Hypothesis Testing in the Minimalist Program
Mark C. Baker
Rutgers University

In many respects, hypothesis testing in the minimalist program is not very different from what it is for anything that aspires to be a scientific theory of an observable natural phenomenon. The normal practices of science should inevitably apply. Nevertheless, it is true that different sorts of natural phenomena present somewhat different opportunities and challenges, and this may cause good scientific practice to take on somewhat different flavors and emphases in different disciplines. For example, language behaviors are ubiquitous in the lives of most of us, so plenty of potentially relevant data is available. On the other hand, these behaviors are only produced by human beings which have intrinsic value and rights, so there are certain kinds of controlled experiments that might otherwise be desirable that we dare not do (intrusive surgical procedures, raising children in unusual environments, etc.). These factors give linguistics a particular character. More subtly, how hypotheses are tested in the minimalist program depends to some degree on what the minimalist program takes language to be, and hence what an adequate theory of language will be. And in this there can be some differences between the minimalist program and other approaches to the study of language that are worth discussing. In this article, then, I say a bit about both facets of the question: both what I take to be good scientific practice applied to natural language, and what may be specific to the minimalist program in light of its characteristic take on the subject.

At its roots, the Minimalist Program—like earlier versions of generative or "Chomskian" linguistics—takes a language to be a system of knowledge that humans use to construct and interpret sentences (and other units, both smaller and larger). As such, it is inherently about the relationship between the finite and the infinite (Chomsky, 1965:4; 1980:220-224; 1995:3).[1] It is concerned with the infinite, because most people can easily create and interpret an unbounded number of distinct sentences and sentence types. For example, it is likely that most of the sentences contained in this article are new to both its author and to you its reader; perhaps many of them have never been used before in the history of the world. That does not necessarily make the article particularly hard to read. At the same time, the Minimalist Program is concerned with the finite, because every person who knows a language learned it from a finite exposure to that language, and each one has finite mental capacity for storing their knowledge of the language. Therefore, we have the challenge of explaining how a finite amount of experience and knowledge can be used to construct and interpret an infinite (unbounded) range of new sentence types. The answer must be some sort of recursive rule system—a generative grammar in the broadest sense. The quest of finding that system underlies the generative program in all its various incarnations, even as its specific terminologies and technologies have changed and developed over time. Chomsky has often summed up this fundamental point by quoting Wilhelm von Humboldt, that language is "the infinite use of finite means" (Chomsky, 1965:8; 1995:4, 15)

Not all approaches to language see it this way, or at least they do not emphasize this part of the project. For example, usage-based approaches to linguistics emphasize that we list much of what we hear, and are rather conservative in how we go beyond the list. Construction grammars emphasize that

[1] Much of this article is about what practicing generative linguists have internalized from what Chomsky has said over the years about how linguistics is best pursued. His influence on these matters is arguably even greater than on specific analyses. He has made the points in many ways over many years, and the references that I give are merely illustrative; they are not necessarily the earliest or the fullest expositions of these views. References to the rather brief discussion in Chomsky 1995 in particular are included to show that they have been carried forward in the Minimalist Program from earlier work.

we learn many specific "constructions" on a continuum from the most narrow and specific to the most general and productive. In certain cases, these approaches might capture well phenomena that generative grammar has tended to neglect or gloss over. But the problem of how to project a finite experience and storage capacity onto an unbounded capacity to produce and interpret new linguistic examples will not go away. It cannot be true that we explicitly store all the sentence types that we use. Computer science teaches us that combinatorial explosion happens very quickly, even with systems of only moderate complexity. If there are more possible chess games than there are particles in the universe, then clearly there are more possible sentences in English or almost any other natural language. So we cannot represent everything we have heard—especially if it is categorized under everything it might count as an exemplar of, at every level of linguistic description. There must be a way of abstracting generalizable knowledge out of experienced data, and of composing listed constructions into new, more complex constructions. In short, there must be some kind of generative grammar.

