8 Toward a quantitative semiotics?*

Bogi Takács
Mikata Kft.

Abstract

This paper will demonstrate that the lack of quantitative, data-based research about nonlinguistic symbol sets that still have internal structure is already adversely affecting linguistics itself. We will survey recent attempts to distinguish between written languages and nonlinguistic inscriptions using information-theoretic, physical and usage-based metrics. We will also provide a list of nonlinguistic corpuses used in current research alongside their shortcomings. Finally, we will propose a possible new framework and a continuum-based concept of language-ness.

1 Introduction and Scope

The aim of this brief theoretical paper is to point out a curious lack, namely that semiotic problem sets are usually not handled using the tools of linguistics. Nonlinguistic symbol sequences that still have some internal structure (henceforth: NSSs) are most often disregarded, even though many such symbol sets exist in all kinds of cultures, dating back to ancient ages. From pottery markings to traffic signs, from Sumerian deity symbols to medieval heraldry, the examples are almost endless. These symbol sets can have all sorts of uses, from the mundane to the religious and even the magical. Some have little internal structure, while others have a relatively complex syntax. They permeate our lives - many readers of this article probably have at least one item on their bodies which features a tag printed with textile care symbols! Yet there is a lack of discussion about the features of these systems in linguistics.

This lack is probably due to two main reasons. First, linguists normally do not concern themselves with NSSs, possibly assuming these symbol sets have no

*Thanks to all participants of the conference for the interesting discussion and pertinent questions, which no doubt led to the improvement of this paper, and which have also inspired a followup related to semiotic universals, currently in preparation. Further thanks to Szymon Pawlas, a similarly inspiring source of visual symbols of all kinds. The paper also benefited from the comments of an anonymous reviewer.

interesting features from a linguistic standpoint. Second, semioticists tend to have an approach that is more closer in spirit to the humanities than to the sciences. Therefore, the usual tools employed in linguistic problems are seldom applied to issues related to NSSs. This especially holds true for quantitative linguistic methods.

We do not propose a sort of theoretical imperialism where the scope of quantitative linguistic methods is widened until it becomes senselessly broad. We also do not consider more hermeneutic, qualitative, "soft" approaches inferior. Not everything should — and could! — be handled using quantitative methods. However, the fact that semiotic content is seldom investigated with quantitative linguistic methods is already adversely impacting conventional linguistics itself, as we will seek to demonstrate using a recent example: an approach using computational methods to find linguistic structure in undecoded ancient inscriptions, especially the Indus Valley symbols and the Pictish symbols.

2 The Linguistic Nature of Symbol Sets

Recently, researchers have stumbled across problems which are generally considered to be beyond the purview of theoretical linguistics. How is it possible to tell if a given symbol sequence is linguistic in nature? Can an algorithm or heuristic be provided to estimate "language-ness"?

The debate has been ignited by Rao et al. (2009a) who titled their paper published in Science "Entropic Evidence for Linguistic Structure in the Indus Script". The title itself is significant, because the paper contains no claims of conclusive evidence, only possibility ("our results increase the probability that the script represents language" — Rao et al., 2009a), and even the authors have pointed this out in the subsequent discussion ("we do not claim to have "proved" any statement regarding the Indus script—our work presents evidence that is supportive of the linguistic hypothesis (in an inductive framework), but does not prove it" — Rao, 2010a). Despite these facts, the paper is still often quoted as the source of "evidence" in the conclusive sense of the term, especially in the popular press (for example, see Mahadevan, 2009).

The goal of Rao et al. (2009a) was to demonstrate that the Indus Valley signs left beyond by the ancient Harappan civilization were of a linguistic nature, as opposed to the theory that the signs were nonlinguistic (Farmer et al., 2004) and possibly used in agricultural rituals (Farmer, 2004). Indus Valley signs tend to be very brief: the average length is just 4.6 symbols, with less than 1% of strings containing over 10 symbols. Furthermore, there are many unique symbols, with estimates ranging from 27% to 50% depending on the method of classification used - some unique symbols seem to arise from combinations of simpler symbols. The symbol inventory

is quite large, with estimates approximately in the 400–600 range (Farmer et al., 2004). These features are highly unusual and have been used to argue both for and against their linguisticity.

