Corpora of Non-Linguistic Symbol Systems

Katherine Wu†, Jennifer Solman‡, Ruth Linehan†, Richard Sproat‡

† Reed College
‡ Center for Spoken Language Understanding,
Oregon Health & Science University

1 Introduction

Humans have been writing for over 5,000 years, but in addition to linguistic symbol systems there have been many non-linguistic systems. Some examples include mathematical symbology, European heraldry, barn stars, Mesopotamian deity symbols and totem poles. While writing represents natural language units such as phonemes, syllables, morphemes or in some cases words, non-linguistic systems represent other, non-linguistic, information. Thus, mathematical symbols represent mathematical operations, functions, variables and the like. Note that it does not matter that one can *read* a mathematical equation using words; the elements of the equation do not *represent* words, or any other linguistic elements.

Within the past few years, two high-profile papers have claimed to provide statistical methods to distinguish writing from non-linguistic symbol systems. The first, (Rao et al., 2009), used bigram conditional entropy to argue that the symbols used by the Indus Valley civilization constituted a writing system. The second, (Lee et al., 2010), used a different technique also based on conditional entropy to argue that Pictish symbols, found on a few hundred standing stones in Scotland, were part of a heretofore unrecognized writing system. Both of these papers were very favorably reported in the popular science press.

The problem is that the techniques reported in the cited papers do not provide evidence that a system is linguistic: for example, they are easily fooled by artificial systems that are generated by non-uniform memoryless random processes. But the deeper and more important point is that in order to test *any* statistical method that purports to distinguish writing from non-writing, one surely needs a set of corpora of clear non-linguistic symbol systems. Few such corpora exist.

The project reported in this paper fills that void by developing electronic corpora of known non-linguistic systems. To date we have developed corpora of the following systems: European heraldry; totem poles; Mesopotamian deity symbols (*kudurrus*) (Seidl, 1989); Vinča symbols (Winn, 1981); Pictish symbols; mathematical equations downloaded from `arXiv.org`; weather icon sequences from 5-day forecasts downloaded from `wunderground.com`, and Pennsylvania German barn stars (also known as "hex signs") (Graves, 1984). Corpus sizes range from several hundred to several tens of thousands of symbols. All corpora are encoded using an XML-markup scheme based in part on the Text Encoding Initiative (`tei-c.org`) conventions. The corpora will be released under an open-source license via the Linguistic Data Consortium.

2 Primary sources

In this section we list the primary sources for the data we have collected:

- **Totem poles**: Primary sources for totem poles are Barbeau (1950); Malin (1986); Stewart (1990).
- **Pictish stones**: An electronic corpus of Pictish stones already exists at the University of Strathclyde `http://www.mathstat.strath.ac.uk/outreach/pictish/database.php`. This in turn was based on a number of sources including Jackson (1984); Mack (1997); Sutherland (1997). The main work done here, over and above what was done in the Strathcyde project, was to reorder some of the symbols in the texts to more accurately reflect what appeared on the stones, and to add XML markup.

- **Vinča**: The only source for Vinča symbols is the work of Shan Winn, the main one of these being his doctoral dissertation (Winn, 1981).
- **Mesopotamian diety symbols**: Our source for the kudurru texts is Seidl (1989).
- **Barn stars (Hex signs)**: The primary scholarly work on barn stars is Graves (1984). The source for our data is W. Farrell's slide collection from the 1950's, housed at the Berks County Historical Society.
- **Heraldry:** 13,207 blazons have been collected from Burke's *General Armory* (Burke, 1884) and the Mitchell Rolls (from the Heraldry Society of Scotland `http://www.heraldry-scotland.co.uk/mitchell-rolls.html`). Blazon is the formal language used to describe heraldic arms, and as such serves as a representation of the symbols and their layout.
- **Weather icons**: All data were downloaded automatically from the `wunderground.com`.
- **Mathematical symbols:** 393,775 LaTeX equations have been downloaded from sources at `arXiv.org`.

One of the difficulties encountered in cataloging any symbol system is determining the set of distinct symbols. In all cases, we relied on the distinctions defined by the previous sources. In the case of electronic sources (such as Weather Underground) we assume whatever symbols are defined by the source. For example the set of weather icon images including `chancetstorms.gif` or `rain.gif` define the symbol types of interest. We also encode texts in the traditional "reading" order, where that is known.

3 XML Markup Schema

For most of our corpora, the main tags and attributes used are as follows:

- `collection`: A collection of entries for a single type of symbol set.
- `entry`: A subset of the corpus including the bibliographical information (`bib`), and a document.
- `bib`: Publication information of the source.
- `document`: Page number, description, the actual text or symbols, and any attributes.
- `docText`: Optional description, a collection of symbols and/or symbol units.
- `symbol`: Title, alternative title, any attribute, and optional description (often used to clarify the relationship of the symbols).
- `symbolUnit`: This is used to represent two or more different symbols that appear as one unit, such as a bear holding a fish on a single segment of a totem pole. A description tag can be used to clarify the relationship among the symbols.

