Inductive General Grammar

Steven Abney

October 8, 2018

> *The only useful generalizations about language are inductive generalizations. Features which we think ought to be universal may be absent from the very next language that becomes accessible .... The fact that some features are, at any rate, widespread, is worthy of notice and calls for an explanation; when we have adequate data about many languages, we shall have to return to the problem of general grammar and to explain these similarities and divergences, but this study, when it comes, will be not speculative but inductive.* (Bloomfield, *Language,* p. 20 [4])

1 Introduction

Eighty-five years after Bloomfield wrote that passage, I believe the time for an inductive general grammar has come, and I would like to put forward a program of linguistic research to develop it. Both the desirability of inductive general grammar, and the means to pursue it, depend on a certain approach to the linguistic enterprise. Even to speak of *inductive* general grammar, with its implicit contrast to *deductive* general grammar, assumes a philosophy of linguistic empiricism, and immediately contrasts with the dominant philosophy within present-day linguistics, which is very much rationalist, even Platonic.

The framework I would like to propose has much more in common with the philosophy of computational linguistics, which is essentially the philosophy of artificial intelligence research generally. A philosophy of linguistics has arisen almost as an afterthought in the context of computational linguistics, but, despite its somewhat accidental nature, it has much to recommend it, and it differs quite starkly from business as usual within mainstream linguistics. However, it is largely unknown, and not always well understood when known, within mainstream linguistics. Nor do I adopt it whole cloth. In particular, present-day computational linguistics has a strong predilection for *tabula rasa* induction that leads to a certain kind of reductionism. In that, I part ways with it.

A philosophy of linguistics is a set of assumptions that provides goals and a foundation for rigorous inquiry. It cannot be proved, by definition. Since it sets the goals for inquiry, one cannot even give empirical evidence to support it. There are grounds for criticism if the assumptions are logically inconsistent or contradict known facts. But otherwise, one either finds the assumptions obvious and compelling, or not. I will try to make the assumptions as clear as I can, and point out the ways in which my assumptions and their logical consequences differ from those of alternative paradigms. I hope that the assumptions become compelling once confusions are dispelled, but I will allow them to stand on their own merits.

2 Empirical particular linguistics

2.1 The linguistic abstraction

Assumption 1 *Language is a mapping between sound and meaning, and the central purpose of a particular grammar is to give an explicit and accurate characterization of that mapping.*

I use **particular grammar** in the classical sense of a synchronic grammar of an individual language, in contrast to **general grammar**, which describes the common properties of, and range of variation across, languages. Particular linguistics is the study of particular grammar and general linguistics, of general grammar.

Corollary 1 *Central questions of particular linguistics are these: For any sentence of the language, what meaning does a native speaker of the language assign to it? And, conversely, for any meaning expressible in the language, what sentences would a native speaker of the language consider to be natural expressions of it?*

To be more precise, what I mean by "a mapping" in Assumption 1 is a many-many relation $R(\mu, \sigma)$, where μ ranges over meanings and σ ranges over sounds, which is to say, spoken sentences. R is equivalent to two functions: $R_1(\mu) = \{\sigma \,|\, R(\mu, \sigma)\}$ is the **production function**, and $R_2(\sigma) = \{\mu \,|\, R(\mu, \sigma)\}$ is the **interpretation function**.[1] $R_1(\mu)$ is what we observe when we ask what sentences a native speaker would consider to be natural ways of expressing meaning μ, and $R_2(\sigma)$ is what we observe when we ask what meaning or meanings a native speaker assigns to sentence σ. A speaker deems σ to be ambiguous just in case $|R_2(\sigma)| > 1$.

This picture is obviously a simplification. It rather blithely assumes that we know what meanings are. It neglects the possibility that different speakers might give different answers to our questions, or that a single speaker might give different answers if asked at different times. It also neglects the possibility that the answers to the two questions might be inconsistent: for example, that a speaker might interpret σ to mean μ but would not include σ among the natural ways of expressing μ. We set such complications aside for the time being.

[1] For the sake of simplicity and familiarity, I have characterized R as a simple relation, so that $R_1(\mu)$ and $R_2(\sigma)$ are sets. This is not meant to preclude a more general account, in which R is a probabilistic relation and $R_1(\mu)$ and $R_2(\sigma)$ are probability distributions.

Let us consider some possible objections to Assumption 1. The first objection is that Assumption 1 simplifies too much. The mathematical relation R is simply a set of meaning-sentence pairs, and language is surely more than that (goes the objection). Let us call the question of how one might rigorously define R the **mathematical question**. The objection is that even a final answer to the mathematical question would leave key questions about language unanswered. The mathematical question treats language at a certain level of abstraction, and it does not address questions that arise at lower levels of abstraction. For example, the **computational question** is the question of how R is computed. What is the range of algorithms by which one could possibly compute meanings from sentences or sentences from meanings? How do they compare in time and space requirements? Are there approximate algorithms that sacrifice perfect accuracy for better profiles of time or space usage? The computational question does not abstract away from algorithms, but it does abstract away from any particular substrate, be it natural or artificial. By contrast, the **psychological question** limits its scope to the human mind; the question is which algorithm, out of all of the possibilities, is the one that humans actually employ. It concerns language at the level of abstraction corresponding to mind, rather than brain. The **neurological question** is the most concrete question; it asks how the human language-processing algorithm is implemented in brain wetware.[2]

Assumption 1 singles out the mathematical question as being *the* linguistic question. To be sure, the term *linguistics* can be used in a broad sense, according to which linguistics encompasses every aspect of language study, including all of the above questions and then some. But the term *linguistics* is more commonly used in a narrower sense, in which linguistics does not encompass but rather contrasts with computational linguistics and psycholinguistics and neurolinguistics. Linguistics in this narrower sense abstracts away from processing altogether, and limits attention to the relation between sound and meaning. This abstraction is so widely adopted in linguistics that I hope I will cause no offence by calling it the **linguistic abstraction**.[3] The linguistic abstraction is legitimate and useful. The question of defining R is crisply defined, and a rigorous definition of R is obviously useful for addressing the computational, psychological, and neurological questions. Assumption 1 is essentially a statement of the linguistic abstraction.