What implications does this have constructing and testing hypotheses? Consider first the notion that theories should aim to be simple and elegant—a goal that has been more prominent in the Minimalist Program than in other approaches. The claim is that it can be a road to truth and understanding to compare two theories and ask which is the simpler and more elegant. On one level, this is simply sound scientific practice; it is Ockham's razor, that one should not multiply hypotheses needlessly. If five assumptions are enough to account for a body of data, then we should not make nine assumptions. One cannot claim that the data support the extra four assumptions, even if they are consistent with them. Linguists of every sort should presumably acknowledge this, and it calls all of us to think carefully about the logical structure of our theories and hypotheses, and what they entail. But the generativist's awareness of the relationship between finite knowledge and unbounded capacity perhaps does—and should—put an extra premium on this consideration. It implies that some serious "data compression" is at work in our knowledge of language, that long and complex sentence structures must be the result of interactions and compositions of simpler pieces and patterns. Often it is hypotheses that we consider "simple and elegant" that provide a significant amount of this data compression, by successfully teasing out these interactions. We should therefore value such hypotheses and seek them actively. (See, for example, Chomsky 1995:5, 29—although when Chomsky discusses the value of simplicity he tends to emphasize how fruitful it has been to pursue it in the past, and linguists who are less impressed by the specific successes of generative grammar might be less moved by this.)

Consider next the relationship of a theoretical hypothesis to observable data. Again, there is a sense in which this is a very general criterion that any aspiring scientific theory must meet. Its descriptions of the world must agree with our observations of the world wherever they overlap—and they should overlap often. In Chomsky's term, linguistic hypotheses must be *descriptively adequate*. Every hypothesis is subject to empirical confirmation and disconfirmation, including the Minimalist Program. This puts a qualification on the values of simplicity and elegance discussed above. The theory should be simple and elegant *relative to the phenomena that are described*. If those phenomena are very complex in absolute terms, then a hypothesis can be simple and elegant relative to a naïve description of the facts, and still be rather complex in absolute terms. For example, it might have nine assumptions rather than five—but also rather than 20. The zeal of generative linguists for simplicity and elegance may have caused us to oversimplify and ignore complexity on some occasions. This is an occupational hazard. But denying the significance of simplicity and complexity relative to the observed data is not likely to be helpful, especially in areas where the issue of projecting finite knowledge onto complex novel behaviors is salient, including syntax, compositional semantics, and complex productive morphology. We need to recognize that "small and finite" need not mean "less than ten." But we also need to recognize that the finite is always impressively small compared to the infinite. Nor has generative grammar been guilty of ignoring detail and complexity across the board, I believe. On the contrary, its historical commitment to completeness and explicitness in linguistic theoretical

descriptions, not presupposing "common sense" but trying to explicate it, has led it to make many, many discoveries about the details of natural languages that were unsuspected by previous descriptive efforts—including island phenomena, binding conditions, scope interactions, and so on (cf. e.g. Chomsky 1995:4, 8). This continues into the present, with no signs of slowing.

There are some further subtleties concerning testing a hypothesis by whether it matches the data, however, since different approaches might differ somewhat as to what counts as data, and how that data should be collected. One issue of current relevance is that people have questioned (again) the value of traditional linguistic elicitation, preferring the study of corpuses of naturally occurring data and/or more formal psycholinguistic-style controlled experimentation. Generativists should welcome these "new" techniques as potential new sources of data that could be helpful. But they can and should resist claims that these sources of data are intrinsically superior to the traditional methods of fieldwork, such as constructing new examples and seeing whether native speakers (possibly including the researcher him/herself) accept them or not, and how they interpret them (Chomsky, 1965:18-21; 1986:36-37). Some might think that if a sentence type does not exist in a large corpus, then it may not be real and almost certainly is not worth talking about, but they forget the logic of the finite and the infinite. If we are right that an infinite number of linguistic structures are possible in most languages, then any finite corpus is small, indeed tiny, compared to the total space of legitimate possibilities. So nothing much follows logically from an observation like "X does not occur in a corpus", as long as native speakers' judgments and interpretations of X are robust and consistent. Of course studying a corpus might be valuable in various ways, but it figures to be a poor and clumsy substitute for direct native speaker judgments across the board.

