Linguistic Society of America

NEWS RELEASE

Embargoed for release: June 2, 2013
Contact: Alyson Reed, LSA Executive Director
areed@lsadc.org; 202-835-1714

New Analysis Contradicts Findings Published in *Science*

(Washington, DC) – New research published in the June 2014 issue of *Language* presents evidence that the methods employed by the authors of articles published in prestigious international science journals are not supported by a more rigorous linguistic analysis. The *Language* article, "A statistical comparison of written language and non-linguistic symbol systems," was authored by Richard Sproat, a Research Scientist at Google, based on work he previously did at the Oregon Health & Science University. A pre-print version of the article is available for review at: http://www.linguisticsociety.org/document/language-vol-90-issue-2-june-2014-sproat.

Sproat's analysis comes in response to a number of papers published in high-profile science publications that have argued that statistical analyses of symbol combinations can provide insights into the origins of written language. One paper, by Rajesh Rao (University of Washington), Iravatham Mahadevan (Indus Research Centre) and colleagues at the TATA Institute in Mumbai, India, appeared in 2009 in the journal *Science*. It argued that a particular statistical measure — bigram conditional entropy — showed that the Indus Valley symbols behave more like those in linguistic texts than those of non-linguistic systems. In another paper in the *Proceedings of the Royal Society*, Rob Lee and colleagues (University of Exeter) claimed that a more sophisticated set of entropic measures put Pictish symbols in the same category as linguistic texts. Both papers (and other subsequent papers by Rao and his colleagues) received a large amount of attention from the news media. In these popular media accounts, the techniques were often presented as demonstrating that the symbol systems in question were written language, though this was not necessarily the intention of the authors.

Understanding statistical techniques for analyzing symbol systems and what they do and do not show is of fundamental importance to language science, as there are many old or ancient symbol systems whose function is largely or completely unknown. Examples include the Easter Island rongorongo inscriptions (19th century), the Pictish symbols of Scotland (6th century onwards), and the Indus Valley symbols (Northern India, Pakistan, 3rd millennium BCE). As part of his work on the question of whether symbol systems such as these exemplify written language, Sproat developed large, structured collections of text, or corpora, from a variety of non-linguistic systems, both ancient and modern, including Mesopotamian deity symbols (Babylonia), Totem poles (Pacific Northwest), Pennsylvania barn stars ("hex signs"), weather forecast icon sequences from www.wunderground.com, and Unicode characters for Asian emoticons. He compared these to corpora developed from fourteen languages representing a variety of different writing-system types, both ancient and modern.

From the point of view of the measures that had been proposed in the previous literature, all of the non-linguistic symbol systems in Sproat's collection or corpora behaved the same as the linguistic systems. However, he also found that a novel measure of the amount of local repetition and a version of one of Lee and colleagues' entropic measures with a different setting than they used could accurately distinguish two different categories of symbol systems. Moreover, his statistical procedure, unlike the earlier ones, classifies both the Pictish and Indus Valley symbols as non-linguistic.

Despite these promising results, Sproat cautions against relying too heavily on statistical measures to analyze ancient symbol systems that have not been deciphered. All statistical measures are heavily influenced by, among other things, the size of the corpus, the length of texts, and what kind of text is involved. Shopping lists, for example, have statistical properties that distinguish them from running prose from a novel. He argues that a truly reliable demonstration that a collection of symbols exemplifies written language requires supporting empirical evidence, such as a credible decipherment or independent archeological evidence of a related culture of active literacy. What is clear, however, is that the previously proposed statistical methods simply do not work for the intended purpose.

Sproat's work was supported in part by the National Science Foundation under grant number BCS-1049308. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation, nor of Google.

####