



Inductive general grammar

STEVEN ABNEY 

RESEARCH

]u[ubiquity press

Abstract

General-linguistic datasets that have become available in recent years promise to enable new progress toward a theory of general grammar. A barrier to success is the incompatibility between the inductive, externalist approach that is natural for exploiting the datasets and the deductive, mentalist philosophy that is currently dominant within linguistics. I argue that the externalist philosophy is viable and that there are reasons to consider it preferable. I argue that the mainstream approach is in some cases unnecessarily concerned with psychological reality, and in other cases too quick to reject required subtheories on the grounds that they belong to “language processing” rather than linguistics, with the result that current grammars give systematically inaccurate answers to questions of the linguistic status of sentences. I suggest that the inductive development of general grammar is already being carried out (though not in those terms) in the field of natural language processing, and that linguistic participation in the effort would be of benefit to both fields.

CORRESPONDING AUTHOR:

Steven Abney

University of Michigan, US

abney@umich.edu

KEYWORDS:

philosophy of linguistics;
general grammar; grammar
induction; machine learning;
corpus-based methods

TO CITE THIS ARTICLE:

Abney, Steven. 2021. Inductive general grammar. *Glossa: a journal of general linguistics* 6(1): 75. 1–22. DOI: <https://doi.org/10.5334/gjgl.1332>

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 1933: p. 20).

Eighty-five years after Bloomfield wrote that passage, resources are at last available that render viable the development of an inductive general grammar. By *inductive* I mean a bottom-up or data-driven approach, in contrast to a top-down or postulate-driven approach. To pursue an inductive approach to general grammar, one requires what we might call *general-linguistic datasets*: broad-coverage annotated language samples from a substantial, typologically diverse range of languages using language-universal systems of annotation. An emphasis should be placed on *language samples*—what I have in mind are not tables summarizing typological properties of languages, but rather significant samples of annotated primary data from a range of languages. Within the last few years, researchers in the field of Natural Language Processing (NLP) have developed the first substantial general-linguistic datasets; the Universal Dependencies Treebank (UDT) offers a prominent example (McDonald et al. 2013).

The fields of linguistics and NLP have proceeded along independent courses for many years, with ever-dwindling connections between them. The divergence between the fields has only accelerated with the rise of so-called deep learning methods in NLP. But the emergence of general-linguistic datasets offers an unusual opportunity for collaboration between the fields, one that promises benefits to both fields.

One barrier to collaboration is the lack of awareness, within linguistics, of the relevant datasets. Another barrier is philosophical. To understand the value of the existing datasets, and to put them to good use, one requires a dedication to good description, systematicity, breadth of coverage, and a flexibility with respect to formalism, that run counter to tendencies within current linguistics. For lack of a better term, I will call the currently dominant linguistic philosophy *internalist*, in contrast to the *externalist* approach that dominates in NLP and enables a general grammar predicated on induction from large datasets.¹ My aim in the present article is to make the case that the externalist approach represents good linguistics, not just good engineering, and in fact that it flows naturally from the most basic assumptions about the primary goals of linguistic inquiry.

To forestall false expectations, I do not intend to catalog the empirical results of research in inductive general grammar, for two reasons: first, because the research that has been done is unlikely to be recognized as *linguistic* research by many readers—and, in fairness, technology development has indeed driven the work more than linguistic inquiry has—and second, because the most interesting work is yet to be carried out. I will focus instead on the more basic philosophical and methodological questions. In part I would like to make the case for a formulation of the goals of linguistics in which general-linguistic datasets play a key enabling role; and in part I would like to make the case that, contrary to initial appearances, much of the work that has been done to date does address linguistic questions and indeed fundamental questions of linguistics, despite the fact that the work appears in NLP venues and not in linguistics journals, and however incidental the linguistic considerations may have been to the researchers' motivations.

To be specific, I will argue below that “disambiguation” should not be dismissed as a processing issue but must be addressed head-on, if we are to answer central questions of linguistics. And I will argue that, instead of allowing uninformed speculation about what is “learnable” or not learnable to guide key decisions in grammar, it is better to work together with machine learning researchers to develop serious models of how language can in principle be acquired, and to test those models on the large scale.

¹ The dominance of the internalist framework began with Chomsky's work, and one might for that reason call it *Chomskyan*, but its influence extends beyond the Chomskyan paradigm, narrowly defined. *Generative grammar* is another designation that is often used, but in the present context that is a misnomer—*generative* has a well-established meaning in formal language theory that has nothing to do with mentalism.

Before discussing those specific issues, I would like to lay out the aims and assumptions of the inductive general grammar paradigm. Calling it *inductive general grammar* is idiosyncratic to me, but the paradigm is, I believe, the consensus view within the field of NLP, and in artificial intelligence research more generally. It is firmly rooted in formal language theory, logic, and machine learning. But it is at odds with the internalist approach, which insists on a “psychological reality” that goes beyond accuracy of prediction, and which dismisses formal languages as epiphenomenal and unworthy of study. I intend to clarify what the differences are between the frameworks and what the motivations are for adopting the inductive approach.

In what follows, I do criticize particular versions of the internalist paradigm, but that is not my primary goal. My purpose is to make a case for a fresh approach to the fundamental questions of linguistics, to make clear how it differs from business as usual, and to give my reasons for considering it superior to the current framework.

2 The linguistic abstraction

2.1 Fundamental questions of particular grammar

The central assumption of the externalist approach is the following.

Assumption 1 (The linguistic abstraction.) *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.

A couple of questions arise immediately. The first is exactly what is meant by an *individual language*. Is it a language as a social construct, language as an abstract grammatical system, language as a psychological faculty, or something else? This question requires space for discussion and I postpone it momentarily.

I have also not spelled out what meanings are. We may for now assume the usual intensions, that is, functions from worlds to model-theoretic constructs, though there are many viable alternatives and little in the following discussion hinges on the choice, except in section 4.2.

Corollary 1 *The fundamental questions of particular linguistics are these: For any sentence of the language, what meaning(s) does a native speaker of the language assign to it? And, conversely, for any meaning expressible in the language, what sentence(s) 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,” by which I really mean expressions of the language, the most basic modality being spoken expressions. R is equivalent to two functions: the *production function* or *expression function* $E_R(\mu) = \{\sigma | R(\mu, \sigma)\}$ takes meanings as input and produces sentences as output, and the *interpretation function* $I_R(\sigma) = \{\mu | R(\mu, \sigma)\}$ takes sentences and produces meanings.² These functions are open to observation, observations that may be characterized as a speaker’s responses to two questions:

[Q1] What expressions are natural ways of expressing meaning μ ?

[Q2] What meaning or meanings does expression σ have?

$E_R(\mu)$ is what we observe when we ask Q1 and $I_R(\sigma)$ when we ask Q2. A speaker deems σ to be ambiguous just in case $|I_R(\sigma)| > 1$.

Question Q1 obviously characterizes linguistic production, and Q2 characterizes linguistic comprehension. Those accustomed to Chomsky’s formulations will expect language acquisition to appear here as well. But at this point, the matter under discussion is particular linguistics. Acquisition is a matter of general linguistics and is addressed later (section 5).

² For the sake of simplicity and familiarity, I have characterized R as a simple relation, so that $E_R(\mu)$ and $I_R(\sigma)$ are sets. This is not meant to preclude a more general account, in which R is a probabilistic relation and $E_R(\mu)$ and $I_R(\sigma)$ are probability distributions.

The reader may also object to the omission of grammaticality judgments as basic sorts of observations. A grammaticality judgment is an answer to the question, “can one say σ ?” for some candidate expression σ . The intended meaning is always implicit in context, and if it would otherwise be unclear, one provides it explicitly: “can one say σ to mean μ ?” In this explicit form, it is evident that a grammaticality judgment is a response to a minor paraphrase of Q1: “is σ a natural way to express the meaning μ ?” As a consequence, the set of well-formed expressions is the same as the set of expressions that R associates with meanings.

