Alex Kara

AR 450

5/2/2013

Latitudes, Landscapes and Limitations:

How Two Incompatible Journals Address Universal Concepts

Archaeological inquiry is a multi-stage process. Before sinking tools into the ground, one has to not only imagine a hypothesis significant enough to attract attention, but also formulate a research design capable of generating strong conclusions. Fieldwork itself requires technical proficiency, physical endurance, and friendly communication. However, publication is the most important phase to both personal careers and archaeology as a whole. Dissemination of the most poorly conducted research provides an educational foundation for future archaeological efforts across the globe. On the other hand, hoarding pure brilliance achieves nothing. In archaeology, as well as most other disciplines, journals are the workhorse of academic discussion because they contain a diverse collection of concise articles catering to a specific interest. These interests take many forms, and each journal maintains a unique flavor by publishing articles that not only cover certain research topics but also subscribe to certain stylistic conventions. The style of individual authors and the nature of their research combines with this inter-journal variation to create the great diversity of articles available to 21st century scholars.

Studying the language of journal articles is a relatively easy way to infer how different authors, studies, and journals contribute to the content of each article. This paper compares word usage among archaeological articles from 2009 in *Latin American Antiquity* and *Arctic Anthropology* by using a correspondence analysis to reveal the relationship between articles, words, and other contextual variables. This method reveals that *Latin American Antiquity* articles focus on interactions between humans while *Arctic Anthropology* articles emphasize the relationship between humans and their environment. However, both the scientific and presentational character of individual articles explains more variation than the journals do alone. This pattern suggests that although each journal presents a certain image by filtering their chosen articles, the nature of those articles is a more significant determinant of their language.

Introduction

These journals have a number of important differences between them, the most important being geography. Latin American Antiquity (LAA) has a regional focus on the American continents below the United States. This vast area contains a multitude of natural settings including the arid plateau of Central Mexico, the lush Maya lowlands, and the cold Andean peaks. Although this region featured a diversity of peoples positioned across a spectrum of social complexity, the majority of its articles focus on the larger civilizations of Mesoamerica and Peru. While all geographical areas exhibit internal diversity, that of the Arctic is relatively limited. Each environmental context described in Arctic Anthropology (AA) is characterized by low biodiversity, poor arable land, and cold. To adapt to these external circumstances, ancient peoples adopted less complex and more mobile organizational patterns. Therefore the cultures discussed in AA do not exhibit the same degree of population density, architectural construction, or artistic splendor as is regularly abundant is LAA.

In addition to including distinct subject material, each journal differs in how they operate as contemporary institutions. LAA is published by the colossal Society for American Archaeology, cited on average 263.31 times per year, and described by Wikipedia as “the flagship profession journal of Latin American archaeology” (Harzing 2007). While ethnographic or ethnohistoric data sometimes appear in this journal, each article in LAA ultimately argues about the past: usually before the arrival of Europeans in 1492. AA includes articles about both past and present societies in the extreme North of North American, Europe, and Asia. However, the 2009 special issue “The Tops of the World” also includes articles on societies from the Southern tip of South America. This smaller journal is published by the University of Wisconsin and is cited 44.13 times per year (Harzing 2007). AA includes more ethnographic data (to support their archaeology and arguments about the past that is) than LAA does. Of course, this pattern may arise because AA includes entirely ethnographic articles. However, I think that the relative similarity of current arctic lifestyles to those from the past could tempt arctic archaeologists to incorporate ethnography more frequently.

All of the differences, both thematic and practical, complicate direct comparisons between the formats and subjects of their articles. For example, that AA features shorter articles could result from a limited budget, a simpler material record, lazy anthropologists, or a combination of these and more conditions. This type of argument also applies to why LAA rarely describes bone tool assemblages. To control for this complexity, I compare these articles across variables that one would expect to remain more consistent across both geographic regions and contemporary institutions.

Article Selection

LAA offered a generous pool of articles to select for this comparison. Half of these articles were chosen for their perceived similarities to those of AA. Hubbe et al. (2009), Williams (2009) and Williams et al. (2009) deal with marine subsistence, a very common theme in AA articles. These three articles interestingly feature less complex societies, specifically the early shell mound builders of southern, coastal Brazil and the village of San Pedro on the eponymous caye in northern Belize. I selected the three additional articles to offer more contrast, each describing perennially warm, landlocked regions that contained much more complex sedentary societies. Scarborough and Valdez (2009) and Garrison and Dunning (2009) both analyze the ancient Maya civilization, which featured a dense and complex population that numbered in the millions. Beresford-Jones et al. (2009) describes a similarly complex culture in the Ica Valley of Peru; however, this society received .3 mm of annual rainfall. These study areas span a broad array of sizes, ranging from the 10km long area in the Ica Valley studied by Beresford-Jones et al. to the 1000km strip of Brazilian coast studied by Hubbe et al. In addition to their regional focus, I selected articles that used a variety of non-excavation methods to study the past. Garrison and Dunning and Scarborough and Valdez both incorporate minimal ethnographic or ethnohistorical accounts to supplement their archaeological data. Williams describes an explicit ethnoarchaeological project. In addition to these observational methods, most authors used a specialized "hard" scientific method to understand the past, including cranial measurements (Hubbe et al. 2009), stable isotope analysis (Williams et al. 2009) and geoarchaeology (Beresford-Jones et al. 2009).

