CHAPTER

11

Integrating Learner Corpus Analysis into a Probabilistic Model of Second Language Acquisition

Yukio Tono

This chapter considers recent research directions that have taken place in the field of learner corpora. I explore how multifactorial analytical techniques that use log-linear analyses can help to identify the extent to which different individual factors (and combinations of factors) influence the output of learners. Additionally, I show how probability-based theories such as the Bayesian approach can be used to explain the second language acquisition process. Using examples, I show how Bayesian probability theory enables statements to be made based on the partial knowledge available (e.g. patterns in corpus data) regarding as yet unobserved L2 competence. Finally, I outline the basic features of the Data Oriented Parsing model (another probabilistic model) and discuss the possibilities of analysing learner language within this framework.

11.1 Introduction

A learner corpus is a collection of texts, normally essays, produced by people who are learning a second or foreign language. Learner corpora have grown into one of the major types of specialized corpora in corpus linguistics. As James (1992: 190) notes 'The really authentic texts for foreign language learning are not those produced by native speakers for native speakers, but those produced by learners themselves.'

Learner corpora are therefore compiled with several different purposes in mind. First, they are used for providing information on learners' common errors for producing reference materials such as dictionaries or grammars (see, e.g. Gillard and Gadsby 1998). Two commercial learner corpora, the Longman Learners' Corpus (LLC) and the Cambridge Learner Corpus (CLC) have been used for these purposes. Second, learner corpora can be used for describing the characteristics of interlanguage produced by second language (L2) learners. The International Corpus of Learner English (ICLE) (Granger 1998) is a project which aimed to identify those features (overuse/underuse phenomena as well as errors) which are common to learners with different first language (L1) backgrounds and those which are unique to each L1. As well as comparing native speakers of different L1s, learner analysis can also carry out comparisons of learners at different levels of

proficiency, see for example, research carried out on the NICT JLE (National Institute of Information and Communications Technology Japanese Learner of English) Corpus (Izumi et al. 2003) or the JEFLL (Japanese EFL Learner) Corpus (Tono 2007). Third, learner corpora can be used to provide classroom language teachers with an opportunity to monitor their students' performance in the framework of Action Research. Action Research is a reflective process of progressive problem solving led mainly by the teacher herself in a specific classroom context. Action Research usually takes place in a collaborative context with data-driven analysis (Johns 1997) or research designed to enable predictions about personal (e.g. teacher or student) and organizational (e.g. school or education board) change. For this kind of analysis, the use of small sets of learner corpora sampled in a specific learning context has become increasingly popular. For more information on types and features of learner corpora, see Pravec (2002).

This chapter will argue that research using learner corpora has come to a turning point methodologically and theoretically. Methodologically, it is necessary to integrate multiple variables concerning learners and learning environments into the analysis of learner corpora. Theoretically, learner corpus researchers need to consider how much contribution they could make to second language acquisition (SLA) theory construction as they investigate the patterns of use in L2 in terms of frequencies and distributions of learner errors and over/underuse phenomena. With this aim in mind, I will first discuss the complex nature of learner corpus research and the importance of multifactorial research design. Then I will introduce how probabilistic analysis can contribute to the formulation of linguistic and acquisition theories and discuss the possibility of theoretical reformulation of problems in SLA in light of probability theories. Finally, I will propose a framework of SLA theory using the concept of the Bayesian network as an underlying theoretical assumption and the Data Oriented Parsing (DOP) model as an attempt to integrate learner corpus findings into a probabilistic model of SLA.

11.2 Recent Challenges for Learner Corpus Research

As more and more learner corpora become available and the findings based on the analysis of such corpora are published, it is becoming clear that there are discrepancies between the approaches taken by learner corpus researchers and mainstream SLA researchers. While researchers working on learner corpora tend to look at overall patterns of language use (and misuse) across proficiency (e.g. native speaker vs non-native speaker; different ability levels of learners) and describe similarities and differences between the groups, SLA researchers have a tendency to set specific hypotheses to test against the data in order to identify causal relationships between variables. SLA researchers usually have specific theoretical frameworks on which their hypothesis is built upon while learner corpus researchers tend not to have such a theory a priori and often pursue data-driven, theory-generating approaches. The corpus-based approach to SLA has emerged among corpus linguists and not SLA specialists, which makes learner corpus research look rather weak in terms of theoretical perspectives. However, learner corpus

researchers have argued that it is worth accumulating facts about SLA processes first before moving onto theory construction.