Rao et al. (2009a) claimed that the conditional entropy of the Indus Valley symbol sequences was similar to that of natural languages, but not to that of other symbol sequences like Fortran code or DNA base pairs. They also used two artificial datasets for comparison, a practice that drew criticism (Farmer et al., 2009). Conditional entropy was chosen because it quantifies "the amount of flexibility in the choice of a token given a fixed preceding token" (Rao et al., 2009a); it was hypothesized that languages behaved differently from NSSs in this respect. Relative conditional entropy — "conditional entropy relative to a uniformly random sequence with the same number of tokens" (Rao et al., 2009a) — was also used, for the same ends. While these concepts have been used in information theory for over half a century, as the authors themselves mention, this was the first time they were used to tackle the Indus Valley issue.

Critics (Shalizi, 2009; Liberman, 2009; Sproat, 2010) immediately generated symbol sequences which had conditional entropies similar to natural languages, but which were provably not linguistic in nature. Moreover, they provided scripts in multiple programming languages to allow immediate replication. The debate soon got very heated, with Sproat (2010) even proposing that the original authors submitted to a general science journal to avoid the paper being peer reviewed by computational linguists!

A similar approach has been used by Lee et al. (2010) to demonstrate that Pictish stone carvings were in fact a written language. (This paper had an even less ambiguous title: "Pictish symbols revealed as a written language through application of Shannon entropy".) They used a two-parameter decision tree where one parameter was entropy-based and the other was related to bigram repetition; they claimed that with this tree it was possible to differentiate between nonlinguistic data, letter-based, syllable-based and word-based writing systems. This paper seemed to attract less attention, though it was discussed by Sproat (2010) and Fournet (2011).

3 Examining the Controls

From our standpoint, the most important problem with these papers is that of nonrepresentative control nonlinguistic corpuses. These controls are supposed to be representative of NSSs in order for us to be able to say that the proposed methods can differentiate between written languages and NSSs. Unfortunately, some of the controls are completely artificial, some are transcribed in a manner that might make them more language-like, and some encode a natural language. The Appendix

contains a list of relevant quotations about the corpus generation methods used in each paper and the corresponding supplementary material; here we will briefly discuss said methods.

Farmer et al. (2004) only made a brief comparison of Indus signs, various natural language scripts, and Scottish heraldic blazons using cumulative frequencies. The choice of blazons was criticized by Vidale (2007) in his thorough critique of the paper, for being culturally inappropriate to serve as a comparison. We would add that depending on the corpus used, the rigidness of blazon texts can vary to the point of them being essentially natural language. While this particular corpus (see the Appendix) seems to be quite rigid, it is not described exactly how the descriptions were separated into units.

Rao et al. (2009a) used two types of artificial datasets in some of the comparisons ("Type 1 and 2"), and DNA base pair sequences, amino acid sequences, and Fortran source code in addition to Type 1 and 2 in other comparisons. While the biological sequences undoubtedly contain information, they are not the output of human cognitive activity and are thus not very useful if we are trying to differentiate between datasets that are undoubtedly human-produced and used to communicate meaning. Fortran as a programming language is probably better-suited to serve as a control.

Another issue is whether the artificially constructed symbol sequences are similar to actual real-life NSSs. In Rao et al. (2009a), whenever real symbol sets were used, there was much less of a difference in relative conditional entropy. In some cases, the difference is less than that between different natural languages (see their Figure 1B). Either we exclude Sanskrit — one of the examples used — from the set of natural languages, or we include Fortran!

Rao et al. (2010) used the same datasets, with the addition of a relatively small music sample from Schmitt and Herzel (1997) — a study estimating block entropies of DNA sequences where written language, Fortran source code and a Beethoven sonata from Ebeling and Nicolis (1992) was used as a comparison. Schmitt and Herzel (1997) only used control samples to demonstrate the high entropies of the DNA sequences (their main focus), not to distinguish between languages and NSSs.

Lee et al. (2010) used "heraldic sematograms, code characters and repetitive lexicographic characters". The issue here is that the "code characters" encode English text, and English is a natural language. The fact that natural languages can be encoded in a manner that significantly changes their relevant parameters is strong evidence that these parameters are not useful in differentiating between full-fledged languages and NSSs. These parameters could still possibly be used in differentating languages encoded in a variety of ways, for example to differentiate between alphabets and syllabaries (as in the original article). Still, one needs to note that the two different systems used in transcribing the Pictish stones give different

results in the authors' system: with one set, characters are classified as syllables, with the other, as words. (The authors argue that one of the transcription systems is wrong.) Sproat (2010) also mentions that his corpus of die tosses — quite obviously not language, and for our purposes, not even a NSS — is classified by the Lee method as a writing system composed from letters.