AA offered a more limited pool of articles, forcing an excursion into the 2010 issue to complete the sample. The 2009 issue, "The Tops of the World", is a special issue featuring articles on ancient habitation of Patagonia and Tierra del Fuego in southern South America. Including this issue broadened the regional scope of the articles, bringing it as close to that of LAA as possible. Nunez (2009) analyzes sedentary settlement along the northern Bothnian coast of modern Finland, the only cultural system in this sample of articles comparable in social complexity to that of LAA. The remaining authors describe much more ephemeral populations that occupied northeastern Siberia (Guzmin 2010), Newfoundland (Renouf et al. 2009), Tierra del Fuego (Mansur et al. 2009), Patagonia (Estevez 2009), and the latter two locations (Piana and

Orquera 2009). AA articles cover a more extreme range of study area sizes. Estevez meticulously excavates a 10m wide site, and Guzmin incorporates data from a 2500 km long area in Siberia. This journal also features articles that include ethnographic data; Mansur and Estevez each incorporate historical accounts and studies on the indigenous populations of the southernmost part of America as a major part of their projects. Some of these articles make use of technical methods such as radiocarbon dating (Kuzmin) or pollen analysis (Renouf et al.) as the main point of their research, while others cite external studies to understand their own excavations (Nunez).

Methodology

In order to compare these articles in a systematic method, the abundances of every distinct word were collected, filtered, and compared between all articles. Each journal article was originally downloaded as a Portable Document File (PDF). To facilitate quantitative analysis, the body of these articles were converted into basic text files (TXT). Each AA article had conveniently undergone Optical Character Recognition (OCR), which enabled them to be copied and pasted into blank a blank TXT. On the other hand, LAA articles had to be “manually” OCR'd using gImageReader, a GUI for the Tesseract OCR engine. Once converted, the files were scanned and analyzed by a script for the R statistical computing environment, accessed via the StatET Integrated Developement Environment (IDE) embedded within the Eclipse multipurpose IDE. All software is free and open source.

The total collection of words and their abundances were amalgamated into two lists representing each journal. Because LAA contained significantly more words, each word's count was converted into a relative frequency within that journal. Next, a master table was created that contained both types of frequencies using an inner join according to each row's respective word.

This step eliminated words that appeared in only one journal, including both local jargon and proper nouns which would confound analysis. A separate list of 500 of the most common English words was first stripped of potentially interesting terms and then was scanned. Any terms in this edited list or with less than 4 characters was removed from the list of journal words. After this initial filter, the ten thousand or so remaining words were further cut down to about 50 by only selecting words that appear at least once every 1000 words within their journal. From this anthropogenically manageable subset, subjective selection eliminated words that were very common or potentially contained regional bias such as landform types or specific artifact classes.

Finally, a list of 22 words was created to incorporate into the analyses. Each of these words has no regional or temporal association *per se*: they represent entities, actions or concepts necessary to understanding all societies. Each article was scanned once again, except this time the script associated the word counts with their respective article instead of the broader journal. Once again, each raw count was converted into a frequency to control for variable article lengths.

In addition to these word counts, each article was characterized according to seven contextual variables (Figure 4). First, a number from 0 to 3 represents each article's incorporation of ethnographic, ethnohistoric, or ethnoarchaeological research. Complete absence of this data earns a 0, a supplementary mention earns a one, a focused, project-level effort to analyze primary sources earns a 2, and actual observational study earns a 3. Second, a number from 0 to 2 represents the application of a more specialized, technical method than sole excavation. Absence or limited inclusion earns a 0, dependence on external sources earns a 1, and actual research with that method earns a 2. Third, the number of words in the article are counted. Fourth, the temporal scope in years of the article is approximated. Fifth, the spacial scale of the article's primary research in years is approximated by measuring the longest edge of any study area. The range between the 10m wide Tunel VII excavated by Estevez and the 2500 km long area of Siberia was scaled down by using the fourth root of the each distance, which created a new range of .18 to 7.07. Estevez and Guzmin did not provide these details for their studies, and these data had to be found in Zurro et al. (2009) and Guzmin (2000) respectively. Sixth, each article's degree of social complexity were ranked. Mobile, egalitarian, hunter-gatherers earn a 0, small scale sedentary societies that feature trade and some political system earn a 1, and very complex societies like the Maya earn a 2. Finally, each article's journal was recorded. These contextual variables provide the background necessary to understanding the raw word counts on their own.

