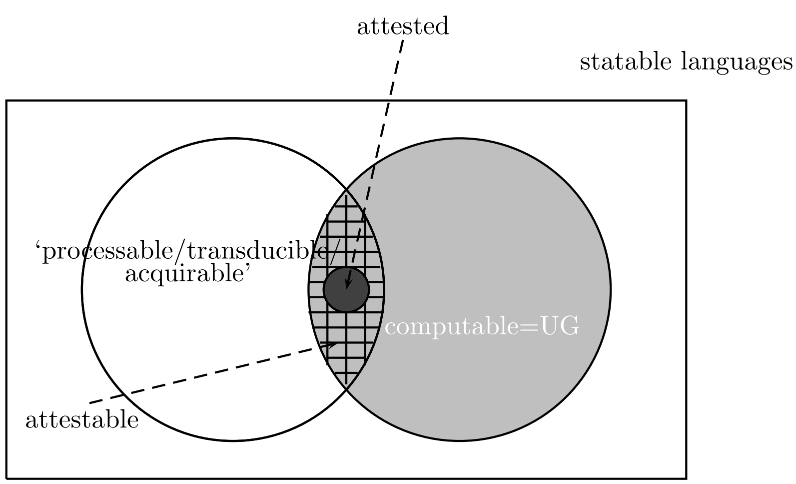
To appear in L. Grestenberger, C. Reiss, H. A. Fellner, and G. Z. Pantillon, editors. *Ha! Linguistic Studies in Honor of Mark R. Hale.* Dr. Ludwig Reichert Verlag, 2022.

Plastics

Charles Reiss
Concordia University

At some point in our realization that the various uses of the term *markedness* in the phonological typology literature were both incoherent and mutually contradictory, Mark Hale and I came up with some version of the following diagram (e.g., Hale and Reiss, 2008, p.4):

(1) What is Universal Grammar a Theory of?



We thought that the point we were trying to make would turn out to be obvious, once we explained the picture. The rectangle represents the set of all possible mathematically definable languages. Linguists are not interested in languages (sets of strings) where each string has a length corresponding to a prime number, for example. That might be interesting for mathematical linguistics, but not for biolinguistics, which is an empirical science concerned with human language. Restricting ourselves to the two big circles, the idea is that the human language faculty has certain properties, and that each human language is defined by some combination of those properties, so the large light gray circle represents the set of human language defined *intensionally* by a model of Universal Grammar. Since the language faculty is distinct from other properties of humans, such as their lifespans, their auditory systems, the size of their memory buffers and so on, a human language can only be attested if it is not only UG-consistent, but also consistent with all those other systems, which we lump together as "performance" systems, corresponding to the white circle. This intersection corresponds to the hashmarked region, the attestable languages. So, a language (defined as a UG-built system) can only be attested if it also is compatible with all the human performance systems (well, in the case of signed vs. spoken language, there are some alternatives). Of course, we have just sketched *minimal* requirements for attestability—they are no guarantee that a language actually *is* attested. In other words, we don't want our theory of UG to over-fit the data of what languages happen to have been attested, which corresponds to the small dark gray circle and depends on things like where graduate students went to do fieldwork or where a plague happened to hit hard. So we don't even want our theory to fit the set of languages that could potentially be attested if we just walked into the next valley—we need a more general model.

We thought the point was pretty clear and not very original, but obviously correct. However, there continues to be a lot of resistance to this idea that the set of languages definable by linguistic theory *must* be larger than the set of languages for which we could ever come across data. One phonologist, in a leaked email, said this of our diagram:

> Phonologists are interested in the patterns we see, and describing all of them, but nothing but them. The object of study cannot be a larger set than the attested data. This is kindy-level foolishness.

Mark and I have often been accused of foolishness, together and individually, but never before the kindy-level kind. To support our view of the purview of UG, we've offered several arguments over the years, and others (e.g., Kaplan 1987; Newmeyer 2005 have made similar points). Here, I'd like to offer one word, a word that should be sufficient to anyone old enough to receive a Festschrift: *plastics*.[1]

If chemists had been interested in the chemicals they saw, describing all of them, but nothing but them, we wouldn't have all the wondrous new materials that decorate our beaches, fill our Amazon boxes, and make it possible for almost everyone to have a comfortable place to sit.[2] Fortunately, chemists were interested in the principles that determined the structure of molecules, and thus were able to create new ones.