Examples have, of course, been put forward that are allegedly grammatical but meaningless; the “autonomy of syntax” examples of Tesnière (1959) or Chomsky (1957) come immediately to mind.³ In fact, though, the proffered examples are *not* meaningless. One has definite judgments about what a noun phrase like *colorless green ideas* refers to, namely, ideas that have the property of being green and also the property of being colorless. The problem is that no such ideas exist, nor could exist in any world we can conceive of. But one certainly judges that, however absurd *colorless green ideas* might be, they do differ from other absurdities, say, *preternatural red cosmologies*. As a technical matter, if a meaning is an intension, and if we assume a one-one correspondence between possible worlds and models, a phrase like *colorless green ideas* does have an intension, though it has an extension only in worlds that violate the relevant lexical axioms. But even if grammaticality judgments are ultimately not reducible to production judgments, any discrepancies are minor, and for the sake of simplicity I put them aside. I return to the issue of grammaticality judgments below (section 4.1).

A more worrisome concern is 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.⁴ What we require is a *grammar*, in the formal sense of a finite definition for the infinite sound-meaning relation. Not only is it non-trivial to formulate a grammar that accurately models R , I argue below that developing an adequate model of R actually requires us to deal with issues that most linguists dismiss as “processing” issues. See sections 4.1 and 4.2.

Another concern is the possibility that different speakers might have different linguistic systems, giving different answers to questions Q1 and Q2, or that a single speaker might give different answers if asked at different times. I have also not addressed the possibility that a speaker’s active and passive linguistic abilities might differ, that is, that the speaker’s answers to Q1 and Q2 might be inconsistent, perhaps interpreting σ to mean μ but not including σ among the natural ways of expressing μ . Again, for the sake of simplicity, I put these possible inconsistencies aside, modeling speakers as being perfectly consistent with each other and with themselves.

The assumption of homogeneity—that is, the assumption that all speakers possess the same linguistic system—is standard; it is adopted both by De Saussure (1986) and by Chomsky (1986), to name two prominent examples. De Saussure and Chomsky disagree about the locus of regularity in language: de Saussure sees it in the social system, whereas Chomsky sees it in the psychology of the individual. But the question is somewhat moot, inasmuch as, as long as we assume homogeneity, language as the possession of the speech community and language as possession of the individual are one and the same.

2.2 Levels of abstraction

Assumption 1 is quite similar to the “Thesis” of Lewis (1975). The original idea is older, stemming from formal language theory, which in turn imports from mathematics and logic the attitude that definitions should be formulated in a maximally general way. In the context

3 Tesnière offers the example *le silence vertébral indispose la voile licite* in support of the claim that syntax is autonomous: *elle [la syntaxe] est autonome* (Tesnière 1959: p. 42). (He generated his example from *le signal vert indique la voie libre* by replacing each word with the next word in the dictionary with the same part of speech.) According to the preface of Tesnière (1959), the manuscript was complete at the onset of Tesnière’s final illness in 1950, and no substantial changes or additions were made to the text in preparing it for publication. If so, Tesnière’s example and his use of the term “autonomy of syntax” appear to predate the better-known presentation of Chomsky (1957).

4 By “satisfactory,” I mean mathematically satisfactory. A set S is *definable* just in case there is an open formula of one variable ϕ such that $x \in S$ if and only if the statement $\phi[x]$ is provable (Boolos & Jeffrey 1980: p. 173). Any definition, by virtue of being a formula, has a written representation of finite length.

of linguistics, that means that the concept *language* should be defined in a way that does not limit it to the specific case of language spoken by humans, but should abstract away from the details of human psychology. Language thus defined views human psychology from the outside, in terms of the input-output relation it computes, rather than the process by which it does the computation. The same externalist philosophy is evident in truth-conditional and model-theoretic semantics.

Chomsky (1986), on the other hand, espouses a form of psychologism in which a system that does not make substantive claims about human psychology is “an epiphenomenon at best” (p. 25), possessing no status in linguistics, on a par with, say, a listing of the pairs of rhyming words in the language. He introduced the term *I-language* for his own, internalist, view, and *E-language* for the externalist approach of Lewis.

The internalist position seems clearly dominant within linguistics. Standard textbooks in syntax adopt it (Carnie 2012), and both proponents of externalism (Soames 1984) and proponents of internalism (Phillips & Lewis 2013) agree that internalism is dominant. Since I take an unequivocally externalist position with Assumption 1, and since an externalist philosophy thoroughly pervades the design of general-linguistic datasets and the work that they support, this issue deserves particular attention.

The chief argument that Chomsky (1986) offers against external models is to claim that, unlike internal models, they do not make “true or false statements about something real and definite” (p. 26). He asserts that “theories of E-languages, if sensible at all, have some different and more obscure status because there is no corresponding real-world object,” and he insists that they are not worth pursuing, “given the artificial nature of the construct [E-language] and its apparent uselessness for the theory of language” (p. 27).

These assertions are really quite baffling.⁵ The fact of the matter is that a central motivation for the externalist approach is to describe real, physically embodied systems at a useful level of abstraction. In adopting an external model, one certainly does not deny that there is a mental mechanism that gives rise to language, but rather, *as a strategy for more effectively studying that mechanism*, one abstracts away from its internal structure in order to focus on the mathematical nature of the association between sounds and meanings that those mental mechanisms give rise to.

The abstraction allows one to more readily compare human languages to other, similar, formal systems, such as animal languages or artificial languages. It also seems more in keeping with pre-theoretic understanding of what a language is. Consider a thought experiment in which a non-human intelligence—an alien or an artificial intelligence—learns, say, English so well that an interlocuter is unable to distinguish the alien’s abilities from those of a native speaker. In such a case it would seem unreasonable to deny that the non-human intelligence *speaks English*, and it follows that, whatever English is, its implementation in a human mental substrate is not a necessary condition.

Nor does an external model preclude a more concrete internal model of the same system. Indeed, as already mentioned, a key motivation for formulating an external model of language is that a clear understanding of the relation that is computed is plainly desirable when one attempts to formulate a more concrete model of *how* that relation is computed. Nor is there only one internal model: it can be useful to identify several different levels of abstraction, and I will explicitly mention a few. A *computational model* is one that considers the algorithms and data structures by which the sound-meaning relation *might* be computed, and what properties

⁵ Chomsky does not give a formal definition of either term, but he seems to take an E-language to be a set of sound-meaning pairs, whereas he appears to take an I-language to be a grammar. In the text, I assume that is a minor inconsistency in presentation. It may, however, be the actual crux of the matter. If so, then Chomsky’s entire discussion hinges on a comparison between apples and oranges: the legitimate comparison is between E-grammar and I-grammar, or between E-observables and I-observables. The sound-meaning relation is the space of E-observables; it is a different object than a grammar, which is a means of generating (a hypothesis about) the population of E-observables. What Chomsky considers the I-observables to be is less clear, though the obvious answer is observations about the time course of processing. Grammars (that is, models accounting for observables) are not the exclusive possession of the internal level of abstraction, and comparing the internal to external approaches by comparing the grammar of the former to the observables of the latter would be simply disingenuous. And if, by I-language, Chomsky means the physical object, the unfairness of the comparison is even more extreme: the physical object is identical for both the external and internal accounts; only the level of abstraction of the description differs.

they have, for example, in time and space usage. A computational model is not committed to a particular substrate—it applies equally to human and artificial systems—but it does provide useful information about the available options in the design of the human system. A *psycholinguistic model*, in turn, is specific to the human mind. The question is which algorithm and data structures humans actually employ, out of all of the possibilities. It concerns language at the level of abstraction corresponding to mind, rather than brain. A *neurolinguistic model* is the most concrete; it asks how the human language-processing algorithm is implemented in brain wetware.⁶

Again, let me emphasize that all of these models are models of the same physical system, but at different levels of abstraction, and that this strategy of stepwise refinement is simply good scientific practice. Since E_R and I_R are largely accessible to observation, we may study R directly, without the necessity of special apparatus or difficulties of interpreting indirect measurements.⁷ 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.

Let us return now to Chomsky’s arguments. He claims that (1) external models are not models of any real-world object, and (2) external models are “apparently useless” for the theory of language. Claim (1) is rebutted in the preceding paragraphs: in matter of fact, external models of language are abstract models of the human linguistic system and are part and parcel of a strategy for effective investigation of the human mind and brain at all levels of concreteness. Claim (2) is equally indefensible, as is made clear by even the most casual awareness of the success of the external approach in logic, formal language theory, and natural language processing.⁸ I conclude that Chomsky’s arguments against external models are simply invalid, and that external models of language, of a sort consistent with Assumption 1 (language as sound-meaning relation), are reasonable and viable.

