8

The multiple bases for linguistic structures

ROBERT BERWICK

8.1 The fundamental tension

A cornerstone of "The cognitive basis for linguistic structures" (CBLS) was to highlight Chomsky's famous distinction between (1) knowledge of language and (2) how that knowledge is put to use, emphasizing that these need not, and indeed probably were not, one and the same thing—if only because knowledge of language makes no reference whatsoever to the computational steps involved in recovering or producing structured representations from serial input or output, while how language is put to use must explicitly address these essentials. CBLS stressed that the constraints on representations inherited from linguists' grammars might well be involved in language processing even though the grammars themselves might not be.

In this way, CBLS highlighted a fundamental tension between two very different motivations behind the study of language—the first aiming to model external behavior, at heart, a theory that, in the limit, sets out to predict the next thing a speaker will say or a listener will perceive; the second aiming to model knowledge of language, at heart, a theory of a person's internal state that enables them to acquire any language. How might we reconcile these two quite different motivations?

8.2 External modeling

It is worth noting that much current work in so-called "corpus linguistics," including its computational counterparts, directly reflects this tension, being aligned with the first motivation rather than the second. This alignment is directly reflected in the methods used in corpus studies to choose among alternative models in contrast to the tradition of modern generative grammar. In corpus studies, the "measure of merit" is how well a particular model, usually a statistical one, predicts the sentences of some corpus. While prediction may be defined in several ways, it often takes the form of so-called *cross-entropy*, a measure of how close the statistical model adheres to the "true" one; informally, how much a model of language reduces the uncertainty about the next word, morpheme, or sentence to appear in a relevant sequence of items. Such a measure already presupposes that the goal of language analysis is to determine as closely as possible the true probability distribution p of sentences over a corpus, estimated by observing some finite set of examples drawn from the true distribution p and yielding some probability model m that is an estimate of p. Cross-entropy is then defined as the information-theoretic difference between the true distribution p and the estimated model distribution m, with smaller cross-entropy being better. Such an approach is inherently external.

Methods incorporating the statistical regularities of a corpus extend at least far back as the work of Markov, Zipf, and Shannon. More recently, some researchers have adopted other models, such as simple recurrent neural networks (SRNs); see, e.g., Elman (1990); Christiansen and Chater (1999) for representative approaches. These recent models also adopt a corpus-matching "figure of merit"; as Christiansen and Chater (1999: 168) note, the approach "that has become standard . . . is to train the network to predict the next item in a sequence given previous context." In this way, such analyses quite directly aim to predict the next word in a sentence given some preceding word sequence, and clearly might embrace whatever contextual or behavioral information might be influencing the likelihood of the "next word"—be it linguistic, word association frequency, cultural, or indeed any factor whatsoever. Thus, given a word sequence such as *It was a bright cold day in April and the clock was striking* . . . , such an approach might ascribe quite distinct values as to the likelihood of the next word being *twelve*, as opposed to *thirteen*, with *twelve* receiving a higher probability in nearly all contexts aside from those where, say, one is modeling a reader conversant and attuned to the first sentence of Orwell's *1984*.

8.3 Internal modeling

By way of contrast, from the outset, the "measure of merit" in traditional generative grammar focused on notions like "simplicity" as related to the notion of capturing law-like generalizations ranging over some representation of knowledge of language, where "simplicity" was couched in terms of the size of the grammar used to describe (or generate) knowledge of language, where the grammar uses some particular notational system. On this account, given a particular grammatical system, one grammar would be better than another to the extent that it expresses the same linguistic data more succinctly than another grammar—that is, better in terms of "compressing" the linguistic data. As is familiar, this grammar size could vary considerably depending on the notational system used. As detailed in Berwick (1982, 1985), well-known results from automata theory establish that depending on the grammatical devices chosen (equivalently, their automata-theoretic counterparts), one can often achieve exponential or far greater compression of the same set of linguistic facts simply by choosing a more expressive grammatical system. Even if the language is describable by a simpler system, say, a language generated by a finite-state grammar, it can turn out that a richer formalism, like a context-free grammar or a generative grammar with "movement," can have a much smaller description.

