The Role of Corpus Linguistics in Redefining SLA[1]

Yukio Tono
Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

1. Introduction

Corpus linguistics has become one of the major areas of research in linguistics and applied linguistics. While there is a mixed view about the status of corpus linguistics as an independent discipline of linguistics or just a methodology (cf. McEnery, Xiao and Tono 2003), there is a growing awareness that the use of corpora will shed light on various aspects of language properties and their use. In this paper, I will argue that whilst language is rule-governed, those rules are not simply categorical in nature, but show various probabilistic nature. Frequency information obtained from corpora will be useful in theorizing human language faculty, integrating probabilistic information into theories of linguistic competence and performance. I will discuss how probabilistic views of language, based on the Bayesian network, will help redefine a theory of second language acquisition.

In recent years, a strong consensus has emerged that human cognition is based on probabilistic processing (cf. Bod, Hay and Jannedy 2003). The probabilistic approach is promising in modeling brain functioning and its ability to accurately model phenomena "from psychophysics and neurophysiology" (ibid: 2). Bod et al also claim that the language faculty itself displays probabilistic properties. I argue that this probabilistic view holds for various phenomena in both L1 and L2 acquisition and could have a significant impact on theory construction in first and second language acquisition. I will briefly outline the nature of this evidence below.

1.1. Variation

Zuraw (2003) provides evidence that language change can result from probabilistic inference on the part of listeners, and that probabilistic reasoning could explain the maintenance of lexical regularities over historical time. Individual variations in SLA can also be explained by probabilistic factors such as how often they are exposed to certain linguistic phenomena in particular educational settings. Variations in input characteristics could be determined by such probabilistic factors as frequencies and order of presentation of language items in a particular syllabus or materials such as textbooks or course modules.

1.2. Frequency

One striking clue to the importance of probabilities in language comes from the wealth of frequency

effects that pervade language representation, processing, and language change (Bod, Hay, and Jannedy 2003:3). This is true for SLA. Frequent words are recognized faster than infrequent words (Jurafsky 1996). Frequent words in input are also more likely to be used by learners than infrequent words (Tono 2002). Frequency affects language processes, so it must be represented somewhere. More and more scholars come to believe that probabilistic information is stored in human brains for helping automatic processing of various kinds.

1.3. Gradience

Many phenomena in language may appear categorical at first glance, but upon closer inspection show clear signs of gradience. Manning (2003) shows that even verb subcategorization patterns should better be treated in terms of gradients, as there are numerous unclear cases which lie between clear arguments and clear adjuncts (ibid: 302). Rather than maintaining a categorical argument/adjunct distinction and having to make in/out decisions about such cases, we might instead represent subcategorization information as a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability (ibid: 303). This type of analysis is readily applicable to cases in L2 acquisition. A strong claim can be made regarding the gradient nature of language in terms of a probability distribution over linguistic phenomena based on the comparison between native speaker's corpora and learner corpora.

1.4. Acquisition

Bod, Hay, and Jannedy (2003) claim that "adding probabilities to linguistics makes the acquisition easier, not harder" (ibid: 6). Generalizations based on statistical inference become increasingly robust as sample size increases. This holds for both positive and negative generalizations: as the range and quantity of data increase, statistical models are able to acquire negative evidence with increasing certainty. In formal L2 classroom settings, it is also very likely that learners will be exposed to negative evidence as well. Instructed knowledge of this type could serve to form a part of probabilistic information in a learner's mind besides actual exposure to primary data, which facilitates the processing of certain linguistic structures more readily than others.

1.5. What does the evidence show?

The evidence above seems to indicate that a probabilistic approach will be very promising in theory-construction in not only linguistics but also second language acquisition. It could be argued that corpus linguistics should provide a very strong empirical basis for this approach. By analyzing various aspects of learner language quantitatively and at the same time integrating the results of the observations into the

probabilistic model of learning, we could possibly produce a better picture of L2 learning and acquisition. In the next section, I will further explore this possibility and introduce one of the most promising statistical approaches, the Bayesian statistics and network modeling as an underlying principle of language acquisition.

2. Integrating probabilities into SLA theory

One of the strengths of corpus linguistics is its data-driven nature. The findings are supported by a large amount of attested language use data. This feature has been increasingly highlighted as more and more electronic texts become available online. This dramatic increase in the size of available corpus data has also changed the way that people use statistics. Traditional mathematical statistics have been replaced by computational statistics, involving robust machine-learning algorithms and probabilistic inferencing on large-scale data. This shift toward more data-centered approaches should also be applied in the formulation of SLA theory by using learner corpora. By working on large amounts of learner data using probabilistic methods, it is possible to create a totally new type of learning model. In the following sections, I will introduce Bayesian network modeling as the basis of such probability theories and discuss how to view the acquisition theory from Bayesian viewpoints.