2.3 Psychological realism

Within the approach that I adopt, there is in principle no conflict between an external model of language, such as one might naturally construct from a treebank, and an internal model, such as the one that Chomsky espouses. A notable feature of externalism is what we might call its catholicism: the externalist approach is perfectly at ease with having multiple models of the same physical system, at different levels of abstraction, or simply for different purposes.

By contrast, Chomsky’s version of internalism rejects external models as ungrounded in reality and useless. When multiple levels of abstraction are introduced, the logic of the argument appears to require that each level of abstraction except the most concrete be rejected: a psychological model rendering a computational model “epiphenomenal” and thus “useless,” and a neurological model in turn rendering psychology epiphenomenal and useless. The natural end of the philosophy is reductionism, and in fact there is more than a hint of that in the recent popularity, within the Chomskyan school, of the idea of “biolinguistics” (Chomsky 2004).

Again, if one adopts an external model of language, any two distinct grammars G_1 and G_2 that generate precisely the same sound-meaning relation are *equivalent definitions* of the language, and thus if one is correct, so is the other. This is a virtue, not a defect: proving the equivalence of superficially very different systems has been a key means of advancing our understanding

⁶ The discussion invites comparison to Marr’s levels of analysis (Marr 1982), but it should be noted that the levels of abstraction in the text differ from Marr’s. Marr’s highest level of abstraction he calls the computational level; he does not consider the external level.

⁷ We may include introspection as a form of observation, provided that there is consistency across judges.

⁸ One might question whether success in those domains really constitutes evidence of efficacy in investigating the human mind. It is true that accurately capturing the sound-meaning relation does not tell us what makes the human mind different from other hypothetical physical embodiments of a speaker of the language, such as alien or artificial intelligences. But it *does* tell us what makes the human mind different from physical systems that do *not* speak the language in question, and that is a great deal indeed.

within the fields of logic and formal language theory, as in mathematics more generally. Different models often have different strengths, and establishing equivalence allows one to translate freely between the models and exploit the strengths of both.

The internalist view, on the other hand, treats grammars like our G_1 and G_2 as standing in competition. Even if G_1 and G_2 provably define the same relation between sounds and meanings, and are thus not even in principle distinguishable by any observations (of the sort that the models account for), the internalist considers them to be *competing models* that make different claims about the human language faculty, and insists that only one of them can be true.

If one insists that G_1 and G_2 make different claims, the only way to determine which is correct is to go beyond the kinds of observations that were used to develop them (conventional linguistic observations), to entirely new classes of observations, such as reaction time data or eye movement data. But most linguists, including those who espouse psychological realism, are unwilling to take the step of devoting themselves in a serious way to psycholinguistics. There is good reason for reluctance: it is simply too daunting to thoroughly master a field other than one's own. In practice, even Chomsky does not stray far from conventionally linguistic observations. Yet, without a thorough mastery of psycholinguistics, any attempt to use psycholinguistic data to argue for one linguistic theory over another opens the researcher to accusations of cherry-picking isolated facts that happen to agree with the researcher's own opinions.

This leaves us with a near-contradiction between the demand for psychological reality and the reluctance to go beyond the scope of conventional linguistic observations. What is the point of insisting that one's theory "makes claims" about psycholinguistics if one has no intention of proving those claims, by doing serious psycholinguistics? In an attempt to resolve the tension, Chomsky proposes that a grammar possesses psychological reality if it constitutes "knowledge of language" that "enters into the expression of thought and the understanding of presented specimens of language" (p. 4)—that is, if the elements of grammar correspond directly to stored data that is put to use in language production and comprehension. He later writes more explicitly of language comprehension being performed by "a parser that incorporates the rules of the I-language along with other elements" (p. 25).

In other words, one can have an internal model without going outside conventional linguistic observations by constructing a *partial* model of language processing. In computational terms, one describes particular data structures, but not the algorithms that operate on them. In this way, what matters is not the form of the grammar or the class of observations it accounts for, but only the linguist's intentions or claims. One and the same grammar might be interpreted as an external model or as a partial model of psychology; the choice lies with the linguist. Soames (1984) uses the term *narrow competence theory* for such a partial model of language processing, in which linguists' grammars are regarded as "the internalized rules employed by competent speakers" (p. 167).

Understood in the "laziest" fashion, in which the linguist simply pursues business as usual, Chomsky's position is still open to charges of making empty claims, of reducing psychological reality to a mere rhetorical flourish. At least some researchers do take it more seriously. Phillips & Lewis (2013), for example, take the position that the grammar describes "a structure building system that combines words and phrases to form sentences and meanings in real-time, in essentially the same manner in comprehension and production" (p. 17). In their version of internalism, the grammar constitutes not a data resource but a *component* of the human language processing system.

But that takes one out of the frying pan into the fire: a *partial* model of psycholinguistics is untestable. The postulated internal grammar, whether it is a rule store or a structure-building component, is not observable. If we relate it only to conventional linguistic observations, then we are back to a distinction without a difference. Our G_1 and G_2 remain indistinguishable, and the claim that one is psychologically real while the other is not becomes a rhetorical flourish, a claim that one has no intention of proving. In order to prove it, one must bite the bullet and actually do psycholinguistics.

Soames comes to a similar conclusion: "if this is right, then proponents of 'the received view' should either stop requiring their theories to maximize linguistic generalizations or stop taking them to represent adequate psychological theories of linguistic competence. ... Which alternative

one chooses will depend on whether one wants to do linguistics or mentalistic psychology” (pp. 170–1). Phillips and Lewis do not concur explicitly, but they go to lengths to sketch a plausible psycholinguistic system that embeds a grammatical component and they cite the psycholinguistic literature heavily. That is, they do “bite the bullet,” implicitly acknowledging that one cannot propose a partial model and stop; one requires some account of the rest of the system, no matter how provisional, if one wishes one’s model to have any empirical content at the psycholinguistic level of abstraction.

Phillips and Lewis nonetheless reject the idea that their model might be consistent with a different model at an external level of abstraction. They offer a novel argument: they claim that an external model is justifiable only if the system is *implementation independent*, by which they mean that “the exact same abstract system can be realized in different ways by multiple lower-level implementations, with no change in the abstract system itself” (p. 14). This definition agrees with remarks I made earlier. It is obvious that languages, including English, can be implemented in a wide variety of ways, as even a passing glance at the NLP literature makes clear. However, Phillips and Lewis assert that a given human speaker has only one method for assembling sentence structures in the course of language processing, not multiple implementations, therefore human language is not implementation independent, therefore an external model is “unwarranted.” But that argument is logically incoherent. Implementation independence as they define it does not imply that any given individual implements the system in multiple ways, or even that different individuals implement it in different ways, but rather that multiple ways *exist* in which the system can be implemented. Thus their argument is invalid and their conclusion unjustified.

Indeed, it is hard to imagine a system description that is genuinely implementation *dependent*, in their sense. Even a description at the level of neurons would be implementation independent, in that it would of necessity leave some aspects of neurons unspecified and thus free to be implemented in multiple ways. If we assume that Phillips and Lewis actually intended *implementation dependence* to mean that all the physical instances of a given system description, in a particular reference set of instances, implement the description in the same way, then speaking English is indeed (more or less) implementation dependent with respect to the set of human speakers of English, but that is a trivial consequence of choosing a homogeneous set of physical instances as reference set. Implementation dependence in that sense is not intrinsic to the system description; it vanishes as soon as we include non-human speakers in our reference set.

Contrary to Phillips and Lewis, Soames argues that the most accurate and insightful model for the sound-meaning relation, and the most accurate and insightful model for psycholinguistic data, may be quite different. Since they are accounts of the same physical system, there will be connections between them, but those connections may be indirect.

I find a couple of analogies to be useful. High-level programming languages are designed to make the programmer’s task as easy as possible, whereas machine code is designed to be as close as possible to the architectural components of the machine and the operations that they support. High-level programs and machine code provide descriptions of one and the same computation, but at very different levels of abstraction, and the connection between them is very indirect: the job of a compiler is to translate one to the other. The analogy to external and internal models is obvious.⁹ One might object that the analogy is imperfect because the brain is not an artifact *designed* to carry out a high-level goal. But in fact, it *is* designed to achieve certain high-level goals: the available neural computing resources are marshalled in the course of evolution and development to solve high-level tasks that are necessary for the survival of the organism. The place of the compiler is taken by something like the weight optimization that takes place in artificial neural networks.