The first analysis simply summed the abundance of each term according to their containing journal. This calculation hoped to reveal discrepancies between the two journals to facilitate their understanding. Next, the word frequencies between individual articles were compared. Finally, the articles, words, and contextual variables were incorporated into a correspondence analysis (CA) according to Borcard et al. (2011). CA is an ordination technique; it orders objects according to their similarity across multiple variables. CA is specifically suited for exploratory analysis of objects characterized by diverse, categorical frequencies. For this reason it has traditionally been popular in ecology, and these same considerations make it useful for this study. After the articles and words had been plotted, the contextual variables were correlated to that ordination to suggest how parameters other than the author(s) affect word usage.

Results

The comparison of word abundances between journals clearly shows that these journals do not exhibit identical word usage (Figure 1). AA authors seem to prefer the words “site” and “sites,” possibly because LAA authors will not shrink to calling their Maya metropolis a “site.”

Long-established ceramic chronologies in the Maya and Inca areas explain the abundance of “period” in LAA. “Social” and “archaeological” are predictably shared between the two journals. However, few distinct words appear and there is no obvious explanations for neither differences nor similarities.

Breaking down these frequencies according to the individual articles instantly reveals the shortcomings of the previous cross-journal comparison (Figure 2, Figure 3). The incredible variation between each journal prevents these authors from being generalized. The journal sums of “years” shows a clear dominance by AA, but Figure 2 shows that the LAA counts are sandwiched between two separate clusters of AA counts. This distribution does not suggest a significant difference across journals. Outliers also skewed the previous results. Hubbe et al. uses “groups” more than the other authors combined, rendering LAA's higher frequency of this word problematic. Although comparing the individual article frequencies reveals the problems with the cross-journal comparison, the variation it exhibits is too overwhelming to understand.

The CA shows a much more interesting picture (Figure 5). This analysis essentially splits the data along axes depending on which axes create the greatest variation. On any given axis, a closer proximity between two terms signals a greater correlation. The best example of this relationship is seen in the top right corner of Figure 5. “Hubbe” and “groups”, the confounding factors in the first journal comparison, are located near each other yet far from other terms. This proximity indicates not only that the abundance of “groups” is mostly due to including Hubbe et al. but also that Hubbe et al. more frequently mentions “groups” than any other word in the analysis, and both can be confirmed by examining Figure 4. The contextual variables are displayed as vectors, and are understood differently than the other labels. Once the authors and terms have been plotted, a variable's vector is oriented in the direction where it achieves the strongest correlation with the author labels, and its length is determined by the strength of that correlation. Hubbe et al. has the highest “science” score of 2 because it describes cranial measurements, and covers one of the largest spaces of 1000km. Therefore, the vectors “science” and “space” are strongly influenced by this article's location, evidenced by the relatively long length and upward direction. On the other hand, Hubbe et al. is one of the shortest articles in the collection, and this characteristic influenced the “length” vector to point away from it. The location of each author(s) influences the contextual vectors, but these variables do not effect the location of the authors in terms of this analysis.

Discussion

The CA proved to be a very useful tool for exploratory data analysis because it was able to not only reveal inter-article lexical patterns in a meaningful presentation, but also relate those patterns to the other contextual variables of each article. (Figure 4). This paper first attempts to interpret a meaning for each axis and then uses that logic to interpret the smaller details and support its claim. CA is an unsupervised ordination technique, meaning that the algorithm examines many possible axes in multidimensional variable space and eventually selects one that results in the greatest amount of variance between the data points. It falls on on the human to interpret what this axis might represent and how it might understand the other variables.

CA1, the X axis, is the generated axis that created the greatest variance between each observation. Although this specific axis was chosen using purely mathematical calculations, it represents a broad theoretical dichotomy among articles both within this sample and beyond. A higher value along CA1 represents a greater emphasis on social organization. The few articles with high values on this axis support this interpretation. Hubbe et al. writes about “postmarital resident practice”, Scarborough about a community-based economy, and Estevez about using meticulous excavation alongside ethnographical reading to infer past social organization. The terms in this area are also sparse, but the greatest three are “groups”, “organization”, and social”, which clearly support this interpretation. The abundance of “analysis” results from the low presence of data in this area and the resulting over-representation generated by Hubbe et al.'s cranial analysis and Estevez's apparent fondness of that term to describe any study of any thing. The vector representing LAA articles is directed in this direction, although with weak correlation. This ordination indicates that, in general, this sample of LAA articles tends to emphasize social dynamics over interactions with the environment.

Data with lower values along CA1 are less concerned with how humans interact with each other and more with how they adapted to their environment. The three lowest articles on this axis support this interpretation. Beresford-Jones et al. describes how deforestation and climatic change may have led to depopulation in the Ica Valley. Renouf et al. describes how two cultures originating from distinct geographies affected their new environment in Newfoundland in different ways. Kuzmin compares human occupation in Siberia to the climatic record. The low-value terms “vegetation”, “species”, “environment”, “resources”, “human”, and “region” directly support this understanding. My Mayanist mind perceived “surface” to be an exception until reading that this term can describe “land surface” and “surface flow” in addition to a plaster floor. The vector representing AA articles is directed in the opposite direction of that of LAA in a negative direction along CA1. This trajectory is explained by the general emphasis on the environment in AA articles, although the nearly parallel direction of the “time” vector raises questions.