Fournet (2011) pointed out that Pictish symbols were forcibly linearized by Lee et al. (2010), which can add artifacts to the analysis. The heraldic controls were also similarly linearized; Sproat (2010) noted that the method used in the latter did not follow established conventions. Lee et al. (2011) responded by claiming that directionality is implied in the Pictish corpus. (In the Indus corpus, the general convention is to assume right-to-left directionality; this was challenged by Farmer et al., 2004, without much apparent impact.)

4 Defining Relevant Concepts

A general characteristic of these papers seem to be that broader meanings of "language", "writing" and "syntax" seem to be used than the common usage of these terms in linguistic discourse. Lee et al. (2010) defined writing in a way that also included "semasiographic characters" = elements of NSSs. This was reiterated in Lee et al. (2011): "Writing communicates information via markings". Language apparently also included animal communication (mentioned as "animal language" in Lee et al., 2010).

There are multiple papers aiming to find syntactic structure in the Indus Valley symbols (Rao et al., 2009a, Yadav et al., 2010, Sinha et al., 2011). Again, syntax seems not to refer to linguistic syntax, but rather internal structure: "Text beginner and ender distributions are unequal, providing internal evidence for syntax." (Yadav et al., 2010) "The script exhibits distinct language-like syntactic structure including equivalence classes of symbols with respect to positional preference, classes of symbols that function as beginners and enders, symbol clusters that prefer particular positions within texts, etc." (Rao et al., 2010) But symbols have preferred positions in many NSSs as well, even in non-linearized ones like vexillology.

The ambiguity stems from "syntax" having a broader and a narrower meaning: in the broader sense, it is used to refer to "a set of rules which govern how the signs are consecutively strung together to form a sequence" (Sinha et al., 2011), as in programming language syntax etc., and in the narrower sense, it is used to refer to linguistic syntax. In linguistics papers, the latter is expected, so the former would ideally be followed by clarifications that syntax as used is not full-scale linguistic syntax. Of the three cited papers, only Sinha et al. (2011) mentioned the broader and narrower meanings of syntax, evidenced by putting "grammar" in scare quotes and

providing their own definition for syntax. (Yadav et al., 2010 contains an explicit discussion of this issue in an arXiv preprint, which surprisingly did not make it into the final published article: "This indicates that the script can certainly be considered as a formal language, but it remains to be seen if these features imply an underlying natural language.")

Strangely, Sinha et al. (2011) also stated that the hypothesis that "the signs are ritual or religious symbols" [...] "implies the absence of any syntactic structure". Many magical or divinatory systems — ostensibly covered under "ritual" — have features which would certainly fall under this broader interpretation of syntax. While we are cautious to bring examples before their structure has been investigated in a similar manner, alchemical symbols would probably qualify, just as various Eastern and Western astrological diagrams.

5 Proposed Metrics of Linguistic Structure

It is probably unlikely that we will have a single metric to differentiate between languages and NSSs, as Rao et al. (2009a) intended and their critics were quick to point out. Even Rao's research group moved away from that approach in their more recent work (Rao et al., 2010; Yadav et al., 2010) and they now aim to use a variety of metrics to provide convergent validity to the notion that Indus Valley inscriptions encode language: "it would be highly unusual for a nonlinguistic system to exhibit a confluence of all of these properties." (Rao et al., 2010) Alas, we do not actually know if this statement is true, since no one has taken the trouble of examining a large amount of nonlinguistic systems along these dimensions.

Lee et al. (2011) used two parameters and a decision tree to quantify the way these parameters are scored to arrive at the final categorization; Yadav et al. (2010) and Rao et al. (2010) used a larger amount of parameters, but they did not provide a way of summarizing them beyond listing them. This is understandable since they do not seek to come up with a general method of categorization, their goal is only to prove that Indus Valley inscriptions are linguistic. But it is probably impossible to reach the latter goal without working on the former problem.

There is a further interesting avenue: there have been attempts to provide validity to the notion that a script encodes natural language using computational modeling. If the model manages to discover regularities in the inscriptions sufficiently to guess missing characters (which are known to researchers), we can say that it managed to capture some of the underlying structure. Rao et al. (2009b) and Yadav et al. (2010) provided the first results.