Or to offer a simpler analogy, consider waves on the surface of a pool. The natural terms to describe the “external” system are wavelengths and amplitudes, whereas the natural terms to describe the “internal” system are locations and momentums of water molecules. The two descriptions (“grammars”) are systematically related but not by any means the same.

⁹ Compilers generally involve program transformations that map very different high level programs to the same machine code; in the same way, equivalent pairs like our G_1 and G_2 may involve grammars that are very different in appearance.

I conclude that it is reasonable (indeed, commonplace) to develop grammars based on conventional linguistic observations, and that this practice is justified, with no further onus of demonstrating “psychological reality,” by considering grammars to be models of the human language system at the external level of abstraction. Given how widespread the practice is, I will henceforth identify the external level of abstraction as the distinctively *linguistic* level of abstraction, and the idea of abstracting away from internal mechanisms to the external sound-meaning relation I will call *the linguistic abstraction*.¹⁰

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 (language as sound-meaning relation) is essentially a statement of the linguistic abstraction.

3 Evaluation

3.1 Fit and prediction

I have characterized a grammar as a definition of the sound-meaning relation R . 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 $E_{\hat{R}}$ regarding the empirical production function E_R , and a hypothesis $I_{\hat{R}}$ regarding the empirical interpretation function I_R .

Assumption 2 (Prediction as measure of quality.) *The acid test of grammar quality is the degree to which the grammar’s 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 $I_{\hat{R}}(\sigma) = I_R(\sigma)$, and if we draw a meaning μ at random from the population of meanings, what is the probability that $E_{\hat{R}}(\mu) = E_R(\mu)$?¹¹

There is a mistake that is widespread in linguistics, namely, the confusing of prediction with *fit* to known examples. Both fit and prediction are measurements of discrepancy between $I_{\hat{R}}(\sigma)$ and $I_R(\sigma)$; the difference between them is whether one measures discrepancies on examples collected *before* formulating the grammar that gives us $I_{\hat{R}}$ or on new “test” examples collected afterwards. The former is fit and the latter is prediction.

The usual practice in linguistics is to collect a range of examples that display some phenomenon, and to devise a grammatical system that “gets the examples right.” When comparing one’s particular grammar to an alternative, one either identifies examples that one’s own grammar gets right and the competitor gets wrong, or one makes an attempt to show that one’s own grammar is “simpler,” or one extends the range of examples under consideration and modifies one’s grammar so that it (and not the competitor) gets the larger set right.

Nowhere in the usual methodology is there any genuine testing of predictions. When determining whether one’s grammar gets certain examples right, one often talks about “the predictions of the grammar,” but in fact all of the comparisons are measuring fit, not prediction. In the usual methodology, the examples under consideration comprise a finite set, and coming up with a theory that fits them is, in a sense, too easy. The examples 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.

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. 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

¹⁰ Soames also refers to external theories simply as *linguistic theories*, in contrast to narrow theories of competence, and Phillips and Lewis admit that “we suspect that [the extensionalist position] is a position that many practicing syntacticians are comfortable with, and it certainly corresponds with much standard practice in linguistics” (p. 13).

¹¹ In statistics, the *population* is the full set of things about which one would like to draw conclusions, and from which one draws samples for statistical study. The probability of a property is the proportion of the population that has that property. For example, the probability of the property $E_{\hat{R}}(\sigma) = E_R(\sigma)$ is the proportion of sentences σ for which $E_{\hat{R}}(\sigma) = E_R(\sigma)$ is true, out of the (infinite) population of all possible sentences of the language.

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 E_R and I_R 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 the 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 $E_{\hat{R}}(\mu)$ and $I_{\hat{R}}(\sigma)$ 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 $E_{\hat{R}}(x) \neq E_R(x)$ or $I_{\hat{R}}(x) \neq I_R(x)$.¹² 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 NLP. The methodology is by no means unique to NLP; 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 \mathcal{X} and an experimental protocol for obtaining observations $y(x)$ for $x \in \mathcal{X}$. 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 a model. The process of constructing a model is viewed as choosing a candidate out of a (usually infinite) space of possible models, and is known as *hypothesis selection*. In machine learning, the choice of model is determined by an *objective function* that typically represents a combination of the fit of the model to the training set and the simplicity of the model by some measure. 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 competing models.¹³ This second step is *hypothesis evaluation*.

In machine learning, hypothesis selection is done mechanically, but apart from that, the prediction task methodology is really just the scientific method. In the scientific method, one proposes a hypothesis (hypothesis selection), and one devises an experiment (prediction task) to test it. The experimental design specifies the form that observations are to take, and thereby determines a population of possible observations. One makes observations—that is, one draws a random sample from the population of potential observations—and one evaluates the hypothesis by comparing its predictions on the test set to the actual observations (hypothesis evaluation). One then refines the hypothesis (hypothesis selection under a different guise) based on accumulated experience, including the test set, which now becomes a training set, and the cycle repeats.

Understood in this way, there is a clear separation between the model and the observations. The model is a hypothesis about the physical system, constructed in the hypothesis selection step, and shaped by considerations of simplicity and fit to previously-collected data. There is no expectation that the constructs of the model are directly observable; they represent

¹² For simplicity I assume what is known as 0–1 loss, where the prediction is either fully correct or fully incorrect. Finer-grained loss functions, for example set differences or divergence between probability distributions, might be used. Technically, training error is the expected loss over the training set.

¹³ 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.

the hidden workings of the system. But they represent the physical system schematically, in the same way that a blueprint represents a building. No model can reveal what the physical system is *in se*; rather, the model is an account of the system as seen through the window of experimental observations. Observations, unlike model-internal constructs, must be objective. To qualify as a well-defined experiment, it must be the case that any two well-intentioned observers, regardless of their theoretical persuasion, agree on what is observed, up to the limits of instrument accuracy. Theory development is a cycle that alternates between refining the experimental design to capture the relevant aspects of the physical system of interest, and seeking models with high predictive accuracy on the experiments. To repeat, the model captures, not everything that is true about the physical system, but specifically *those aspects of the physical system that are revealed in the observations*. The experimental design embodies a choice of which aspects of the physical system matter, which aspects of the system are to be modeled. The linguistic abstraction (Assumption 1) identifies production and interpretation observations as that which makes the brain a *linguistic* system.

The failure of internalist linguistics to distinguish fit and prediction, or hypothesis selection and hypothesis evaluation, is closely tied to an underappreciation of systematic observations and proper experiments. The point of an experiment is to measure predictive accuracy, not only in order to adjudicate between competing theories, but also, through systematic testing of predictions, to provide information that can be used to direct theory development. Without experiments, one has only fit to particular examples that have come to one's attention, and simplicity, to guide hypothesis selection.

A few comments about simplicity may be useful. Simplicity is really a family of notions that are formalizations of Occam's Razor. The central idea is that, although fit to the known facts is a good criterion for choosing a hypothesis, one obtains a better criterion by also factoring in the simplicity of the candidate hypotheses. During hypothesis selection we would like to choose the *best* hypothesis, in the sense of the one with the highest predictive accuracy, which is to say, the lowest generalization error. But generalization error is not available during hypothesis selection, so we cannot directly measure predictive accuracy. Fit and simplicity provide proxies for predictive accuracy, and, as a rule of thumb, combining them works better (yields hypotheses with lower generalization error) than using fit alone.

In Bayesian statistics, the role of simplicity is filled by prior probability, the connection being that prior probability is naturally assigned in such a way that prior probability decreases monotonically with complexity, or, conversely, that we can *define* complexity to be a monotone decreasing function of prior probability. Log prior probability is widely used as a quantitative measure of simplicity.

Let me close this section by addressing a couple of possible objections. First, in response to my assertion that conventional linguistic practice fails to distinguish fit and prediction, one might object that, although most linguists eschew the rigor of machine-learning style experimentation, they are well aware of the importance of not simply getting the known examples right, but doing so in a way that "extends without further stipulation to other phenomena."¹⁴ Such an objection misses the point. The predictions of a model concern unobserved instances of *the modeled phenomenon*, not other phenomena. In order to consider prediction, one must have an idea of a population of instances of the phenomenon, a population that can be sampled to construct a test set and to measure predictive accuracy. That is largely absent from current linguistic practice.

