# Last Words

# Ancient Symbols, Computational Linguistics, and the Reviewing Practices of the General Science Journals

Richard Sproat*
Center for Spoken Language
Understanding

## 1. Introduction

Few archaeological finds are as evocative as artifacts inscribed with symbols. Whenever an archaeologist finds a potsherd or a seal impression that seems to have symbols scratched or impressed on the surface, it is natural to want to "read" the symbols. And if the symbols come from an undeciphered or previously unknown symbol system it is common to ask what language the symbols supposedly represent and whether the system can be deciphered.

Of course the first question that really should be asked is whether the symbols are in fact writing. A writing system, as linguists usually define it, is a symbol system that is used to represent language. Familiar examples are alphabets such as the Latin, Greek, Cyrillic, or Hangul alphabets, alphasyllabaries such as Devanagari or Tamil, syllabaries such as Cherokee or Kana, and morphosyllabic systems like Chinese characters. But symbol systems that do not encode language abound: European heraldry, mathematical notation, labanotation (used to represent dance), and Boy Scout merit badges are all examples of symbol systems that represent things, but do not function as part of a system that represents language.

Whether an unknown system is writing or not is a difficult question to answer. It can only be answered definitively in the affirmative if one can develop a verifiable decipherment into some language or languages. Statistical techniques have been used in decipherment for years, but these have always been used under the assumption that the system one is dealing with *is* writing, and the techniques are used to uncover patterns or regularities that might aid in the decipherment. Patterns of symbol distribution might suggest that a symbol system is *not* linguistic: For example, odd repetition patterns might make it seem that a symbol system is unlikely to be writing. But until recently nobody had argued that statistical techniques could be used to determine that a system *is* linguistic.[1]

It was therefore quite a surprise when, in April 2009, there appeared in *Science* a short article by Rajesh Rao of the University of Washington and colleagues at two research institutes in India that purported to provide such a measure (Rao et al. 2009a). Rao et al.'s claim, which we will describe in more detail in the next section, was that

* Center for Spoken Language Understanding, Oregon Health & Science University, 20000 NW Walker Rd, Beaverton, OR, 97006, USA. E-mail: rws@xoba.com.

1 People *have* used the existence of quasi-Zipfian distributions in symbol systems to argue for their status as writing; such claims figure in the work of Rao and colleagues. But because it has been long known that Zipfian distributions hold of many things besides language, such arguments are easy to dismiss.

one could use *conditional entropy* as evidence that the famous symbol system of the third millenium BCE Indus Valley civilization was most probably writing, and not some other kind of system.

That the Indus symbols were writing is hardly a novel claim. Indeed, ever since the first seal impression was found at Harappa (1872–1873 CE), it has been the standard assumption that the symbols were part of a writing system and that the Indus Valley civilization was literate. Over the years there have been literally hundreds of claims of decipherment, the most well-known of these being the work of Asko Parpola and colleagues over the last four decades (Parpola 1994). Parpola, who argues that the Indus Valley people spoke an early form of Dravidian, has produced interpretations of a small set of symbols, but nothing that can be characterized as a decipherment.

The first serious arguments against the idea that the Indus symbols were part of a writing system were presented in work that Steve Farmer, Michael Witzel, and I published in Farmer, Sproat, and Witzel (2004), which reviews extensive support for that view from archaeological evidence and comparisons with other ancient symbol systems. Although our arguments were certainly not universally acknowledged—least of all among people who had spent most of their careers trying to decipher the symbols—they have been accepted by many archaeologists and linguists, and established a viable alternative view to the traditional view of these symbols. It was against this backdrop that the Rao et al. (2009a) paper appeared.

Taken at face value, Rao et al.'s (2009a) paper would appear to have reestablished the traditional view of the Indus symbols as the correct one, and indeed that is how the paper was received by many who read it. A number of articles appeared in the popular science press, with *Wired* declaring "Artificial Intelligence Cracks Ancient Mystery" (Keim 2009). The Indian press had a field day; they had studiously ignored the evidence reported in our paper, presumably because it led to the unpalatable conclusion that India's earliest civilization was illiterate. But Rao et al.'s paper, which appeared to demonstrate the opposite, was widely reported.