Every researcher seems to agree that there is at least a possibility that we could use a quantitative method to distinguish between written languages and NSSs. Even

Sproat (2010), the most vocal critic of the current attempts, made this clear: "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."

We have chosen to divide the proposed metrics into information-theoretic, physical and usage-based categories. The metrics are listed with the original authors' spelling and naming conventions, which leads to minor discrepancies and ambiguities; this was intended, we refer readers to the original papers for more detail. References with asterisks refer to the supplementary material.

Information and related metrics:

- Bigram probability (Yadav et al., 2010)
- Block entropies (Ebeling and Nicolis, 1992; Schmitt and Herzel, 1997; Rao et al., 2009a*, 2010; Rao, 2010b)
- Conditional entropy (Rao et al., 2009a; Rao, 2010b)
- Conditional probabilities of text beginners and text enders / Syntactic structure (Yadav et al., 2010; Rao et al., 2010; Rao, 2010b)
- Connectivity analysis (Sinha et al., 2011)
- Cumulative frequency distribution of signs (Farmer et al., 2004; Yadav et al., 2010)
- Degree and strength distribution analysis (Sinha et al., 2011)
- Di-gram entropy [with adjustments] (Lee et al., 2010)
- Di-gram repetition factor (Lee et al., 2010)
- Directed network construction and comparison with random sequences (Sinha et al., 2011)
- Entropy (Rao et al., 2009a*; Yadav et al., 2010)
- Log-likelihood significance test (Yadav et al., 2010)
- Mutual information (Yadav et al., 2010)
- Network of significant links (Sinha et al., 2011)
- Percentage of unique and rare signs (Farmer et al., 2004)

- Perplexity (Rao et al., 2009a* — only to justify the model used –; Yadav et al., 2010)
- Segmentation tree construction (Sinha et al., 2011)
- Sign-repetition rates (Farmer et al., 2004)
- Zipf-Mandelbrot law (Farmer et al., 2004 — only to reject –; Rao et al., 2009a*, 2010; Rao, 2010b; Yadav et al., 2010)

Physical characteristics of writing:

- Linearity (Rao et al., 2010; Rao, 2010b)
- Directionality (Rao et al., 2010; Lee et al., 2011; Vidale, 2007 also mentions this in passing)
- Use of diacritical marks or ligatures (Rao et al., 2010; Rao, 2010b; Vidale, 2007)

Usage-based characteristics:

- Diverse usage (Rao et al., 2010)
- Use in foreign lands (Rao et al., 2009b, 2010 — also tested foreign inscriptions with a log likelihood test; Rao, 2010b)

Many of these metrics are not satisfied for every writing system used to encode natural language, and there are also NSSs which satisfy many of these criteria. The question is exactly how many.

We would also like to propose a set of metrics of our own that we are currently working on; results will be presented elsewhere. A visual-complexity metric for individual signs could probably not only distinguish between writing and NSSs, but also among various sorts of writing systems, and among different groups of NSSs. For instance, while this remains to be substantiated, religious and magical systems could have different visual complexity. In the literature, there are probably as many ways of quantifying visual complexity as of quantifying linguistic structure, ranging from user reports of perceived complexity (Harper et al., 2009) to compression-based methods (Székely et al., 2000), so current approaches will need to be surveyed first. Many of these studies and their results will not be directly applicable to language and similar NSSs, as they often have to do with the visual complexity of natural scenes (Oliva et al., 2004).

To allow us to make statements based on empirical evidence, we are also planning on creating both vectorized and raster corpuses of signs. This avenue could lead to practical application in user interface research as well; as mentioned in the discussion, another area where there seems to be a need for quantitative semiotic approaches.

6 Discussion and Further Possibilities

Unfortunately we do not know enough about NSSs. The real opposition is not between natural languages or random sequences — we can isolate these two groups quite handily, for instance using cryptographic methods (Rao et al., 2009a). Instead we probably have a continuum of language-ness at hand: at the maximum we have natural languages, at the minimum, completely nonlinguistic sequences, and inbetween we have various NSSs which still have some sort of internal structure that may have some linguistic characteristics. Right now we are interested in the middle of the continuum.