Predictive accuracy on the test set provides an unbiased estimate of *generalization*, which is predictive accuracy on the population as a whole. Whether or not aspects of the model extend to other phenomena is a matter of overall *simplicity*. Simplicity and generalization are quite distinct. As mentioned above, they are related by Occam's Razor, which can be paraphrased as stating that the simplest model is most likely to make the best predictions. However, Occam's Razor is a rule of thumb, not a logical necessity, and it is relevant for hypothesis selection, not hypothesis evaluation.

14 The quote is from an anonymous reviewer, to whom I am grateful for thoughtful comments.

Second, I have assumed that predictive accuracy is defined in terms of error rates when sampling from the distribution of possible observations. Plainly stated, we are counting tokens rather than types, when counting errors. Some readers may feel that the number of *types* that are correctly handled is more significant. I suspect that the underlying idea is that types more closely align with pieces of theory—components, constraints, rules, what have you—and that one should measure progress in terms of the proportion of pieces of the theory that are “working right.” If that diagnosis is correct, the issue is again one of conflating hypothesis selection, where simplicity matters and counting pieces of the model makes sense, with hypothesis evaluation, where prediction matters and the structure of the model is immaterial.

To be clear, I do not wish to suggest that one can substitute scores on an experiment for good scientific judgment. But hypothesis selection unguided by evaluation is equally defective. The essence of the scientific method is the cycle of selection and evaluation, and either on its own is incomplete.

3.2 Systematicity and linguistic datasets

As noted above, Assumption 1 lends itself readily to experimentation. Q2 (the interpretation question) in particular immediately suggests the following experimental design: collect a corpus of sentences and ask native speakers what each sentence means. The task for a hypothesis—that is, for a grammar—is to predict the native speaker’s answer. The general-linguistic datasets mentioned in the introduction enable precisely this sort of experiment.

To achieve any significant degree of predictive accuracy on the sentence-interpretation task, a grammar must cover all phenomena that occur in sentence σ , for some significant proportion of randomly drawn sentences σ . This leads immediately to an emphasis on large, systematic grammars, and a need for large, labeled data sets to develop them. Hence, there is a natural connection between Assumption 1 (the linguistic abstraction) and an inductive methodology, with a central role for general-linguistic datasets. The connection is not one of logical necessity, but I nonetheless use the terms *externalist* and *inductive* as more or less synonymous designations for the overall approach.

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 (Bos et al. 2017) and a collection of meaning banks that use Abstract Meaning Representation (AMR) (Banarescu et al. 2013); 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 (McDonald et al. 2013). As for an example of a semantic treebank, I may cite the LinGO Redwoods Treebank (Oepen et al. 2002).

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 NLP; and even if systematic grammars have lost their prestige within the currently dominant school of linguistics, such work does continue, especially within documentary linguistics as well as in “feature grammar” frameworks such as LFG and HPSG. Work within these latter frameworks straddles the line between NLP and linguistics. 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 (Copestake & Flickinger 2000). The work has been extended to multiple languages, and the underlying formalism is intended to be language-general.¹⁵

The development of general-linguistic datasets is motivated by a philosophy that differs from the internalist philosophy, but contributing to general-linguistic efforts does not require any radical change in standard linguistic *practice*. The development of such datasets is very much

¹⁵ Although the LinGO project is largely conducted under the assumptions that I espouse, it remains susceptible to the criticisms I raise below in sections 4.1 and 4.2.

in the spirit of language documentation and description, and the expansion of the existing datasets to include a larger variety of languages, particularly endangered languages, is an area where collaboration between NLP and linguistics is particularly urgent. The main requirement, I think, is a willingness to view grammars as tools for description and prediction, rather than as canonical expressions of exclusive truth. The inductive approach values description above speculative theorization. It is not wedded to a particular grammatical formalism, but rather is happy to use any formalism whose meaning, in the sense of what it says about the sound-meaning relation, is clear, and it is happy to convert among formalisms, so as to employ whichever is most convenient for a given task.

Some readers may worry that the inductive approach emphasizes systematicity to the exclusion of narrower, “piecemeal” inquiry into individual phenomena. That would be problematic, in that most linguistic inquiry is of the narrower sort and thus would appear not to be evaluable, under the proposed assumptions. In fact, the linguistic concepts that one uses to define phenomena are for the most part characterizable in terms of the sound-meaning relation, and thus contribute to predictions, even if a particular piece of work does not describe a complete grammar.¹⁶ In standard linguistic practice, the main desideratum is to “get the facts right,” the facts generally boiling down to what the valid expressions of the language are and what they mean. The main differences among the frameworks reside in how one thinks of the ultimate goals, the big picture into which the narrower work fits.

The main practical consequence is how one responds to uncertainties in how to apply the concepts. We can agree that, in order to capture the relation between sound and meaning, it is useful to subdivide expressions into phrases and words, words into morphemes, and morphemes into phones, and we may agree even about the syntactic categories and phonological features to use, and yet still face uncertainty about how to analyze particular cases or particular classes of cases. The structuralists dealt with those uncertainties by defining discovery procedures, which were intended to determine a unique analysis for every example in the corpus, which is to say, in the training set. Now operationally defined concepts of this sort do have a role in data annotation—one cannot get consistent annotation without clear operational definitions—but they do not solve the problem. Arguably, discovery procedures replace one problem with an even harder problem: if it is difficult to make a principled decision in individual cases, surely it is even harder to spell out a way of automating those decisions. One can characterize discovery procedures as an attempt to automate science, which is unnecessary and impossibly difficult.

In contrast to operational concepts, theoretical concepts are elements in a grammar, which is an explicit account of the sound-meaning relation. In the internalist approach, the correct decision in cases of uncertainty is the one that corresponds with the internal mechanisms that human speakers use. The problem with that approach is that it again replaces one problem with an even harder one: if it is difficult to determine which choice gives the best account of the linguistic facts, it is even harder to determine which mechanisms human speakers use. Psycholinguistic results are very difficult to interpret, and do not provide definitive information nor particularly fine-grained information.

In the inductive/externalist approach, the meaning of a theoretical concept is determined entirely by its consequences for the sound-meaning relation. In deciding on the best course in a particular case, one is relying on one’s understanding of the consequences for predictions about the sound-meaning relation, and one is making an informed guess about which of the options will ultimately yield the best predictions about the full (infinite) range of the sound-meaning relation. Simplicity can certainly play a role: in extrapolating from the collected examples, the simpler account has the best chance of being right. But “being right” is ultimately measured by accuracy in predicting the sound-meaning relation in its entirety. This approach relies on good investigator judgment, not on the blind application of discovery procedures, nor promissory notes for a future day when psycholinguistics has been solved, nor empty claims of psychological reality. What is at stake is not ultimate truth, but a judgment about which choice will yield a system that makes optimal predictions.

¹⁶ In addition, it is possible to define experiments that are more narrowly targeted than the ultimate sentence-interpretation experiment. The only requirement is that one can relate the phenomena of interest to well-defined observables for which one can draw samples.

3.3 Critical experiments

Some are likely to counter by holding up cases in which predictions are not tested by drawing samples from a population, but rather in which theory dictates the choice of specific, rare, critical observations.

A famous example is the empirical confirmation of the general theory of relativity by Dyson et al. (1920). 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 gravitational 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 prediction-testing in 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 consensus 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 is 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.

4 The psychological reality of externalism

As we have discussed, internalism considers the sound-meaning relation to be superficial and “apparently useless,” and demands psychological reality. Taking psychological reality seriously all but forces one to go beyond conventional linguistic observations into the realm of psycholinguistics.

By contrast, the linguistic abstraction of Assumption 1 leads immediately to a sentence-interpretation experiment that requires no observations that fall outside of conventional linguistics. I pointed out that devising a grammar that does well on large, systematic data sets is highly non-trivial, without additional requirements of psychological reality.

I would like to show now that, although the externalist sentence-interpretation task requires no unconventional *observations*, it does push us to go beyond conventional linguistic *theories*. In particular, doing it well requires one to address issues that are usually considered to belong to language processing. In other words, despite the externalism of Assumption 1, the logic of the problem itself imposes a kind of organic psychological reality.

