Culture in the Mirror of Language: A Latent Semantic Analysis Approach to Culture

Eyal Sagi (ermon@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Stefan Kaufmann (kaufmann@northwestern.edu)
Brady Clark (bzack@northwestern.edu)
Department of Linguistics, Northwestern University
2016 Sheridan Road, Evanston, IL 60208 USA

Abstract

In the social sciences, culture is often explored via the use of knowledgeable informants and direct observation. In this paper we present a novel approach for cultural investigation that focuses on the statistical analysis of texts. We use Latent Semantic Analysis to generate a semantic space representing one or more cultures based on a corpus of texts. By comparing the vector representation of texts within this corpus it is possible to gain insights into cultural change. We demonstrate this method by exploring the divergence of British and American societies following the Revolutionary War. Possible uses of this method for exploratory and experimental research are discussed.

Keywords: Culture, Language, Corpus, Statistics, Latent Semantic Analysis, Cultural Change, Cultural Comparison

Introduction

The concept of culture has been increasingly criticized over the past two decades as being problematic and obsolete (Brumann, 1999). These criticisms are normally waged from a post-modernistic viewpoint, such as that presented by Abu-Lughod (1991, 1999). While those criticisms raise many valid issues, at this point in time there seems to be no alternative suggested. Culture, or the idea of the existence of cultural influences on behavior, therefore, seems irreplaceable at the moment, and still plays a vital role in the scientific investigation of human existence and behavior. In this paper we will present a novel approach to the scientific study of culture that is based on the statistical analysis of texts. Because this approach makes few assumptions about the nature of the underlying concept of culture or its permanence, it is perhaps somewhat less susceptible to criticisms such as those brought forth by Abu-Lughod.

The notion that cultures exist and that cultural affiliations provide important insights into the behavior of individuals is deeply rooted in our understanding of society. At the same time, this concept seems to defy a scientific definition and while science has been able to investigate some aspects of culture, mostly under the paradigms of anthropology, sociology, and psychology, the concept as a whole seems elusive (cf. Atran et al., 2005; Ross, 2004). One of the issues such a definition must overcome is the apparent fuzziness and amorphousness of the boundaries that separate one culture from another, if indeed such boundaries are to be found.

This is the point of contention where much of the post-modernistic criticism of the scientific use of the concept is directed. Brumann (1999), while conceding that the boundaries are not clear, argues that culture is still a useful notion that provides valid insights into human existence and the differences that are found between various communities. He goes on to describe his view of how the cultural can be explored, as an abstraction describing the distribution of features across groups of people. He further suggests that while cultural boundaries are flexible and uncertain, this distribution itself can be viewed as representing culture and that culture can therefore be investigated using statistical tools. Brumann's view of culture as described by the distribution of certain features is not unique. There are several other lines of research where culture is viewed in a similar manner. Most notable, perhaps, is the argument that cultures, like species, evolve. This argument has been presented in several ways.

Richerson and Boyd (2002) hypothesize that cultural evolution and genetic evolution are entwined and 'co-evolve'. This notion of co-evolution argues that from a genetic standpoint the species evolves to support cultural structures, and that this evolution later provides an increasingly specific fit to the cultural structures that come into existence. At the same time, culture itself evolves to fit within the cognitive and behavioral limitations imposed by the biological nature of the species and to accommodate these limitations. This argument is not without problems, as the idea of co-evolution, while appealing, lacks in empirical evidence, and seems problematic due to the vast difference in timescale between the hypothesized cultural evolution and Darwinian genetic evolution.

Henrich and Boyd (2002) support a more general view of evolution as a mechanism for cultural change, based on the idea of memes presented by Dawkins (1976). Unlike the idea of co-evolution, this notion is more generic and tries to identify a mechanism of cultural change apart from other mechanisms. Henrich and Boyd argue that change in culture can still be investigated using the mechanisms of Darwinian evolution, even though the mechanism of cultural transmission itself is based on non-discrete representations, is inaccurate, and that its products are not exact replications.

This argument bears striking similarities to Brumann (1999)'s notion of cultural boundaries as based on statistical distributions of 'cultural units', where culture is defined by examining entire populations, rather than individuals.

Overall, there seems to be little dispute over this view of culture as a fuzzy entity with uncertain boundaries. This idea is also supported by the 'naïve' and intuitive notion of culture, as can be seen by the American Heritage dictionary definition of culture: "The predominating attitudes and behavior that characterize the functioning of a group or organization" (American Heritage dictionary, 2000, definition 1d). Under this definition culture is seen as a collection of attitudes and behavior that are characteristic of the group, rather than as a requirement of it.