The problem is how can this be achieved? There are groups of researchers who set specific hypotheses within a certain framework of SLA theories and use learner corpora where appropriate to test these hypotheses. Such people do not always have sufficient knowledge about corpus design criteria and may lack technical skills in extracting necessary linguistic observations from the corpus. Other groups of people working in SLA research construct their own corpus data which suit their needs and methodological assumptions. It is often the case, however, that their corpora are limited in size from a standard corpus linguistic point of view and therefore generalization beyond the corpus under examination is difficult. Most of the studies so far exhibit such methodological shortcomings.

It is also difficult to relate corpus findings to actual educational settings. Corpus building/analysis projects such as ICLE largely ignore the educational contexts in each country and they assume that the findings could be applicable for advanced learners of English in general. This is reasonable as long as the performance of advanced learners is relatively stable and less vulnerable to a specific learning environment in each country. In the case of younger or less advanced learners, however, observed data are heavily dependent upon the nature of input and interactions in the classroom. For example, the order of acquisition could be a reflection of the order of instructions given to learners. It would be ideal to relate corpus findings to specific input sources such as textbooks used in the class or classroom observation data (cf. the SCoRE Project in Singapore; Hong 2005). Therefore, one future direction will be to integrate educational and contextual as well as linguistic variables into corpus analysis, together with specific SLA hypotheses in mind. In dealing with multiple variables, a careful statistical treatment should be made to ensure internal and external validity. In the following section, I will illustrate such an approach in further detail.

11.3 Multifactorial Analyses and Beyond

Tono (2002) investigated the acquisition of verb subcategorization patterns by Japanese-speaking learners of English. This issue is related to the acquisition of argument structure and has been discussed with reference to both L1 and L2 acquisition (Pinker 1989; Juffs 2000). Tono compiled a corpus of free compositions written by beginning to intermediate level learners of English in Japan, called the JEFLL (Japanese EFL Learner) Corpus. The JEFLL Corpus consists of English compositions written by approximately 10,000 students, ranging from the first year of junior high school to the third year of senior high school (Year 7 to 12), totalling a little less than 700,000 tokens. Composition tasks were strictly controlled: all compositions were written in the classroom without the help of a dictionary. Additionally, a twenty-minute time limit was imposed. Students chose one of six topics (based on three narrative and three argumentative composition questions).

What makes Tono's (2002) design multifactorial? The primary goal of his research is to identify relative difficulties in acquiring different verbs in terms of the use of their subcategorization frame (SF) patterns.[1] A simple picture of the design would be illustrated as follows:

(1) Independent variable: proficiency level (defined for the purpose of the study as 'years of schooling')
Dependent variable: use of verb SF patterns

However, a more wide-reaching analysis would take into account (a) whether only one verb or multiple verbs are being examined, (b) types of verb SF, and (c) distinctions between use and misuse. Thus, (1) should be rewritten as (2):

(2) Independent variables:
a. proficiency level
b. types of verbs
c. types of SF patterns
Dependent variable: frequencies of use vs misuse of SF patterns

Tono's study also took into account L1 influence, L2 inherent factors and input from the classroom. L1 influence was defined as the similarities in SF patterns between the English verbs under study and their translation equivalents in Japanese. For L2 inherent factors, Levin's (1993) verb semantic categories were used to classify the verbs in the study. The influence of classroom input was defined as the amount of exposure to specific verbs and their SF patterns in the textbooks, measured by their frequencies in the English textbook corpus. Thus, the overall relations among multiple variables are shown in (3):

(3) Independent variables:

- Learner factor: (a) Proficiency level (Factor 1)
- L1 factors: (b) Degree of similarity/difference in SF patterns between L1 Japanese and L2 English (Factor 2)
(c) Frequencies of SF patterns in English (Factor 3)
(d) Frequencies of SF patterns in Japanese (Factor 4)
- Input factor: (e) Frequencies of SF patterns in the textbooks (Factor 5)

Dependent variable: Frequencies of use vs misuse of SF patterns (Factor 6)

Since many of the factors described in (3) are categorical or nominal data, we need to deal with multi-way frequency tables. For this, Tono employed log-linear analysis. The term 'log-linear' derives from the fact that one can, through logarithmic transformations, restate the problem of analysing multi-way frequency tables in terms that are very similar to ANOVA. Specifically, one may think of a multi-way frequency table as reflecting various main effects and interaction effects that add together in a linear fashion to bring about the observed table of frequencies.