Higher values along CA2 indicate broader scales of analysis and less attention to detail. Contrary to my own struggle to recognize this pattern, the clearest evidence of this relationship can be seen in the “space” contextual vector, which is directed upwards with strong correlation. Kuzmin, Nunez, and and Hubbe et al. are the highest values on this axis and also study the largest area. On the other hand, Renouf et al. studies the second smallest site: the 3km long Port Au Choix. I interpret this discrepancy to be the result of that article's focus on how different cultures alter regional vegetation, which is described using similar terminology as the other authors with high CA2 values. However, instead of describing “remains” or a “structure,” finer scale behaviors that are relevant to this theme are “trampling” and “clearing”, which were filtered due to global infrequency. Words such as “groups”, “cultural”, “site(s)”, “human”, and “years” contribute to this area's vast scope. “Vegetation” is another outlier due to its overrepresentation by Renouf et al. I attribute the strong correlation between the “science” contextual variable and large spatial scales to sampling bias: I could have just as easily selected micromorphology articles and achieved the opposite result. In addition to spatial scale, this direction corresponds to slightly larger time scales, evidenced by both the “time” vector's ordination and the words “years” and “period.” While “years” may seem like a short amount of time, one has to remember that this point represents only the specific word and not the concept, and this word is most frequently similarly to how Renouf et al. describe “over 10,000 years” of sedimentation in Bass Pond.

The many terms that suggest an association between negative CA2 values and more detailed studies are spread over three categories. While all terms that are positive along CA2 describe a collection of things or abstract concept, many negative valued terms represent physical entities such as “surface”, “structure”, “species”, and “resources” that are less important to analyzing areas over 1000km long. Not surprisingly, all terms associated with actual excavation such as “excavation”, “remains”, and “archaeological” have negative values because methodological details are more difficult to apply to larger-scale issues such as Holocene migration through Siberia. Finally, terms about social interactions have low CA2 values due to similar reasons as the previous classifications. These words include “social”, “organization”, and “structure,” with the latter term representing specifically a social structure. Not surprisingly, the vector representing ethnographic, ethnohistoric, and ethnoarchaeological research points almost directly down with more correlation than that of any other contextual variable vector. This correlation exists because this type of observational research facilitates analyzing individual social events, which are not discussed in articles discussing more regional patterns. Finally, the “length” vector is also directed downward with slightly less powerful correlation. This variable is explained by by the simple observation that regional studies tend to be shorter reviews of past work while including text on methodology and results requires more ink.

The center of this 2 dimensional ordination exhibits data that are not strongly correlated with any of the previously discussed themes. Garrison and Dunning and Williams et al. both juxtapose descriptions of a society's physical environment with their social and political structure while balancing detail with scope. “Region” can be used to describe a space of any size whether its a 250km stretch of Siberia or the corner of a hut in Tunel VII, and “data” is unsurprisingly located nearby because of its universal necessity. Surprisingly, the “socComp” vector that represents the social complexity of an article's subject is barely correlated to any direction. This does not indicate that social complexity has no influence on word choice, only that social complexity does not depend on any originally plotted data.

Conclusion

There are innumerable factors that shape the specific vocabulary of any one or more authors. This infinite pool of influences further combines with the thousands of distinct words within each text and their syntactical permutations, and this configuration frustrates understanding the linguistic dynamics underlying the textual manifestation of perceived scientific progress that is the academic journal. By reducing its analysis to include twelve articles between two journals, this paper presents an ambitious glimpse at these otherwise opaque patterns. This subset was further filtered to include only the most statistically powerful and thematically meaningful terms, which simple comparisons still failed to explain. Only the mathematical calculation of the most meaningful dimensions in these data coupled with manually entered variables permitted their interpretation by a human.

The automatic ordination revealed that an article's focus on either the social or environmental interactions of a population most determined that piece's word choice. A modest correlation between this emphasis and the respective article’s publication suggests that more subtle, theoretical differences between LAA and AA compliment more obvious distinctions in physical geography, inclusion of ethnographic articles, and past social complexity. Article length, spatial scale, and incorporation of ethnography exhibited stronger correlation to the plotted data, indicating that although the two journals do share some differences, the objectives of individual articles outweigh the publication's influence on its language. The components identified and evaluated by this paper may be essential to understanding the mechanics behind academic language use, but these conclusions are hardly even a foundation for being able to model and predict dictional properties emerging from certain parameters. Future studies will need to analyze exponentially larger samples using techniques to both explore patterns in data and test emerging hypotheses. Only a sustained, coordinated effort in this direction will allow one to accurately explain why Hubbe et al. refuses to use any of the dozens of thesaurus entries under “group.”