2.1. Bayes' theorem

Let me briefly describe Bayes' theorem and how it is useful for theory construction in SLA.2 Bayes' theorem is a probability rule, currently widely used in the information sciences to cope with uncertainty from known facts or experience. It serves as a base theory for various problem solving algorithms as well as data mining methods. Baye's theorem is a rule in probability theory that relates to conditional probabilities. Conditional probability is the probability of the occurrence of an event A, given the occurrence of some other event B. Conditional probability is written P(A|B), and is read "the probability of A, given B". Bayes' theorem relates the conditional and marginal probabilities of stochastic events A and B and is formulated as in (1):

(1) $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Each term in Bayes' theorem has a conventional name as in (2):

(2) a. $P(A)$ is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B.

b. $P(A|B)$ is the conditional probability of A, given B. It is also called the posterior pro bability because it is derived from or depends upon the specified value of B.

c. $P(B|A)$ is the conditional probability of B given A.

d. $P(B)$ is the prior or marginal probability of B, and acts as a normalizing constant.

Bayes' theorem can also be interpreted in terms of likelihood, as in (3):

(3) $$P(A \mid B) \propto L(A \mid B)P(A)$$

Here L(A|B) is the likelihood of A given fixed B. The rule is then an immediate consequence of the relationship P(B | A) = L(A | B). In many contexts the likelihood function L can be multiplied by a constant factor, so that it is proportional to, but does not equal the conditional probability P. With this terminology, the theorem may be paraphrased as in (4):

(4) $$posterior = \frac{likelihood \times prior}{normalizing}$$

In words, the posterior probability is proportional to the product of the prior probability and the likelihood.

An important application of Bayes' theorem is that it gives a rule regarding how to update or revise the strengths of evidence-based beliefs in light of new evidence a posteriori. This is a kind of probabilistic formulation of our daily activities. In a sense, we make a judgment about everything at every moment in our lives; things we are going to do next, things we are going to say, how we evaluate the things we see or hear etc. Every human judgment, whether conscious or unconscious, is influenced by our prior probability of the events (i.e. past experiences or personal beliefs), adjusted by some likelihood of the events, given the new data (i.e. likelihood), which yields a posteriori probability (i.e. new ideas or something learned). Therefore, Bayes' theorem can be viewed as a probabilistic model of human learning. The architecture of human cognitions will be modular and need specifications in their own right, but the overall learning algorithm can be explained in Bayesian terms.

There are a growing number of researchers in different disciplines of sciences who adopt the Bayesian model as a theoretical basis. While Bayes' rule itself is quite simple and straightforward, it is very flexible in the sense that the same rule and the procedure can be adapted to varied sample sizes, from a very small to a huge set. Unlike frequentist probability, Bayesian probability deals with a subjective level of knowledge (sometimes called 'credence', i.e. degree of belief). This is intuitively more likely as a model of human learning, because we all have personal beliefs or value systems on which every decision is based. Some of these subjective levels of knowledge are formed via instructions in specific social and educational settings in a country. The levels of knowledge about what is appropriate in what situations are also partially taught and partially learned through experiences. In Bayesian terms, every time pepole are exposed to new situations, they learn from new data and revise their posterior probability including their belief system. I argue that the exactly same process is also applied to the acquisition and the use of second language.

2.2. Bayesian theory in SLA

How could we realize Bayesian modeling in SLA? The overall picture is rather simple. Since Bayes' theorem itself is a formulation of 'learning from experience', in other words, obtaining posterior probability by revising prior probability in light of new attested data, we could give a model of language learning using recaptured Bayes' rule in (3), as in (5):

(5) (a revised system given the new data) $\propto$
(likelihood) × (an old system of language)

What is promising is that corpus-based approaches will suggest a very interesting methodological possibility in providing input for these empty arguments in the model in (5). For example, if we compile a corpus of learners at different proficiency levels, we could formulate the model in such a way that probability scores for given linguistic items obtained at a certain proficiency level (Stage x, for instance) serves as prior probability and the scores obtained at the next level (Stage x+1) will serve as the condition for posterior probability, as in (6).