Table 11.1 The results of log-linear analysis for the verb *get* (from Tono 2002).

Verb	The final model by log-linear analysis
get	643, 543, 532, 432, 61, 1

Log-linear analysis can also be used for evaluating the relative importance of each independent variable against the dependent variable. By taking a model fitting approach with backward deletion using the saturated model, we could reduce the number of independent variables to the most parsimonious sets of variables. Table 11.1 shows the results of log-linear analysis performed on the use of the verb *get*. Each number in the table consists of the factors which interact with each other. For example, 643 signifies that there is a significant interaction between factors 6, 4 and 3.

The table shows that the verb *get* can be best explained by a model consisting of the three-way interaction effects of Factors 6-4-3, 5-4-3, 5-3-2 and 4-3-2 and the two-way interaction effects of Factors 6 and 1, with the main effect being Factor 1. This shows that there is a significant effect of school year (Factor 1) and interaction effects between school year and use/misuse (Factors 6 and 1). Factor 6 is also related to frequencies of subcategorization frame (SF) patterns of the Japanese-equivalent verb *eru* (Factor 4) and degrees of matching in SF patterns between English and Japanese equivalents (Factor 2). This kind of analysis makes it possible to identify which factors play a significant role in explaining the complex interactions of multiple variables. The results show the following interesting findings. See Tono (2002) for further details:

(4) a. There is a significant relationship between school years and frequencies of use of SF patterns in major verbs.
b. Frequencies in SF patterns have a strong correlation with frequencies in English textbooks, which means students use more verbs with various SF patterns as they are exposed to more variations in the textbooks.
c. Despite the findings in (b), there is no significant correlation between frequencies of 'correct' use of SF patterns and frequencies in textbooks. The amount of exposure does not ensure correct use of the forms.
d. The factors affecting the use/misuse of the SF patterns are mainly cross-linguistic factors, such as degrees of similarities in SF patterns between the target language and the mother tongue, or frequencies of SF patterns in the first language (Japanese).

Following Tono (2002), other studies have emphasized the value of multi-factorial corpus analysis as a methodological innovation (see, e.g. Gries 2003). Second language acquisition is a multi-faceted phenomenon, composed of complex factors related to learner's cognitive and affective variables, environmental variables

(types of formal instruction, school settings, curriculums, educational policies of the country), as well as linguistic (L2) or cross-linguistic (L1 vs L2) factors. Therefore it is sensible to adopt an integrative approach which takes multiple variables into account when tackling specific problems in SLA using corpus-based methodologies. Also since the primary information from learner corpora is based on frequencies and distributions of language features in interlanguage, it would be important to consider how the findings can fit into existing SLA theory. Below I argue that corpus-based approaches can help to shed light on theoretical perspectives in SLA. The following section outlines a new framework in learner corpus research, bridging the gap between purely descriptive studies using learner corpora and theoretical perspectives in SLA.

11.4 Motivating Probabilities

In recent years, a strong consensus has emerged that human cognition is based on probabilistic processing (cf. Bod et al. 2003). The probabilistic approach is promising in modelling 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 also holds for various phenomena in L2 acquisition and could have a significant impact on theory construction in SLA. I will briefly outline the nature of this evidence below.

11.4.1 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 (historic) time. Individual variations in SLA can also be explained by probabilistic factors such as how often learners 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.

11.4.2 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 et al. 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 have come to believe that probabilistic information is stored in human brains to assist automatic processing of various kinds.

11.4.3 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 comparisons between native speaker's corpora and learner corpora.

11.4.4 Acquisition

Bod et al. (2003: 6) claim that 'adding probabilities to linguistics makes the acquisition easier, not harder'. Generalizations based on statistical inferences 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.