Although the linguistic abstraction is almost universally adopted, there is a widespread reluctance to abstract away from processing entirely. If we abstract away from processing altogether, and consider the definition of R to circumscribe the full scope of inquiry, then any structures we propose serve exclusively as a means to connect μ to σ. If two competing candidates for structure are equally useful at connecting μ and σ, then there are no grounds for choosing between

[2] The discussion invites comparison to Marr's levels of analysis, but it should be noted that the levels of abstraction in the text differ from Marr's.

[3] The distinction between competence and performance will likely come to mind. However, I find that the terms *competence* and *performance* are used in a variety of inconsistent ways by different writers, and even in different passages of Chomsky, who introduced them. They have also accumulated a good deal of connotational baggage. I find it simpler to avoid them.

them.

By contrast, within the dominant paradigm of linguistics, postulated structures are considered to be illegitimate unless they are "real" in the sense of somehow being "grounded" in the mind or brain. **Psychological realism** of this sort represents at least a partial rejection of the linguistic abstraction. One abstracts away from processing, but one requires a correspondence between one's structures and the data structures employed by the processor.

Let us consider what motivates the linguistic abstraction in the first place. Since R_1 and R_2 are largely accessible to observation,[4] we may study R directly, without the necessity of special apparatus or difficulties of interpreting indirect measurements. Understanding the nature of R is a natural preliminary to inquiry into the computational question of how R might be computed, which in turn provides a foundation for tackling the psychological question of determining which algorithm and data structures the mind actually uses. The psychological algorithm and data structures are not directly observable, but the results of the computational inquiry allow one to determine what "signature" each candidate would produce on observable variables, and comparison of predictions to experimental results allows one to draw conclusions about which candidate is most likely.

The unidirectional course of development just sketched is an idealization, and real scientific inquiry is unlikely to be so simple. Nonetheless, imposing a requirement of psychological realism introduces a circularity that makes the picture incoherent. How can we know whether a structure corresponds to a psychological data structure without answering the psychological question first? How can we determine what the psychological data structures are without also addressing the question of psychological algorithms? At least if computer science is any guide, data structures and algorithms go hand in hand, and choosing one determines the other, or at least restricts one's options. What remains then of the linguistic abstraction? If one does not wish to commit oneself to a serious study of psychology, then one appears to be largely limited to asserting that one's structures "make claims" about psychology without having any intention of proving those claims.

I suspect that many who adhere to psychological realism do so, not because they see any advantage to making empty claims, but because they recoil even more violently from the alternative. If we reject psychological realism, then what is at stake in comparison of alternative candidates for linguistic structures is not truth but utility. Utility may suffice for engineering (so the reasoning goes), but not for science.

However, utility is not merely a poor substitute for truth. For a psycholinguist studying language processing, the availability of multiple alternative representations of linguistic structure, differing not in truth but in utility, constitutes an advantage, not a defect. To see why, let us consider an analogy. Suppose we have an object defined by a set of points, and as our analogues of competing

[4] We may include introspection as a form of observation, provided that there is consistency across judges.

structures, let us consider the representation of the object in Cartesian coordinates versus its representation in polar coordinates. It is pointless to ask which representation is correct—they are equivalent, so either they are both correct or neither is correct. The more pertinent question is which is more useful when doing a particular computation. For some computations, Cartesian coordinates are more useful, and for others, polar coordinates are more useful. If a biological system does something well, it probably adopts the more useful representation for the particular computation at hand. Utility is not a dirty word; utility may in fact provide the best route to truth.

Let us consider one last potential objection to Assumption 1. It might seem that the goal, as I have formulated it, admits a trivial solution: to define R, one need only list the pairs (μ, σ). An enumeration would indeed be possible if the domain were finite; but, critically, the domain is *not* finite. R is an infinite set of pairs. Defining it is non-trivial because a satisfactory *definition* must be finite. A finite definition of an infinite set is a **generator**. What we require is a generator for R, and a generator for a language is a **grammar**.

2.2 Hypotheses, fit, and prediction

To be precise, a grammar generates a relation $\hat{R}$ that models or approximates R. In common practice, a grammar G is represented as a set of rules or constraints or the like, and there is a **grammar notation definition** $\mathcal{F}$, also known as a **grammar formalism**, that determines $\hat{R} = \mathcal{F}(G)$. In this way, a particular choice of grammar provides a **hypothesis** $\hat{R}_1$ regarding the empirical production function R_1, and a hypothesis $\hat{R}_2$ regarding the empirical interpretation function R_2.

Assumption 2 *The acid test of grammar quality is the degree to which its predictions are correct.*

In particular, if we draw a sentence σ at random from the entire infinite population of sentences, what is the probability that $\hat{R}_2(\sigma) = R_2(\sigma)$, and if we draw a meaning μ at random from the population of meanings, what is the probability that $\hat{R}_1(\mu) = R_1(\mu)$?

Here it is important to avoid a mistake that is widespread in linguistics, namely, evaluating one's grammar (or model or theory) by its **fit** to the examples that one examined while developing the grammar. Confronted with alternative grammars with comparably good fit to familiar examples, common practice is to resort to claims about "simplicity" or "elegance" or "independent motivation." Now I think there is an appropriate place for considerations of simplicity, but it is not here.[5] Certainly it is premature to appeal to simplicity before even considering the accuracy of the grammar's predictions.

Confusion between fit and prediction is so pervasive in the linguistic literature that genuine grammatical predictions are rarely tested. Fit to a working body of data is not prediction, and it is important not to confuse the two. The

[5] I return to the issue of simplicity below, in section 3.5, and again in section 3.6.

data under consideration comprise a finite set, and coming up with a theory that fits them is too easy. What matters in the end is, not the theory's fit to the finite set of motivating examples, but rather, how well the theory predicts the entire infinite relation between sounds and meanings, the population of sound-meaning pairs.