Appendix

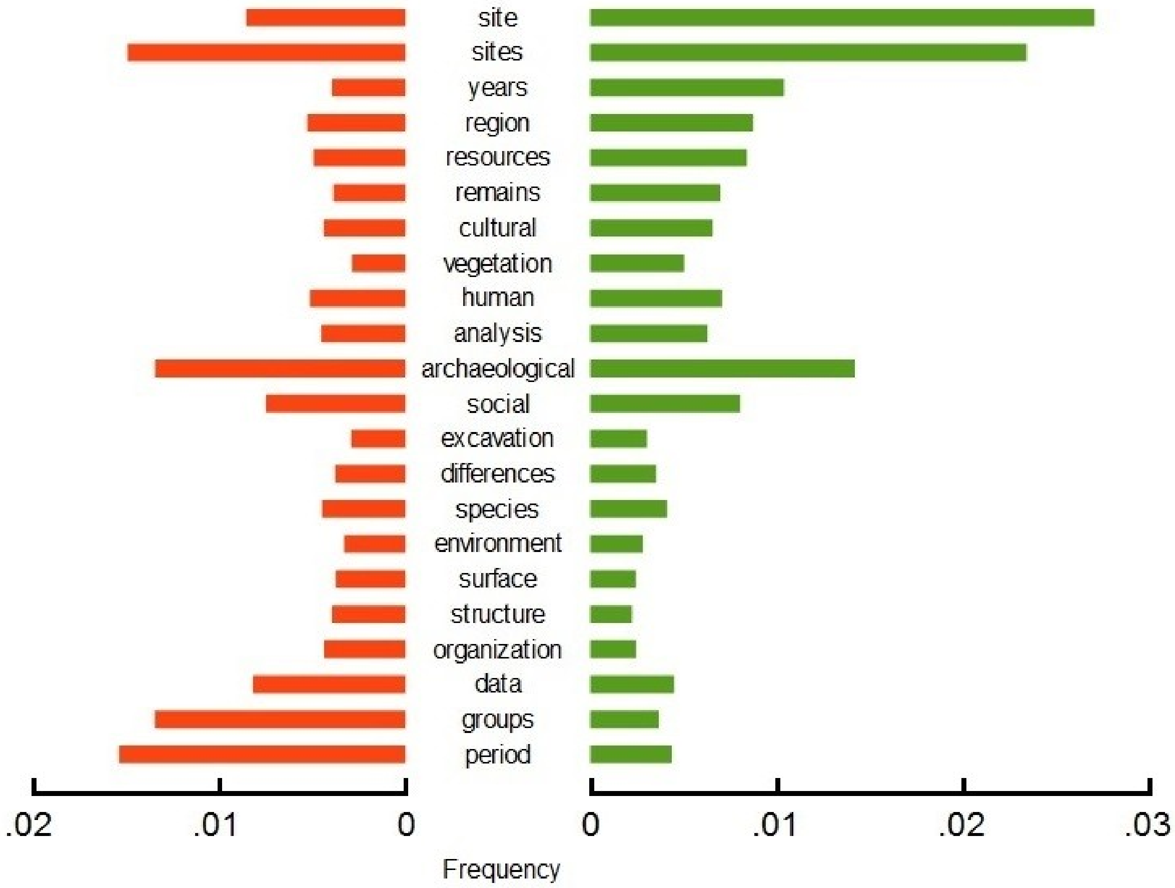


Figure 1. Summed frequency of terms between LAA (red bars) and AA. (green bars)