(6) (Language at Stage x+1) $\propto$ (Likelihood) × (Language at Stage x)

While a real picture would be much more complex than above, Bayesian reasoning can still provide a very interesting possibility of describing a model of SLA from a probabilistic viewpoint. In order to illustrate this point, let us take the acquisition of verb subcategorization frame (SF henceforth) patterns as an example. Suppose we are interested in the occurrence of a particular SF pattern of a verb, we might try to represent subcategorization information as a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability. For instance, we might estimate the probability of SF patterns for a verb get as in (7):

(7) $P(\text{NP}_{[\text{SUBJ}]} \_\_ \mid \text{V= } get)$ = 1.0
$P(\text{NP}_{[\text{SUBJ}]} \_\_ \text{NP}_{[\text{OBJ}]} \mid \text{V= } get)$ = 0.377
$P(\text{NP}_{[\text{SUBJ}]} \_\_ \text{ADJ} \mid \text{V= } get)$ = 0.104
$P(\text{NP}_{[\text{SUBJ}]} \_\_ \text{PP} \mid \text{V= } get)$ = 0.079
$P(\text{NP}_{[\text{SUBJ}]} \_\_ \text{NP}_{[\text{OBJ}]} \text{PP} \mid \text{V= } get)$ = 0.056
$P(\text{NP}_{[\text{SUBJ}]} \_\_ \text{NP}_{[\text{OBJ}]} \text{NP}_{[\text{OBJ}]} \mid \text{V= } get)$ = 0.053
(Note: other constructions are omitted; probabilities are derived from the British National Corpus.)

So, for instance, the probability of choosing the SF pattern "get up" can be described by modeling the probability that a VP is headed by the verb "get", and then the probability of certain arguments

surrounding the verb (in this case, SUB + get + PART[up]), as in (8):

(8) P(VP → V[*get*] PART[*up*]) = P(VP[*get*] | VP) × P(VP[*get*] → V PART | VP[*get*]) × P(PART[*up*] | PART, VP[*get*]).

The probabilities in (8) can be computed from corpora and the formal grammatical description can be given by yet another stochastic language model called a DOP model, described in 4.2.

If such probabilistic descriptions for the choice of verb SF patterns can be extracted from learner corpora, this information can then be integrated into a general probabilistic inference system. Suppose we wish to reason about the difficulty in acquiring verb SF patterns by L2 learners of English. Let M be the misuse of a particular subcategorization frame pattern of the given verb, allowing "yes" and "no." For explanatory purposes, let possible causes be J: the match in subcategorization pattern between English and L1 Japanese, with P(J = yes) = 0.5, and T: Textbook influence (Whether the same subcategorization pattern occurs in the textbook as a source of input), with P(T = yes) = 0.2. We adopt the following hypothetical conditional probabilities for the correct use:

(9) P(M = yes | J = no, T = no) = 0.7
P(M = yes | J = no, T = yes) = 0.4
P(M = yes | J = yes, T = no) = 0.3
P(M = yes | J = yes, T = yes) = 0.1

The left hand diagram in Figure 1 shows a directed graphical model of this system, with each variable labeled by its current probability of taking the value "yes".

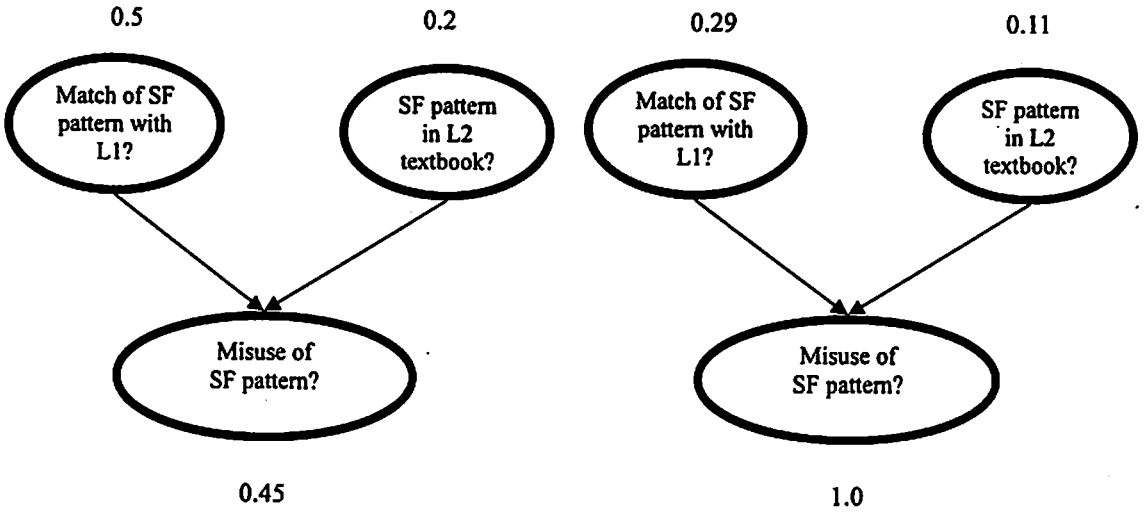

Figure 1: Directed graphical model representing two independent potential causes of the misuse of a verb SF pattern, with probabilities of a 'yes' response before and after observing the misuse.