By these accounts, culture can be seen as *an average* of social knowledge and behavior taken over a large group of people which are perceived as being part of the specific culture. Cultural knowledge, then, is not stored at the individual level (except, perhaps, a few special individuals which are *extremely* versed in the culture), but rather is an aggregate of the cultural knowledge of the group. This conceptualization of culture poses a problem to anthropologists, whose work is based on the extraction of cultural knowledge from individuals and on their own experiences of the knowledge and behavior relevant to the investigated culture, as well as others attempting to uncover the regularities underlying it.

The psychological methodology, which usually involves averaging over a large group of subjects, seems well suited for dealing with the concept of culture as presented so far, although it has shortcomings of its own in this case. An average, by its very nature, dilutes the effects of the individual by aggregation and results in a focus on the commonalities of the sample at the expense of individual variability. While this works well where the commonalities are strong compared to individual variability, this is not always the case where culture is concerned. Culture is based on knowledge, and the variability of knowledge among individuals is large. The cultural knowledge, then, is difficult to 'extract' by averaging over only a few individuals, and the psychological process itself becomes increasingly complicated with the size of the sample.

The Cultural Consensus Model (CCM), presented by Romney et al. (1986) attempts to overcome this problem by identifying the 'culturally knowledgeable' individuals in the sample. In this way, the CCM helps to alleviate many of the problems found in the averaging and aggregation of data, especially the dilution of information due to variability, while augmenting the amount of relevant information found in the sample. The CCM achieves this through the assumption that the culturally appropriate knowledge shows little variability among individuals. From this assumption, it follows that individuals who share this cultural knowledge should show little variability, while those who lack the cultural knowledge would have greater variability in their responses. By weeding out this extraneous variability the CCM reduces the overall variability in the sample and achieves its greater focus.

This method, however, is not without its limitations. Perhaps the most apparent limitation is the assumption of the existence of a *single* 'culturally correct' answer for every question. While some cultural knowledge (perhaps most) might exhibit this type of unimodal distribution, there might be no such culture-wide accepted answer, in which case a bimodal, or multimodal distribution might occur in the data. When this occurs, the CCM might fail altogether, or point to one of these answers as the 'culturally-correct' one ignoring the others.

An obvious example of such a case would be a result of cultural change over a generation. While the elders of a tribe are usually held to hold most of the cultural knowledge, and should therefore be expected to be 'chosen' by the CCM as the most appropriate source of cultural knowledge, their knowledge is mostly based on experiences of the past. If the culture has undergone significant change in the decade prior to the research, for example, there might be a discrepancy in the cultural knowledge between these elders, who might be less flexible and less open to change, and the younger generation, who might be embracing this change. This will result in the CCM accepting the elders' notion of the culture over that of the younger generation, and thus the 'freezing' in time of the culture that is not appropriate.

A related drawback of the CCM is the assumption it makes about the relative 'difficulty' of the various pieces of cultural knowledge. Some parts of the cultural knowledge are less accessible than others – Shamanic rituals, for example, as well as the beliefs that underlie them, are usually kept as a secret of the shamans, and are rarely imparted on third parties. Applying the CCM to data collected from the tribe as a whole, then, would tend to dismiss those as part of the variability of non-cultural knowledge, unless the shamans themselves are considerably more versed in the culture than other members of the tribe.

The CCM is based on the assumption that cultural knowledge is *shared knowledge*, and that as such it should elicit similar responses. This assumption is prevalent throughout most of the literature on culture, and is one of the cornerstones on which evolutionary theories of culture, for example, rely. Brumann (1999) also argues that culture should be conceived of as something that is common between individuals, although he goes to great lengths to avoid making claims as to *what* it is that is shared. It seems, however, that no matter what is shared, part of it must be knowledge.

Up to this point, cultural knowledge was assumed to 'exist' only in the minds of the individuals who are a part of the culture (to a greater or lesser degree). However, cultural transmission, which is a requirement for culture to continue its existence across generations, does not happen by the explicit transmission of the information itself, but rather through the experiences of the individual as it passes through childhood and adulthood. These *social* and *cultural* experiences may involve the observation and mimicking of

behavior, but are usually believed to involve a high degree of linguistic interaction, as is the case with most social behaviors. In fact, there are some linguistic repositories whose *expressed purpose* is to transmit cultural values and ideas from one generation to the next, such as the Jewish 'Hagadah'. Moreover, in Dawkins (1976)'s conceptualization, memes use language as their primary mode of transmission. Therefore, it is generally accepted that most cultural knowledge is transferred, in one form or another, through the use of language[1].

