SSLA, **24**, 209–213
DOI: 10.1017.S027226310200205X

FREQUENCY IN PRODUCTION, COMPREHENSION, AND ACQUISITION

Robert Bley-Vroman
University of Hawai'i

Although there are certainly observable frequency effects in language, in most cases, there are alternative approaches to explanation that more directly relate to the essential characteristic of language—that it is a system relating form to meaning. For example, for both word choice in production and ambiguity resolution in comprehension, meaning-based approaches can often provide equally satisfying, or more satisfying, explanations. In the meaning-based approach, the statistical structure of the language can affect the development of linguistic knowledge (for example, by influencing acquisition order or providing evidence for developing grammars); however, linguistic knowledge is not itself knowledge of the statistical structure of language. An example is provided of how frequency may relate to grammaticality judgments of nonnative speakers acquiring multiple *wh*-questions.

As Ellis correctly points out, in any sample of language production, certain words and combinations of words are more likely to occur than others; some constructions are more common than others; some words occur more frequently in certain constructions than in others. In comprehension, when confronted with ambiguity, hearers may find one interpretation more likely than others. This much is hardly controversial, but what are we to make of it?

COLLOCATION AND KNOWLEDGE OF FREQUENCY IN PRODUCTION

The words *profound* and *ignorance* seem to go together. In the 200,000 words of Charles Darwin's *The Origin of Species*, over half of the uses of the word

Address correspondence to: Robert Bley-Vroman, Department of Second Language Studies, University of Hawai'i, Honolulu, HI 96822; e-mail: vroman@hawaii.edu.

profound (or *profoundly*) are together with the word *ignorance* (or *ignorant*). (Almost all the others refer to the depths of the sea.) Here is a typical example: "how profoundly ignorant we are in regard to the normal and abnormal action of the reproductive system" (Darwin, 1872, p. 400). Additionally, about a quarter of the uses of the word *ignorant* are together with *profound*. That is, in Darwin's *Origin*, you can make a very good guess that if you see the word *profound*, you will also see *ignorant*; and, if you see *ignorant*, there is a good chance that *profound* will be there with it. In fact, the statistical association of *profound ignorance* (or *profoundly ignorant*) in this work is an exceptionally strong one, among the strongest one ever finds in corpus linguistics. There is clearly a sense in which these two words go together or "collocate." What is the explanation for the association of these two words? Two general approaches suggest themselves. The first is based on frequency; the second is based on meaning.

The first explanation is this: Darwin's knowledge of English (his representation of English) includes statistical information about co-occurrence, as Ellis suggests. In English, the word *profound* frequently occurs with *ignorance*. (Perhaps, more generally, *profound* frequently occurs with mental states and attitudes: *admiration*, *insight*, *shock*, *commitment*, and *dissatisfaction*, to take examples from the Brown University corpus.) The lexical statistics of Darwin's production in *The Origin of Species* reflect Darwin's knowledge of likelihood of co-occurrence, which presumably originally emerged from Darwin's exposure to English. In this approach, the statistical structure of a corpus is essentially its own explanation. Words go together because they go together, and speakers of a language know that they go together, so they put them together.

However, there is another approach to explaining why the word sequence *profound ignorance* occurs frequently in Darwin's work. It is simply this: Darwin believes that our ignorance of the matters under discussion is profound. From this perspective, language provides the means of expressing thoughts and intentions. The phrase *profound ignorance* is Darwin's favored phrase for expressing this communicative intent. The chief reason that *profound* modifies words like *ignorance* or *admiration* more often than it modifies words like *roof* or *telephone* is because of what *profound* means. It makes sense to talk of profound ignorance; it makes less sense to speak of a profound telephone. This is a matter of human cognition and the use of language to express meaning, rather than of calculating word transition probabilities based on the analysis of a corpus. In this approach, the statistical facts are secondary and derivative. Language itself is thought of as a system of expression of ideas and intentions. The statistics of language use follow from the interaction of this system with the communicative intents of the users in particular contexts. To be sure, language production does have a statistical structure, but it is derivative and with little direct explanatory force.