Let me describe how to obtain probability scores in Figure 1 in more detail. Suppose you observe the learner corpus data and found the correct use of the SF pattern get up, and you wish to find the conditional probabilities for J and T, given this correct use. By Bayes' theorem,

$$(10) \qquad P(J, T \mid M = yes) = \frac{P(M = yes \mid J,T)P(J,T)}{P(M = yes)}$$

The necessary calculations are laid out in Table 2. Note that, owing to the assumed independence, P(J, T) = P(J)P(T). Also P(M = yes, J, T) = P(M = yes |J, T)P(J,T), and when summed this provides P(M = yes) = 0.45.

Table 2. Calculations of probabilities for possible causes J and T

J [*P*(J)]	no [0.5]		yes [0.5]		
T [*P*(T)]	no [0.8]	yes [0.2]	no [0.8]	yes [0.2]	
P(J,T)	0.4	0.1	0.4	0.1	1
P(M=yes|J,T)	0.7	0.4	0.3	0.1	
P(M=yes,J,T)	0.28	0.04	0.12	0.01	0.45
P(J,T|M=yes)	0.62	0.09	0.27	0.02	1

By summing the relevant entries in the joint posterior distribution of J and T we thus obtain P(J = yes | M = yes) = 0.27 + 0.02 =0.29 and P(T=yes | M = yes) = 0.09 + 0.02 = 0.11. These values are displayed in the right-hand diagram of Figure 1. Note that the observed misuse has induced a strong dependency between the originally independent possible causes.

We now extend the system to include the possible misuse of another SF pattern of the given verb, denoted by M2, assuming that this verb pattern is not dealt with in the textbook and that

$$(11)\ P(M2 = yes \mid T = yes) = 0.2$$
$$P(M2 = yes \mid T = no) = 0.8$$

so that P(M2 = yes) = P(M2 = yes | T = yes)P(T=yes) + P (M2 = yes | T = no)P(T = no) = 0.2 × 0.2 + 0.8 × 0.8 = 0.68. The extended graph is shown in Figure 2:

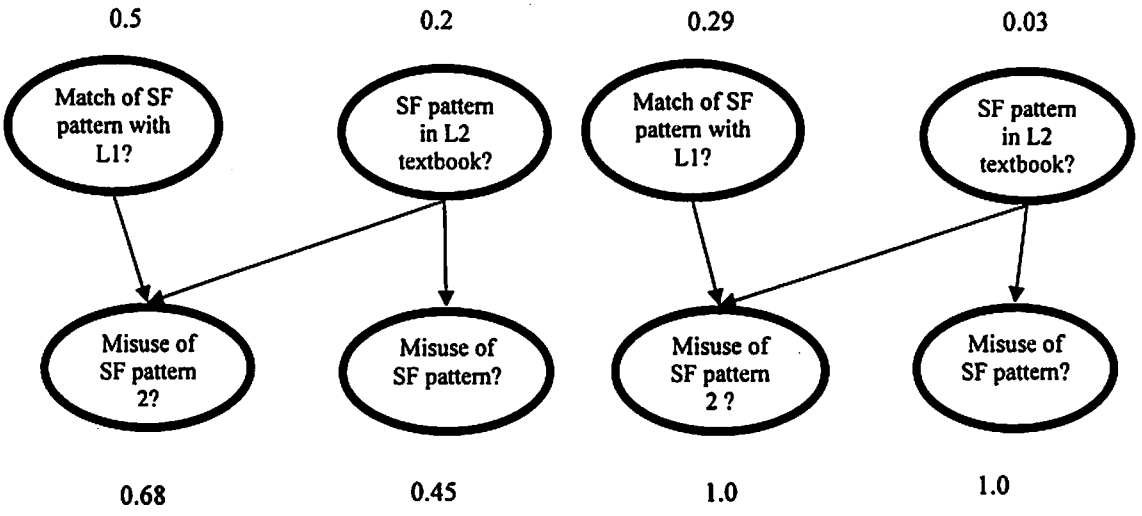

Figure 2: Introducing another misuse of an SF pattern into the system, before and after observing that neither of the patterns are correctly used.

Suppose we now find that the other SF pattern is wrongly used (M2 = yes). Our previous posterior distribution P(J, T | M=yes) now becomes the prior distribution for an application of Bayes' theorem based on observing that the second SF pattern has failed. The calculations are displayed in Table 3.