The work has also apparently attracted attention beyond the popular science press and those with some sort of axe to grind on the Indus Valley issue, for in March 2010 there appeared in the *Proceedings of the Royal Society, Series A*, a paper that used similar techniques to Rao et al.'s (2009a) in order to argue that ancient Pictish symbols, which are found inscribed on about 300 standing stones in Scotland, are in fact a previously unrecognized ancient writing system (Lee, Jonathan, and Ziman 2010). A trend, it seems, has been established: We now have a set of statistical techniques that can distinguish among ancient symbol systems and tell you which ones were writing and which ones were not.

The only problem is that these techniques are in fact useless for this purpose, and for reasons that are rather trivial and easy to demonstrate. The remainder of this article will be devoted to two points. First, in Section 2, I review the techniques from the Rao et al. (2009a) and Lee, Jonathan, and Ziman (2010) papers, and show why they don't work. The demonstration will seem rather obvious to any reader of this journal. And this in turn brings us to the second point: How is it that papers that are so trivially and demonstrably wrong get published in journals such as *Science* or the *Proceedings of the Royal Society*? Both papers relate to statistical language modeling, which is surely one of the core techniques in computational linguistics, yet (apparently) no computational linguists were asked to review these papers. Would a paper that made some blatantly wrong claim about genetics be published in such venues? What does this say about our field and its standing in the world? And what can we do about that? Those questions are the topic of Section 3.

## 2. The Fallacies

Rao et al.'s (2009a) paper is a typical short paper in *Science* consisting of a page of text and figures, and a link to a longer description that details the techniques and data. The main paper—which is presumably all that most people would read—contains a convincing-looking plot, their Figure 1A, here reproduced as Figure 1. The plot purports to show that *bigram conditional entropy* , defined as

$$H(Y|X) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(y|x) \tag{1}$$

can distinguish between non-linguistic symbol systems and linguistic symbol systems, and that the Indus Valley symbols behave like linguistic symbol systems.

The plot looks very convincing indeed, but what does it mean?

Several aspects of the plot require explanation. First the horizontal axis, labeled as "number of tokens," represents the bigram conditional entropy of subsets of each corpus starting with the subset consisting of the 20 most common symbols, the 40 most common symbols, the 60 most common symbols, and so forth. What we see for each corpus is that the conditional entropy grows over these successive subsets until it approaches the conditional entropy of the corpus as a whole.

Second, the corpora represent small samples of various languages including English (sampled both as words and letters), Sumerian (cuneiform symbols), Old Tamil (largely consonant–vowel combinations in the Tamil alphasyllabary), the Indus Valley corpus due to Mahadevan (1977), and two types of non-linguistic systems (though see subsequent discussion). The sample sizes are small because the Indus corpus against which all other symbol systems are compared is very small. The average length of an Indus "inscription" (in Mahadevan's corpus) is only about 4.5 symbols; the total size of

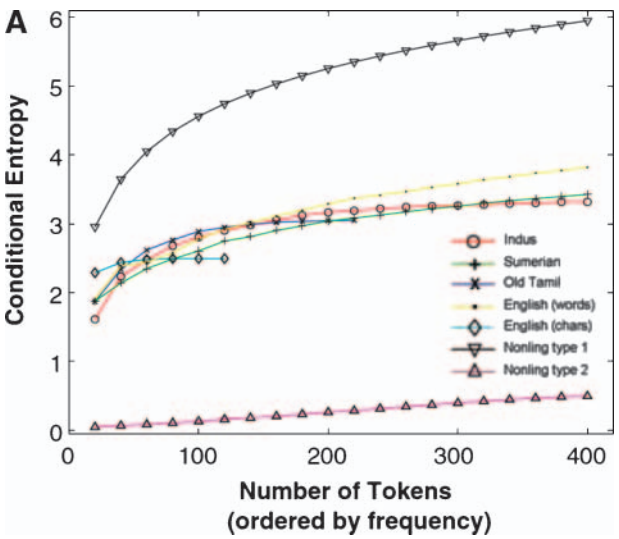


**Figure 1**
Conditional entropies for a variety of linguistic scripts and other symbol systems. From: Rao, Rajesh, Nisha Yadav, Mayank Vahia, Hrishikesh Joglekar, R. Adhikari, and Iravatham Mahadevan. 2009. Entropic evidence for linguistic structure in the Indus script. *Science*, 324(5931):1165. Figure 1A, page 1165. Reprinted with permission from AAAS. Poor quality of the figure is due to poor quality in the original.

the Mahadevan corpus is 7,000 tokens (and about 400 types). Though Rao et al. (2009a) make a point of stressing that they use sophisticated smoothing techniques (a modified version of Kneser-Ney), one must remember that with such small data sets, smoothing can only do so much for you.