8.4 Some examples

An example is given immediately below, but it is worth examining some other cases here. For instance, Berwick (1982) shows that the *finite* language consisting of just sixteen sentences—eight "active voice" noun–verb strings over a fixed, finite lexicon, e.g., *John has eaten ice-cream, John will eat ice-cream*, etc., plus the corresponding eight "passive voice" strings, e.g., *the ice-cream will be eaten,* etc.—has, quite naturally, a description in terms of a finite-state grammar or a finite-state automaton, because the number of strings is finite. In this case, we can measure the size of the description directly in terms of the number of states and transition arcs in the automaton, or the number of symbols in the finite-state grammar. But this finite language also has a description in terms of the set of "active voice" forms *plus* the addition of a single transformational rule that maps active voice forms to passive voice forms. The bottom line is that this second method of expressing the same facts takes only half the space of the purely finite-state description. To consider another example that will play a role below, extending a result of Ginsburg and Lynch (1976), Berwick (1982) demonstrates that the *only* time that a finite-state description will be just as succinct, or perhaps more so, than the corresponding context-free description of the same finite-state language is when the linguistic relationships in the language can be stated only in terms of what word in the language can follow what other word, that is, in terms of bigrams. If there is *any* relationship that must be stated in terms of whole phrases, for example, the common linguistic fact in English that a subject phrase must "agree" in number and person with a verb phrase, then a description in terms of machinery more powerful than a finite-state machine—a context-free grammar or any more powerful device—will be *exponentially* more succinct than the corresponding finite-state one. Such a result provides a formal "succinctness" counterpart to the well-known results of Chomsky (1956) that finite-state grammars are inadequate in the *weak* generative capacity sense as descriptions of natural languages: finite-state grammars are also inadequate in terms of their ability to describe natural languages compactly.

Finally, a common confusion that arises here in the context of comparing descriptions using grammars (or automata) of increasing power should be noted and dismissed. It has from time to time been suggested, even quite recently (see, e.g., Perruchet and Rey 2005), that since the human brain is finite, it clearly is describable as a finite-state automaton and so it is worth wondering whether one or another device like a context-free grammar (realizable as a push-down stack automaton of arbitrary depth) or some other grammatical framework even deserves consideration. More specifically, such proposals, as in Perruchet and Rey, often advert to the well-known point that the sentences that are "easily parseable" comprise a finite-state language, as in the well-known case of center-embedded sentences, concluding that the best description of internalized knowledge of language must therefore be similarly finite-state. But this conclusion is fallacious. Any physically realizable computational system is finite in this sense, but that does not mean that its best description is finite-state; in general, one arrives at an (at least exponentially) more succinct description if one talks about a laptop computer *as if* it were a Turing machine or a random-access machine, or some other kind of general-purpose computer with unlimited memory. In particular, in the case of constructing a parser for a finite-depth number of center-embedded structures, it has long been known, since at least Chomsky (1963), that it is better to decompose one's description into two parts: (1) a finite-state control; and (2) a (truncated, i.e., depth-limited) push-down stack store. Though the combination is formally still finite-state, this "minimally augmented" finite-state device is much simpler (more succinct) than a description that combines (1) and (2) into a single finite-state automaton. Furthermore, not only does it crisply represent the idealization that it is possible *in principle* for language to contain an unlimited number of such dependencies, it is also easier to extend to the case of "adding" more memory to the push-down store, in line with the augmentation that one might envision of adding one or two more units of short-term memory.