Table 3. Calculations of probabilities after observing another misuse of SF patterns

J [*P*(J)]	no [0.5]		yes [0.5]		
T [*P*(T)]	no [0.8]	yes [0.2]	no [0.8]	yes [0.2]	
P(J,T|M=yes)	0.62	0.09	0.27	0.02	1
P(M2=yes|J,T, M=yes)	0.8	0.2	0.8	0.2	
P(M2=yes,J,T|M=yes)	0.496	0.018	0.216	0.004	0.734
P(J,T|M=yes, M2=yes)	0.676	0.025	0.294	0.005	1

We obtain P(J=yes|M=yes, M2=yes) = 0.299, P(T=yes|M=yes, M2=yes) = 0.03. Thus, observing another misuse of SF patterns has decreased the chance of the influence of English textbooks on the use of verb SF patterns. This ability to withdraw a tentative conclusion on the basis of further information is extremely difficult to implement within a system based on logic, even with the addition of measures of uncertainty. In contrast, it is both computationally and conceptually straightforward within a fully probabilistic system built upon a conditional independence structure. Although the example shown above is rather limited in scope, it is possible that we can add more variables to the model and apply exactly the same procedure to obtain probabilistic inference from the observed data.

The above example has heuristically argued for the explanatory power of probabilistic models based on Bayesian reasoning. I have informally introduced the idea of representing qualitative relationships between variables by graphs and superimposing a joint probability model on the unknown qualities.

When the graph is directed and does not contain any cycles, the resulting system is often called a Bayesian network. Using the terms introduced earlier, we may think of this network and its numerical inputs as forming the knowledge base, while efficient methods of implementing Bayes' theorem form the inference engine used to draw conclusions on the basis of possibly fragmentary evidence.

The above example assumes that the random variables involved are discrete. However, the same formula holds in the case of continuous variables (or a mixture of discrete and continuous variables), as long as, when M is continuous (e.g. instead of yes/no, the accuracy rate of SP patterns in a certain learner group), we interpret P(M) as the probability density of M. See Pearl (1988) for more work related to this.

3. Advantages of the Bayesian SLA model

So far we have seen how Bayesian networks can be adopted for describing phenomena in SLA. By using Bayesian reasoning, we could possibly define the whole framework of SLA as one realization of an expert system. An expert system consists of two parts, summed up in the equation:

(12) Expert System = Knowledge Base + Inference Engine.

The knowledge base contains the domain specific knowledge of a problem. It is a set of linguistic descriptions of a language. For this, I assume a DOP model, which will be described in detail in 4.2. The inference engine consists of one or more algorithms for processing the encoded knowledge of the knowledge base together with any further specific information at hand for a given application. It is similar to what cognitive scientists call "declarative vs. procedural" knowledge.

It is definitely important to define the knowledge base properly, for the knowledge base is the core of an expert system. For this, probabilistic information from a large amount of learner data will be very useful. Linguistic features with probability scores will be stored in the knowledge base in each of the language domains such as phonology, morphology, syntax, semantics and lexicon. The probabilistic data of linguistic features from each stage of learners will be used to form the input for the Bayesian network described in the previous section. Additional information or variables will constantly change the posterior probability, which will subsequently be used as the prior distribution for the new input.

This whole picture of obtaining new posterior probability assumes to be similar to what is happening cognitively in the brain. The Bayesian network model of SLA will have a strong explanatory power for human cognition. The following section will describe how a probabilistic view should be dealt with precisely in a formal language theory and introduce the basic notion of a Data Oriented Parsing (DOP) model as a candidate for such a theory. This model will help better integrate probabilistic information of a language into the acquisition model based on Bayesian reasoning.

4. Implementation: a Data Oriented Parsing model

As we integrate Bayesian network modeling into SLA theory construction, it is necessary to look for a framework for linguistic description. Some people (Manning 2003, for example) claim that probabilistic syntax can be formalized within existing formal grammatical theories such as the stochastic Optimality Theory. I argue that it would be desirable to look for a more data-driven approach as a theoretical framework. For this purpose, the Data Oriented Parsing (DOP) model proposed by Bod (1993, 1998) seems to be very promising. Here I will outline the basic features of DOP and discuss the possibilities of analyzing learner language within this framework.

4.1. Probabilistic grammars

Before explaining DOP, let me briefly describe probabilistic grammars in general. Probabilistic grammars aim to describe the probabilistic nature of a large number of linguistic phenomena, such as phonological acceptability, morphological alternations, syntactic well-formedness, semantic interpretation, sentence disambiguation, and sociolinguistic variation. (Bod 2003: 18)

The most widely used probabilistic grammar is the probabilistic context-free grammar (PCFG). PCFG defines a grammar as a set of phrase structure rules implicit in the tree bank (phrase structure trees) with probabilistic information attached to each phrase structure rule. Let us consider the following two parsed sentences in (13). We will assume that they are from a very small corpus of phrase structure trees:

(13) a. (S (NP John) (VP (V gave) (NP Mary) (NP flowers))).
b. (S (NP Mike) (VP (V gave) (NP flowers) (PP (P to) (NP Mary)))).