Prediction and control are arguably the primary goals of science, and of them, prediction is more basic, to the extent that control can be seen as the use of predictions to choose interventions that have particular desired effects. It is commonplace to talk about the predictions of one's grammar on one's data, but that usage actually involves a certain abuse of terminology. The data under discussion were used to *design* the grammar. The data are already known, and one cannot meaningfully "predict" what one already knows. At that point, it is too late for prediction; only hindsight is possible.

Coming up with a theory that fits any finite data set is not hard. What is hard is coming up with a theory that accurately predicts the rest of the infinite relation. One of course cannot examine the entire relation, but there is a straightforward way to get a good estimate of accuracy. Draw a fresh, random sample of pairs from the relation and measure accuracy on the sample. Borrowing terms from machine learning, the set of examples used to develop the grammar is the **training set** and the freshly-drawn sample is the **test set**. Hindsight is 20/20: it is well known that accuracy on the training set overestimates true accuracy. But the grammar's accuracy on the *test* set does give an unbiased estimate of accuracy on the population as a whole. Not only is the estimate unbiased, its confidence interval can be measured and made arbitrarily small by drawing a large enough test set.

To be clear, we know the values of the empirical functions R_1 and R_2 for a finite set of instances (the training set), and we can query their values for freshly drawn instances (the test set), but we do not know, and will never know, their values over their entire domain. In this, R differs from $\hat{R}$. We *do* know the entirety of $\hat{R}$, in the sense that it is rigorously defined by the grammar, and we can compute $\hat{R}_1(\mu)$ and $\hat{R}_2(\sigma)$ for any given meaning μ or sentence σ.

The grammar's fit to the known data is measured by its **training error**, which is the proportion of training instances x that it gets wrong, that is, where $\hat{R}_i(x) \neq R_i(x)$. The quality of the grammar's predictions is measured by its **test error**, which is the proportion of test instances that it gets wrong, and the test error is an unbiased estimate of its **generalization error**, which is the proportion of the population that it gets wrong.

The idea of evaluating grammars based on their predictive performance, as measured by test error, is fundamental to the methodology used in computational linguistics. The methodology is by no means unique to computational linguistics; it is used in all branches of artificial intelligence and data science, as well as applications of machine learning and predictive modeling in a wide range of subject areas. The methodology revolves around **prediction tasks** like the task of predicting human sentence production or interpretation. One defines the domain of instances x and an experimental protocol for obtaining observations $y(x)$. The protocol may take different forms, ranging from a psycholinguistic experiment with sophisticated apparatus and careful instructions to the subject, to the stylebook used to annotate a treebank, to personal judgments made by the author of a paper when labeling example sentences. Instances of the appropriate sort are collected and annotated with the results of observation, producing **labeled data** consisting of pairs $\langle x, y(x)\rangle$. One set of labeled data is designated as the training set, and it is used to construct or refine models. A second, independent, representative sample of instances is drawn and annotated to produce a test set, and performance on the test set provides an objective measure of the relative quality of the models.[6]

2.3 Practical steps

Serious attention to prediction has an important consequence. To achieve any significant degree of predictive accuracy, a grammar must cover all phenomena that occur in sentence σ (or meaning μ), for some significant proportion of randomly drawn sentences σ (resp., meanings μ). This leads immediately to an emphasis on large, systematic grammars, and a need for large, labeled data sets to develop them.

For a grammar defining relation $\hat{R}$, the relevant labeled data sets are samples of pairs (μ, σ). It is useful, but not essential, to include postulated syntactic structures as well. A collection of sentences with syntactic structures is a **treebank**, one with meanings is a **meaning bank**, and one with both is a **semantic treebank**. All three types of resource exist. Examples of meaning banks include the Groningen Meaning Bank [5] and a collection of meaning banks that use Abstract Meaning Representation (AMR) [3]; the latter are notable for the size and activeness of the community involved in their development. Syntactic treebanks do not provide everything we need, but they do provide at least a crude approximation to meaning, and they have the advantage of being available for a wider range of languages. The largest family of treebanks sharing a single, language-universal labeling scheme, is the Universal Dependencies Treebank (UDT) collection [7]. As for an example of a semantic treebank, I may cite the LinGO Redwoods Treebank [8].

As the existence of these resources makes clear, developing and evaluating grammars on the basis of their predictive accuracy is not an unreachable ideal, and the approach I describe is not merely a proposal. It is the standard approach within computational linguistics; and even if systematic grammars have lost their prestige within the currently dominant school of linguistics, such work does continue, especially within "feature grammar" frameworks such as LFG and HPSG. Work within these frameworks straddles the line between computational linguistics and linguistics proper. Practioners have developed relevant resources, such as the Redwoods Treebank just mentioned, and have used them to develop a rigorous, broad-coverage grammar, the English Resource Grammar. The work has been extended to multiple languages, and the underlying

[6] As a practical matter, the test set is not always drawn after the hypotheses to be compared have been constructed, but using the test set in any way in the construction of the hypothesis is an egregious violation of the methodology.

formalism is intended to be language-general.[7]

Before continuing to a discussion of general grammar, however, let us first consider how an approach that adopts Assumptions 1 and 2 differs from the dominant paradigms in linguistics and computational linguistics.

3 Contrasted with the dominant paradigms

3.1 I-language, E-language, and the scientific method

The dominant paradigm in linguistics is one of essentialism, to adopt the term used by Scholz, Pelletier, and Pullum [9]. Essentialists object to Assumption 1 on the grounds that the proper subject matter of linguistics, as they see it, is not a mathematical relation, but a psychological (or even "biological") reality. In their terms, the only subject of interest is "I-language," whereas the mapping between sound and meaning is "E-language," a mere epiphenomenon and distraction.

Rejection of E-language would appear to be a rejection of the linguistic abstraction. Now there is nothing wrong with working at a lower level of abstraction, but that is not what essentialists do. In practice, they do abstract away from processing and hence adopt the linguistic abstraction, despite their nominal rejection of it. As a result, as I have already observed, it is hard to see what content any claims of psychological reality could have.