8.5 Succinctness in grammatical theory

It seems less well known that the earliest statement about the role of succinctness in grammatical theories may be found in one of the first works of modern generative grammar, Chomsky (1951), where this criterion is expressed as follows: "Given the fixed notation, the criteria of simplicity are as follows: . . . the shorter grammar is simpler, and among equally short grammars, the simplest is that in which the average length of derivation of sentences is least" (1951: 6). As also shown in Berwick (1982, 1985), the criterion to "compress" grammar size (relative to the descriptive machinery available) so as to prefer shorter grammars reduces to the usual scientific criterion of using the grammar to express *generalizations* with respect to some set of data, *D*, where a *generalization* may be defined as any set of statements that is shorter than the original length of *D*. A familiar example, as noted in Lasnik (2001), is that of the eight basic auxiliary verb sentences of English, e.g., from *John ate* (with 0 auxiliaries); to *John has, John will eat, John is eating* (with 1 auxiliary, either a form of *have*, a modal like *will*, or a form of *be*); to *John has been eating, John will be eating*, etc. (2 auxiliaries); to, finally, *The ice-cream has been being eaten* (3 auxiliaries). These eight separate examples may be described via a *single* grammatical rule roughly in the form,

(1) Auxiliary → (Modal)(Have)(Be)

where the parentheses are a notational device denoting optionality. Since there are then three binary options in the rule (select the item in parentheses or not), this yields eight possible sentences. This rule thus compresses the eight original auxiliary verb sequences described as eight separate rules into a single rule, an enormous gain in succinctness.

But there is much more to this original formulation. In a particularly prescient way, this original statement regarding the criterion for selecting grammars also lends itself to a more modern interpretation that connects to the most recent work on the formal inference, a fact that seems to have gone unnoticed in the literature. Given a set of linguistic data, *D*, say a set of sentences along with their structural descriptions, and a grammar, *G* in some presupposed notational framework generating those sentences, we let $|G|$ denote the cardinality (size) of the shortest encoding of the grammar in terms of the number of symbols it takes to write the grammar down (its length). Further, following the 1951 formulation given above, we let $|D|_G$ denote the total length of derivations of the sentences with respect to this grammar. This component tells us how much the original data has been "compressed" by the grammar (note it is tacitly assumed that the grammar can generate all the sentences that make up the data. If a particular sentence is an "exception" to the grammar, in the sense that the grammar cannot generate it, then we must add in the total length of that sentence to the "data size" component, without any compression. For example, referring back to the auxiliary verb examples, if it had been the case that one particular auxiliary pattern was not describable by the grammar, say, *John could have been being eaten*, then we would have to add the total length of that sentence as is into the overall sum). Given this formulation, then the criterion for finding the "best" grammar reduces to the problem of minimizing the sum, $|G|+|D|_G$. where G ranges over the space of possible grammars in the notational system. More generally, this criterion for finding the "best" grammatical description of a set of linguistic data is called the *minimum description length* (MDL) principle (Rissanen 1989), and over the past twenty years has been applied to problems of grammatical inference. De Marcken (1995) shows how MDL may be used to learn how to discover morphological units, while Brent (1999) applied to it the related problem of word segmentation. Still others have used MDL within specific linguistic frameworks. For example, Villavicencio (2003) applied MDL within Head-driven Phrase Structure Grammar (HPSG), to develop a learning model for child language data. Hsu, Tomblin, and Christiansen (2009) used MDL to decide whether to simply memorize each example of a verb construction type, such as a dative alternation (*send the library a book/ send a book to the library*), as opposed to replacing a list of examples with a rule—the same situation as with auxiliary verb sequences described above, with the key criterion as to whether to use a rule rather than a memorized form being whether the total description length is thereby shortened. However, they apparently do not recognize that the principle they apply is in fact in complete accord with the simplicity metrics of traditional generative grammar.