Are there various different dimensions of language-ness? How could we formulate these? We need to get a lot of descriptive work done simply to assess various NSSs and build corpora. Sproat (2010) points this out quite strongly: "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." But we also need to reach beyond this to see if we can make predictions about NSSs using the tools of linguistics, or the tools of other sciences.

The work described above has aimed to produce a method which sorts symbolic systems into categories. But do we even need sharply-defined categories? In most cases, they have served us quite well, but in borderline cases — which are the most interesting for us here — they have led to extremely heated debates. This conceptualization has also led to a general neglect of said borderline cases in linguistics. A parallel could be brought from inside linguistics, a classic syntax debate: the syntax of idioms and regular constructions was neglected for a long time and considered peripheral to core syntax, and when people devoted effort to investigating them (as in the classic paper of Fillmore et al., 1988), this led to several new and fruitful approaches, and the general Construction Grammar paradigm.

We also need to consider the case that what we are looking for already exists. In fact there are multiple approaches which label themselves as quantitative or — especially — computational semiotics, mostly in user interface design and intelligent systems design, but there the "computational" refers not to a way of analysis, but rather to semiotic processes involving a computer (for example see Mehler, 2003).

Similarly, there also existed the Computational Semiotics (COSIGN) conference series, running from 2000 to 2004, whose aim was "to explore the way in which meaning is understood by, or produced with, computers." (http://www.cosignconference.org/). There are probably countless papers using some form of quantitative analysis on NSSs, in a variety of fields from archeology to anthropology, but these generally do not tend to use the tools of (computational) linguistics. Because at present there is no overarching paradigm, the commonalities go unnoticed.

At the outset, corpuses which have known purposes would be more useful than corpuses whose purposes are unknown. It serves for great publicity to investigate Indus Valley signs, Pictish stones, undeciphered codices like the Voynich manuscript or the Rohonc codex, and similar unsolved historical mysteries; but before we have a reliable baseline of diverse NSS corpuses for comparison, we are only shooting in the dark. Vidale (2007) listed no less than ten corpuses which could serve as good comparisons to the Indus Valley signs in particular: "graphic non-linguistic systems of symbols from Central and South Asia of the 3rd–2nd millennium B.C." Unfortunately, no systematic quantitative comparison was provided by the author beyond a simple "number of signs" column and a brief description of each corpus; though it was noted that many of these inscriptions consisted mostly of isolated marks or repetitive designs. The recent Indus Valley studies have not attempted to use these NSSs as controls, either, even though Rao et al. (2010a) cited Vidale (2007), so we can assume this research group was familiar with the paper's contents. Some of the listed corpuses in Vidale (2007) seem to be of low quality ("the signs are not copied with the necessary detail"), or hard to access, which can cause difficulty. So far, the control corpuses used by all research groups seem to have been chosen not for their comparative value, but for their ease of access.

Protolinguistic corpuses should also be used, especially when investigating ancient systems. Protolinguistic writing has characteristics that probably differentiate it from more developed writing systems and nonlinguistic systems alike. Ancient undecoded scripts could be similar to either! (Fournet, 2011 also points out that "it is typologically probable that an archaic writing system will be defective in one way or another. " Both linguistic corpuses and NSSs as comparisons might miss the point in different ways.)

Now that Fortran has been used in studies, it would also be interesting to see a comparison between various programming languages; more low-level languages would be expected to score lower on languageness metrics than more high-level languages. The border between high-level programming/scripting languages and natural languages is increasingly blurred; some domain-specific languages like Inform 7 (Nelson, 2011) are essentially subsets of natural language.

Human linguists can make reasonably accurate guesses as to whether particular symbol sequences represent written language, hotly-debated examples like the Indus

Valley symbols notwithstanding. This is also one of the reasons linguists seem to view Lee et al. (2010) with strong skepticism, as evidenced in Fournet (2011) and the general online discussion: linguists (and archeologists!) tend to find the claim that Pictish stones represent writing highly counterintuitive. If that intuition could be quantified somewhat, everyone would benefit.

Appendix: Control samples used in the referenced studies

Farmer et al. (2004):

Scottish heraldic blazons:
"2,069 coats-of-arms, encoded as blazons, from the Mitchell Rolls, the Heraldic Society of Scotland, http://www.heraldry-scotland.co.uk/index.htm (838 distinct terms in a corpus of 18,300 total terms)."