Likewise, most of the investigations of culture, whether they are anthropological, sociological, or psychological, rely on language as the means of communication – usually in the form of interviews or questioning. It seems that the underlying assumption, although rarely discussed, is that culture *can* be transmitted through language. Linguistic interaction, being the most common social interaction in most societies, is likely to be the most common vehicle through which culture is normally expressed. Some make an even stronger claim – Wierzbicka (1992), for example, states that "to many, it is axiomatic that language is a mirror of culture, as well as being a part of culture" (page 373).

Language, then, can be seen as a vehicle for culture, and as such can be used to transmit cultural knowledge. Written language is also a repository of cultural knowledge, as it is a repository of general knowledge, where part of that knowledge is cultural. Moreover, it seems that some texts have greater cultural content than others – Myths, for example, tell stories that are clearly cultural in origin, and that rely to a great degree on the behavioral and moral codes of their culture of origin (Bierlein, 1994).

If texts and literature hold cultural knowledge, then there might be a method by which this knowledge can be elicited and examined, giving rise to a different approach for cultural investigation. This is not an entirely new idea, as the field of anthropological linguistics has been probing cultures through their texts for decades, but anthropological exploration of texts is methodologically similar to the anthropological field work – It is systematic in its own way, but relies heavily on qualitative analysis and is therefore subject to many of the criticisms waged at anthropology in general, and anthropological field work in particular.

The field of linguistics has developed a number of methods whose purpose is to extract 'linguistic meaning' out of corpora (see Manning and Schütze, 2002, chapter 15 for a short review of several of these), most of which are statistical in nature. One of the most successful of these methods, *Latent Semantic Analysis* (LSA*)*, involves the generation of a compressed *semantic space* out of a corpus. Within this space, texts with similar meaning will be represented by vectors whose 'distance' as measured within the space is correspondingly small while texts that are less similar will be represented by vectors that are farther apart. Texts (as well as individual words) can be judged for semantic relatedness by examining this measure of distance (usually measured by the cosine of the angle formed between the two vectors, *cos* θ). This semantic space, therefore, can be used to compare documents, and parts of documents, for similarity of content.

This measure is somewhat similar to the CCM's measure of consensus. This similarity is more than accidental, as both methods employ statistical techniques that are based on *least-square fitting* of the original data. These techniques transform the original set of vectors into a different space where the translated distances between vectors correspond to the degree of difference between them.

The main difference between the two methods is in the unit of analysis – While the CCM uses sets of responses by individuals, LSA examines texts for patterns of word co-occurrence. In this respect, LSA has a certain advantage: The CCM uses a factorial design exclusively as a means for uncovering a *single* set of specific answers that correspond to a set of questions. Therefore, the CCM explicitly assumes the existence of a single uniform culture to which the sample it is analyzing belongs. LSA makes no such a-priori assumptions. Instead, LSA takes a corpus and analyzes it, searching for statistical patterns and regularities. It is therefore possible to use LSA not only to identify the 'core culture' of a sample, but also to test whether the sample consists of a single culture or a cluster of several cultures.

Nevertheless, because it makes fewer assumptions, LSA is probably less focused and less sensitive. Therefore, it is likely that LSA would require considerably more data than the CCM. Fortunately, texts are often abundant in post-literacy cultures and written responses are likewise easy to obtain.

The Method

Latent Semantic Analysis (LSA) is a collective term for a family of related methods, all of which involve building numerical representations of words based on occurrence patterns in a training corpus. The basic underlying assumption is that co-occurrence within the same contexts can be used as a stand-in measure of semantic relatedness (see Firth, 1957; Halliday and Hasan, 1976; Hoey, 1991, for early articulations of this idea). The success of the method in technical applications such as information retrieval and its popularity as a research tool in psychology, education, linguistics and other disciplines suggest that this hypothesis holds up well for the purposes of those applications.

The relevant notion of "context" varies. The first and still widely used implementation of the idea, developed in Information Retrieval and originally known as Latent Semantic Indexing (Deerwester et al., 1990), assembles a term-document matrix in which each vocabulary item (term) is associated with an *n*-dimensional vector recording its distribution over the *n* documents in the corpus. In contrast, the version we applied in this work measures co-occurrence in a way that is more independent of the characteristics of the documents in the training corpus, building instead a term-term matrix associating vocabulary items with vectors representing their frequency of co-occurrence with each of a

[1] It should be noted that cultural transmission through language is often done implicitly – e.g., using stories and myths.

list of "content-bearing" words. This approach originated with the "WordSpace" paradigm developed by Schütze (1996). The software we used is a version of the "Infomap" package developed at Stanford University and freely available (see also Takayama et al., 1999). We describe it and the steps we took in our experiments in some detail below.