Third, the curves labeled as "Type 1" and "Type 2" non-linguistic systems are explained as follows:

> Two major types of nonlinguistic systems are those that do not exhibit much sequential structure ('Type 1' systems) and those that follow rigid sequential order ('Type 2' systems). For example, the sequential order of signs in Vinča inscriptions appears to have been unimportant. On the other hand, the sequences of deity signs in Near Eastern inscriptions found on boundary stones (*kudurrus*) typically follow a rigid order that is thought to reflect the hierarchical ordering of the deities. (Rao et al. 2009a, page 1165)

On the face of it, it is not too surprising, given these descriptions, that the Type 1 system shows rapid growth in the conditional entropy, whereas Type 2 stays close to zero. The problem is that there is little evidence that either of these types accurately characterized *any* ancient symbol system. So for example, the Vinča symbols of Old Europe were certainly not random in their distribution according to the most authoritative source on the topic (Winn 1981).[2] Indeed, Gimbutas (1989) and Haarmann (1996) even proposed that they represented a pre-Sumerian European script; although that is highly unlikely, it is also unlikely they would have proposed the idea in the first place if the distribution of symbols seemed random. Similarly, it is apparently not the case that the deity symbols in *kudurrus* were arranged in a rigid order (see subsequent discussion): Clearly it is not only computational linguists who should be bothered by the claims of this paper. In fact, as one learns only if one reads the supplementary material for the paper, the data for Type 1 and Type 2 were artificially generated from a rigid model (Type 2) and a random and *equiprobable* model (Type 1).

Various on-line discussions, starting with Farmer, Sproat, and Witzel (2009), criticized Rao et al. (2009a) for their use of artificial data.[3] So, in subsequent discussion, including a recently published paper (Rao 2010) that largely rehashes the issues of both the *Science* paper and another paper in *PNAS* (Rao et al. 2009b),[4] Rao backs off from these claims and talks about the Type 1 and Type 2 curves as the limits of the distribution. The take-home message appears to be that in principle symbol systems could vary as widely as being completely rigid or completely random and equiprobable. It is therefore surprising, the story goes, that the Indus symbols seem to fall right in that narrow band that includes unequivocal writing systems. The problem with this argument is that it is highly unlikely that there were *ever* any functional symbol systems that had either of these properties, and one can argue this point on basic information theoretic grounds. A symbol system that was completely rigid—had an entropy of 0—would convey no information whatsoever. If whenever symbol $x$ occurred, sym-

2 Rao et al. (2009a) mis-cite Winn to claim that the Vinča sequences were random.

3 We also summarized our criticisms of the paper in a letter to the editor of *Science*. This was rejected for publication with the note "we receive many more letters than we can accommodate." This seemed an odd excuse given that the letter would presumably be published online rather than in print—so space would not be an issue, and the letter pertained directly to flows in a paper published in the magazine, which one would think would be of importance.

4 Rao et al. (2009b) has one advantage over Rao et al. (2009a) in that they actually *do* show something: They use Markov models to show that there is structure, which they term "rich syntactic structure," in the Indus texts. That there is structure—the system is not random—has of course been known for decades; see Farmer, Sproat, and Witzel (2004) for discussion of this point. And given the average length of the Indus texts of around 4.5 glyphs, one wonders just how "rich" the syntax could have been.

bol $y$ always followed, there would be little point in having more than just symbol $x$, except perhaps for decorative purposes. Even in language one finds pockets of such predictability: The word sequence *Monty Python's Flying* will hardly ever be followed by anything other than *Circus*. For a *whole system* to be so rigid would be unexpected. The other extreme—random and equiprobable—seems equally unlikely in general, if only because symbols represent things, and the things they represent typically do not occur with equal probability. So although Rao is technically correct that his Types 1 and 2 do represent the logical extremes of the distribution, it is not likely that any meaningful symbol systems were ever created that had either of these properties.