Further, given the duality between description length and probabilities (Shannon 1951), one can show that the MDL principle does the same work as Bayesian inference methods that attempt to find the most likely grammar G given the data D (Grünwald 2007). Under a Bayesian formulation, we have so-called prior probability assessments of grammars, $pr(G)$, as well as that of the observed linguistic data, $pr(D)$. Given these two probabilities, on a Bayesian view one then attempts to find a grammar G in the space of grammars spanned by the linguistic theory that maximizes the *posterior* probability of G *given* the linguistic data D, $pr(G|D)$. We do this using Bayes' rule, to reformulate the posterior probability as $pr(G)pr(D|G)/pr(D)$. But to find the maximum of this quantity over all possible Gs, it suffices to maximize just $pr(G)pr(D|G)$, ignoring the fixed value of the denominator, $pr(D)$, which in turn (by the duality of description length and probabilities, and taking logarithms), turns out to be the same as minimizing $|G|+|D|_G$ (for details, see de Marcken 1996). It is this Bayesian formulation that has been advanced in several recent attempts to "rationalize" grammar construction, in, for example, child language corpora (Perfors, Tenenbaum, and Regier 2011; see also Dunbar, Dillon, and Idsardi this volume), though apparently without recognizing the connection of the Bayesian viewpoint to the original formulation of simplicity measures in generative grammar. So recast, we have indicated one concrete way to bridge an apparent "disconnect" between modern statistical approaches to induction and the traditional generative linguistic viewpoint, as well as one way to invest linguistic theory with modern statistical tools, a topic to which we return below. It remains to apply this methodology to other current linguistic frameworks, e.g., modern generative grammar in the so-called "Minimalist Program."

8.6 Linguistic theory and modern statistical tools

Putting the topic of Bayesian inference to one side, however, it is still the case that the goal of "corpus matching" need not necessarily align with traditional linguistic notions, for example, those of conventional phrase structure, and it is exactly in such situations that one can illuminate potential interactions between different knowledge sources that conspire to yield the distribution of sentences that are actually observed. To see this in a particularly simple case of the interaction between syntactic and semantic information, following an example of de Marcken (1996), note that a sequence of words such as *walked on ice* has a conventional linguistic analysis as a verb phrase, consisting of the verb *walked* followed by the prepositional phrase *on ice*, in turn a preposition followed by a noun phrase. This may be justified by constituency tests such as topicalization, e.g., *On ice, I walked.* However, one cannot similarly front *walked on* as a single phrase, leaving behind *ice*. Finally, it is clear that *walked on ice* has the properties of a verb phrase, since it may be conjoined with other verb phrases and take verbal modifiers. The important point, though, is that the linguistic analysis is at odds with an observed *statistical* regularity, in part due to the semantics of English, that verbs such as *walk* are more closely linked to prepositions such as *on* than to nouns like *ice*, a fact that can be quantified by observing that the bigram frequency of *walked–on* is quite high as compared to *on–ice* (de Marcken 1995). Thus, a statistical method that attempts to describe language in terms of bigram properties will "greedily" chunk *walk* together with *on*, while a linguistically oriented representation will tend to keep them apart in separate phrases—once again, illustrating the tension between observed linguistic behavior and underlying linguistic representations.

De Marcken indicates two ways to potentially solve this problem, both of some relevance to the issues raised by CBLS and to the matter of resolving the tension between the two views of language described at the outset, since they involve how grammars relate to observed surface regularities. We describe only the first here, which involves replacing a conventional context-free grammar with a system based on X-bar theory or, more radically, bare phrase structure as in Minimalist approaches (the second involves eliminating the context-free grammar entirely in favor of a different representation). De Marcken notes that one might augment a context-free grammar representation for *walked on ice* to correspond more closely to current linguistic frameworks. In this case, drawing on the notions of X-bar theory (or beyond, in the case of Minimalist approaches), de Marcken was the first to introduce the notion of *head* explicitly into a context-free parsing and learning framework. The verb phrase (VP) is replaced by the (complex) symbol <VP, verb>, where *verb* (= *walked*) is the *head* of the phrase (in fact, the symbol VP could just as easily be replaced by <XP, verb> or, as in Minimalist frameworks, <XP, *walked*>). Similarly, the prepositional phrase is relabeled as <XP, *preposition*>, and the noun phrase, <XP, *noun*>. If we use these rules, we have in effect "promoted" the head information up to the phrase level where it can be "seen" by the verb: the preposition *on* is now visible to the verb, and so any collocation regularity is at least expressible in such a system. In this way, de Marcken shows that by aligning the underlying grammatical knowledge into a format closer to that assumed in some current linguistic theories, one arrives at a representation that turns out to be easier to learn. In fact, incorporating lexical head information into parsing has proved to be an important line of inquiry into modern statistically based parsing models for corpora (Collins 1996).