4.1 “Disambiguation”

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 (I_R), and which sentences are natural expressions of a given meaning (E_R). 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 $E_R(\mu)$. In short, the prediction tasks represented by E_R and I_R are fundamental and elementary tasks of linguistics. Even so, no contemporary linguistic theory models them or even attempts to model them.

Readers may find that comment puzzling, so let me give an example to make the point clearer. 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 single best interpretation. Since disambiguation is a matter of processing, it is beyond the scope of linguistic theory.

Such a response sounds reasonable at first, but it is actually nonsensical. The function I_R 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 I_R 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 an unambiguous 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 indeed have an effect on the mapping R .¹⁷ Let us represent that dependency by writing R^y for the sound-meaning mapping that obtains in context y . 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 I_R and E_R , or I_{R^y} and E_{R^y} for any other context y . That is, a conventional grammar does not define a relation \hat{R} between sound and meaning, but that relation does not make good predictions about R^y for any choice of y . Specifically, $I_{R^y}(\sigma)$ is a singleton set, for most contexts y and sentences σ , but $I_R(\sigma)$ is usually a large set. If one responds to that discrepancy by dismissing the singularity of $I_{R^y}(\sigma)$ 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 $I_{R^y}(\sigma)$ is a singleton for most sentences σ in most contexts y . If sentences were genuinely as ambiguous as conventional grammars have it, communication would be impossible.¹⁸ Consider: a speaker has a meaning μ in mind and chooses a sentence $\sigma \in E_R(\mu)$ to express it. The hearer receives σ and must choose a meaning

¹⁷ For discussion of the role of context in elicitation of meaning judgments, see Matthewson (2004).

¹⁸ Treebanks rely critically on the singularity of $I_{R^y}(\sigma)$ 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.

$\mu' \in I_R(\sigma)$. A communication failure occurs if $\mu' \neq \mu$. Assuming the hearer chooses uniformly at random from $I_R(\sigma)$ and that $I_R(\sigma)$ 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 two on average, and certainly far smaller than the rate of ambiguity predicted by conventional grammars.

Rather than dismissing the discrepancy between I_{R^*} and I_{R^*} 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 $I_R(\sigma)$ nor for production judgments $E_R(\mu)$ —but to assert that a conventional grammar does provide an account of what we might call a *context-free* sound-meaning relation:

$$R^*(\mu, \sigma) \leftrightarrow \exists \gamma R^\gamma(\mu, \sigma).$$

This provides context-free interpretation and production functions:

$$E_{R^*}(\mu) = \bigcup_{\gamma} E_{R^\gamma}(\mu) \quad I_{R^*}(\sigma) = \bigcup_{\gamma} I_{R^\gamma}(\sigma)$$

These in turn correspond to the following questions:

- Q1* Given meaning μ , what sentences σ exist such that σ can be naturally used to express meaning μ in at least one context?
- Q2* Given sentence σ , what meanings μ are such that there is at least one context in which σ can mean μ ?

I think this is in fact the right way to understand the predictions of conventional grammars. But there are two issues. First, answering question Q1* or Q2* is not a simple matter of making a native-speaker judgment. It takes considerable training and ingenuity to come up with relevant contexts. Untrained native speakers cannot answer Q1* or Q2*, and even with training and ingenuity, there is no guarantee that one has discovered all possibly relevant contexts. Second, Q1 and Q2 are the more fundamental questions. One poses Q1 (resp., Q2) repeatedly in the course of answering Q1* (resp., Q2*). Moreover, $E_{R^*}(\mu)$ cannot be computed from $E_{R^*}(\mu)$ nor can $I_{R^*}(\sigma)$ be computed from $I_{R^*}(\sigma)$. Namely, in going from E_{R^*} to E_{R^*} or from I_{R^*} to I_{R^*} , one throws away information by pooling across contexts. The information about which reading goes with which context cannot be determined, given only E_{R^*} or I_{R^*} . And the fact remains that conventional grammars provide no model for E_{R^*} or I_{R^*} .

One reason why conventional grammars focus on the context-free functions instead of the more fundamental context-specific functions is that, to model the context-specific functions well, it appears that one needs to consider world knowledge and probabilities.¹⁹ This cedes the point with which I began the discussion—that contemporary grammars fail to model E_R and I_R —but it does provide a way of viewing conventional grammars in which they at least make an indirect contribution to our our understanding of R .²⁰

Without making any particular claims about mental processes, we may linguistically model $I_R(\sigma)$ in two derivational steps. In the first step, a set of meanings (or structures) $I_{R^*}(\sigma)$ is generated, modeling $I_{R^*}(\sigma)$. In the second step, the meanings that are salient in context are selected to yield $I_{R^*}(\sigma)$, either by a model of pragmatics (loosely speaking), or by a context-specific weighting.²¹ The advantage of this approach is that world knowledge and probabilities are needed only for the second step.

¹⁹ 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.

²⁰ One may justify avoidance of probabilities and world knowledge, and modeling of R^* , as an effort to “cut nature at its joints.” Structuring one’s theory so that it captures what one believes to be “nature’s joints” is perfectly legitimate; but the fact remains that a model of R^* is only a partial theory of R . If R is the fundamental linguistic relation, a theory of R^* is only a partial theory of linguistics. Partial theories risk being untestable, though in this case the necessary observations (Q1* and Q2*) are for the most part obtainable, even if they are not the most fundamental observations.

²¹ The idea of a weighting as second step comes from Collins (1999). He proposes dividing the parsing problem into a *model* consisting of grammar plus weights, and a *parsing algorithm* that, for a given sentence, constructs the tree that is optimal according to the model.

Although the two-step approach can be metaphorically thought of as a disambiguation procedure, I would like to emphasize that there is nothing intrinsically computational about it, as defined. Note that *derivation* and *generation* in the preceding paragraph are mathematical terms, not computational terms. They concern only the definition of sets and functions. It is possible that the two derivational steps correspond to two separate steps of mental processing, but it is not necessary; that is a question for psycholinguistics.

4.2 Reversibility

To this point, we have focussed on sentence interpretation. Evaluating a model's predictions on the production task is also important, though more difficult. Just as a systematic approach to the interpretation task requires a model that incorporates components that are conventionally relegated to language processing, addressing the production task requires an attention to issues that are conventionally dismissed as processing matters. An adequate account of the relation between sounds σ and meanings μ must satisfy at least the following criteria.²²

- i. It must define $I_{\hat{R}}(\sigma)$ for all sentences σ ,
- ii. The function $I_{\hat{R}}(\sigma)$ must accurately predict human interpretation judgments,
- iii. It must define $E_{\hat{R}}(\mu)$ for all meanings μ that higher cognition may produce, and
- iv. The function $E_{\hat{R}}(\mu)$ must accurately predict human production judgments.

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

Meanings are obviously less tangible than sentences. The manner in which meanings are represented will necessarily be theory-internal, since direct observations are impossible, but the meaning representation must capture those properties of meanings that are subject to observation. If we take an externalist approach to representing meaning—if, roughly speaking, we seek to characterize the representations implicated by the “laws of thought,” in the words of Boole (1854)—we largely recapitulate the history of the development of the predicate calculus. The programme, from the time that Leibniz (1677/1951) coined the term until the complete first-order predicate calculus was defined (Frege 1879), aimed to boil thought down to its essence, abstracting away from the details of mental processes to the logical structure of thoughts and inference. In short, externalism suggests that one represent meanings using a formal language such as the predicate calculus that supports logical inference or a similar abstract model of thought processes.

To say it plainly, the μ in $R(\mu, \sigma)$ should not actually be a model-theoretic construct, but rather a formula of some appropriate logical language, which itself has a model-theoretic interpretation. Formulae μ serve as abstract representations of thoughts, and there is a space of formulae \mathcal{M} that may arise in the course of human higher cognition. The production task involves sampling formulae μ from \mathcal{M} and testing the predictions of the grammar regarding $E_{\hat{R}}(\mu)$.

Although standard accounts in semantics assign model-theoretic denotations directly to natural-language sentences, it is straightforward to recast a standard account as defining a mapping from a sentence to an expression in the chosen meaning-representation language. Let us write f for the translation function from parse tree τ to meaning representation $\mu = f(\tau)$. Then $I_{\hat{R}}(\sigma)$ is the set of $f(\tau)$ where τ ranges over parse trees that represent possible readings of σ . Inversely, we may define $E_{\hat{R}}(\mu)$ as the set of σ such that $f(\tau) = \mu$ for some parse τ of σ .

