get “trapped” in such local maxima. Indeed, there is no guarantee that estimation is possible at all: the parameter vector  $\theta$  may simply be *non-identifiable* from the kind of data available. For example, it is logically possible that universal grammar permits two different languages with exactly the same marginal distribution over phonological forms, even though the two languages associate each phonological form with different semantic interpretations.

## 5 Pseudo-likelihood estimation

The previous section introduced maximum likelihood estimation of  $\theta$  for both fully visible and partially hidden data. Unfortunately, it seems that directly maximizing the likelihood (5) is computationally infeasible even for fully visible data (and since the EM technique reduces the partially hidden data case to the fully visible data case, it too is infeasible). The standard algorithms for maximizing this likelihood are iterative, and require the calculation of the expected value of each feature  $E_{\theta}(f_j)$  for a variety of different parameter vectors  $\theta$  (see Berger, Della Pietra, and Della Pietra (1996) and Jelinek (1997) for an introduction to these algorithms). Informally, the cause of the infeasibility is

that maximum likelihood estimation requires us to select the parameter vector  $\theta$  that maximizes the weight  $V_\theta(x_i)$  on the observed datum  $x_i$  relative to the sum  $Z_\theta$  of the weights on *all* possible linguistic structures  $x \in \Omega$  (see Figure 3 and equations 1, 2 and 5). Because  $\Omega$  is infinite, we cannot calculate the partition function  $Z_\theta$  or the feature expectations  $E_\theta(f_j)$  by directly enumerating  $\Omega$ . Indeed, even calculating the probability  $P_\theta(X = x)$  of a single linguistic structure  $x$  seems infeasible, since it too crucially involves  $Z_\theta$  (see equation 2).

If  $\Omega$  and the feature vector  $\mathbf{f}$  have a suitably simple structure, then it may be the case that  $Z_\theta$  and  $E_\theta(f_j)$  can be calculated analytically. For example, if  $\Omega$  is the set of trees generated by a context free grammar and the feature  $f_i$  maps an  $x \in \Omega$  to the number of times the  $i$ th production is used in a derivation of  $x$ , then  $Z_\theta$  and  $E_\theta(f_j)$  can be calculated without an explicit enumeration of  $\Omega$  (Abney, McAllester, and Pereira, 1999; Chi, 1999). However, this calculation depends crucially on the context free or Markovian independence properties of Probabilistic Context Free Grammars. It seems that such context free systems cannot describe the true set  $\Omega$  of possible linguistic structures (Shieber, 1985), yet these context free properties are what makes the direct calculation of  $Z_\theta$  and  $E_\theta(f_j)$  feasible. Indeed, precisely because the Lexical Functional Grammars used in this research are capable of capturing the non-local, context-sensitive dependencies of natural language, the methods that can be used to calculate  $Z_\theta$  and  $E_\theta(f_j)$  for PCFGs do not extend to LFGs.

Never the less, we believe that there may be techniques for calculating or approximating  $Z_\theta$  for LFGs that avoid explicit enumeration. Abney (1997) points out that  $E_\theta(f_j)$  can be approximated using Monte Carlo sampling techniques that do not enumerate all of  $\Omega$ . While this is in principle correct, a “back of the envelope” calculation suggests that the particular Hastings Metropolis sampling scheme that Abney proposes is computationally impractical for all but small grammars (see Johnson et al. (1999) for further discussion).

However, note that the full joint distribution over phonological forms and their parses is not actually required for natural language processing tasks. For example, as explained above, comprehension and parsing only requires the *conditional* distribution  $P(X|W)$  of linguistic structures given their phonological forms (see equation (3) and Figure 2 above). Crucially, estimating these conditional distributions is often computationally feasible, even though estimation of the joint distribution is infeasible.

Consider the case where the data is fully observed:  $D$  consists of parses  $D = (x_1, \dots, x_n), x_i \in \Omega$  as above. Each parse is associated with a phonological form  $w_i = W(x_i)$ . Making the same independence assumptions as before, the conditional likelihood or *pseudo-likelihood*  $PL_D$  of the data  $D$  and the corresponding maximum likelihood estimate  $\hat{\theta}$  of  $\theta$  are:

$$Z_\theta(w) = \sum_{x \in \Omega(w)} V_\theta(x)$$

$$P_\theta(X = x | W = w) = \frac{V_\theta(x)}{Z_\theta(w)}$$

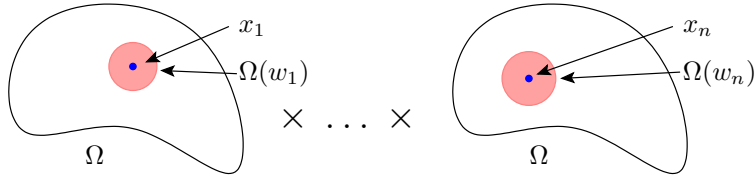


Figure 5: Maximum pseudo-likelihood estimation from fully observed data

$$\begin{aligned}
 PL_D(\theta) &= \prod_{i=1}^n P_\theta(X = x_i | W = w_i) \\
 \hat{\theta} &= \operatorname{argmax}_\theta PL_D(\theta)
 \end{aligned}
 \tag{7}$$

Whereas the likelihood  $L_D$  is a product of (unconditional) probabilities (5), the pseudo-likelihood  $PL_D$  is a product of conditional probabilities (7). Ultimately, pseudo-likelihood differs from likelihood in that pseudo-likelihood only involves  $Z_\theta(w)$  in place of the infeasible  $Z_\theta$  in the likelihood. It is straight forward to show that at the maximum of (7), the sum of the conditional expectations of each feature must be same as sum of their empirical values, where  $E_\theta(f|W)$  is the expectation of  $f$  with respect to the conditional distribution  $P_\theta(X|W)$ :

$$\sum_{i=1}^n E_{\hat{\theta}}(f_j | W = w_i) = \sum_{i=1}^n f_j(x_i), \quad j = 1 \dots m.$$

Moving to pseudo-likelihood makes a crucial difference in the kinds of expectations that must be computed in the standard algorithms for maximizing  $\theta$ ; they now involve the generally feasible conditional expectations  $E_\theta(f_j|W)$  rather than the infeasible unconditional expectations  $E_\theta(f_j)$ .

It turns out that this idea of directly estimating a conditional distribution (rather than the joint) has been independently discovered at least twice. Besag (1975), who coined the name ‘pseudo-likelihood’, uses it in a computational vision setting in which one part of an image serves as the conditioning environment for another part of the image (here, the phonological form corresponds to one part of the linguistic structure, and everything else in the structure corresponds to the other part). Berger, Della Pietra, and Della Pietra (1996) and Jelinek (1997) both describe optimizations in their algorithms which replace joint probabilities with conditional probabilities in exactly the manner described here (but they do not acknowledge that this means they are estimating a conditional rather than a joint distribution).

Figure 5 graphically depicts maximum pseudo-likelihood estimation. Informally, maximum pseudo-likelihood estimation adjusts  $\theta$  to make the weight  $V_\theta(x_i)$  of each training datum as large as possible relative to  $Z_\theta(w_i)$ , i.e., the sum of the weights of all parses  $\Omega(w_i)$  of the phonological form  $w_i$ . As remarked earlier,  $\Omega(w_i)$  is finite and of manageable size for LFGs, so  $Z_\theta(w_i)$  and

the conditional expectations required for maximizing the pseudo-likelihood can be calculated using direct enumeration of  $\Omega(w_i)$ .

While pseudo-likelihood estimation is consistent *for the conditional distribution*, it is not hard to see that maximizing  $PL_D$  will not always correctly estimate the joint  $P_\theta(X)$  (Chi, 1998). Suppose there is a feature  $f_j$  which depends solely on the phonological form  $W(x)$  of a linguistic structure  $x$ , i.e.,  $f_j(x') = f_j(x)$  for all  $x \in \Omega$  and  $x' \in \Omega(W(x))$ ; we call such features *pseudo-constant*. (For an example of a pseudo-constant feature, let  $f_j(x)$  be the number of times the word *eat* occurs in  $x$ ). If  $f_j$  is pseudo-constant, then it is easy to show that the pseudo-likelihood does not depend on the value of the parameter  $\theta_j$  associated with  $f_j$ , so maximum pseudo-likelihood estimation provides basis for choosing a value for  $\theta_j$ . In fact, in this case *any* value of  $\theta_j$  gives the *same* conditional distribution  $P_\theta(X|W)$ , so  $\theta_j$  is irrelevant to the problem of choosing good parses.