Similarly, it is important to realize that any fieldwork-style interaction is essentially a kind of informal psycholinguistics experiment, seeing how a "subject" responds behaviorally to a stimulus presented by the fieldworker. Knowing this, the sensitive fieldworker should be alert to the possible influence of factors that a good psycholinguist would naturally control for, including the overall length of the stimulus, attentional factors, the order of presentation, and so on. The real question, then, is when will "upgrading" an informal psycholinguistic experiment into a formal one add value. The answer is clearly "sometimes". For some subtle matters where there are confounding factors to be carefully controlled for, using psycholinguistic techniques can be valuable. But there are many things which can be quickly and accurately determined without going through that laborious process—for example the badness for many of **The child seems sleeping*. Doing a laborious psycholinguistic experiment on this would arguably be a waste of valuable tax dollars. And sometimes data from a formal psycholinguistic experiment will be positively misleading—for example, when it avarages over a number of subjects who may have different internalized grammars, or when it is impossible to control for all relevant factors and still get enough stimuli to run the desired statistics. In such situations, traditional fieldwork interviews in which a large amount of data is collected from a single speaker who the interviewer gets to know well will give more accurate results. So psycholinguistic data is not intrinsically different from or better than linguistic data, and sometimes one tool will be the best, sometimes another (see Chomsky 1980:199-202).

The bottom line here is that the fact that people are able to construct and interpret novel sentences spontaneously is itself a very important and instructive datum that needs to be accounted for. Traditional grammaticality judgment tasks, elicitation tasks, and interpretation tasks draw on this crucial capacity rather directly. Hence they are valid and ecologically natural things to do. Of course, mistakes can be made in this method. For example, when linguistic elicitation is not double-blind, it is possible for researchers to unconsciously skew the results in directions that favor their hypotheses. But we also know what to do about that: we correct the mistakes by the normal scientific practices of replication by others, peer review, looking for converging lines of evidence, etc. There is no inherent need to abandon or discount the technique. (We can even experiment with a style of elicitation which is double-blind, in

which say a graduate assistant asks a native speaker for judgments on some paradigm without knowing what the research leader who constructed it predicts.)

Generative practice also differs from some others in how it gets hypotheses that are considered worth testing. Some positivistic approaches strongly favor "bottom up" approaches, where one starts with the data, organizes it and reflects on it in certain ways, and looks for a theoretical hypothesis to emerge. That is a fine way for a generativist to proceed as well. But it is not the only way. In the Minimalist Program, it is considered equally appropriate to proceed in a "top down" fashion, where one first considers the logical consequences of some theory that one finds appealing, and then goes on to test whether those consequences match up with observable data. Most typically, the practicing generative linguist works in cycles, first going "bottom up" from relatively obvious data that we already have to a promising first hypothesis, then working "top down" to find new consequences of that hypothesis and checking them out, often against "second order" (more complex, less common) data, then working "bottom up" again to find new hypotheses that help explain the mismatches, and so on. Some individual linguists might make their best contributions working "bottom up" and others working "top down". In part, the difference in openness to more "top down", deductive approaches may reflect different readings of scientific practice, some inspired primarily by British empiricism, whereas generativists after Chomsky embrace Cartesian rationalist approaches as well. In particular, generative linguistics values a "Galilean" approach to science, in which prestige is attached to an abstract theory that is explicit and detailed enough to make novel predictions about phenomena which have not yet been observed, which are investigated and found out to be correct (or not, in which case one learns something and tries again) (Chomsky, 1980:2-10, 218-219). But there is also a natural connection here to the generativist's goal of explaining how people project a finite experience onto an unbounded range of behavior. From this perspective, people who learn a particular language are often functioning in a kind of "top down" fashion, extrapolating their abstracted knowledge of the patterns and rules of the language into "hypotheses" of what someone meant when they uttered a novel sentence and what they themselves can say and be understood by others. So when generative linguists are functioning in "top down" mode, they are simply mimicking this natural process that native speakers must be doing on an on-going basis. We are seeing if our explicit theoretical hypotheses about a language project onto new structures in the same way that speaker's tacit practical knowledge of the language does.