In the standard approach, semantic theory assigns a unique meaning $\mu = f(\tau)$ to each parse tree τ , but it pays no attention to the *range* of f . The reversibility problem arises if the range of f differs from the set of meaning representations \mathcal{M} that higher cognition might produce. The issue is not one of “ineffable” thoughts; let us assume that every meaning $\mu \in \mathcal{M}$ is related to at least one expression σ by the true sound-meaning relation R , hence that $E_{\hat{R}}(\mu)$ is nonempty for any $\mu \in \mathcal{M}$. The issue is also not one of gross incompleteness of the grammar. Let us suppose that, for any meaning $\mu \in \mathcal{M}$ that is expressible as sentence σ , the grammar interprets σ as μ' , where μ and μ' are logically equivalent. The problem is that conventional grammar-construction in semantics concerns itself exclusively with $I_{\hat{R}}$ and pays no attention to $E_{\hat{R}}$. Thus it is possible to

²² I omit the context superscript y to avoid clutter; the reader may supply it.

have $R(\sigma, \mu)$, $I_{\hat{R}}(\sigma) = \mu'$, and $\mu \leftrightarrow \mu'$, and yet have $E_{\hat{R}}(\mu) = \emptyset$. The grammar incorrectly predicts that μ is inexpressible. Standard approaches make no effort to avoid this situation, so we must assume that it arises; it does indeed arise for practical NLP systems that interpret sentences by translating them to a logical calculus.

The reader might suppose that an easy fix is available. Namely, we have $\hat{R}(\sigma, \mu')$ and $\mu' \leftrightarrow \mu$, so let us expand \hat{R} to include $\hat{R}(\sigma, \mu)$. That is, to make a prediction regarding $E_{\hat{R}}(\mu)$, return any sentence σ such that $I_{\hat{R}}(\sigma) = \mu'$ and $\mu' \leftrightarrow \mu$. But that proposal is not viable: we cannot enumerate all expressions μ' that are logically equivalent to μ , because logical equivalence is undecidable. This issue has been discussed in the NLP literature (Shieber 1993). Even if the undecidability of logical equivalence were not an issue, the proposal does not give us a tractable way of computing predictions. Blindly enumerating sentences will not do; we must rather invert the model's semantic-interpretation function f . But conventional proposals for $f(\tau)$ are not invertible. The usual formulations of f involve beta reductions, for example, and the inverse of beta reduction is infinitely ambiguous.

I do not believe the problem to be insoluble. The human production function $E_R(\mu)$ is presumably such that one need not search the space of μ' that are logically equivalent to μ , but it may be such as to require a more limited kind of equivalence. To give a concrete example, we might have $R(\sigma, \mu) \leftrightarrow R(\sigma, \mu')$ if μ and μ' have the same Conjunctive Normal Form (CNF) representation (Abney 2018). Having the same CNF representation implies logical equivalence, but not all pairs of logically-equivalent formulae have the same CNF representation. That is, the reversibility problem may have a solution that tells us something interesting about the sound-meaning relation.

To summarize, Assumptions 1 and 2, despite their externalism, have a kind of psychological reality, in that constructing a model that does well on the associated prediction tasks takes us beyond the traditional boundaries of linguistics into areas usually relegated to processing. This variety of psychological reality is intrinsic to the tasks themselves, when they are approached systematically.

5 Inductive General Grammar

Let us now turn from 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 1 *The empirical learning function $L : D \rightarrow R$, where D is primary data and R is the sound-meaning relation, 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* \hat{L} that maps D to a particular grammar G , and the sound-meaning relation predicted by G models the output of L . That is:

$$R = L(D)$$

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

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

My formulation is more or less the standard approach in machine learning, though somewhat adapted to the linguistic context. In machine learning, there is a target function f , and D provides information that the learning algorithm uses to construct a hypothesis \hat{f} that approximates f . In my formulation, the target is R , and the learning algorithm \hat{L} constructs a grammar G that represents the hypothesis \hat{R} . The target and hypothesis are relations rather than functions, but they can each be thought of as a pair of functions: E_R and I_R for the target R , $E_{\hat{R}}$ and $I_{\hat{R}}$ for the hypothesis \hat{R} .

The natural measure of the quality of a general grammar is as follows.²³

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

Assumption 3 takes learning to be the crux of general linguistics (a.k.a. universal grammar), and is in that respect in agreement with the Chomskyan paradigm. 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 the field of machine learning.

Instead of applying the results of machine learning, the strategy in the Chomskyan approach has been to avoid learning by making 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 (Chomsky 1981). 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

²³ For the sake of concreteness, what I have in mind is something along the following lines. General-grammar quality may be measured as:

$$\mathbb{E}_{R \sim \text{Pr}(R), D \sim \text{Pr}(D;R)} \text{loss}(\hat{R}, R')$$

where $R' = L(D)$, $\hat{R} = \mathcal{F}(\hat{L}(D))$, and:

$$\text{loss}(\hat{R}, R) = \mathbb{E}_{\sigma, \mu \sim \text{Pr}(\sigma, \mu; R)} \left[\mathbb{1}_{E_{\hat{R}}(\mu) \neq E_R(\mu)} \right] + \left[\mathbb{1}_{I_{\hat{R}}(\sigma) \neq I_R(\sigma)} \right]$$

Note that R is the language of the speech community, R' is the language learned by the child, and \hat{R} is the language that the general grammar *predicts* the child should learn. If we assume that the child learns the community language correctly, which is to say that $R' = R$ on average (more precisely, that the expected loss of R' is zero), then we can simplify the measure of quality to:

$$\mathbb{E}_{R \sim \text{Pr}(R), D \sim \text{Pr}(D;R)} \text{loss}(\hat{R}, R)$$

This brings the formulation more in line with the usual problem statement in machine learning.

The formula for grammar quality assumes a probability distribution $\text{Pr}(R)$ over languages. That assumption is somewhat problematic. It is clear that the “convenience samples” consisting of all languages currently spoken on Earth, or all languages for which we have documentation, are biased samples, with strong correlations within families. It is desirable to try to reduce that bias by sampling across as many families as possible, but I am not in a position to offer a definitive method for reducing sampling bias. More fundamentally, it is unclear how exactly the population of natural languages should be defined.

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. Lappin & Shieber (2007) give additional examples and discuss the issue in considerably greater detail. In short, the parametric grammar strategy is not only speculative, based on presumptions rather than investigations, the presumptions on which it is based are ill-informed and at least partially erroneous.

The alternative is an empirical investigation of what is learnable and of what hypotheses regarding general grammar are most effective, as measured by their ability to actually learn languages. That alternative may seem far beyond our current abilities, but it is not. Work on learning languages is currently being carried out in NLP, and has spawned a substantial literature. A key venue is the Conference on Natural Language Learning (CoNLL), which has sponsored a number of “shared tasks” on the learning of entire languages, to push research forward through friendly competition on common ground (Zeman et al. 2018). The shared tasks in fact were a key driver behind the development of the Universal Dependencies Treebanks (UDT), mentioned earlier. The main shared task is called *multi-lingual dependency parsing*. It is a supervised learning task, hence not immediately relevant to our interests, but it can be modified to provide an approximate version of the induction of a general grammar.

In outline, multi-lingual dependency parsing has the following form. Given a hypothesized general grammar, in the form of a learning algorithm \hat{L} , a test set of languages is selected at random from the UDT, \hat{L} is used to learn a model for each test language, and the success of learning is measured by the prediction accuracy of the model. As a practical matter, parsing is used as a proxy for the interpretation function, since meanings are not explicitly represented in the UDT. If we accept a parse tree as an approximate meaning representation, a parsing model does qualify as a grammar in a certain form: it computes the function $I_{\hat{R}}(\sigma)$, which determines \hat{R} .

There are three learning tasks that have the overall form just sketched, but which differ in how much information is made available to the learning algorithm. The most common but least interesting task is *supervised learning*, already mentioned, in which the learner is given parsed training data in the test language. That is, in supervised learning, \hat{L} learns a parser based on parsed examples.