Rao et al. (2009a) / Rao (2010b):

Type 1 nonlinguistic system (e.g., Vinča system):
"[They] involve signs that may occur in groups but the ordering of signs is not important (as it appears to have been, for example, in the Vinča system (5)). To enable comparison with the Indus texts, we assumed a Type 1 nonlinguistic system with the same number of signs as in the Indus corpus above and created a dataset of 10,000 lines of text, each containing 20 signs, based on the assumption that each sign has an equal probability of following any other."

Type 2 nonlinguistic system (e.g., Sumerian deity symbol system on kudurrus):
"[They] exhibit ordering of signs but the order is rigid. For example, in the Sumerian deity sign system found on boundary stones (kudurrus) (6), the ordering of deity signs appears to follow the established hierarchy among the various deities. As in the case of Type 1 systems above, we assumed a Type 2 nonlinguistic system with the same number of signs as in the Indus corpus above and created a corpus of 10,000 lines of text, each containing 20 signs, based on the assumption that each sign has a unique successor sign (variations of this theme where each sign could be followed by, for example, 2 or 3 other signs produced similar results)."

DNA — Sequence from human chromosome 2:
"We used the first one million nucleotides in human chromosome 2 obtained from the Human Genome Project (http://www.ncbi.nlm.nih.gov/genome/guide/human/), made available as a text file by Project Gutenberg (http://www.gutenberg.org/etext/11776).

Roughly similar values for conditional entropy were obtained when sequences from other chromosomes were used." [...] "The tokens were the 4 bases A, T, G, and C."

Protein — Sequences from Escherichia coli:
"The entire collection of amino acid sequences for the bacteria E. coli was extracted from the E. coli genome obtained from the NCBI website
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=U00096.2. This yielded a dataset containing a total of 374,986 amino acids comprising the sequences." [...] "The tokens were the 20 amino acids."

Programming language:
"We used a representative computer program in the programming language FORTRAN for solving a physics problem (fluid flow) using the finite element method. The program contained 28,594 lines of code (including comments). We removed the comments and used for our analysis the remaining code sequence containing 55,625 occurrences of tokens (examples of tokens include: if, then, else, integer, x, =, 50, etc.) " [...] "The tokens were the various programming language constructs (if, then, else, write, call, etc.), operators (=, +, -, etc.), and user-defined variables and constants (maxnx, maxny, reynld, len, 80, 17, etc.). For the analysis, we used the top 417 most frequently occurring tokens."

Rao et al. (2010):

As in Rao et al. (2009a), with the addition of "Music" from Schmitt and Herzel (1997): "Beethoven Sonata no. 32" [...] "The piece of music was encoded by Ebeling & Nicolis (1992) using a dynamic partitioning: the symbols were attributed to the change in pitch (lower or higher than the previous note or constant)."

Schmitt and Herzel (1997) also used Fortran source code, but did not specify the sampling method. Rao et al. (2010) apparently used a different Fortran corpus.

Lee et al. (2010):

Heraldic sematograms:
"A normal distribution of arms from the Heraldic Arms of British Extinct peerages (1086–1400) was used (Burke 1962). The charges (symbols) on the shield were used as characters for analysis. The colour of the charge was also used for analysis. A simplified set of characters was also generated using only the base symbols, e.g. (i) all the different lion charges such as rampant or passant are classified as a 'lion' character and (ii) all different cross charges such as bourdonny and fleuretty

are classified as 'cross' in the base-symbol categorization. Each arms was read as observed symbols from bottom to top. Text size was 400–1200 symbols."

Code characters:
"A range of English texts was transposed using morse code and a three-character code for the letters. Text size was 400–75 000 characters."

The repetitive sequences are not discussed in similar detail beyond "non-concordant letter, syllable and word character texts that are repetitive".

References

Since the Indus Valley debate is very recent, many of the authors also participated in online discussion beyond their publications in peer-reviewed venues. In addition to papers from those venues, we also use these online references to represent the authors' position and their response to criticisms more accurately. All hyperlinks were current as of March 30 2011.

Ebeling, W. and G. Nicolis (1992). Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons and Fractals, 2(6)*, 635-650.

Farmer, S. (2004). Mythological functions of Indus inscriptions. Paper presented at the Sixth Harvard Indology Roundtable.

Farmer, S., R. Sproat and M. Witzel (2004). The collapse of the Indus-Script Thesis: the myth of a literate Harappan civilization. *Electronic Journal of Vedic Studies, 11(2)*.