In particular it is important to remember that *random* is not the same thing as *random and equiprobable*: at least some of the discussion of Rao et al.'s (2009a) paper (and the Lee, Jonathan, and Ziman [2010] paper we examine subsequently) seems to depend upon the confusion of these two quite distinct notions. If one allows that symbols have a quasi-Zipfian distribution—something that is surely true of linguistic symbol systems, but of many other things too—then one finds curves that look very similar to what Rao et al. find for their "linguistic" systems in their *Science* paper. Thus, as I argued in a contribution to Liberman (2009), one can "get a very good fit to [Rao et al.'s] results for the Indus corpus with a model that has 400 elements with a perfect Zipf distribution, with $\alpha = 1.5$, and conditional *independence* for the bigrams." Similarly in my invited talk at EMNLP'09 (Sproat 2009), I showed that one could replicate their results with an artificially generated corpus that only matched the *unigram* frequencies from the Mahadevan corpus and again had conditional independence for the bigrams. It is not hard to understand why the plot for a randomly generated corpus with a roughly Zipfian distribution should "look like" language using Rao et al.'s methods. There are no constraints on what symbols can follow others, so for the $n$ most frequent symbols there is a large amount of uncertainty. But as one's sample grows to the $2n$ most frequent, the $3n$ most frequent, and so forth, the gain in uncertainty decreases simply because the next $n$ symbols have a smaller overall probability and thus their incremental contribution to the uncertainty is smaller. Furthermore at *no* point will the entropy be maximal: because the distribution of symbols *is not equiprobable*.

In subsequent discussions Rao—for example, Rao (2010)—has defended his position by arguing that conditional entropy and other such measures are not intended to be definitive, but merely suggestive and, when combined with other evidence that points in the same direction, supportive of the conclusion that the Indus system is writing: Simply put, it is an issue of weight of evidence. The problem is that for that argument to work there must at least be some weight: If conditional entropy measures of a particular form correlate more with language than they do with non-linguistic systems, if even weakly, then that might count as evidence for the conclusion. In other words, one wants a measure that can tell one, with better than chance accuracy, that the system in question is (or is not) linguistic. But this has not been demonstrated: Nobody has done the legwork of putting together the needed corpora of ancient linguistic and non-linguistic symbol systems, and demonstrated that one can in fact use such measures to do a better than chance job of classifying systems. The simple experiments involving randomly generated texts discussed earlier do not leave one with much optimism that this will be the case. But one has to admit that it is an open question. But it is *the* question that has to be asked, and the fact that none of the reviewers of the *Science* article thought to ask it speaks to the reviewing practices of that journal, at least as it relates to our field.

We turn now to Pictish symbols. The Picts were an Iron Age people (or possibly several peoples) of Scotland who, among other things, left a few hundred standing stones inscribed with symbols, with "texts" ranging from one to a few symbols in

length. Lee, Jonathan, and Ziman's (2010) paper attempts to use measures derived from entropy to ascertain whether these symbols are part of a linguistic writing system. Similarly to Rao et al.'s (2009a) work, they compare the symbols to a variety of known writing systems, as well as symbol systems like Morse code, and European heraldry, and randomly generated texts—by which, again, is meant *random and equiprobable*. As their title "Pictish symbols revealed as a written language through application of Shannon entropy" suggests, they are much bolder than Rao et al. (2009a) in what they think they have demonstrated.

As with Rao et al.'s (2009a) paper, there are a number of things in Lee, Jonathan, and Ziman (2010) that should bother people other than computational linguists: They characterize Egyptian hieroglyphs as a "syllabic" writing system (it was a consonantal and thus essentially a segmental writing system); they linearize their corpus of European heraldry by reading bottom to top, which follows no conventions that I am aware of; and they refer credulously to the supposed "script" examples from Chinese Neolithic pottery, which few Sinologists take seriously. But again, we focus here on the issues that relate to computational linguistics.

Lee, Jonathan, and Ziman's (2010) techniques are substantially more complicated than Rao et al.'s (2009a), and we do not have space to describe them fully here. One reason for the complication is that they recognize the problem imposed by the very small sample sizes of the corpora (a few hundred symbols in the case of Pictish), and seek a method that is robust to such small sizes. They develop two measures, $U_r$ and and $C_r$, defined as follows. First, $U_r$ is defined as

$$U_r = \frac{F_2}{\log_2(N_d/N_u)} \tag{2}$$

where $F_2$ is the bigram entropy, $N_d$ is the number of bigram types, and $N_u$ is the number of unigram types.[5] $C_r$ is defined as

$$C_r = \frac{N_d}{N_u} + a\frac{S_d}{T_d} \tag{3}$$

where $N_d$ and $N_u$ are as before, $a$ is a constant (for which, in their experiments, they derive a value of 7, using cross-validation), $S_d$ is the number of bigrams that occur once, and $T_d$ is the total number of bigram tokens; this latter measure will be familiar as $\frac{n_1}{N}$, the Good-Turing estimate of the probability mass for unseen events. To illustrate the components of $C_r$, Lee, Jonathan, and Ziman show a plot (their Figure 5.5), reproduced here as Figure 2. According to their description this shows