Even below the molecular level, at the level of elemental atoms, we see that physicists do not act like the object of study simplistically matches the "attested data". If physicists had been so naive, they would never have invented synthetic elements like tennessine and copernicium. In some cases, like technetium and plutonium, the elements were *first* synthesized, and then discovered in nature. Again, none of this would be possible if physicists had been as literal-minded about what constitutes "attested data" as some phonologists are.

I said I would offer one *word*, so in order to keep going, I'll use a number, instead: ten to the billionth power. Basic math gives us another reason to reject

[1] If you aren't old enough, see https://youtu.be/eaCHH5D74Fs.

[2] See https://www.vice.com/en/article/bn5e4m/white-plastic-chairs-are-taking-over-the-world-128.

our anonymous critic's view of the relationship between the "attested data" and the object of study, the human language faculty. If we assume that segments are just sets of binary phonological features like $\pm$Round, then a Univeral Grammar with just three features, say F, G, H intensionally defines eight (2^3) possible segments, including $\{+F, -G, +H\}$, and so on. Building on the work of Pat Keating and others, Mark Hale has spent a lot of time discussing the evidence that even surface segments may lack specification for some features, and for his 2007 book, he developed what has become the most popular transcription system in the world for such underspecified segments, at least for the Marshallese vowels. Given underspecification, feature combinatorics turns to powers of 3, not 2, so with the same three features we get twenty-seven (3^3) segments, including $\{+F, -G, +H\}$, $\{-F, +G\}$, and so on. To make this more concrete, recall that using four features in his treatment of the Marshallese vowels, Mark posits the four segments in (2) with specification for the features High and Atr, but no specification for Back or Round:

(2) Marshallese vowels only have specification for [ATR] and [high]

- {+Hi, +Atr}: ☕
- {+Hi, −Atr}: ☎
- {−Hi, +Atr}: ☯
- {−Hi, −Atr}: ⚽

Expanding from the four features in Mark's discussion to a UG with, say, twenty features, we get a system that intensionally defines 3^{20}, about 3.5 billion, segments. Given this set of possible segments, if we consider each language as just a segment inventory, then a language can be characterized as the answer to a string of 3^{20} Yes/No questions like *Is* [u] *in the language? Is* [θ] *in the language? Is* [☕] *in the language?*, etc. There are $2^{3^{20}}$ ways to answer that list of questions—Yes, Yes, No, Yes, No, . . . , or maybe No, Yes, No, Yes, Yes, . . . , etc. This yields about 10^{10^9}, ten to the billionth power, segment inventories or languages. This is a number that Gallistel and King (2009) would call "essentially infinite", being greater than the number of elementary particles in the universe, which they put at around 10^{85}.

So, we have two simple conclusions. First, science, including linguistic science, is never about the attested data. The second is that as soon as one analyzes languages in terms of a base of features, syntactic categories, rule orderings, and so on, simple combinatorics gives us an intensionally defined set of languages for which each particular language has an infinitesimal probability of attestation. The object of study for linguistics is that base of categories, that set of universal building blocks that comprise all individual languages. Ideally that base is small, especially if we want to argue that it is encoded in the genome, but intensionally it defines a set of languages that is beyond astronomically large. Whether or not Mark and I are generally guilty of kindy-level foolishness—or any other kind of foolishness—we seem to be right that the characterization in extension of the object of study, UG, *must* constitute a larger set than the set of languages from which we draw our evidence.

References

Gallistel, C. R., and Adam Philip King. 2009. *Memory and the computational brain: why cognitive science will transform neuroscience*. Chichester, West Sussex, UK: Wiley-Blackwell.

Hale, Mark, and Charles Reiss. 2008. *The phonological enterprise*. Oxford: Oxford University Press.

Kaplan, Ronald M. 1987. Three seductions of computational psycholinguistics. In *Linguistic theory and computer applications*, ed. P. Whitelock, M. M. Wood, H. L. Somers, R. Johnson, and P. Bennett, 149–188. London: Academic Press.

Newmeyer, Frederick J. 2005. *Possible and probable languages: a generative perspective on linguistic typology*. New York: Oxford University Press.