Farmer, S. R. Sproat and M. 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) http://www.safarmer.com/Refutation3.pdf

Fillmore, C., P. Kay, and C. O'Connor (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language, 64*, 501–38.

Fournet, A. (2011). A linguist's comment on 'Pictish symbols revealed as a written language through application of Shannon entropy'. *Proceedings of the Royal Society A, 467*, 305-308.

Harper, S., E. Michailidou and R. Stevens (2009). Toward a definition of visual complexity as an implicit measure of cognitive load. *ACM Transactions on Applied Perception, 6(2)*, 10:1-10:18.

Lee, R., P. Jonathan and P. Ziman (2010). Pictish symbols revealed as a written language through application of Shannon entropy. *Proceedings of the Royal Society A, 466*, 2545-2560.

Lee, R., P. Jonathan and P. Ziman (2011). Reply to Fournet: Pictish symbols revealed as a written language through application of Shannon entropy. *Proceedings of the Royal Society A, 467*, 309-313.

Liberman, M. (2009). Conditional entropy and the Indus script. http://languagelog.ldc.upenn.edu/nll/?p=1374

Mahadevan, I. (2009). The Indus 'non-script' is a non-issue. *The Hindu, Sunday, May 03 2009*. http://www.hindu.com/mag/2009/05/03/stories/2009050350010100.htm

Mehler, A. (2003). Methodological aspects of computational semiotics. *SEED Journal, Special Issue on Computational Intelligence and Semiotics*, 71-80.

Nelson, G. (2011). Natural language, semantic analysis, and interactive fiction. In: K. Jackson-Mead, J. R. Wheeler, and E. Short, (eds.), *Interactive fiction theory reader*. Published online: http://www.ifwiki.org/index.php/IF_Theory_Book

Oliva, A., M. L. Mack, M. Shrestha and A. Peeper (2004). Identifying the perceptual dimensions of visual complexity of scenes. Paper presented at the 26th Annual Meeting of the Cognitive Science Society Meeting, Chicago.

Pereira, F. (2009). Falling for the magic formula. http://earningmyturns.blogspot.com/2009/04/falling-for-magic-formula.html

Rao, R. P. N., N. Yadav, M. N. Vahia, - H. Joglekar, R. Adhikari, and I. Mahadevan (2009a). Entropic evidence for linguistic structure in the Indus script. *Science, 324*, 1165.

Rao, R. P. N., N. Yadav, M. N. Vahia, - H. Joglekar, R. Adhikari, and I. Mahadevan (2009b). A Markov model of the Indus script. *PNAS, 106(33)*, 13685-13690.

Rao, R. P. N. (2010a). Rebuttal of Sproat, Farmer, et al.'s supposed "refutation". http://www.cs.washington.edu/homes/rao/IndusResponse.html

Rao, R. P. N. (2010b). Probabilistic analysis of an ancient undeciphered script. *IEEE Computer, 43(4)*, 76-80.

Rao, R. P. N., N. Yadav, M. N. Vahia, - H. Joglekar, R. Adhikari, and I. Mahadevan (2010). Entropy, the Indus Script, and language: a reply to R. Sproat. *Computational Linguistics, 36(4)*, 795-805.

Schmitt, A. O. and H. Herzel (1997). Estimating the entropy of DNA sequences. *Journal of Theoretical Biology, 1888*, 369-377.

Shalizi, C. (2009). That word does not exist in any language. http://cscs.umich.edu/ crshalizi/weblog/611.html

Sinha, S., A. Izhar, R. K. Pan, and B. K. Wells, (2011). Network analysis of a corpus of undeciphered Indus civilization inscriptions indicates syntactic organization. *Computer Speech and Language, 25*, 639-654.

Sproat, R. (2010). Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics, 36(3)*, 1-12.

Székely, A. and E. Bates (2000). Objective visual complexity as a variable in studies of picture naming. *CRL Newsletter, 12(2)*, 3-33.

Vidale, M. (2007). The Collapse melts down: a reply to Farmer, Sproat and Witzel. *East and West, 57 (1-4)*, 333-366.

Yadav, N., H. Joglekar, R. P. N. Rao, M. N. Vahia, R. Adhikari and I. Mahadevan (2010). Statistical analysis of the Indus script using *n*-grams. *PLoS ONE, 5(3)* e9506.