This view does not deny the existence of collocation on a linguistic-conceptual level. It may well be that, for Darwin, "profound ignorance" forms a conceptual unit associated with the particular phrase *profound ignorance*. Why he

should favor this phrase (rather than, say, *great ignorance*) is a complex matter. He may have found it particularly apt or appealing. Perhaps he heard it used sometime and was struck by it. Perhaps it was indeed a commonly used phrase among his contemporaries. It cannot, though, be reduced to a distillation of the statistical properties of language production.

FREQUENCY IN LANGUAGE COMPREHENSION

Ellis provides a useful summary of probabilistic effects in language comprehension. Often, language production contains examples that are linguistically ambiguous. The word *plane* can refer to a carpenter's tool, an airplane, or a flat, two-dimensional surface. This much is certainly part of linguistic knowledge. Furthermore, in context, a hearer normally understands which sense of *plane* is intended. If you hear *the plane left* you are more likely to think of the airplane interpretation. How does this happen, and what is the role of knowledge of frequency? Again, one can imagine two general approaches.

First, assuming that linguistic knowledge includes information about the statistical structure of language, the hearer could make guesses based on known frequencies of words and on interword transition probabilities (and more complex statistical relationships, no doubt). Perhaps the "airplane" sense of *plane* is more frequent overall, or more frequently occurs with the verb *leave*, or both. In the Brown University corpus, *plane* occurs 163 times: in the sense of "airplane" 72 times (44%); as a "two-dimensional surface" 84 times (51%); as a carpenter's tool six times; and as a type of tree once. Assuming that these probabilities are indicative of frequency overall, then, *ceteris paribus*, a hearer might guess a two-dimensional surface first, an airplane second, then a carpenter's plane, and finally a tree.

Ellis suggests that if the word *plane* is followed by *left* there may be a greater frequency of the "airplane" interpretation, presumably because if we look at a large body of text, we will see that the verb *left* (or a form of the verb *leave*) is more likely to occur with an airplane as subject than with a carpenter's tool as subject. Interestingly, an analysis of the 56 million words of the Bank of English shows that the word *left* is not among the more significant collocates of *plane*, in any of the senses of *plane*. The word *left* occurs far too rarely with *plane* (just four times in over 50 million words) and far too often with other words—and conversely—to make any reliable association between *left* and particular senses of *plane*. Nonetheless, we never find *left* with any other sense of plane except airplane. (Just to get a feel for the frequencies we are talking about: It would take several years of being exposed to English for many hours a day to be exposed to 50 million words.) The key aspects of this approach to explanation are that the learner's knowledge of language incorporates information about complex statistical probabilities of occurrences of words and combinations, and that this information is used in language understanding.

The alternative perspective, as in the case of language production, is based

on the notion that language is a syntactic system that represents meaning. Users do not just make guesses based on frequency of word combinations. Resolution of ambiguity depends on the syntactic parser and on what is most likely in context, given what sense the human hearer can make of the words. In the phrase *the plane left*, the airplane interpretation initially makes the most sense. The hearer knows what airplanes are, what carpenter's tools are, and what planar surfaces are. The hearer knows that *plane* can be any of these. The hearer also has a representation of the verb *leave*. If *plane* occurs as subject of *leave*, the hearer figures that the "airplane" interpretation is by far more likely than the others, because airplanes are the sort of things, given their nature, that can "leave," whereas carpenter's tools or planar surfaces are not. (I deliberately ignore the issue of garden-path parsing also raised by some of Ellis's examples, especially as regards choice of active or passive interpretation of *left*; these problems would take us far afield.) Note that this approach is not directly dependent on a knowledge of frequency of words or senses, or of frequency of combinations of words. It relies primarily on the concept of language as part of a system of human understanding. "Likelihood" relies on much more than statistical analysis.