But putting that criticism aside, the notions of truth and reality that provide the argument in favor of I-language deserve closer examination. I believe that the natural impetus to say something *true* about *reality* can be satisfied without rejecting the linguistic abstraction. Although *truth* and *reality* appear to be synonyms, there is a subtle distinction, in that discussions of truth generally focus on epistemology, and discussions of realness focus on causal explanation. I suggest that the connection is this:

Definition 1 *A theoretical construct is* **real** *if it is an element of a true account of a causal mechanism. An account is* **true** *to the degree that its predictions are accurate over the population.*

The only means we have at our disposal for evaluating the truth of a postulated account of a causal mechanism is the scientific method. The scientific method is this: replace appeals to authority with objective observations, propose hypotheses about the hidden causal mechanisms giving rise to the observations, and evaluate those hypotheses by testing their predictions on future observations. The adoption of the scientific method reflects an implicit recognition that we can never know the truth about contingent matters in any absolute sense. The best we can do is to compare alternative explanations on the basis of their predictions.

[7] Although the LinGO project is largely conducted under the assumptions that I espouse, it remains susceptible to the criticisms I raise below in sections 3.2 and 3.3.

In the scientific method, an explanation is identical to a hypothesis about, or a model of, the causal mechanism that gives rise to the observables. The causal mechanism itself is hidden; its true form is unknown and unknowable. The model as a whole represents the causal mechanism that produces the observations, so, if we accept Definition 1, the elements of the model are real to the degree that the model's predictions on representative test data are accurate. It is quite possible that two models that are very different in structure make equally good predictions, in which case the elements of both models are equally real, despite their apparent incompatibility. The true form of reality, if there is such a thing, is forever unknowable. When we say that something is real, we mean that it accurately captures aspects of the actual causal mechanism. Very different models may be equally real in this sense. And this sense is the strongest sense of *real* that the scientific method makes available to us.

What essentialists refer to as *I-language* is the body of linguistic hypotheses, the proposals about the hidden mechanisms that give rise to what we can observe. What essentialists refer to as *E-language* is the body of observations.[8] To dismiss E-language is to dismiss the scientific method. Without E-language the only way to arrive at I-language is via authority, aesthetics, or mysticism.

In the scientific method, the only access we have to I-language is through E-language. The true form of the causal mechanisms is unknowable, and in fact there is no guarantee that a unique "true form" even exists. If I-language is "more real," in that it models the causal mechanisms, E-language is "truer," in the sense that it is more directly accessible. Our confidence in the truth of a particular model cannot be greater than our confidence in the observations by which we evaluate it, and is often a good deal less, in that it also depends on our imaginativeness at coming up with alternatives to compare the model to.

We do not discover the hidden mechanisms by denigrating the observations. Essentialists have a tendency to be casual, even cavalier about observational data; for example, they feel justified in dismissing or "correcting" some inconvenient data on the grounds that it represents "performance errors." To make genuine progress in uncovering causal mechanisms, one requires objective data. The observations must be done in such a way that any honest observer, regardless of theoretical persuasion, would agree that the recordings of the observations are accurate. Errors do occur, and there are other sources as well of variability in linguistic judgments, but that does not justify "correcting" the judgments by fiat.

It is legitimate to adopt a theory-internal representation in which postulated performance errors have been corrected, and to derive the actual observations via an error process that is part of the model. To evaluate the model, one must describe the error process rigorously, and compare the resulting predictions about the "errorful" data to the actual observations. Such an approach would be legitimate, but the conventional practice of considering the "corrected" data to the exclusion of the actual observations is not legitimate.

[8] I am being somewhat equivocal about whether *E-language* refers to the (infinite) empirical relation R or to a corpus of actual observations. But I do not think that those who use the term generally distinguish between the relation and the corpus.

Even with an explicit error process, variability is likely to remain. The appropriate way to deal with it is to quantify the variability (for example, by measuring the inter-annotator agreement rate), and to refine the experimental protocol (the annotator instructions) to reduce the variability. Variability in itself does not nullify the legitimacy of the prediction-task methodology: observations in the natural sciences are routinely presented with error bars to quantify uncontrolled variability. In our context, the major consequence of variability is that one needs relatively large test sets to compare models if the differences in predictive accuracy are small.

3.2 "Disambiguation"

There is another way that Assumption 1 represents a significant departure from the dominant paradigm. Under Assumption 1, the central human judgments to be modeled are interpretation judgments and production judgments: what does a given sentence of the language mean (R_2), and which sentences are natural expressions of a given meaning (R_1). These judgments correspond to basic exercises in beginning linguistics classes. Drawing a syntax tree largely determines its meaning, and translating an English sentence to predicate calculus more directly represents an interpretation judgment. Conversely, translating a predicate calculus formula μ to English represents the computation of at least one element of $R_1(\mu)$. Even so, the dominant paradigm makes no effort to model interpretation judgments or production judgments.

Readers may find that comment puzzling, so let me give an example to make the point clearer.[9] What is the linguistic status of the following example?

(1) the dog barks

This is not a difficult or questionable example. Example (1) is unambiguously a grammatical sentence of English, which states that a particular member of the species *canis familiaris* vocalizes in the manner that is typical for its species. And yet, almost no conventional grammar predicts this judgment correctly. Virtually every conventional grammar asserts that the example is several ways ambiguous. For example, (1) has a reading as a noun phrase referring to members of a certain class of sailing vessels (having square-rigged fore- and mainmasts but a fore-and-aft rigged mizzenmast) that are associated with dogs. Perhaps their cargo consists of members of *canis familiaris,* or perhaps their cargo consists of andirons or various other hardware items that are used to stop movement. The point is this: a human confidently judges the sentence to be unambiguous, and identifies a single interpretation as correct, whereas the grammar categorizes the sentence as ambiguous, and provides no guidance even about which is the most common or natural interpretation.