Informally, the relationship between maximum likelihood and pseudo-likelihood estimation is the same as the relationship between the joint  $P(X, W)$  and the conditional  $P(X|W)$ , which are related by the marginal  $P(W)$ :

$$P(X, W) = P(X|W) P(W).$$

The parameter vectors estimated by maximum likelihood estimation model the joint; they describe both the condition distribution of parses given phonological forms as well as the marginal distribution of phonological forms  $P(W)$ , while pseudo-likelihood estimation focuses on the conditional  $P(X|W)$  and ignores the marginal.

Interestingly, from a cognitive modularity perspective, the conditional and the marginal distributions seem to correspond to two different kinds of information. As noted above, the conditional distribution  $P(X|W)$  is precisely the information required for disambiguation in sentence comprehension, which seems to be purely linguistic knowledge. The marginal distribution  $P(W)$ , on the other hand, describes the distribution of phonological forms, which seems to involve world knowledge and contextual information at least as much as it involves linguistic knowledge. Thus pseudo-likelihood estimation may be more compatible with a modular view of language, since it seems to focus on more purely linguistic knowledge than does maximum likelihood estimation.

We now briefly describe some of the more practical details of pseudo-likelihood estimation. Despite the assurance of consistency, pseudo-likelihood estimation is prone to over fitting when a large number of features is matched against a modest-sized corpus of training data. One particularly troublesome manifestation of over fitting results from the existence of features which, relative to the training data, we call “pseudo-maximal”. A feature  $f$  is *pseudo-maximal* for a phonological form  $w$  if and only if for all  $x' \in \Omega(w)$   $f(x) \geq f(x')$  where  $x$  is any correct parse of  $w$ , i.e., the feature’s value on every correct parse  $x$  of  $w$  is greater than or equal to its value on any other parse of  $w$ . Pseudo-minimal features are defined similarly. It is easy to see that if  $f_j$  is pseudo-maximal on *each sentence* of the training corpus then the parameter assignment  $\theta_j = \infty$  maximizes the corpus pseudo-likelihood. (Similarly, the assignment  $\theta_j = -\infty$

maximizes pseudo-likelihood if  $f_j$  is pseudo-minimal over the training corpus). Such infinite parameter values indicate that the model treats pseudo-maximal features categorically; i.e., any parse with a non-maximal feature value is assigned a zero conditional probability.

Of course, a feature which is pseudo-maximal over a finite training corpus is not necessarily pseudo-maximal for all phonological forms in those the language. This is an instance of over fitting, and it can be addressed, as is customary, by adding to the objective function a *regularization* term that promotes small values of  $\theta$ . In Johnson et al. (1999) we added a quadratic to the log pseudo-likelihood, which corresponds to multiplying the pseudo-likelihood itself by a normal distribution. Specifically, we multiplied the pseudo-likelihood by a zero-mean normal in  $\theta$  with diagonal covariance and with standard deviation  $\sigma_j$  for  $\theta_j$  equal to 7 times the maximum value of  $f_j$  found in any parse in the training data. Thus instead of choosing  $\hat{\theta}$  to maximize the pseudo-likelihood (7), in the experiments reported in Johnson et al. (1999) and Johnson and Riezler (2000) we actually selected  $\hat{\theta}$  to maximize:

$$\log PL_D(\hat{\theta}) - \sum_{j=1}^m \frac{\theta_j^2}{2\sigma_j^2} \quad (8)$$

Interestingly, this way of regularizing has a Bayesian interpretation. In Bayesian estimation one seeks a parameter vector  $\theta$  that maximizes the *posterior probability*  $P(\theta|D)$  of the parameter vector  $\theta$  given the training data  $D$ . According to Bayes theorem, this can be done by maximizing the product of the *prior probability*  $P(\theta)$  of the parameter vector and the likelihood  $P(D|\theta)$  of the data given the parameter vector. If one sets the prior probability  $P(\theta)$  to be proportional to  $\exp(-\sum_{j=1}^m \theta_j^2/2\sigma_j^2)$  and makes the same independence assumptions concerning the data as above, then it is possible to show that the Bayesian estimate for  $\theta$  is precisely the  $\theta$  that maximizes (8).

In these experiments, the set of possible linguistic structures  $\Omega$  was defined by a hand-written LFG for English, which was specifically designed at Xerox PARC to generate the sentences in two corpora of business appointment dialogs and “Homecenter” printer/copier documentation, consisting of 500 and 1000 parsed sentences respectively. Even though the grammar included all standard linguistic constraints, the sentences in the corpora were often highly ambiguous, with an average of 8 parses per sentence. The training data consisted of the correct parse for each sentence (which was identified manually) together with the set of all alternative (i.e., incorrect) parses of the sentence generated by the grammar. Using a cross-validation framework, we showed that a model trained by maximum pseudo-likelihood correctly disambiguated approximately 58% of ambiguous test sentences, whereas a model that treated each parse as equally likely would correctly disambiguate only 25% of the ambiguous test sentences.