Even if the explanation for the statistical properties of human language production is ultimately a matter of human cognition and of the function of language as the expression of thought and intention, nonhuman devices for dealing with natural language may still be able to take advantage of statistical structure, compensating in part for the inability of such devices to model the complexity of human cognition. Indeed, there are many demonstrations, some referred to by Ellis, of how nonhuman systems use statistical structure as a surrogate for understanding. They really can do quite well.

In summary, for human beings, knowledge of statistical properties of language may not be the chief means of comprehending and producing language. Indeed, the observed statistical structure of a corpus may be not so much evidence of a stochastic knowledge system as an indirect reflection of a system of understanding and expression, operating in context.

FREQUENCY IN ACQUISITION

Suppose we concede that human knowledge of language may not make direct, central use of frequency in production and comprehension. Might there, nonetheless, be a place for frequency in second language acquisition? Here, the arguments for frequency effects are much stronger. In order for something to be acquired, it must be encountered (or deduced from something encountered). Something that does not occur, or occurs only rarely, is, *ceteris paribus*, less likely to be encountered and "noticed" than something that occurs frequently. Put somewhat differently, the target of acquisition is not knowledge of frequency; rather, the (epiphenomenal) statistical structure of the input may affect acquisition.

As an example of how this might work, consider the interpretation of Yoshi-

naga's research on the acquisition of multiple *wh*-questions, as reported by Bley-Vroman and Yoshinaga (2000). In English, multiple *wh*-questions are grammatical when the subject and an argument of the verb are questioned, whether the argument is a direct object or an expression of location, for example, *Who saw what?* or *Who sat where?* A subject and an adjunct adverbial, however, cannot be questioned in a multiple *wh*-question, for example, **Who came why?* or **Who came how?* (Note that multiple *wh*-questions are intended here, not echo questions.) Interestingly, learners of English whose native language is Japanese (in which all these sorts of questions—both argument and adjunct—are equally grammatical) found examples of the type *Who saw what?* to be grammatical, but, in an analysis of judgment patterns, grammatical types such as *Who sat where?* clustered with ungrammatical **Who came why?* and **Who came how?* Why is this? Bley-Vroman and Yoshinaga speculated that subject-object questions might be more frequent in the input than subject-location types. Bley-Vroman (2001), in an analysis of the Bank of English, found that subject-object questions are in fact overwhelmingly the most frequent, with subject-location types being extremely rare, and conjectured that rarity was the reason that subject-location types clustered with ungrammatical examples.

Interestingly, native speakers of English, as anticipated, readily accepted both types, which together form a single cluster in an analysis of native speaker judgments. That is, native speakers are able to treat both types as grammatical, even though one is overwhelmingly more frequent than the other. Perhaps, as proposed by generative theorists of language acquisition, they are deducing the properties of *wh*-questions from other, more robust and prevalent features of the input. At the least, native speakers are categorizing the input based on abstract characteristics, in this case, on the complement-adjunct distinction, rather than on superficial features, in this case, on type of *wh*-word.

None of this is to suggest that frequency is all that matters in second language acquisition. Many things that are encountered only once or very rarely may strike the learner as salient, be noticed and processed deeply, and be incorporated into linguistic knowledge. The mechanisms that are hidden behind the word "salient" remain largely mysterious. The best we can say, given how profoundly ignorant we are in regard to these mechanisms, is that the more often something occurs in the input, the more opportunities there will be for it to be noticed.

REFERENCES

Bley-Vroman, R. (2001, March). *Input frequency, learnability, and parametric variation: The case of multiple* wh*-questions in learners of English*. Paper presented at the third North American Symposium on Corpus Linguistics and Language Teaching, Boston, MA.

Bley-Vroman, R., & Yoshinaga, N. (2000). The acquisition of multiple *wh*-questions by high-proficiency non-native speakers of English. *Second Language Research*, *16*, 3–26.

Darwin, C. (1872). *The origin of species* (6th ed.). London: John Murray.