Table 4 gives the rules together with their frequencies in the Treebank.

Table 4: The rules implicit in the sample Treebank and their probabilities

Rule	Frequency	PCFG Probability
S → NP VP	2	2/2 = 1
VP → V NP NP	1	1/2 = 1/2
VP → V NP PP	1	1/2 = 1/2
PP → P NP	1	1/1 = 1
NP → John	1	1/6 = 1/6
NP → Mike	1	1/6 = 1/6
NP → Mary	2	2/6 = 1/3
NP → flowers	2	2/6 = 1/3
V → gave	2	2/2 = 1
P → to	1	1/1 = 1
Total	14	

This table allows us to derive the probability of randomly selecting the rule S → NP VP from among all rules in the Treebank. The rule S→ NP VP occurs twice in a sample space of 10 rules; hence its probability is 2/10 = 1/5. We are usually more interested in the probability of a combination of rules (i.e., a derivation) that generates a particular sentence. For this, we compute the probability by dividing the number of occurrences of rules involved in the derivation of a certain sentence by the number of occurrences of all rules. Note that this probability is actually the conditional probability P(structure A | structure B), and thus the sum of the conditional probabilities of all rules given a certain non-terminal to be rewritten is 1. The third column of Table 4 shows the PCFG probabilities of the rules derived from the Treebank.

Let us now consider the probability of the derivation for John gave Mary flowers. This can be computed as the product of the probabilities in Table 4, that is, 1 (S→NP VP) × 1/6 (NP→John) × 1/2(VP→NP NP) × 1(VP→gave) × 1/3(NP→Mary) × 1/3(NP→flowers) = 1/108. Likewise, we can compute the probability of John gave flowers to Mary: 1 × 1/6 × 1/2 × 1 × 1/3 × 1 × 1/3 = 1/108.

What is important in these probabilistic formalisms is that the probability of a whole (i.e., a tree) can be computed from the combined probabilities of its parts. The problem of PCFG is its derivational independence from previous rules, since in PCFG, rules are independent from each other. For example, if we consider a larger Treebank, it surely contains various derivational types of prepositions: P → to; P → for; P → in, and so forth. The probability of observing the preposition to, however, is not equal to the probability of observing to given that we have first observed the verb give. But this dependency between give and to is not captured by a PCFG.

Several other formalisms, such as head-lexicalized probabilistic grammar (Collins 1996; Charniak 1997) and probabilistic lexicalized tree-adjoining grammar (Resnik 1992), tried to capture this dependency and the Data Oriented Parsing model (Bod 1993, 1998) is one of such models. A DOP model captures the previously mentioned problem dependency between different constituent nodes by a subtree that has the two relevant words as its only lexical items. Moreover, a DOP model can capture arbitrary fixed phrases and idiom chunks, such as to take advantage of (Bod 2003: 26).

4.2. A DOP model

I have no room to elaborate on a DOP model in detail here, but let me provide a simple example of how a DOP model works. If we consider the examples in (13a): John gave Mary flowers, we can derive from this treebank the following subtrees:

(14) (S (NP John) (VP (V gave) (NP Mary) (NP flowers)))
(S (NP) (VP (V gave) (NP Mary) (NP flowers)))
(S (NP John) (VP (V) (NP Mary) (NP flowers)))

(S (NP John) (VP (V gave) (NP) (NP flowers)))
(S (NP John) (VP (V gave) (NP Mary) (NP)))
(S (NP) (VP (V) (NP Mary) (NP flowers)))
(S (NP) (VP (V gave) (NP) (NP flowers)))
(S (NP) (VP (V gave) (NP Mary) (NP)))
(S (NP John) (VP (V) (NP) (NP flowers)))
(S (NP John) (VP (V) (NP Mary) (NP)))
(S (NP John) (VP (V gave) (NP) (NP)))
(S (NP) (VP (V) (NP) (NP flowers)))
(S (NP) (VP (V) (NP Mary) (NP)))
(S (NP) (VP (V gave) (NP) (NP)))
(S (NP John) (VP (V) (NP) (NP)))
(S (NP) (VP (V) (NP) (NP)))
(VP (V) (NP Mary) (NP flowers))
(VP (V gave) (NP) (NP flowers))
(VP (V gave) (NP Mary) (NP))
(VP (V) (NP) (NP flowers))
(VP (V) (NP Mary) (NP))
(VP (V gave) (NP) (NP))
(VP (V) (NP) (NP))
(V gave)
(NP John)
(NP Mary)
(NP flowers)

Note: In actual DOP models in Bod (2003), all the subtrees are written in tree diagrams.