Attributes include `reliability` (of `symbol`) and `type` (of `document`, e.g. "totem pole"). As an example of a marked up text, consider the following example of a totem "text" from Malin (1986); see Figure 1. In this example the tag `symbolUnit` is used to represent a group of symbols that are ligatured together.

```
<document type="totemPole" origin="Haida">
    <description>Grizzly Bear Pole of Yan, a house frontal pole</description>
    <page> p16 </page>
    <docText>
       <symbol><title>3-Skils</title></symbol>
       <symbol><title>Grizzly-Bear</title></symbol>
       <symbol><title>Bear-Mother</title></symbol>
       <symbol><title>Cub</title></symbol>
       <symbol><title>Cub</title></symbol>
       <symbolUnit>
           <symbol><title>Supernatural-Grizzly-Bear</title></symbol>
           <symbol><title>Frog</title></symbol>
           <description>Supernatural Grizzly Bear holding a Frog</description>
       </symbolUnit>
       <symbol><title>Grizzly-Bear</title></symbol>
    </docText>
</document>
```

Figure 1: Grizzly Bear Pole of Yan, a house frontal pole (Malin, 1986).

Corpus	# Texts	# Tokens	# Types	Mean text length
Totem poles	325	1,798	477	5.5
Pictish stones	283	984	104	3.5
Vinča	591	804	185	1.4
Mesopotamian deity symbols	69	939	64	13.6
Weather icons	11,588	57,940	16	5.0
Barn stars	222	746	28	3.4

Table 1: Number of texts, type and token counts, and mean text length for the corpora collected so far.

The basic structure of our XML markup was designed originally around the totem pole collection, which is one of the collections that is simpler to describe in an XML format. As the set of corpora was expanded, the XML was elaborated as well to accommodate complexities in the additional collections. For example, Pictish symbols are not arranged in a single symbol string but instead appear arranged in rows on the front and back of slabs of stone. Accordingly we introduced the `line` and `side` tags. Similarly the distribution of Vinča is often circular, warranting a `circle` tag.

The above-described XML schema covers most of the corpora we have developed so far. A more elaborate schema is under development for European heraldry, because unlike most of the other symbol systems under consideration, heraldry makes meaningful use of *two dimensions*. We started with pyBlazon (`http://web.meson.org/pyBlazon/`), a Python module for parsing blazon descriptions. The 13,207 blazons that we referenced in Section 2 were those blazons (from a much larger set) that could be parsed using pyBlazon. The output of pyBlazon is an XML representation. pyBlazon's XML representation is rather verbose and has many levels of embedding that seem largely unnecessary. We have been working on simplifying it so as to represent all the critical information with a minimum of structure.

4 Corpus Statistics

Table 1 lists the basic statistics of the corpora we have developed to date. Note that this only includes corpora where we have a reasonably finalized XML encoding.

5 Future Work

In future work we will use our corpora, along with a wide range of already available language corpora, to investigate whether there are indeed statistical methods that are useful in distinguishing writing from non-writing. In previous work (Sproat, 2010) we have argued that at least some previously proposed statistical techniques (Rao et al., 2009; Lee et al., 2010), fail to provide measures that are informative on the issue of whether a symbol system is linguistic or not. For example, Lee et. al.'s measure falsely classifies kudurru symbols as linguistic. Furthermore, as we will report in future work, this measure also classifies totem pole symbols and weather icon sequences as linguistic. It thus remains to be seen if there are any statistical tests that will prove useful in deciding if an unknown system is linguistic or not. The work here will allow for the first rigorous investigation of whether this type of test is possible.

6 Acknowledgments

This work was funded in part by National Science Foundation grant BCS-1049308, as well as REU supplement BCS-1137631. We thank the Berks County Historical Society for help in locating Farrell's slides.

References

Barbeau, M., 1950. Totem Poles. Vol. 1–2 of Anthropology Series 30, National Museum of Canada Bulletin 119. National Museum of Canada, Ottawa.

Burke, B., 1884. The general armory of England, Scotland, Ireland, and Wales; comprising a registry of armorial bearings from the earliest to the present time. Harrison & Sons, London.

Graves, T., 1984. The Pennsylvania German Hex Sign: A Study in Folk Process. Ph.D. thesis, University of Pennsylvania.

Jackson, A., 1984. The symbol stones of Scotland: a social anthropological resolution of the problem of the Picts. Orkney Press.

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

Mack, A., 1997. Field guide to the Pictish symbol stones. The Pinkfoot Press, Balgavies, Angus, updated 2006.

Malin, E., 1986. Totem Poles of the Pacific Northwest Coast. Timber Press, Portland, OR.

Rao, R., Yadav, N., Vahia, M., Joglekar, H., Adhikari, R., Mahadevan, I., 2009. Entropic Evidence for Linguistic Structure in the Indus Script. Science 324 (5931), 1165.

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

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

Stewart, H., 1990. Totem Poles. University of Washington, Seattle, WA.

Sutherland, E., 1997. The Pictish Guide. Birlinn Ltd.

Winn, S. M. M., 1981. Pre-writing in southeastern Europe: The sign system of the Vinča culture, ca. 4000 B.C. Western Publishers.