> [a p]lot of $S_d/T_d$ (degree of di-gram repetition) versus $N_d/N_u$ (degree of di-gram lexicon completeness). . . . Dashes, sematograms—heraldry; filled diamonds, letters—prose, poetry and inscriptions; grey filled triangles, syllables—prose, poetry, inscriptions; open squares, words—genealogical lists; crosses, code characters; open diamonds, letters—genealogical lists; filled squares, words—prose, poetry and inscriptions. (Lee, Jonathan, and Ziman 2010, page 8)

Note that the non-linguistic system of heraldry (given their assumptions of how to "read" heraldic "texts") seems to have a much lower number of singleton bigrams than would be expected given the corpus size, clearly separating it from linguistic systems.

5 Unfortunately, a complication in Lee, Jonathan, and Ziman's (2010) paper is that their formulation of bigram entropy in their Equation (2.2) is apparently wrong.

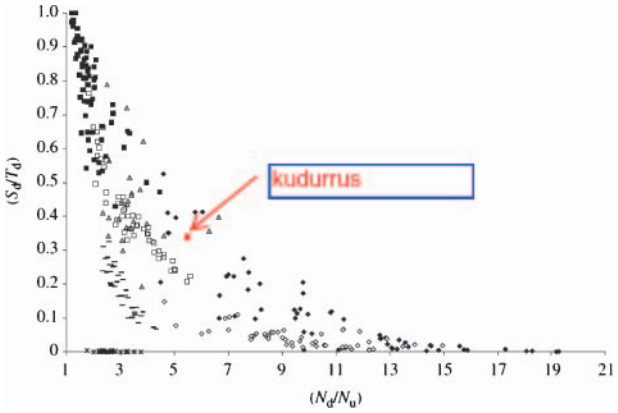


**Figure 2**
Reproduction of Figure 5.5, page 8, from Lee, Rob, Philip Jonathan, and Pauline Ziman. "Pictish symbols revealed as a written language through application of Shannon entropy." *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences*, pages 1–16, 31 March 2010. Used with permission of the Royal Society. See text for explanation.

Lee, Jonathan, and Ziman (2010) use $C_r$ and $U_r$ to train a decision tree to classify symbol systems. If $C_r \geq 4.89$, the system is linguistic. Subsequent refinements use values of $U_r$ to classify the system as segmental ($U_r < 1.09$), syllabic ($U_r < 1.37$), or else logographic.

All very impressive looking, but does it really work? In order to put the Lee, Jonathan, and Ziman (2010) theory to a serious test, I looked to another symbol system, namely, Mesopotamian deity symbols from *kudurrus* (boundary stones) catalogued in Seidl (1989). A small corpus was developed from the stones for which the depictions in Seidl's book were clear enough to read. The corpus contains only 545 tokens, with 59 types (the full set of types described by Seidl comprises 66). The Mesopotamian deity symbols are pictographic, a property shared with many scripts, including Egyptian and Luwian hieroglyphs and Mayan glyphs; and there are other script-like properties, including the fact that the symbols are often arranged linearly (Figure 3), and some symbols are "ligatured" together. Yet we know that these symbols were not part of a writing system.

Unfortunately the corpus is far too small for a meaningful comparison with the results of Rao et al. (2009a), though one point is clear from even a cursory examination

**Figure 3**
The linearly arranged symbols of the major deities of Aššurnaṣirpal II. From `http://upload.wikimedia.org/wikipedia/commons/8/87/Ashurnasirpal_II_stela_british_museam.jpg`, released under the GNU Free Documentation License, Version 1.2.