11.4.5 What Does the Evidence Show?

The evidence above seems to indicate that a probabilistic approach will be very promising in theory-construction, not in only linguistics but also second language acquisition. It could be argued that corpus linguistics can provide a very strong empirical basis for this approach. By analysing 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, Bayesian statistics and network modelling as an underlying principle of language acquisition.

11.5 Integrating Probabilities into SLA Theory

One of the strengths of corpus linguistics is its data-driven nature: findings are supported by a large amount of attested language use data. This feature has been

increasingly highlighted as electronic texts become increasingly available online. The 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 towards more data-centred 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 modelling as the basis of such probability theories and discuss how to view acquisition theory from Bayesian viewpoints.

11.5.1 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. Bayes' 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 (5):

(5) $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Each term in Bayes' theorem has a conventional name as in (6):

(6) 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 probability* 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 (7):

(7) $$P(A|B) \propto L(A|B)P(A)$$

Here $L(A|B)$ is the likelihood of *A* given fixed *B*. The rule is then an immediate consequence of the relationship $P(B \mid A) = L(A \mid B)$. 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 (8):

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

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 judgement 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 and so on. Every human judgement, 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 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 have adopted 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 people are exposed to new situations, they learn from new data and revise their posterior probability including their belief system. I argue that exactly the same process is also applied to the acquisition and the use of second language.

11.5.2 Bayesian Theory in SLA

How could we realize Bayesian modelling in SLA? The overall picture is 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 based on Bayes' rule in (7), as in (9):

(9) (a revised system given the new data) ∝
(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 (9). 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) serve as the prior probability, while the scores obtained at the next level (Stage x + 1) will serve as the condition for posterior probability, as in (10).

(10) (Language at Stage x+ 1) ∝ (Likelihood) × (Language at Stage x)

While the real picture would be much more complex than above, Bayesian reasoning can still provide the interesting possibility of describing a model of SLA from a probabilistic viewpoint. In order to illustrate this point, let us now come back to the example of verb SF pattern acquisition. 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 (11):

(11)

$P(NP_{[SUBJ]}$	$\mid V=$ *get*)	= 1.0
$P(NP_{[SUBJ]}$	$NP_{[OBJ]} \mid V=$ *get*)	= 0.377
$P(NP_{[SUBJ]}$	ADJ $\mid V=$ *get*)	= 0.104
$P(NP_{[SUBJ]}$	PP $\mid V=$ *get*)	= 0.079
$P(NP_{[SUBJ]}$	$NP_{[OBJ]}$ PP $\mid V=$ *get*)	= 0.056
$P(NP_{[SUBJ]}$	$NP_{[OBJ]}$ $NP_{[OBJ]} \mid V=$ *get*)	= 0.053

(Note: Probabilities are derived from the British National Corpus. Other constructions are omitted.)

So, for instance, the probability of choosing the SF pattern *get up* can be described by modelling 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 (12):

(12) $P(VP \rightarrow V[get]\ PART[up]) = P(VP[get] \mid VP) \times$
$P(VP[get] \rightarrow V\ PART \mid VP[get]) \times P(PART[up] \mid PART, VP[get])$.

The probabilities in (12) 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 Section 11.6.

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 Japanese 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:

(13)
$$\begin{aligned} &P(M = \text{yes} \mid J = \text{no}, T = \text{no}) &&= 0.7 \\ &P(M = \text{yes} \mid J = \text{no}, T = \text{yes}) &&= 0.4 \\ &P(M = \text{yes} \mid J = \text{yes}, T = \text{no}) &&= 0.3 \\ &P(M = \text{yes} \mid J = \text{yes}, T = \text{yes}) &&= 0.1 \end{aligned}$$

The left-hand diagram in Fig. 11.1 shows a directed graphical model of this system, with each variable labelled by its current probability of taking the value 'yes'.