These subtrees form the underlying grammar by which new sentences are generated. Subtrees are combined using a node substitution operation similar to the operation that combines context-free rules in a PCFG, indicated by the symbol "o". Given two subtrees T and U, the node substitution operation substitutes U on the leftmost nonterminal leaf node of T, written as T o U. For example, John gave Mary flowers can be generated by combining three subtrees from (14) as shown in Figure 3:

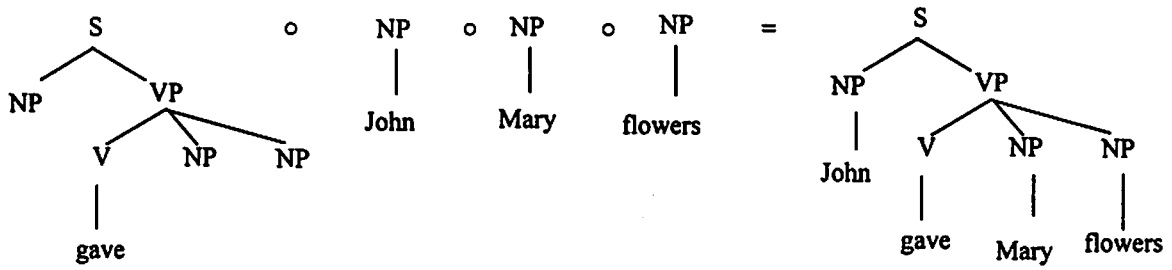

Figure 3: Generating *John gave Mary flowers* by combining subtrees from (18)

The events involved in this derivation are listed in Table 5.

Table 5. The probability of a derivation is the joint probability of selecting its subtrees

Event
(i) selecting the subtree [$_S$ NP [$_{VP}$ [$_V$ gave] NP NP]]] from among the subtrees with root label S,
(ii) selecting the subtree [$_{NP}$ John] from among the subtrees with root label NP,
(iii) selecting the subtree [$_{NP}$ Mary] from among the subtrees with root label NP,
(iv) selecting the subtree [$_{NP}$ flowers] from among the subtrees with root label NP.

The probability of (i) in Table 5 is computed by dividing the number of occurrences of the subtree [S NP [VP [V gave] NP NP]]] in (14) by the total number of occurrences of subtrees with root lable S: 1/16. The probability of (ii) is equal to 1/3, and the probabilities of (iii) and (iv) are also equal to 1/3 respectively.

The probability of the whole derivation is the joint probability of the four selections in Table 4. Since in DOP each subtree selection depends only on the root label and not on the previous selections, the probability of a derivation is the product of the probabilities of the subtrees, in this case 1/16 × 1/3 × 1/3 × 1/3 = 1/432.

The DOP model outlined here does not exhibit a one-to-one correspondence between derivation and tree, as PCFG does. Instead, there may be several distinct derivations from the same tree. The probability that a certain tree occurs is the probability that any of its derivations occurs. Thus the probability of a tree is the sum of the probabilities of its derivations. Bod (2003) points out that this means that in DOP, evidence for a tree accumulates: the more derivations a tree has, the larger its probability tends to be (ibid: 30). For further detail, see Bod (1992, 1998, 2003). He argues that language users store arbitrarily large sentence fragments in memory, and continuously and incrementally update its fragment memory given new input. Although language users cannot remember everything, they will initially store everything, as they process language input, and calculate the frequencies in order to accumulate the frequency information of subtrees. Thus, the DOP model fits very well with the Bayesian way of thinking described earlier. Another advantage is that the DOP model can effectively handle dependency problems that other models have. Since the model proposes that language users store sentence fragments in memory and that these fragments can range from two-word units to entire sentences, language users do not always have to generate or parse sentences from scratch using the rules of the grammar. Sometimes they can productively reuse previous heard sentences or sentence fragments. This will make language processing and language learning very easy, dealing with the problem of how to incorporate idiom principles (Sinclair 1991) or lexicalized sentence stems (Pawley and Syder 1983) into a stochastic model of language.

5. Integrative perspectives of SLA as a stochastic language model

I have shown that learner corpus research has come to a turning point, where a statistical linguistic analysis of learner language needs to be integrated into a formal theory of language acquisition. As an

example, I have proposed a DOP model as a promising model of a probabilistic grammar. I have also argued that the entire SLA process can be explained by Bayesian reasoning.