of the texts: Rao et al.'s claim that *kudurru* texts are rigidly ordered is clearly false (which we also showed in Farmer, Sproat, and Witzel [2004]); if nothing else, some symbols repeat within the same text, with different symbols following each repetition. Turning now to Lee, Jonathan, and Ziman's (2010) method, I computed $C_r$ and $U_r$ for the *kudurrus*, yielding values of $C_r = 8.0$ and $U_r = 1.55$. For the Pictish symbols, Lee, Jonathan, and Ziman computed values for $C_r$ and $U_r$ under various assumptions of what the symbol type set was, with the largest values being $C_r = 6.16$ and $U_r = 1.45$. The values for the *kudurru* texts are different than what they calculate for the Pictish stones, but crucially they are different in the direction that, given their decision tree, suggests that *kudurrus* are writing. In particular, $C_r \geq 4.89$ and $U_r \geq 1.37$, yielding classification of the system as a logographic writing system. It is worth noting also that the values for $N_d/N_u$ and $S_d/T_d$ are 5.58 and 0.35, respectively, which puts them firmly in the "linguistic" range, as shown by the superimposed point in Figure 2.

More embarrassingly, a set of 75 "texts" consisting of "symbols" derived by successive tosses of seven six-sided dice, as suggested by Liberman (2010), with individual text lengths ranging between 3 and 14, with a total of 638 "symbols," is revealed by the application of Shannon entropy to be a syllabic writing system. For this system $C_r = 12.64$ and $U_r = 1.18$.

Lee, Jonathan, and Ziman's method thus fails a crucial test: It misclassifies as writing systems whose true classification—as a non-linguistic system, as a randomly generated and meaningless sequence—is known. Again, the reasons for this failure seem clear enough. First, the tiny sample sizes of many of the texts they use make it unlikely that one can derive reliable statistics in the first place. And second, even if we allow that Lee, Jonathan, and Ziman's measures reveal something about the structures of the systems they are examining, the source of the structure could in principle be many things. Perhaps it would have been too much to expect that a reviewer would have known about the Mesopotamian deity symbols and suggested that Lee, Jonathan, and Ziman should check those with their methods. But it would have been reasonable to expect that someone should have asked them whether they can detect a truly random but *non-equiprobable* system.

In summary, what neither the Rao et al. work on the Indus symbols, nor the Lee, Jonathan, and Ziman work on Pictish symbols have shown is that one can distinguish structure that derives from linguistic constraints from structure that derives from some other kind of constraints. Furthermore, they fail for rather trivial reasons—reasons that should have been caught if competent reviewers had been assigned to these papers.

I must stress that I do not wish to argue that it is *impossible* that one could come up with a sound statistical argument to show that a particular symbol system is not linguistic. If one took a large sample of known linguistic and non-linguistic symbol systems, and showed that a particular set of measures could reliably distinguish between them with very high accuracy, then such measures could presumably be applied in the case of unknown systems such as the Indus or Pictish systems. Then, and only then would one have a clear and unequivocal demonstration of anything. But it is patently clear that the papers we have critiqued here do not even come close to this.

## 3. What Can We Do about This?

The situation described in this article surely presents a problem for the field of computational linguistics. Although entropy and related concepts clearly predate computational linguistics, they are central to statistical language processing and are used widely in the field. Such measures certainly can tell us some things about a corpus of symbols,

but there is no evidence that they can tell us what Rao et al. (2009a) or Lee, Jonathan, and Ziman (2010) think they can tell us. Yet, with the publication of these papers, and their promotion by the all-too-eager popular science press, non-specialists might easily believe that "artificial intelligence" methods can provide crucial evidence for a symbol system's status as writing. One can only expect that more such papers will appear.

Such work represents a misuse of the methods of the field of computational linguistics, so in principle it should be of interest to practitioners in that field to try to do something about this. At the very least, it would be useful if one could convince general "peer" reviewed publications such as *Science* or the *Proceedings of the Royal Society* to include qualified computational linguists among the peer reviewers of any such publications in the future. This was essentially Pereira's plea (Pereira 2009). Such a situation would hardly be tolerated in other fields, yet in publications like *Science* it seems to be common when it comes to issues having to do with language.