Let me describe how to obtain probability scores in Fig. 11.1 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,

(14)
$$P(J, T \mid M = \text{yes}) = \frac{P(M = yes \mid J,T)P(J,T)}{P(M = yes)}$$

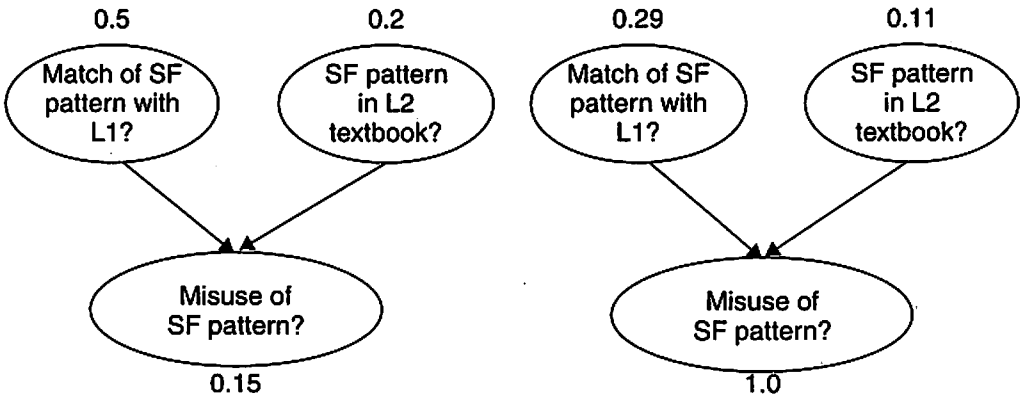

Figure 11.1 Directed graphical model representing two independent potential causes of the misuse of a verb SF pattern, with probabilities of a 'yes' response before and after observing the misuse.

The necessary calculations are laid out in Table 11.2. 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 $P(M = yes) = 0.45$.

By summing the relevant entries in the joint posterior distribution of J and T we thus obtain $P(J = yes \mid M = yes) = 0.27 + 0.02 = 0.29$ and $P(T = yes \mid M = yes) = 0.09 + 0.02 = 0.11$. These values are displayed in the right-hand diagram of Fig. 11.1. 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

(15)
$$\begin{aligned} P(M2 = yes \mid T = yes) &= 0.2 \\ P(M2 = yes \mid T = no) &= 0.8 \end{aligned}$$

so that $P(M2 = yes) = P(M2 = yes \mid T = yes)P(T=yes) + P(M2 = yes \mid T = no)P(T = no) = 0.2 \times 0.2 + 0.8 \times 0.8 = 0.68$. The extended graph is shown in Fig. 11.2.

Table 11.2 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

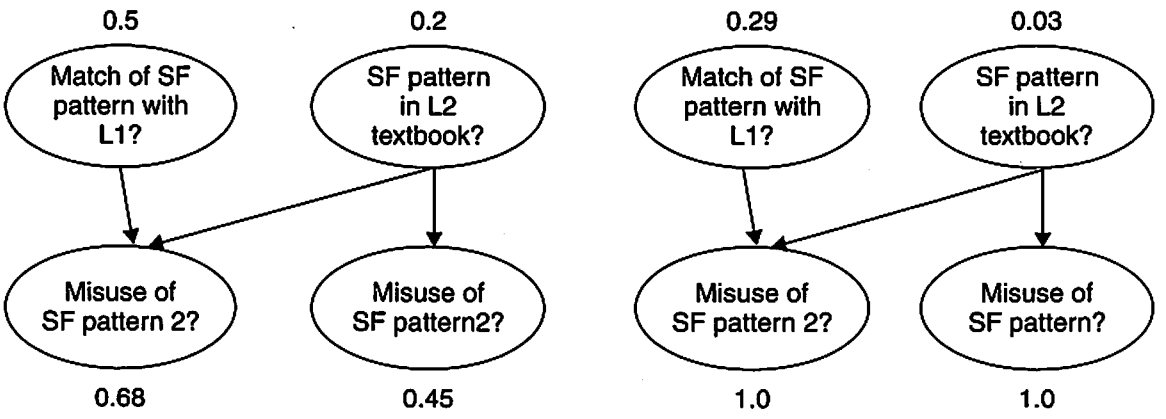

Figure 11.2 Introducing another misuse of an SF pattern into the system, before and after observing that neither of the patterns is correctly used.

Suppose we now find that the other SF pattern is incorrectly 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 11.3.

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,[3] 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.

Table 11.3 Re-calculations of probabilities after observing another misuse of the SF pattern.