Since this paper gave just a brief sketch of a new probabilistic model of SLA and how learner corpora play an important role there, it would be desirable to make a specific proposal as to how probabilistic data from learner corpora can form input for Bayesian network models. For this, not only a specific model of L2 acquisition of, for example, verb SF patterns, but also the general picture of L2 acquisition processes in the context of stochastic theories of human learning needs to be explored. Second language learners' use of a particular expression in a language depends on their intended meanings, various contextual as well as situational settings. All the choices made are affected by the prior probability of L2 learners' knowledge or belief. Once a certain string of a sentence is produced, then that string becomes a part of prior probability, leading to the prediction of what is going to be said next. This knowledge of a language is also influenced by many other factors, including learners' L1 knowledge, age, sex, cognitive maturity, L2 instructions, motivation, and exposure to L2 input, among others. There are too many variables to consider, thus it would be extremely difficult to estimate the true value of L2 competence of a particular L2 learner. That is why we use Bayesian modeling in describing a complicated system of L2 acquisition. Instead of hoping to identify an exact state of abilities in L2, Bayesian methods enable statements to be made about the partial knowledge available (based on data) concerning 'state of nature' (unobservable or as yet unobserved) of L2 competence in a systematic way using probability as the yardstick.

The strength of Bayesian network modeling is that the same mathematical procedure can be applied to very small data or to larger, multilevel data. We can start from a very simple model in SLA and build up the model into a relatively complex one, without changing any statistical assumptions. The same Bayesian methods will always work on new data.

It is also important to place Bayesian networks in a wider context of so-called highly structured stochastic systems (HSSS). Many disciplines, such as genetics, image analysis, geography, marketing, predictions of El Niño effects among others, adopt this as a unifying system that can lead to valuable cross-fertilization of ideas. Within the artificial intelligence community, neural networks are natural candidates for interpretation as probabilistic graphical models, and are increasingly being analyzed within a Bayesian statistical framework (see Neal 1996). In natural language processing, Hidden Markov models can likewise be considered as special cases of Bayesian networks (Smyth et al. 1997). By defining a model of SLA as a stochastic system, using probabilistic information from a large body of learner language production data at different levels of proficiency or developmental time frames, we could possibly describe and explain the SLA process from an innovative viewpoint.

6. Future directions

In this paper, I have argued that learner corpus research should be able to make a significant contribution

to a probabilistic view of SLA theory based on Bayesian reasoning. This research is still at the preliminary stage of theory construction, and I need to further exploit the possibility of applying Bayesian networks for SLA modeling, but it would suffice to say that such an attempt will have a great potential and that learner corpora would play a significant role there. Further studies need to be done to provide specific procedures of building a stochastic model of SLA.

Notes

This paper is based on my article in Baker (2009) with a slight modification.

References

Baker, P. (ed.) (2009), *Contemporary Corpus Linguistics*. Continuum.

Bod, R. (1992), 'Data-Oriented Parsing', in *Proceedings of COLING 1992*, Nantes, France, pp. 855-859.

Bod, R. (1998), *Beyond Grammar: An Experience-based Theory of Language*. Stanford, Calif.: CSLI Publications.

Bod, R. (2003) 'Introduction to elementary probability theory and formal stochastic language theory', in Bod, Hay, and Jannedy (2003) (eds.) pp. 11-38.

Bod, R., Hay, J. and Jannedy, S. (2003) *Probabilistic Linguistics*. Cambridge, Mass: MIT Press.

Charniak, E. (1997) 'Tree-bank grammars', in *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI '96)*, Menlo Park, Calif., pp.1031-1036.

Collins, M. (1996) 'A new statistical parser based on bigram lexical dependencies', in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, Calif., pp. 184-191.

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., and Spiegelhalter, D.J., (2007) *Probabilistic Networks and Expert Systems*. New York: Springer.

Jurafsky, D. (1996) 'A probabilistic model of lexical and syntactic access and disambiguation.' *Cognitive Science* 20, 137-194.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-Based Language Studies*. London: Routledge.

Neal, R. (1996) *Bayesian Learning for Neural Networks*. New York: Springer Verlag.

Pawley, A. and Syder, H. (1983) 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency', in J. Richards and R. Schmidt (eds.), *Language and Communication*. London: Longman, pp. 191–226.

Pearl, J. (1988) *Probabilistic Inference in Intelligent Systems*. San Mateo, California: Morgan Kaufmann.

Sinclair, J.McH. (1991) *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Smyth, P., Heckerman, D., and Jordan, M.I. (1997) 'Probabilistic independence networks for hidden Markov probability models.' *Neural Computation*, 9, 227-269.

Tono, Y. (2002) *The Role of Learner Corpora in Second Language Acquisition and Foreign Language Learning: The Multiple Comparison Approach*. An unpublished Ph.D. thesis. Lancaster University.

Zuraw, K. (2003) 'Probability in language change' in Bod, Hay, and Jannedy (2003) (eds.), pp. 139-176.