The mental posture of relating what the linguist is doing to what a person learning and speaking the language is doing leads to perhaps the most distinctive way of testing hypotheses in the Minimalist Program. This is the learnability test (see, e.g., Chomsky 1965:25-27, 55-56, 1986:38); Chomsky often calls it the test of *explanatory adequacy*, as opposed to the test of "mere" *descriptive adequacy*. While our theories may be abstract and have complex interactions, it must be possible in principle for every healthy child in a normal linguistic environment to learn the equivalent of those theories from the data they are exposed to. And that is far from trivial. When generative linguists try to choose between two theories which make the same predictions for the obvious surface facts, they typically do so by intentionally checking more obscure facts, where the two theories make distinctively different predictions. They do this most efficiently by targeted elicitation. But children do not do this: they do not construct sentences to test them for grammaticality; rather they and their care-givers are obsessed by communication. Indeed, the data that some successful language learners is exposed to is unsystematic, not tagged for its relevance, "noisy" in that it contains speech errors, and relatively simple (compared, say, to the sentence structures attested in the *Wall Street Journal* corpus). So how do children succeed in learning their language without using an investigative tool that generative linguists find indispensible?

One answer that generative grammarians give is that the child does not have to learn which hypothesis is correct; it is given to the child innately in some fashion (e.g. Chomsky 1965:47-59, 1980:232-234, 1995:4-6). But by itself this only works for features that all human languages have in common, given that as far as we know any normal child can learn any human language when raised in

the right environment. So if the hypothesis concerns a feature of one language that distinguishes it from another, that hypothesis must in principle be learnable from evidence that any normal child would encounter. For relatively simple and salient properties, like word order, case-marking, and agreement, this criterion is easily met. But languages also differ in much less obvious ways, which linguists only detect in obscure sentences of significant complexity. For example, the conditions under which a question or a relative clause can be formed are rather different in English, Mohawk, and Sakha. But the sentence types are complex, so children wouldn't necessarily expect to hear them, and they should not infer from their not hearing them that such sentences are impossible. Nevertheless, children do learn these differences, because adults know them. We conclude that they must learn differences like these *indirectly*: they must be consequences that can be inferred from interactions of universal-innate knowledge and some more obvious feature(s) of the language that a child can reliably learn. This gives us a substantive additional constraint on our hypotheses. A theory of a given domain might successfully predict the grammatical and ungrammatical structures within that domain, but not relate them in any way to simpler, more learnable features of the language. Such hypotheses must be rejected, in favor of hypotheses that posit the right kinds of connections between simpler data and more complex data (Chomsky 1986:38, 151-152). (For my theory of Mohawk syntax that tries to meet this condition while explaining strange island effects in Mohawk, see Baker 1991, 1996.) This condition exists now mostly in the form of thought experiments and plausibility arguments, because we do not know that much about what the limits of children's capacities for direct learning really are. But even as a thought experiment it is a useful constraint on hypotheses. And as we learn more details of how language acquisition actually happens across an interesting range of languages, we can hope to test generative hypotheses more and more by studying actual language development as well as hypothetical language acquisition (for some steps in this direction, see Snyder 2007).