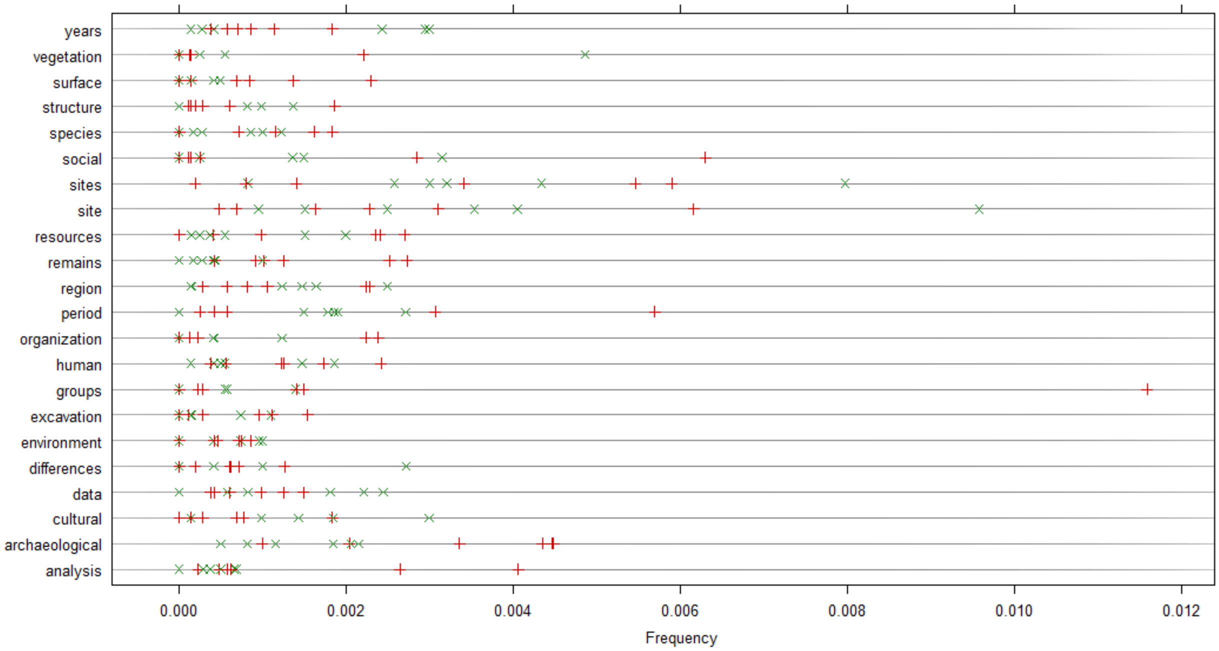


Figure 2. Frequency of terms according to their article, plotted by term. Red crosses represent LAA articles while green X's represent AA articles.

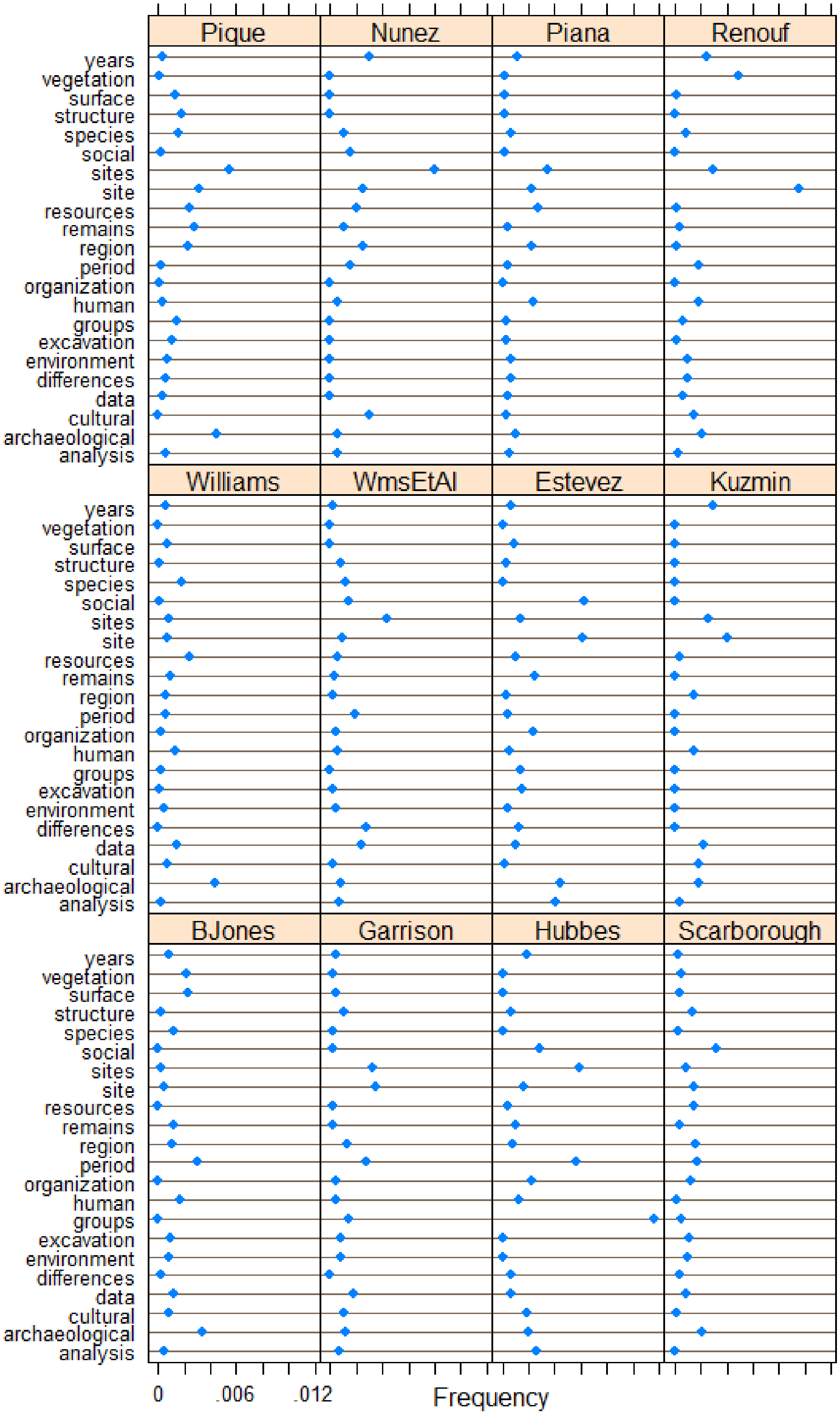
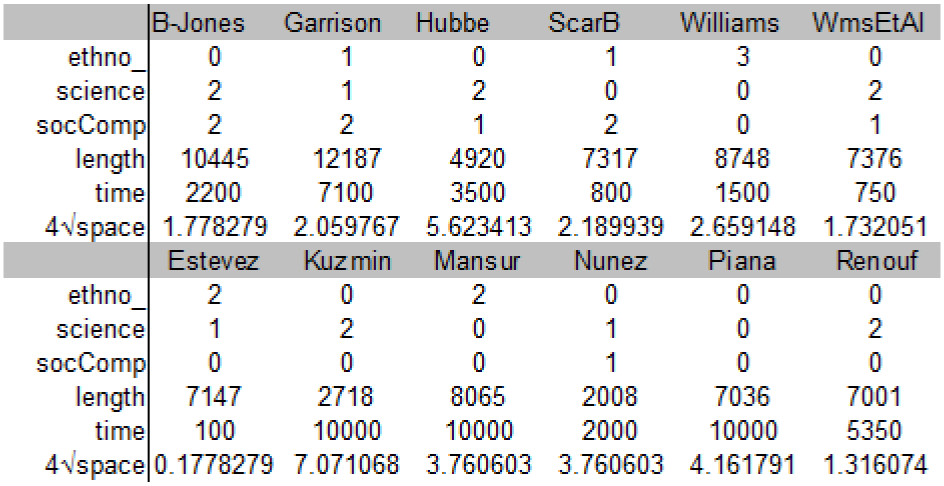