One possible response is that example (1) is in fact grammatically ambiguous, but that there is a disambiguation procedure that humans use to choose a

[9] For more detailed discussion, see Abney [1].

single best interpretation. Since disambiguation is a matter of processing, it is beyond the scope of grammatical theory.

Such a response sounds reasonable at first, but it is actually nonsensical. The function R_2 is not intrinsically computational, any more than anything else in linguistics. It assigns a particular set of interpretations to each sentence, but it says nothing about how that assignment is to be computed. Under the linguistic abstraction, any rigorous (and accurate) definition of the function will do. Interpretation judgments are centrally important linguistic judgments, regardless of how they are mentally computed. It is completely immaterial whether R_2 is mentally computed in two steps—first enumerating a long list of possibilities, followed by disambiguation—or directly computed in a single step. Discriminating between those two possibilities is outside our purview, if processing is outside our purview. The linguistic fact remains: (1) is robustly judged to be, unambiguously, a well-formed sentence of English.

A subtler version of the "performance" argument is the following. One might assert that the string *the dog barks* is grammatically ambiguous even if it is not *perceived* as ambiguous. After all, if one sets up the right context, one can get human judges to perceive the alternative interpretations, even if they do not spring to mind in the neutral context. In this view, the grammatical ambiguity of the string reflects its status in an account of competence, whereas the perceived lack of ambiguity is a performance error. I think this argument is also mistaken, even within the conventional paradigm, but it is instructive to consider why.

We have not previously discussed contexts. Obviously, the context can have an effect on the mapping R. Let us represent that dependency by writing R^γ for the sound-meaning mapping that obtains in context γ. We can continue to use R without a superscript to represent the mapping in the null or default context. Then my original point stands: grammars developed in the dominant paradigm fail to model R_2 and R_1, or R_2^γ and R_1^γ for any other context γ. That is, a conventional grammar *does* define a relation $\hat{R}$ between sound and meaning, but that relation does not make good predictions about R^γ for any choice of γ. Specifically, $R_2^\gamma(\sigma)$ is a singleton set, for most contexts γ and sentences σ, but $\hat{R}_2(\sigma)$ is usually a large set. If one responds to that discrepancy by dismissing the singularity of $R_2^\gamma(\sigma)$ as a "performance error," one is essentially claiming that the theory is right; the data are wrong. I trust that the absurdity of such a claim is self-evident.

In fact, it is neither an error nor an accident that $R_2^\gamma(\sigma)$ is a singleton for most sentences σ in most contexts γ. If sentences were genuinely as ambiguous as conventional grammars have it, communication would be impossible.[10] Consider: a speaker has a meaning μ in mind and chooses a sentence $\sigma \in R_1(\mu)$ to express it. The hearer receives σ and must choose a meaning $\mu' \in R_2(\sigma)$. A communication failure occurs if $\mu' \neq \mu$. Assuming the hearer chooses uni-

[10] Treebanks rely critically on the singularity of $R^\gamma(\sigma)$ given only the limited discourse context available in the treebank itself. The existence of large treebanks for numerous languages makes it clear that human judgments concerning *the* unique correct interpretation for a given sentence are very robust.

formly at random from $R_2(\sigma)$ and that $R_2(\sigma)$ contains n meanings on average, the probability of successful communication is $1/n$. Although misunderstandings clearly do occur, they occur relatively rarely, much less than half the time, from which we can conclude that n is significantly less than 2, and certainly far smaller than the rate of ambiguity predicted by conventional grammars.

Rather than dismissing the discrepancy between R_2^γ and $\hat{R}_2$ as a performance error, a better response, within the conventional paradigm, is to concur with my original statement—that a conventional grammar provides a model neither for interpretation judgments nor for production judgments—but to observe that a conventional grammar does provide a model for a different sort of judgment, what we might call **professional grammaticality judgments.** A professional grammaticality judgment is the answer to the question, Can sentence σ mean μ in some context γ? I include the qualification "professional" because the judgments in question may require linguistic expertise, and considerable ingenuity, to construct a suitable context γ.

Professional grammaticality judgments fit comfortably within the prediction-task methodology. Moreover, they are useful, even if our primary goal is to model interpretation and production judgments. In practice, one usually knows not just that some γ exists, but the identity of particular classes of γ that yield a positive judgment, providing us with facts of the form $\mu \in R_2^\gamma(\sigma)$ that contribute to our knowledge of R.

One reason for modeling professional grammaticality judgments instead of the more important interpretation judgments is that, to model interpretation judgments well, it appears that one needs to consider world knowledge and probabilities.[11] If one wishes to avoid tackling world knowledge and probabilities, one can at least model professional grammaticality judgments and contribute in a more limited way to our understanding of R. That reasoning is consistent with the position that I espouse, according to which the central questions are the interpretation and production functions.

To round out the discussion, I would like to point out that there is an alternative version of "disambiguation" that is consistent with the linguistic abstraction. Without making any particular claims about mental procedures, we may linguistically model $R_2(\sigma)$ in two derivational steps. In the first step, a set of meanings (or structures) $\hat{R}_2^*(\sigma)$ is generated. In particular, we may define

$$\hat{R}_2^*(\sigma) \triangleq \bigcup_\gamma \hat{R}_2^\gamma(\sigma).$$

That is, $\hat{R}_2^*(\sigma)$ is designed to model professional grammaticality judgments. In the second step, a weighting is applied to the members of $\hat{R}_2^*(\sigma)$, and $\hat{R}_2(\sigma)$ is defined to contain those meanings whose weight is within ϵ of the maximum weight in the set. The advantage of this approach is that world knowledge and probabilities are needed only for the weighting. It is possible that the two

[11] The success of probabilistic parsers trained on treebanks shows that one can do a good job of modeling interpretation using only probabilities and not world knowledge, but ultimately both are surely necessary.

derivational steps correspond to two separate steps of mental processing, but it is not necessary; that is a question for psycholinguistics.

Since I just used the terms, let me emphasize that **derivation** and **generation** are mathematical terms, not computational terms. They concern only the definition of sets and functions. They contrast with **algorithm** and **procedure**, which concern computation or mental processing.