In any case, such "bigram regularities" as indicated by *walk on* would at first seem to be literally impossible on many generative linguistic accounts, as noted by Moro (2008). Writing about the results of recent fMRI experiments demonstrating the reality of "chunking" of word sequences into whole phrases, Moro observes that the indefinite extendibility of a phrase—as is familiar, a noun phrase like *the cat* can be arbitrarily stretched out in terms of words, *the cat that killed the rat*; *the cat that killed the rat that ate the malt*...—implies that the "distance" between two linguistic elements like *the cat* and, e.g., its corresponding verb, say, *ran*, can similarly be arbitrarily extended. If so, there can be no *absolute* requirement that there be two, three, four, or any particular number of words between two linguistically relevant items, what Moro calls a "rigid dependency." In short, if there is phrase structure at all, then the only predicate expressible is whether one phrase is adjacent to another one or not; there are no predicates that "count" two, three, four, . . . phrases, and so no human language that expresses a rule in terms of counting, beyond the notion of one, which reduces to "adjacent to in terms of phrases." In particular, there can be no human language that, say, forms the negation of a rule by inserting a special morphological item exactly three words from the start of a sentence; however, there can be languages with rules that carry out manipulations with respect to the first *phrase* of a sentence, as in the attested examples of so-called "verb-second" languages—like German or Dutch—where verbal morphology can be placed immediately after, hence adjacent to, the first phrase of sentences.

That would leave the apparent evidence of bigram and other, higher-order "statistical regularities" somewhat mysterious. However, this puzzle can be readily resolved as soon as one realizes that the door is left open for the variation in any particular language to interact in such a way as to result in what might be called *derived* regularities, that is, corollaries that result from the interaction of a particular language's more basic constraints. Consider again the *walked on ice* example. English is known as a "head first" language, so in verb phrases the verb comes first, e.g., *walk* in *walked on ice*; and in prepositional phrases, the preposition comes first, e.g., *on*, in *on ice*. Taken together, this leads to the *derived* fact that one will tend to find verbs that take prepositional phrase adjuncts or arguments that follow the bigram format *verb–preposition*, as has already been noted. In fact, we can take such an example even further, and suppose that there are general lexical association factors, familiar from much other psychological work, that would admit the influence of frequency in word sequences such as *cotton clothes* or *walked on*, thereby directing an underlying processor to use such information if possible. Of course, as noted since at least CBLS, such associations are not infallible: thus the typical grouping of *cotton clothes* as adjective–noun can lead one "down the garden path" in *The cotton clothes are made of grows in Mississippi.* Nevertheless, this kind of "multiplying out" of consequences that follow from the more general principles of an internalized grammar to arrive at a transformed set of operating rules for practical perception or production can be applied more generally. In computer science, it often goes under the general rubric of *compilation*: the notion that the programmer will write the instructions for some algorithm in a higher-level language, and those statements will be mapped, often through a succession of intermediate steps, into the step-by-step instructions that the underlying computer must actually follow to arrive at the desired result. Importantly for our discussion, the end result might not bear much resemblance to the original instructions, in the sense that we could easily recover the original instructions by "reverse engineering." Indeed, depending on the actual physical hardware, the "target machine" we intend to run the program on—say, a parallel computer vs a serial one with a very different kind of basic instruction set—the end results could look quite different. If we now think of knowledge of language as the "higher-level language" and the resulting machine instructions as the "knowledge put to use," we arrive at an operational version of Chomsky's original division, and one that is central to the concerns of CBLS. In this case, of course, we have much less understanding of the "target machine"—the neural wetware—on which the original grammar "runs." But the distinction stands: in order to parse or produce sentences efficiently, it seems quite reasonable that the grammar—the knowledge attained after acquiring a language—could look quite different from the "actual" parser and the operations it uses to analyze language. Just as in the case of computer compilers, the parser's actions could include optimizations tailored to the particular language, extra-language contextual information, and low-level (but as yet little known) properties of the neural system.