In the remaining two tasks, test-language trees are used *only* to evaluate hypothesized parsers; the learning algorithm is not given access to them. In *unsupervised learning*, the learning algorithm is given unparsed test-language text, and nothing else. The learning algorithm instantiates assumptions about the parameters of variation in language and how to identify structure “from scratch” (Klein 2005; Spitzkovsky et al. 2012). The methods that have been explored are not implementations of proposals from linguistics, but they at least broadly accord with common linguistic assumptions about language learning algorithms.

The third learning task is *transfer learning*. As in unsupervised learning, the only test-language information that the learning algorithm receives is unparsed text, but the learner is also given a second set of treebanks (the training treebanks), containing languages other than the test language. The training treebanks are used to *induce* a general-linguistic model in the form of a “universal parser,” obtained by generalizing over the rules of the particular grammars for the languages in the training set. The universal parser is then applied to unparsed test-language text to create a crude treebank, which can in turn be used to bootstrap a language-specific parser for the test language (McDonald et al. 2011; 2012; Naseem et al. 2012).²⁴ These methods are called transfer methods because general knowledge about language induced from the training set is being transferred and applied to the test set.

A small modification to transfer learning gives us a concrete instance of the inductive general grammar paradigm. Namely, instead of expecting the construction of a universal parser to be done automatically by the learning algorithm, let us view the training treebanks as raw material for the linguist to use in inducing a general grammar. There is, after all, no motivation for treating the induction of a general grammar (in the form of a universal parser or in any

²⁴ In addition to the cited works, the shared task of CoNLL 2018 (Zeman et al. 2018) provided no within-language training data for a handful of target languages, making that portion of the shared task a transfer learning task.

other form) as a learning task. It is more properly a matter of theory formation, to be performed by the linguist. The induction algorithms used in the current transfer-learning systems may, however, provide useful machine assistance in the construction of a model of general grammar.

From a linguistic perspective, there is plenty of room to criticize the NLP language-learning efforts. The use of parsers in lieu of particular grammars may make the work seem remote from linguistic sensibilities, and the lack of semantics or real-world context for learners are significant limitations that have been accepted in the interest of tractability. None of the current unsupervised or transfer-learning algorithms yield learned parsers that are as good as parsers created by supervised learning from a treebank for the target language. Even so, clear progress has been made in improving the predictive accuracy of learned parsers, giving reason to believe that research within the inductive general grammar paradigm is possible and practical. Two results of interest are that, at least across the range of proposals that have been evaluated so far, transfer methods significantly outperform unsupervised methods, and that hybrid methods that incorporate traditional typological universals into transfer perform even better than transfer alone (Naseem et al. 2012; McDonald et al. 2012).

In conclusion, the inductive approach that Bloomfield foresaw is not only reasonable and possible, but something very much like it is already being pursued, albeit by NLP researchers rather than linguists. I believe current efforts would benefit by greater involvement of linguists, particularly in the development of treebanks for an even broader and more typologically diverse range of languages, in improving the annotation systems that have become standards in NLP, in addressing linguistic questions overtly rather than incidentally to technology-driven efforts, and, most importantly, in developing a general grammar with the assistance of large datasets and induction algorithms. I hope I have also made a credible case that pursuing the inductive approach to general grammar is of benefit not only to NLP but also to linguistics, and may represent our best strategy for addressing the fundamental questions of linguistics.


Acknowledgements

I am grateful to Patrice Beddor for encouragement and comments on an early draft. I have also benefited greatly from thoughtful and detailed comments provided by anonymous reviewers.

Competing interests

The author has no competing interests to declare.

Author affiliation

Steven Abney  orcid.org/0000-0002-7467-6690
University of Michigan, US

References

- Abney, Steven. 2018. A bidirectional mapping between English and CNF-based reasoners. *Proceedings of the Society for Computation in Linguistics (SCIL)* 1.
- Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer & Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the linguistic annotation workshop*, 178–186. Association for Computational Linguistics.
- Bloomfield, Leonard. 1933. *Language*. New York: Holt.
- Boole, George. 1854. *An investigation of the laws of thought*. London and Cambridge: Walton and Maberley and Macmillan and Co.
- Boolos, George S. & Richard C. Jeffrey. 1980. *Computability and logic*, second edition. Cambridge University Press.
- Bos, Johan, Valerio Basile, Kilian Evang, Noortje Venhuizen & Johannes Bjerva. 2017. The Groningen Meaning Bank. In Nancy Ide & James Pustejovsky (eds.), *Handbook of linguistic annotation*, 463–496. Berlin: Springer. DOI: https://doi.org/10.1007/978-94-024-0881-2_18
- Carnie, Andrew. 2012. *Syntax: A generative introduction*, third edition. Wiley-Blackwell.
- Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton. DOI: <https://doi.org/10.1515/9783112316009>

- Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris Publications.
- Chomsky, Noam. 1986. *Knowledge of language*. New York: Praeger.
- Chomsky, Noam. 2004. *Biolinguistics and the human capacity*. Budapest: Text of a lecture delivered at MTA.
- Collins, Michael. 1999. *Head-driven statistical models for natural language parsing*. University of Pennsylvania dissertation.
- Copestake, Ann & Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd conference on language resources and evaluation (LREC-2000)*, European Language Resources Association.
- De Saussure, Ferdinand. 1986. *Course in general linguistics*. Chicago and La Salle, Illinois: Open Court.
- Dyson, Sir F. W., A. S. Eddington & C. Davidson. 1920. 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*.
- Frege, Gottlob. 1879. *Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Halle a/S: Verlag von Louis Nebert.
- Klein, Dan. 2005. *The unsupervised learning of natural language structure*. Palo Alto, California: Stanford University dissertation.
- Lappin, Shalom & Stuart M. Shieber. 2007. Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics* 43. 393–427. DOI: <https://doi.org/10.1017/S0022226707004628>
- Leibniz, Gottfried. 1677/1951. Preface to the general science. In *Leibniz: Selections*. New York: Charles Scribner's Sons.
- Lewis, David. 1975. Languages and language. In *Language, mind, and knowledge*, 3–35. Minnesota studies in the philosophy of science, Volume 7.
- Marr, David. 1982. *Vision*. W.H. Freeman and Co.
- Matthewson, Lisa. 2004. On the methodology of semantic fieldwork. *International Journal of American Linguistics* 70. 369–415. DOI: <https://doi.org/10.1086/429207>
- McDonald, Ryan, Joakim Nivre, Yvonne Quirnbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló & Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the conference of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics.
- McDonald, Ryan, Slav Petrov & Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of empirical methods in natural language processing (EMNLP)*, 62–72. Association for Computational Linguistics.
- McDonald, Ryan, Oscar Täckström, Slav Petrov, Keith Hall & Joakim Nivre. 2012. Advances in cross-lingual syntactic transfer. In *XLiTe: Cross-lingual technologies, NIPS 2012 workshop*, Neural Information Processing Systems.
- Naseem, Tahira, Regina Barzilay & Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the meeting of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics.
- Open, Stephan, Dan Flickinger, Kristina Toutanova & Christopher D. Manning. 2002. LinGO Redwoods: A rich and dynamic treebank for HPSG. In *Proceedings of the first workshop on treebanks and linguistic theories (TLT)*, Sofia, Bulgaria: BulTreeBank Group.
- Phillips, Colin & Shevaun Lewis. 2013. Derivational order in syntax: Evidence and architectural consequences. *Studies in Linguistics (STIL)* 6. 11–47.
- Shieber, Stuart M. 1993. The problem of logical-form equivalence. *Computational Linguistics* 19. 179–190.
- Soames, Scott. 1984. Linguistics and psychology. *Linguistics & Philosophy* 7. 155–179. DOI: <https://doi.org/10.1007/BF00630811>
- Spitkovsky, Valentin I., Hiyan Alshawi & Daniel Jurafsky. 2012. Three dependency-and-boundary models for grammar induction. In *Proceedings of EMNLP-CoNLL*, Association for Computational Linguistics.
- Tesnière, Lucienne. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.
- Zeman, Daniel, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre & Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies*, Association for Computational Linguistics.

TO CITE THIS ARTICLE:

Abney, Steven. 2021. Inductive general grammar. *Glossa: a journal of general linguistics* 6(1): 75. 1–22. DOI: <https://doi.org/10.5334/gjgl.1332>

Submitted: 01 June 2020

Accepted: 10 May 2021

Published: 03 June 2021

COPYRIGHT:

© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Glossa: a journal of general linguistics is a peer-reviewed open access journal published by Ubiquity Press.