We now turn to the more realistic situation (in terms of language acquisition) where the training data consists of phonological forms alone. Whereas maximum likelihood estimation from partially visible data is conceptually straight forward—one adjusts  $\theta$  to maximize the likelihood of the phonological forms



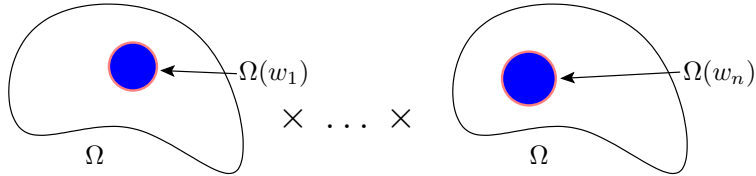


Figure 6: A straight forward application of maximum pseudo-likelihood estimation from partially visible (phonological form) data fails

that constitute the training data  $D$ —it turns out that similar approach based on pseudo-likelihood fails. Specifically, conditioning the marginal  $P(W = w_i)$  in the likelihood (6) on the phonological form results in a constant-valued likelihood that does not vary with  $\theta$  or  $D$ , so estimation fails.

Intuitively, the problem is that we are trying to maximize the sum of the weights  $V_\theta(x)$  placed on the  $x \in \Omega(w_i)$  relative to the sum of the weights of exactly the same set  $\Omega(w_i)$ , as depicted in Figure 6. Standard maximum likelihood estimation from partially visible data (as performed by the EM algorithm) maximizes the sum of the weights placed on  $\Omega(w_i)$  relative to the sum  $Z_\theta$  of the weights placed on all possible linguistic structures  $\Omega$ .

We noted earlier that maximum likelihood estimation is infeasible because the partition function  $Z_\theta$  and the expectations  $E_\theta(f_j)$  involve summing over all possible linguistic structures  $\Omega$ . In Riezler et al. (2000) we developed a method for maximum likelihood estimation from partially visible data that exploits a data-oriented approximation to  $Z_\theta$  and  $E_\theta(f_j)$  in which we replace the summation over  $\Omega$  with a summation over the finite set  $\Omega(D')$  consisting of all possible parses the phonological forms that constitute the training data  $D' = (w_1, \dots, w_n)$ . More precisely, the likelihood function  $L''_{D'}$  maximized in these experiments is:

$$\begin{aligned} \Omega(D') &= \bigcup_{i=1}^n \Omega(w_i) \\ Z_\theta(D') &= \sum_{x \in \Omega(D')} V_\theta(x) \\ P_\theta(W = w | W \in D') &= \frac{Z_\theta(w_i)}{Z_\theta(D')} \\ L''_{D'}(\theta) &= \prod_{i=1}^n P_\theta(W = w_i | W \in D') \end{aligned}$$

Figure 7 depicts the likelihood function that this data-oriented approach maximizes. Unlike with standard maximum likelihood estimation, computation of the data-oriented partition function  $Z_\theta(D')$  and the corresponding expectations is feasible. The data-oriented approach can be viewed as a version of

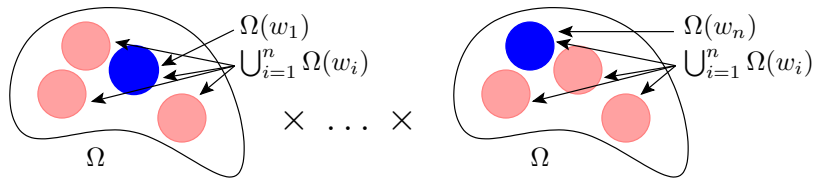


Figure 7: An data-oriented approach to estimation from partially visible (phonological form) data

pseudo-likelihood in the following way. Recall the key idea behind pseudo-likelihood: namely, that one can define a likelihood function by conditioning one part of the structure on another part of that structure. In pseudo-likelihood estimation from fully visible (parsed) data we take each sentence to be an observation and condition each linguistic structure on its phonological form. In this data-oriented approach, we take the entire data set  $D'$  to be an observation, and condition each phonological form on the fact that it occurred in  $D'$ .