8.7 Accounting for probabilistic factors in language

Given that there seem to be obvious probabilistic influences in language *use* that arise from a variety of sources, be they lexical bigram collocation frequencies, prosodic information, or the like, how can we combine this evidence in a probabilistically well-founded way? To be concrete, consider our *cotton clothes* example again, where a parser might have to decide whether *cotton* is an adjective or a noun, so determining whether *clothes* begins relative clause or not. We could imagine that we have (ad hoc) "scores" for each hypothesis, features that rate the evidence as to which one of these choices is correct, given some context. For example, it might be that the preceding sentence was about the wool that clothes are made of; and so the "value" of this feature ought to boost the choice of *cotton* as an adjective; on the other hand, the high frequency of *cotton* preceding nouns as an adjective pulls in the other direction. We could also pose this information as a set of constraints: for example, that 80 percent of the time, given the sequence *cotton clothes*, then *cotton* is an adjective; further, that in 10 percent of all sentences that contain *grows* following *clothes*, then *cotton* is a noun, and so forth. How should we combine such feature scores, or constraints, so as to adjudicate between the two possible outcome labels for *cotton*? Should we simply compute some weighted average? Should we just use the most reliable feature score? Note that the features or constraints themselves might overlap or not even be independent, and that the feature scores, being ad hoc, could wind up as arbitrary numbers, and so not correspond to a valid probability distribution at all, which must add up to one over all outcomes.

However, since the late 1950s, it has been known how to do just precisely that kind of scoring combination in a statistically sound way, via a general approach first advanced by Jaynes (1957), though the methodology seems to have been adopted in computational linguistics only later, in the 1990s, with the research of Berger and Della Pietra (1996) and Ratnaparkhi (1996), among others. The basic idea, known as *multinomial logistic regression*, or sometimes *maximum entropy classification/modeling*, is that we can just add up the scores over all features and convert them to probabilities if we first apply a *logistic* transform. We can then use this transformed score, somewhat modified as indicated below, as a combined probability distribution over all features. Suppose that *score* (x,y) denotes the score for a particular classification outcome y (e.g., *cotton* is an adjective; or that the parser should select a relative clause next), given that a feature value is x (e.g., *clothes* follows *cotton*). In general, the score will be the weighted sum of the values returned by each feature, $\Sigma w_i\, f_i(x, y)$, where the weights will be transformed and determined later by training on a given set of data. The logistic transform maps the weighted sum to $1/Z(\lambda_i)\ exp\ \Sigma_i\, \lambda_i\, f_i\, (x, y)$, replacing the weights w with λ, and the key result that can be established is that this yields a valid probability distribution over classification labels c and sentences s, $pr(c,s)$. Further, while there are many distributions that one might pick for $pr(c,s)$, it turns out that if we choose the weights λ_i so as to maximize the probability of a particular classification label c given the sentences s, then this distribution will be the one that has the maximum entropy, that is, the one that is as nearly uniform as possible, while still obeying the constraints imposed by the observed regularities, e.g., that 80 percent of the time, *cotton* is an adjective when followed by *clothes*.

In brief, we can legitimately combine different information sources, even those from other cognitive domains or arising from nonlinguistic constraints on the external world, to estimate patterns of actual language use. For example, maximum entropy models give state-of-the-art performance in real-world applications like part-of-speech tagging in large corpora. A question remains as to the extent of such influences, beyond simple word association patterns like *cotton clothes*. Nonetheless, there is nothing principled that bars the infiltration of such information sources into one's model of language use, while retaining the advantages of the linguist's conventional notion of knowledge of language represented as a grammar.