3.3 Reversibility

Ambiguity is a familiar problem; only my characterization of it as a linguistic rather than computational problem is unconventional. What is not widely familiar is a complementary problem that afflicts the production function under standard linguistic approaches.

An adequate account of the relation between sounds σ and meanings μ must satisfy at least the following criteria:

1. It must define $\hat{R}_2(\sigma)$ for all (spoken) sentences σ,
2. The function $\hat{R}_2(\sigma)$ must accurately predict human interpretation judgments,
3. It must define $\hat{R}_1(\mu)$ for all meanings μ that higher cognition may produce, and
4. The function $\hat{R}_1(\mu)$ must accurately predict human production judgments.

The ambiguity problem is the failure of standard accounts to satisfy criterion (2). A much less familiar problem, which I will call the **reversibility problem**, is the failure of standard accounts to satisfy criterion (3).

The dominant approach to semantics takes as its goal the definition of a model-theoretic value function $[\![\tau]\!]^{M,g}$, where τ ranges over syntactic parse trees, M is a model, and g an assignment of values to variables. Such an approach ignores production entirely. If nothing else, we require a characterization of the space of meanings $\mathcal{M}$ that may be passed across the interface between higher cognition and the linguistic transducer.

Model-theoretic constructs are unsuitable for $\mathcal{M}$: at the very least we require a recursive characterization, effectively a logical calculus, for naming model-theoretic constructs. Fortunately, standard accounts can be transparently restated as translations to a logical calculus. Let us write f for the translation function; $f(\tau) = \mu$ is a logical expression representing the meaning assigned to τ, in the sense that the value of μ under the standard interpretation of the logical calculus, $[\![\mu]\!]^{M,g}$, is equal to $[\![\tau]\!]^{M,g}$. Then $\hat{R}_2(\sigma)$ is the set of $f(\tau)$ where τ ranges over parse trees that represent possible readings of σ. Inversely, we may define $\hat{R}_1(\mu)$ as the set of σ such that $f(\tau) = \mu$ for some parse τ of σ.

Unfortunately, with this clarification, a problem emerges. The standard approach assigns a unique translation $\mu = f(\tau)$ to each parse tree τ, but it pays no attention to the *range* of f. In particular, if μ' is logically equivalent to μ, but $\mu' \neq \mu$, then $\mu' \neq f(\tau)$, and it is in fact quite possible that μ' is absent from the range of f altogether. Such an expression μ' is not ill-formed; it may be produced by higher cognition as input to $R_1(\cdot)$. And yet $f^{-1}(\mu')$ is empty, hence $\hat{R}_1(\mu') = \emptyset$. That is, the semantic account incorrectly predicts that μ' cannot be expressed as a natural-language sentence. Standard approaches make no effort to avoid expressions like μ', so we must assume that they exist; they do indeed arise for practical computational-linguistic systems that interpret sentences by explicitly translating them to a logical calculus.

The reader might suppose we could simply enumerate expressions that are logically equivalent to μ' and collect $f^{-1}(\mu'')$ for each logically equivalent expression μ'', but logical equivalence is undecidable, making that proposal unsatisfactory. There are also problems with computing $f^{-1}(\cdot)$. The computation of $f(\tau)$ typically involves simplifications, in particular beta reductions, but the inverse of beta reduction is infinitely ambiguous.

These issues have been discussed in the computational linguistic literature. Shieber [10] provides a clear statement of the problem; see also Abney [2] and references cited there.

3.4 The scientific method and systematicity

I have argued that the scientific method leads one inevitably to broad-coverage, systematic grammars, inasmuch as only broad-coverage grammars can provide good predictions, as opposed to merely good fit to a body of motivating examples. Proponents of the dominant paradigm will counter by holding up cases where the scientific method does not apparently require systematicity, cases in which theory dictates the choice of specific, rare, critical observations.

A famous example is Dyson, Eddington, and Davidson's empirical confirmation of the general theory of relativity [6]. During the solar eclipse of May 29, 1919, they observed the apparent location of stars whose light passed very close to the sun. The Newtonian model predicted a small deflection of the light because of the gravitional influence of the sun on the photons, whereas the theory of general relativity predicted a slightly larger deflection. The observed deflection was closer to the predictions of general relativity than to the predictions of the Newtonian model, and this confirmation of Einstein's theory made front-page news in the popular press.

The example is instructive, but not in the way it is usually thought. It is instructive, not because it is a paradigmatic example of the application of the scientific method, but precisely because it is *not* business as usual. A critical experiment arises only if we have two theories that are nearly equivalent—not equivalent in form, but in making the same predictions in nearly all cases—and if there is broad concensus that one of the two is correct. The Newtonian model had already been systematically confirmed by countless observations over the course of 250 years, and Einstein intentionally constructed his model of general relativity to make exactly the same predictions as the Newtonian model under all but the most extreme circumstances.

However satisfying (and news-worthy) the dramatic critical experiment may be, it is not business as usual. The bread and butter of science is the 250 years of systematic, large-scale observations that gave us such confidence in the Newtonian model. Few linguistic models have had their predictions seriously examined, and none have had their predictions confirmed as systematically as those of the Newtonian model. As long as that as true, conducting a critical experiment to distinguish between Theory A and variant theory A′ will be of interest only to those who are already committed to Theory A for essentially subjective reasons such as ideology or familiarity.

3.5 Reductionism

My thesis to this point is that predictive power is a necessary condition for explanatory adequacy: the ability to predict outcomes on freshly drawn samples is the acid test for a hypothesis about the hidden mechanism of language. But let me place emphasis on *necessary condition:* predictive power is not a *sufficient* condition for explanation. Adopting the prediction-task methodology does not guarantee inerrancy. Despite its advantages, it does present its own characteristic pitfalls.

Machine-learning methods have proven very effective at prediction tasks, and their very effectiveness has tempted many computational linguists into a *tabula rasa* philosophy in which any human contribution, apart from the design of the learning algorithm, is suspect. That philosophy is most evident in the current trend to deep learning and deep reinforcement learning. It is considered a virtue to build in as few subject-matter assumptions as possible when constructing a learner for a prediction task. Ideally, according to this philosophy, a single undifferentiated network will suffice for any task, whether it be language, vision, credit-card risk prediction, or autonomous driving.