Since the above described estimation procedure does not require manually annotated data for training but merely data consisting of phonological forms alone, large sets of training data can easily be provided. In our experiments we parsed a large corpus of newspaper text with a German LFG grammar (developed in the ParGram project at the University of Stuttgart), and extracted all parses for sentences which were assigned at most 20 parses by the grammar. This resulted in a training corpus of approximately 36,000 sentences and 250,000 parses. The rationale behind the employed restriction of the ambiguity of the training data is to simplify the estimation problem by restricting the entropy of the distribution over the training parses. A further attempt to regularize the estimation procedure is a initial regularization of parameter values with the effect of focussing the search in maximization on a proper subspace of the parameter space. Together, these regularization techniques serve to make EM a manageable estimation tool for highly precise statistical disambiguation. An evaluation of disambiguation performance on LFG parsed newspaper sentences with on average 25 parses per sentence showed the following results: The task of matching full  $c/f$ -structure pairs to the manually selected pair could be performed correctly in over 60% of the test cases; a disambiguation of the predicate-argument structures of the parses of the test sentences (which is sufficient for many application purposes) could be performed correctly in over 90% of test cases.

## 6 Conclusion and further directions

Because log-linear models make no assumptions about relationships between features, they provide a general framework for defining probability distributions over linguistic structures from virtually any linguistic theory (Abney, 1997).

Maximum likelihood estimation is an optimal method for estimating the parameter vectors for such models from data, but precisely because log-linear models are so general, maximum likelihood estimation is typically computationally infeasible because it requires us to calculate expectations over all possible linguistic structures. This leads us to develop techniques based on pseudo-likelihood (Besag, 1975) for estimating parameter vectors from fully visible (parsed) data (Johnson et al., 1999; Johnson and Riezler, 2000) and partially visible (phonological form) data (Riezler et al., 2000).

This work is still in its infancy, and many interesting avenues remain to be explored. As we explained above, in order to study language acquisition the set of possible linguistic structures  $\Omega$  should be the set of *universally* possible structures, but currently we lack a “universal grammar” with which to conduct our experiments. As we explained in the text, it seems that basic conceptual issues still need to be addressed before such a universal grammar can be constructed. Sticking to the framework used in the experiments described in this paper where  $\Omega$  is the set of possible linguistic structures of a particular language, we believe there is interesting empirical linguistic research to be done in investigating the trade-off between the “hard” grammatical constraints incorporated in the grammar that determines  $\Omega$  and the “soft” preferences that can be encoded using features  $f_j$  in the statistical model. The grammars we used in our experiments were not written with our statistical models in mind, and we might obtain a more robust system with broader coverage by removing some of the grammatical constraints from the grammar and re-expressing them as features in the statistical model.

Turning to more mathematical issues, it would be valuable to investigate other ways for estimating the partition function and the expectations required for maximum likelihood estimation from both parsed and phonological form data. Techniques for approximating these quantities have been developed in other fields (e.g., mean field approximations), and it may be possible to apply them in computational linguistics as well (Saul and Jordan, 1999).

A problem left unaddressed in our applications is efficient searching for most probable parses. This question becomes crucial if higher coverage is desired and traded in for more superficial parses and for higher ambiguity. Clearly, for such cases it is desirable to adapt techniques such as Viterbi’s algorithm (Viterbi, 1967) to searching efficiently for most probable parses in probabilistic LFG grammars. Here a closer look at generalized dynamic-programming techniques as developed for graphical models (Frey, 1998) seems promising.

Finally, we believe that there may be other ways of applying pseudo-likelihood to language learning besides the ways described in this paper. The pseudo-likelihood estimation approach from visible (parsed) data is highly unrealistic in one respect: such a learner *learns nothing from unambiguous sentences* in its training data, even though such sentences are intuitively most informative of all. This is because the pseudo-likelihood we used conditioned on phonological form; i.e.,  $P(X|W)$ . Suppose instead we adopt a “generation-oriented” pseudo-likelihood, where we condition on the semantic interpretation  $S(x)$  of each linguistic structure  $x$ , so the likelihood is the product of terms

$P(X = x_i | S = s_i)$ . Such a learner would learn from each sentence in its training data whose semantic interpretation can be expressed in more than one way universally.