Returning to more widely shared territory, it follows from this reasoning that generative hypotheses can also be tested typologically. The test of descriptive adequacy discussed above concerns whether a hypothesis about phenomenon X in language Y scales up to other phenomena in language Y; the typological test concerns whether it also succeeds in saying correct things about relevant phenomena in language Z, ideally for all Z. Two subtypes can be distinguished. First, for features of language Y that one wants to say are innate and universal, those should be demonstrably consistent with the observable facts of all other languages, correctly analyzed. These should be valid absolute universals in something like the Greenbergian sense (Greenberg, 1963)—although the useful hypotheses might be at a higher level of abstraction, hence somewhat harder to match up with the easily observed facts of a language than Greenberg's universals were (e.g. Chomsky 1980:218-219, 1986:43). Second, for complex features of language Y that are hypothesized to be acquired indirectly via easily-learned feature A, we can test whether those features generally occur together in other languages that have feature A—whether our hypotheses give the right "parametric clusters." These correspond roughly to Greenbergian implicational universals of the form "A implies X"—although again the right level of analysis may be more abstract than the one often used in existing typological literature.

In practice, the typological test is pursued in two somewhat different ways. One is the *macro-comparative* approach, in which one tests hypotheses of a universal or parametric nature against a set of languages from different families and different parts of the world (Baker, 2008, 2010). This is not so different from what functionalist-typologists do in practice, although with a somewhat different idea of what is an attractive hypothesis for testing. The other is the *macro-comparative* approach, in which one tests hypotheses against a set of closely related languages or dialects, such as the Romance languages or the Germanic languages (Kayne, 2005). Within the Minimalist Program, there have been some debates about which of these approaches is the best to pursue, but those debates are about short term tactics, about what is most productive to do *first*. As far as I know everyone agrees that the two approaches complement each other in principle, and that both should be pursued as much as is feasible. For

example, the macro-comparative approach will often be the strongest test of a universal hypothesis, whereas the micro-comparative approach may be the strongest test of a parametric hypothesis.

Another practical question related to hypothesis testing is when should a hypothesis that has encountered counterevidence be abandoned. Generative linguists have sometimes given the impression of holding on to hypotheses too long, in the face of significant counterevidence. No doubt this is true in some cases (although it is not necessarily distinctive; researchers of all kinds tend to have a stubborn streak). But simple and hard-to-avoid assumptions about the nature of language convince us that language is a complex system where many factors interact, and that some linguistic knowledge must be relatively abstract. Therefore, we cannot know immediately whether the empirical problem that has come to light is a problem with the specific hypothesis being tested, or with some other feature of the system.

The inescapable fact underlying this is that we cannot really test individual hypotheses one at a time, even though this is what we strive to approximate. Ultimately, given that language is a complex interacting system, we can only test *sets* of hypotheses, which work together to derive *sets* of predictions. When the predictions are supported, great. When those predictions are falsified, we must change one or more of the hypotheses, but we cannot immediately know for sure which one to change.[2] In the most typical case, some of the predictions will be supported and a few will be falsified. Then we might be encouraged that our system of hypotheses has some important features of the truth, although there are some corrections to be made in it somewhere too. Since the interactions need not be linear ones, one cannot necessarily measure how many of the hypotheses in the set are likely to be correct from how many of the predictions are correct, so literal counting of counterexamples is unlikely to be helpful. More generally, a mere fact is not enough to overturn a theory. Rather, it is a fact that has no analysis under that theory but that does have an analysis under a rival theory that truly threatens the first theory—especially if a pattern of such facts begins to accumulate. So generative theories are defeatable, but generally not by mere counterexamples; those can and often enough should be attributed to anomalies or to unknown factors. Rather, generative theories can be defeated by better theories, ones that fit better a wider range of observed facts.