The performance improvements enabled by deep learning methods provide considerable heft to the *tabula rasa* philosophy, and they are indeed impressive. However, they provide a textbook example for my caveat that predictive performance is not sufficient for explanation. Deep learning methods are arguably the most powerful currently known **function approximation** methods. But they approximate functions by using enormously large training sets to estimate models with enormously large numbers of parameters. However effective they are at prediction, they do not provide a satisfying understanding of causal mechanisms. After all, we have always had the ability to produce a black box that predicts human judgments with perfect accuracy: procreation. A neural network may be just as effective as a human at prediction, but as long as it remains a black box, it does not bring us any closer to an understanding of the causal mechanism.

One conclusion that this argument leads to is that we presuppose a particular level of abstraction when we discuss explanation or causal mechanism. Suppose for argument's sake that we can construct a deep neural network that exactly models the human brain, in the sense that its units can be placed in one-one correspondence with individual neurons, and its units behave in all relevant respects just as neurons do. There is a sense in which doing so completely describes the causal mechanism. But that account would not be linguistically satisfying. The linguistic structures are abstractions that “make sense” of the neural computations. Linguistic explanation does not reside in a full and accurate description of neurons; it resides in an account of the abstract computations that the neurons implement.

A brief discussion of abstraction might be useful. First, it is tempting to consider neurons “more real” than linguistic objects, because they are lower on the scale of abstraction, hence, more concrete. However, that is misleading.

Consider the objects of everyday experience, and compare them to, say, atoms. In this case, I think, the intuitions are reversed. One feels that everyday objects are more real and atoms are more “abstract;” we are less certain that they exist. But we are being led astray by casual usage. An abstraction is a simplified representation of a complex system. It abstracts away from many of the details and attempts to capture the most important objects and relationships. Plainly, a characterization of an everyday scene that lists all the atoms and their properties is much more concrete (and complex) than a characterization in terms of everyday objects. If by “real” we mean more concrete, then atoms are clearly more “real.” But they are epistemologically more remote.

Everyday objects, as collections of atoms, are imposed by our perceptual systems; they are no more “real,” in the sense of concrete, than any random collection of atoms scattered around the scene. What distinguishes the objects that we perceive is not that they are more “real,” but that they are more *useful.* The system of everyday objects is an abstraction and simplification that we impose on the world, but if we use this particular abstract system to make predictions about what will occur, the predictions are vastly better than if we use random groupings of atoms. The objects that we impose on the world have no special claim to “physicality,” but they are special among equally simple systems of objects in the quality of the predictions they enable. In that sense they capture important aspects of the causal structure of the world. I suggested earlier that we should use the term *real* to mean *accurately capturing causal structure.* If so, then realness is orthogonal to concreteness. Elements at all levels of abstraction may capture causal structure.

In these terms, neurons are not more real than linguistic objects. They are more concrete, and they may be epistemologically more proximate, in that we can more directly observe neurons than we can most linguistic objects. But neurons themselves are abstractions that can be reduced to molecules and atoms; like everyday objects, they are not real in any absolute sense. They are real to the extent that they capture causal structure sufficiently well to make good predictions. Linguistic objects are real in exactly the same sense.

Many ways of abstracting the behavior of the brain are possible, and each of them is real to the extent that it enables us to make good predictions. Neurolinguistics attempts to predict measurements on the neurons themselves. Psycholinguistics seeks to understand the time course of linguistic computation, but abstracts away from the actual implementation in neurons. Computational linguistics seeks to understand (among other things) the space of *possible* solutions to the computational problems that one faces in language processing, abstracting away from the particular choice of algorithm embodied in the human mind.

And linguistics simpliciter seeks to characterize only the functions that are computed, abstracting away entirely from how they are computed.

3.6 Simplicity

The issue of simplicity has arisen a couple of times now, and it merits a closer examination. There are at least three roles that simplicity may play. The first is this: when choosing the body of assumptions to adopt in one's own work, one naturally chooses those assumptions one finds simplest. This choice has a great deal to do with familiarity: one naturally finds the assumptions that one has been trained in to be simplest. Accordingly, this notion of simplicity is highly subjective and of little use for persuading someone who is not already committed to the same body of assumptions.

A second notion of simplicity arises in machine learning and statistical inference, and by extension in scientific inference. It is really a family of notions that are formalizations of Occam's Razor. The central idea is the heuristic that, of two hypotheses that have equally good fit to the training data, the simpler is most likely to have low generalization error. In Bayesian statistics, the same role is played by the prior probability, the connection being that the prior probability is naturally assigned in such a way that prior probability decreases monotonically with complexity, or, conversely, that we can *define* simplicity to be a monotone decreasing function of prior probability. (Negative log probability is widely used as a measure of complexity.) This sense of simplicity is thoroughly objective, and very useful in learning, but not of much practical use in theory formation.

The third place where simplicity arises is in the discussion of reductionism. A reductionist account may be true in the sense of making excellent predictions, and yet be unsatisfactory because it is too low-level and complex to be humanly comprehensible. This lies somewhere between the first two on the objectivity-subjectivity scale. It discriminates only very coarsely between theories. It provides no guidance at all for a choice between two theories at the same level of abstraction.

4 Inductive General Grammar

Let us now move beyond individual languages to general linguistics.

Assumption 3 *The main goal of general grammar is to characterize human language acquisition by providing a learning function that maps samples of primary data to particular grammars.*

"Primary data" is the term commonly used in linguistics; "training data" is the standard term in machine learning. I am being intentionally vague about exactly what form primary data takes. Clearly it includes examples of well-formed sentences, but presumably also something more. Psychologically most realistic would be a representation of the physical context of utterance, but a more practical approximation might be the meaning or partial meaning for a subset of the training sentences, making the problem a semi-supervised learning problem.