## References

- Abney, Steven. 1997. Stochastic Attribute-Value Grammars. *Computational Linguistics*, 23(4):597–617.
- Abney, Steven, David McAllester, and Fernando Pereira. 1999. Relating probabilistic grammars and automata. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 542–549, San Francisco. Morgan Kaufmann.
- Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.
- Besag, J. 1975. Statistical analysis of non-lattice data. *The Statistician*, 24:179–195.
- Bresnan, Joan. 1982. *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, Massachusetts.
- Bresnan, Joan. 1998. The lexicon in optimality theory. In Paula Merlo and Susan Stevenson, editors, *Proceedings of the 11th Annual CUNY Conference on Human Sentence Processing, Rutgers University*. To appear.
- Bresnan, Joan, Ronald M. Kaplan, Stanley Peters, and Annie Zaenen. 1982. Cross-serial dependencies in dutch. *Linguistic Inquiry*, 13:613–635.
- Chi, Zhiyi. 1998. *Probability Models for Complex Systems*. Ph.D. thesis, Brown University.
- Chi, Zhiyi. 1999. Statistical properties of probabilistic context-free grammars. *Computational Linguistics*, 25(1):131–160.
- Chomsky, Noam. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.
- Dempster, A., N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.
- Frey, Brendan J. 1998. *Graphical Models for Machine Learning and Digital Communication*. The MIT Press, Cambridge, Massachusetts.
- Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58.

- Gold, E.M. 1967. Language identification in the limit. *Information and Control*, 10:447–474.
- Hobbs, Jerry R. and John Bear. 1995. Two principles of parse preference. In Antonio Zampolli, Nicoletta Calzolari, and Martha Palmer, editors, *Linguistica Computazionale: Current Issues in Computational Linguistics: In Honour of Don Walker*. Kluwer, pages 503–512.
- Horning, James Jay. 1969. *A study of grammatical inference*. Ph.D. thesis, Stanford.
- Jelinek, Frederick. 1997. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts.
- Johnson, Mark. 1998. Optimality-theoretic Lexical Functional Grammar. In Paula Merlo and Susan Stevenson, editors, *Proceedings of the 11th Annual CUNY Conference on Human Sentence Processing, Rutgers University*. To appear.
- Johnson, Mark, Stuart Geman, Stephen Canon, Zhiyi Chi, and Stefan Riezler. 1999. Estimators for stochastic “unification-based” grammars. In *The Proceedings of the 37th Annual Conference of the Association for Computational Linguistics*, pages 535–541, San Francisco. Morgan Kaufmann.
- Johnson, Mark and Stefan Riezler. 2000. Exploiting auxiliary distributions in stochastic unification-based grammars. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 154–161, San Francisco. Morgan Kaufmann.
- Kaplan, Ronald M. 1995. The formal architecture of LFG. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, editors, *Formal Issues in Lexical-Functional Grammar*, number 47 in CSLI Lecture Notes Series. CSLI Publications, chapter 1, pages 7–28.
- Kaplan, Ronald M. and Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*. The MIT Press, chapter 4, pages 173–281.
- Kaplan, Ronald M. and Annie Zaenen. 1995. Long-distance dependencies, constituent structure and functional uncertainty. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, editors, *Formal Issues in Lexical-Functional Grammar*, number 47 in CSLI Lecture Notes Series. CSLI Publications, chapter 3, pages 137–165.
- Maxwell III, John T. and Ronald M. Kaplan. 1993. The interface between phrasal and functional constraints. *Computational Linguistics*, 19(4):571–590.

- Pereira, Fernando C.N. and David H.D. Warren. 1983. Parsing as deduction. In *The Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 137–144, M.I.T., Cambridge, Massachusetts.
- Pollard, Carl and Ivan Sag. 1994. *Head-driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.
- Pollard, Carl and Ivan A. Sag. 1987. *Information-based Syntax and Semantics*. Number 13 in CSLI Lecture Notes Series. Chicago University Press, Chicago.
- Riezler, Stefan, Detlef Prescher, Jonas Kuhn, and Mark Johnson. 2000. Lexicalized stochastic modeling of constraint-based grammars using log-linear measures and em training. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00), Hong Kong*.
- Saul, Lawrence and Michael Jordan. 1999. A Mean Field learning algorithm for unsupervised neural networks. In Michael Jordan, editor, *Learning in Graphical Models*. The MIT Press, Cambridge, Massachusetts, pages 541–554.
- Shieber, Stuart M. 1985. Evidence against the Context-Freeness of natural language. *Linguistics and Philosophy*, 8(3):333–344.
- Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, IT-13:260–269.