Illustration 1: Figure 3. Frequency of terms by article, plotted by term



	B-Jones	Garrison	Hubbe	ScarB	Williams	WmsEtAl
ethno_	0	1	0	1	3	0
science	2	1	2	0	0	2
socComp	2	2	1	2	0	1
length	10445	12187	4920	7317	8748	7376
time	2200	7100	3500	800	1500	750
4√space	1.778279	2.059767	5.623413	2.189939	2.659148	1.732051

	Estevez	Kuzmin	Mansur	Nunez	Piana	Renouf
ethno_	2	0	2	0	0	0
science	1	2	0	1	0	2
socComp	0	0	0	1	0	0
length	7147	2718	8065	2008	7036	7001
time	100	10000	10000	2000	10000	5350
4√space	0.1778279	7.071068	3.760603	3.760603	4.161791	1.316074

Figure 4. Contextual variables according to their article

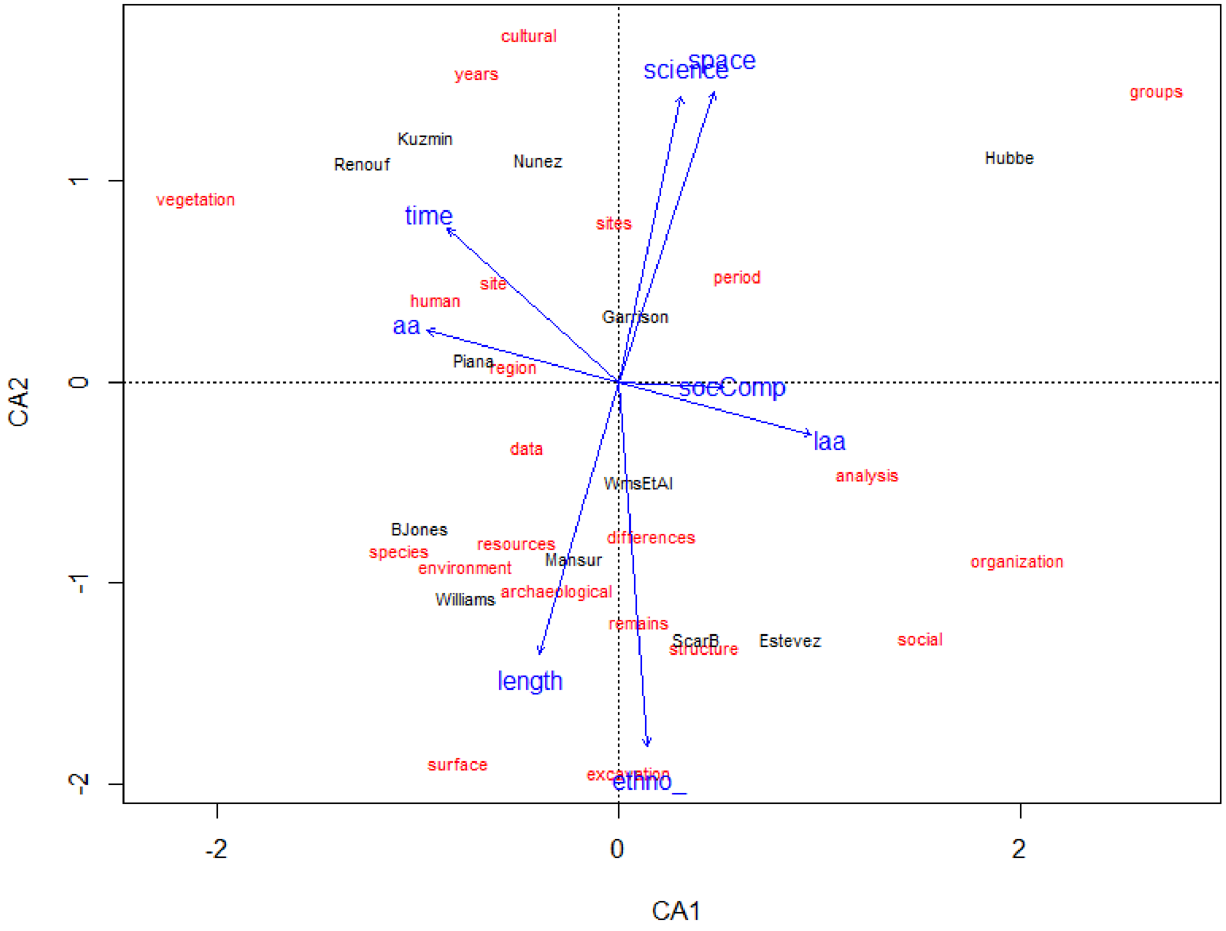


Figure 5. Correspondence analysis for terms and articles with superimposed contextual variable vectors. Black labels are articles, red labels are authors, and blue arrows are the contextual vectors.

Literature Cited

Beresford-Jones, David G., and Susana Arce T, Oliver W. Whaley, and Alex J. Chepstow-Lusty
2009 The role of Prosopis in Ecological and Landscape Change in the Samaca Basin, Lower Ica Valley, South Coast Peru from the Early Horizon to the Late Intermediate Period. *Latin American Antiquity* 20: 303–332.

Borcard, Daniel, Francois Gillet, and Pierre Legendre
2011 Numerical Ecology with R. Springer, New York City

Estévez, Jordi
2009 Ethnoarchaeology in the Uttermost Part of the Earth. *Arctic Anthropology* 46: 132-143.

Garrison, TG, and NP Dunning
2009 Settlement, Environment, and Politics in the San Bartolo-Xultun Territory, El Peten, Guatemala. *Latin American Antiquity* 20: 525–552.

Harzing, A.W. (2007) Publish or Perish, available from http://www.harzing.com/pop.htm

Hubbe, Mark, WA Neves, EC de Oliveira, and A Strauss
2009 Postmarital Residence Practice in Southern Brazilian Coastal Groups: Continuity and Change. *Latin American Antiquity* 20: 267-278.

Kuzmin, Yaroslav V
2000 Radiocarbon Chronology of the Stone Age Cultures on the Pacific Coast of Northeastern Siberia. *Arctic Anthropology* 37, no. 1: 104–115.