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

11.5.3 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:

(16) 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 the following section. 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.

As the knowledge base is the core of an expert system, it is important to define it properly. To achieve this, probabilistic information obtained from a large amount of learner data will be 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 is assumed 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 DOP model as a candidate for such a theory. This model will help to better integrate probabilistic information of a language into the acquisition model based on Bayesian reasoning.

11.6 Implementation: A DOP Model

As we integrate Bayesian network modelling 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. I argue that it would be desirable to look for a more data-driven approach as a theoretical framework. For this purpose, the DOP model proposed by Bod (1992, 1998) seems to be promising. Here I will outline the basic features of DOP and discuss the possibilities of analysing learner language within this framework.

11.6.1 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 (17). We will assume that they are from a very small corpus of phrase structure trees:

(17) 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 11.4 gives the rules together with their frequencies in the Treebank.

Table 11.4 allows us to derive the probability of randomly selecting the rule S → NP VP from among all the rules in the Treebank. For example, the rule S→ NP VP occurs twice in a sample space of 10 possible rules, so 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

Table 11.4 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**	

the sum of the conditional probabilities of all rules given a certain non-terminal to be rewritten is 1. The third column of Table 11.4 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 11.4, 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), have tried to capture this dependency and the DOP model (Bod 1992, 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).

11.6.2 A DOP Model

There is no space 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 example in (17a): *John gave Mary flowers*, we can derive from this treebank the following subtrees:

(18) (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 (18) as shown in Fig. 11.3.

The events involved in this derivation are listed in Table 11.5.

The probability of (1) in Table 11.5 is computed by dividing the number of occurrences of the subtree $[_S$ NP $[_{VP}$ $[_V$gave] NP NP]] in (12) by the total number of occurrences of subtrees with root label S: 1/16. The probability of (2) is equal to 1/3, and the probabilities of (3) and (4) are also equal to 1/3 respectively.

The probability of the whole derivation is the joint probability of the four selections in Table 11.4. 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 $1/16 \times 1/3 \times 1/3 \times 1/3 = 1/432$.

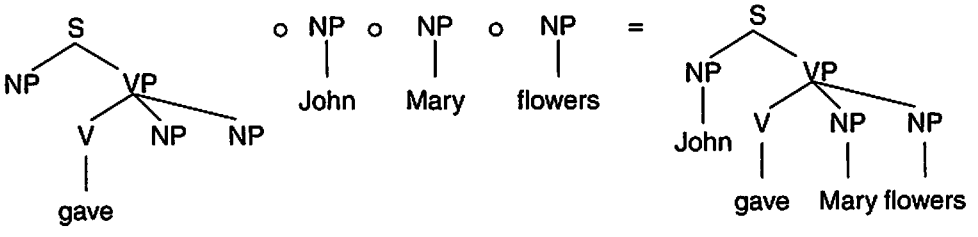

Figure 11.3 Generating *John gave Mary flowers* by combining subtrees from (18).

Table 11.5 The probability of a derivation is the joint probability of selecting its subtrees.

Event
(1) selecting the subtree [$_S$ NP [$_{VP}$ [$_V$ gave] NP NP]] from among the subtrees with root label S,
(2) selecting the subtree [$_{NP}$ John] from among the subtrees with root label NP,
(3) selecting the subtree [$_{NP}$ Mary] from among the subtrees with root label NP,
(4) selecting the subtree [$_{NP}$ flowers] from among the subtrees with root label NP.

The DOP model outlined here does not exhibit a one-to-one correspondence between derivation and tree, as with the PCFG. 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 previously 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.

11.7 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 modelling 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 modelling 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 analysed 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.

11.8 Future Directions

In this chapter, 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 we need to further exploit the possibility of applying Bayesian networks for SLA modelling, 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

1 Verb subcategorization patterns are sometimes expressed in several different terms with slightly different connotations, such as verb patterns, verb complementation patterns, verb valency patterns or argument structures.
2 For a general introduction to Bayesian networks, see Cowell et al. (2007) and Sivia (2006).
3 Cycles in graph theory are loops starting from one node, forming arches among several nodes, and coming back to the same node again. This is not allowed in one-directional cause-effect relationships and is thus a necessary condition for a Bayesian network.