Assumption 3 needs some unpacking.

Definition 2 *The* **empirical learning function** $L : D \to R$ *is the function computed by humans in the natural course of language acquisition.*

Language acquisition is a change in brain state in response to exposure to primary data D. Given the inextricability of D from all other aspects of the environment that affect learning and development, and given the stochastic nature of changes at the level of brain state, we cannot suppose that the mature brain state is a function of D, in the sense of being uniquely determined by D. But it is reasonable to assume a family of brain states that all correspond to being a speaker of a particular language, and in accordance with Assumption 1, we represent the language as a sound-meaning relation R.

A general grammar provides a hypothesis concerning L. In keeping with standard practice, I assume that it does so indirectly. Namely, a general grammar provides a **grammatical inference function** $\mathcal{L}$ that maps D to a particular grammar G, and the sound-meaning relation predicted by G models the output of L. That is:

$$\begin{aligned} R &= L(D) \\ \hat{R} &= \mathcal{F}(G) \quad \text{where} \quad G = \mathcal{L}(D) \end{aligned}$$

Accordingly, we may represent a general grammar formally as a pair $\mathcal{G} = (\mathcal{F}, \mathcal{L})$, where $\mathcal{F}$ is the grammar notation definition previously discussed and $\mathcal{L}$ is the grammatical inference function. The natural measure of the quality of a general grammar is the following.

Assumption 4 *The measure of quality of a general grammar* $(\mathcal{F}, \mathcal{L})$ *is the expected predictive accuracy of* $G = \mathcal{L}(D)$*, namely, the expected rate of agreement between* $R = L(D)$ *and* $\hat{R} = \mathcal{F}(G)$.

The first expectation is taken over languages and primary data samples, and the second is taken over sound-meaning pairs for a given language.

The dominant paradigm agrees that learning is the crux of general linguistics (a.k.a. universal grammar), and I think the only controversy that Assumption 3 might raise is the question of the nature of the output of L. It is thus surprising that the field has shown so little interest in the sizeable body of theoretical and practical knowledge of learning in general and language learning in particular that comes from computer science and statistics. Given the constant press coverage of language-related advances in artificial intelligence, the reason can hardly be lack of awareness.

Instead of applying the results of machine learning, the strategy has been to avoid learning by make maximally pessimistic assumptions about learnability and adopting a strategy of reducing what must be learned to a minimum. In particular, a **parametric grammar strategy** is adopted, in which the aim is to express variation across languages in terms of a relatively small number of abstract parameters. Concretely, instead of taking the form of a collection of rules, a particular grammar is to take the form of a lexicon and a vector of parameter settings.

In point of fact, reducing the number of possible grammars is not guaranteed to make the learning problem easier. More important than the raw number of possible hypotheses is the relationship between observable properties and the hypothesis space. For example, if the hypothesis space is the uncountably infinite set of lines on the plane, then one of the oldest and simplest learning algorithms, the perceptron algorithm, is guaranteed to find either the correct hypothesis or one that is indistinguishable from it on the basis of the training instances. On the other hand, if there are only four possible instances corresponding to the corners of the unit square, yielding only $2^4 = 16$ distinguishable hypotheses to choose among, the perceptron algorithm cannot be guaranteed to succeed, inasmuch as some hypotheses (those in which diagonally opposite corners of the square are grouped together) are not expressible as dividing lines on the plane.

For these reasons, I think it is fair to characterize the parametric grammar strategy as speculative, in Bloomfield's terms. Nonetheless, the idea of factoring out commonalities across particular grammars is attractive. My personal preferences accord with the **inductive general grammar** that Bloomfield foresaw. A practical approach is to develop concrete particular grammars with good predictive accuracy, for the languages for which treebanks or meaning banks are available, and then to develop more succinct "higher-level" grammars, analogous to higher-level programming languages, from which the concrete particular grammars may be generated, with no loss of predictive accuracy. The translation from high-level grammar to concrete grammar is analogous to the translation that a compiler performs from high-level language to machine code.

But however one arrives at a general grammar, the proof of the pudding is in the eating, by which I mean that Assumption 4 provides neutral grounds for deciding whether extant proposals regarding universal grammar, or the inductive approach just sketched, provide a better account of the human language capacity.

References

[1] Abney, Steven. Data-intensive experimental linguistics. *Linguistic Issues in Language Technology (LiLT),* 6(2). 2009.

[2] Abney, Steven. A bidirectional mapping between English and CNF-based reasoners. *Proc. of the Society for Computation in Linguistics (SCIL),* Vol. 1, Article 7. DOI: 10.7275/R5PZ571N. 2018.

[3] Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, Nathan

Schneider. Abstract meaning representation for sembanking. *Proc. of the Linguistic Annotation Workshop,* pp. 178–186. Sofia. 2013.

[4] Bloomfield, Leonard. *Language.* Holt, New York. 1933.

[5] Groningen Meaning Bank. Bos, Johan, Valerio Basile, Kilian Evang, Noortje Venhuizen, Johannes Bjerva. The Groningen Meaning Bank. In: Nancy Ide and James Postejovsky (eds) *Handbook of Linguistic Annotation,* pp. 463–496. Springer, Berlin. 2017.

[6] Dyson, Sir F. W., A. S. Eddington, and C. Davidson. A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society A.* Jan. 1, 1920.

[7] McDonald, Ryan; et al. Universal dependency annotation for multilingual parsing. *Proc. of the Conference of the Association for Computational Linguistics (ACL).* 2013.

[8] Oepen, Stephan, Dan Flickinger, Kristina Toutanova, Christoper D. Manning. LinGO Redwoods. A Rich and Dynamic Treebank for HPSG. In *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT).* Sozopol, Bulgaria. 2002.

[9] Scholz, Barbara; Pelletier, Francis; Pullum, Geoffrey. Philosophy of Linguistics. *Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/linguistics. Accessed 2018 Oct 8.

[10] Shieber, Stuart M. The problem of logical-form equivalence. *Computational Linguistics* 19(1), 179–190. 1993.