2010 Holocene Radiocarbon-Dated Sites in Northeastern Siberia: Issues of Temporal Frequency, Reservoir Age, and Human-Nature Interaction. *Arctic Anthropology* 47, no. 2: 104–115.

Renouf, M.A.P., Trevor Bell, and Joyce Macpherson
2009 Hunter-Gatherer Impact on Subarctic Vegetation: Amerindian and Palaeoeskimo Occupations of Port au Choix, Northwestern Newfoundland. *Arctic Anthropology* 46: 176–190.

Nunez, Milton
2009 The Sea Giveth, The Sea Taketh: The Role of Marine Resources in Northern Ostrobotnia, Finland, 4000–2000 B.C. *Arctic Anthropology* 46: 167–175.

Piana, Ernesto Luis, and Luis Abel Orquera
2009 The Southern Top of the World: The First Peopling of Patagonia and Tierra del Fuego and the Cultural Endurance of the Fuegian Sea-Nomads. *Arctic Anthropology* 46: 103–117.

Piqué, Raquel, and María Estela Mansur
2009 Between the Forest and the Sea: Hunter-Gatherer Occupations in the Subantarctic Forests in Tierra del Fuego, Argentina. *Arctic Anthropology* 46: 144–157.

Renouf, M.A.P., Trevor Bell, and Joyce Macpherson
2009 Hunter-Gatherer Impact on Subarctic Vegetation: Amerindian and Palaeoeskimo Occupations of Port au Choix, Northwestern Newfoundland. *Arctic Anthropology* 46: 176–190.

Scarborough, Vernon L, and Fred Valdez
2009 An Alternative Order: the Dualistic Economies of the Ancient Maya. *Latin American Antiquity* 20: 207–227.

Williams, Eduardo
2009 The Exploitation of Aquatic Resources at Lake Cuitzeo, Michoacan, Mexico: an Ethnoarchaeology study. *Latin American Antiquity* 20 607–627.

Williams, J.S., C.D. White, and F.J. Longstaffe
2009 Maya Marine Subsistence: Isotopic Evidence from Marco Gonzalez and San Pedro, Belize. *Latin American Antiquity* 20: 207-227.

Zurro, Debora, Marco Madella, Ivan Briz, and Assumpcio Vila
2009 Variability of the phytolith record in fisher-hunter-gatherer sites: An example from the Yamana society. Quarternary International 193: 184-191