A final practical question concerns the gold-standard of hypothesis confirmation that we should be aiming for, versus acceptable levels of confirmation for intermediate results gleaned along the way. Studying complex interacting systems is hard, so we must be both idealistic and practical in how we go about it. We must have a standard of truth and practice that we are aiming for, and stick to it. But we must also have intermediate standards for research that has not gotten all the way there yet, but could be a serious step on the way. Otherwise we will be unable to make progress, given that the object of study is too complex for anyone to figure out in a single step. For example, it seems to me that many of the recommendations proposed by Croft (this volume) would be valid for the Minimalist Program as well, as expressions of the gold standard that the whole community is aiming for over time. I certainly agree that proposals about universal grammar need to be tested crosslinguistically against many unrelated languages. I also agree that theoretical constructs should in principle be justifiable for each language they apply to using evidence gleaned from within that language. But those are high standards that cannot be applied immediately to every intermediate result or we will put an impossible burden on linguistic research. For example, I think it would be a mistake to lay down a standard of the form “no

[2] And in dealing with this, it makes perfect sense to make different choices of which phenomena are “figure” and which are “ground” in different studies. Picking up on one of Croft’s (this volume) examples, in one study one might use different sorts of direct objects to study the phenomenon of passive, and in other one might use the passive as a “test” to study what is a direct object. In studying complex systems, anything can be instrument and anything can be object as we try to bootstrap our way to better overall knowledge. I believe that this is well-understood by practicing generative linguists, although it may not be clearly expressed in introductory textbooks.

proposal about a language universal will be taken seriously unless it has been tested on at least 50 unrelated languages"—even though that standard may seem very modest to some. For my money, a hypothesis that has a great deal of logical coherence, that makes distinctive correct predictions about complex new examples that were engineered to test it, and that can be justified by facts internal to five unrelated languages could be more promising than a shallow hypothesis that is statistically significant over 50 languages (although even this may be too much to expect for an average journal article, and needs to be broken down further.) Promising hypotheses present themselves in different ways, and the gold-standard testing of them is a job for the whole community, not a single researcher. Perhaps this is more evident to the minimalist than to other kinds of linguists, if we are focusing on the more complex parts of language, where several things interact, but I imagine that a distinction between the gold standard of truth and a reportable intermediate result is useful to all linguists.

In conclusion, this article has reviewed what generative linguists take the essence of language to be, and how this affects the hypotheses they entertain and how they test them. Some of the methods are general to any scientific endeavor, including empirical support or falsification, favoring hypotheses with fewer assumptions, the ability to make testable correct predictions, and generality (e.g., over many languages). Others are more distinctive, involving particular choices about how to do science, and a particular idea of what a linguistic theory needs to be. These include accepting more abstract hypotheses, which are connected to data but indirectly via other hypotheses, accepting directly elicited data, appreciating top-down deductive reasoning in the development and testing of hypotheses, the criterion that a hypothesis must be learnable, and perhaps a different sensibility as to what constitutes a useful intermediate result. It is a good overall package—needing to be practiced well rather than badly, to be sure, but not in need of fundamental change. Linguists of all kinds are invited to use as much of it as they will.

References


Baker, M. (1991). On some subject/object non-asymmetries in Mohawk. *Natural Language and Linguistic Theory, 9*(4), 537-576.

Baker, M. (1996). *The polysynthesis parameter*. New York: Oxford University Press.

Baker, M. (2008). The macroparameter in a microparametric world. In T. Biberauer (Ed.), *The limits of syntactic variation* (pp. 351-374). Amsterdam: John Benjamins.

Baker, M. (2010). Formal generative typology. In B. Heine & H. Narrog (Eds.), *Oxford Handbook of Linguistic Analysis*. New York: Oxford University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass.: MIT Press.

Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.

Chomsky, N. (1986). *Knowledge of language: its nature, origin, and use*. New York: Praeger.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, Mass.: MIT Press.

Croft, W. (To appear). Hypothesis formation. In S. Luraghi & C. Parodi (Eds.), *A Companion to Syntax*.

Greenberg, J. (1963). *Universals of language*. Cambridge, Mass.: MIT Press.

Kayne, R. (2005). Some notes on comparative syntax, with special reference to English and French. In G. Cinque & R. Kayne (Eds.), *The Oxford Handbook of Comparative Syntax* (pp. 3-69). New York: Oxford University Press.

Snyder, W. (2007). *Child language: the parametric approach*. New York: Oxford University Press.