Part of the problem may be that computational linguistics has relatively low visibility. It is not clear that the editors of publications like *Science* even know that there are people who spend their lives doing statistical and computational analyses of text; or, if they do, that computational linguists have knowledge that is relevant to judging papers like the ones under discussion here. The time is ripe for changing that. As the results of computational linguistic research, in the form of things like machine translation or automatic speech recognition systems, become more widely known and used, computational linguists have an opportunity to educate the wider community—and we should take every opportunity to do so. For example the fact that *n*-gram language models are used with a high degree of success in speech recognition systems depends upon the fact that such language models are typically built from data consisting of millions or even billions of tokens. Such points need to be stressed more fully in dealings with the press or the science magazines, so that people do not get the impression that one can derive reliable results by such techniques from corpora consisting of only a few hundred or few thousand symbols. Despite a famous XKCD cartoon[6] that characterizes computational linguistics as a field that is "so ill-defined" that people can "subscribe to any of dozens of contradictory models and still be taken seriously," there are core methods that are backed up by solid empirical data. Yet, as with any science, there are good ways and bad ways to apply such methods.

Ultimately we may be fighting a losing battle. It is more exciting to learn that a statistical method can tell you that such-and-such an ancient symbol system was writing, than to learn that in fact the proposed methods do not work. But at least one has a duty to try to set the record straight.

**Acknowledgments**

I thank Steve Farmer, Brian Roark, Robert Dale, and a reviewer for *Computational Linguistics* for useful comments on earlier versions of this article.

**References**

Farmer, Steve, Richard Sproat, and Michael Witzel. 2004. The collapse of the Indus-script thesis: The myth of a literate Harappan civilization. *Electronic Journal of Vedic Studies*, 11(2):19–57.

Farmer, Steve, Richard Sproat, and Michael Witzel. 2009. A refutation of the claimed refutation of the nonlinguistic nature of Indus symbols: Invented data sets in the statistical paper of Rao et al. (Science, 2009). `www.safarmer.com/Refutation3.pdf`.

Gimbutas, M. 1989. *The Language of the Goddess: Unearthing the Hidden Symbols of*

6 `http://xkcd.com/114/`.

*Western Civilization*. Thames and Hudson, London.

Haarmann, Harald. 1996. *Early Civilization and Literacy in Europe: An Inquiry into Cultural Continuity in the Ancient World*. Mouton de Gruyter, Berlin.

Keim, Brandon. 2009. Artificial intelligence cracks 4,000-year-old mystery. *Wired*, 23 April. www.wired.com/wiredscience/2009/04/indusscript/

Lee, Rob, Philip Jonathan, and Pauline Ziman. 2010. Pictish symbols revealed as a written language through application of Shannon entropy. *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences*, pages 1–16, 31 March 2010.

Liberman, Mark. 2009. Conditional entropy and the Indus script. Language Log, 26 April. http://languagelog.ldc.upenn.edu/nll/?p=1374.

Liberman, Mark. 2010. Pictish writing? Language Log, 2 April. http://languagelog.ldc.upenn.edu/nll/?p=2227.

Mahadevan, Iravatham. 1977. *The Indus Script: Texts, Concordance and Tables*. Archaeological Survey of India, Calcutta and Delhi.

Parpola, Asko. 1994. *Deciphering the Indus Script*. Cambridge University Press, New York.

Pereira, Fernando. 2009. Falling for the magic formula, April 26. http://earningmyturns.blogspot.com/2009/04/falling-for-magic-formula.html.

Rao, Rajesh. 2010. Probabilistic analysis of an ancient undeciphered script. *Computer*, April:76–80.

Rao, Rajesh, Nisha Yadav, Mayank Vahia, Hrishikesh Joglekar, R. Adhikari, and Iravatham Mahadevan. 2009a. Entropic Evidence for Linguistic Structure in the Indus Script. *Science*, 324(5931):1165.

Rao, Rajesh, Nisha Yadav, Mayank Vahia, Hrishikesh Joglekar, R. Adhikari, and Iravatham Mahadevan. 2009b. A Markov model of the Indus script. *Proceedings of the National Academy of Sciences*, 106(33):13685–13690.

Seidl, Ursula. 1989. *Die babylonischen Kudurru-Reliefs. Symbole mesopotamischer Gottheiten*. Universitätsverlag Freiburg, Freiburg.

Sproat, Richard. 2009. Symbols, meaning and statistics. Invited talk at EMNLP, Singapore. http://www.fask.uni-mainz.de/lk/videoarchive/videos/2009-08-06-emnlp-2009-richard-sproat.html.

Winn, Shan M. M. 1981. *Pre-writing in Southeastern Europe: The Sign System of the Vinča Culture, ca. 4000 B.C*. Western Publishers, Calgary.