Word and Document Vectors

The information encoded in the co-occurrence matrix, and thus ultimately the similarity measure depends greatly on the genre and subject matter of the training corpus (Takayama et al., 1999; Kaufmann, 2000). In our case, we used the entire available corpus as our training corpus. The word types in the training corpus are ranked by frequency of occurrence, and the Infomap system automatically selects (i) a vocabulary W for which vector representations are to be collected, and (ii) a set C of 1,000 "content-bearing" words whose occurrence or non-occurrence is taken to be indicative of the subject matter of a given passage of text. These choices are guided by a stoplist of (mostly closed-class) lexical items that are to be excluded. The vocabulary W consisted of the 20,000 most frequent non-stoplist words. The set C of content-bearing words contained the 50th through 1,049th most frequent non-stoplist words. This method may seem rather blunt, but it has the advantage of not requiring any human intervention or antecedently given information about the domain.

The cells in the resulting matrix of 20,000 rows and 1,000 columns were filled with co-occurrence counts recording, for each pair $\langle w, c\rangle \in W \times C$, the number of times a token of c occurred in the context of a token of w in the corpus.[2] The "context" of a token w_i in our implementation is the set of tokens in a fixed-width window from the 15th item preceding w_i to the 15th item following it (less if a document boundary intervenes). The matrix was transformed by Singular Value Decomposition (SVD), whose implementation in the Infomap system relies on the SVDPACKC package (Berry, 1992; Berry et al., 1993). The output was a reduced 20,000 × 100 matrix. Thus each item $w \in W$ is associated with a 100-dimensional vector $\vec{w}$.

Once the vector space is obtained from the training corpus, vectors can be calculated for any multi-word unit of text (e.g. paragraphs, queries, or documents), regardless of whether it occurs in the original training corpus or not, as the normalized sum of the vectors associated with the words it contains. In this way Infomap calculates a *vector* representing the overall content of each document in the corpus. These vector representations of documents can provide a starting point for various types of quantitative analyses. In this paper we focus on one such analysis – an investigation of cultural change and divergence based on the similarity between documents at different points in time. This similarity is measured as the correlation between a particular document vector and a baseline vector representing texts produced at a fixed earlier period.

The Analysis

One of the advantages of working with texts, as opposed to people, is that texts are preserved over time, and can therefore be used as historical representations of their culture of origin at that period in which they were written. Project Guttenberg (http://gutenberg.org/) for example, is a repository of texts ranging from the 16th century and earlier to the 20th century. As such, these texts can be used to examine cultural change across that range in time, especially where there are many such texts, as is the case with British and American literary works.

Cultural change in British and American societies is especially intriguing as these two cultures share a common origin, namely the culture of the British Isles, but have since diverged into distinct cultures. This divergence occurred within the period covered by Project Gutenberg and makes for an interesting test case of cultural evolutionary theories. Taking the analogy from the evolution of natural species, the two cultures share a common ancestral origin – The British culture of the 17th and 18th centuries, with external influences and a parting of ways afterwards. According to evolutionary theory, once separated, the two cultures faced different constraints and challenges, and should therefore have become increasingly different.

This comparison yields three independent predictions: Firstly, cultural change should result in documents that are increasingly different from those produced during the baseline period. Secondly, because the United States broke off from the British hegemony, American texts should be less similar to baseline documents from earlier periods in British history than their contemporary texts which originate from the British Isles. Finally, the divergence of the two cultures should result in an increase in the semantic distance between documents produced by the two cultures.[3]

To test these predictions, we examined the correlations of a baseline vector, representing 226 British texts from the 17th and 18th century, with document vectors from British and American texts for periods of 25 years starting from 1775 and ending in 1900. Based on our three predictions,

[2] Two details are glossed over here: First, the Infomap system weighs this raw count with a $tf.idf$ measure of the column label c, calculated as follows:

$$tf.idf(c) = tf(c) \times \Big(log(D+1) - log\big(df(c)\big)\Big)$$

where tf and df are the number of occurrences of c and the number of documents in which c occurs, respectively, and D is the total number of documents. Second, the number in each cell is replaced with its square root, in order to approximate a normal distribution of counts and attenuate the potentially distorting influence of high base frequencies.

[3] It should be noted that while other factors, such as language contact and technological innovation, could affect language change, it could be argued that these factors are themselves tied to cultural change and that the resulting change in language proceeds hand in hand with a related cultural change. For instance, language contact is usually tied to cultural contact in the sense that both are the result of the interaction of people of different cultures.

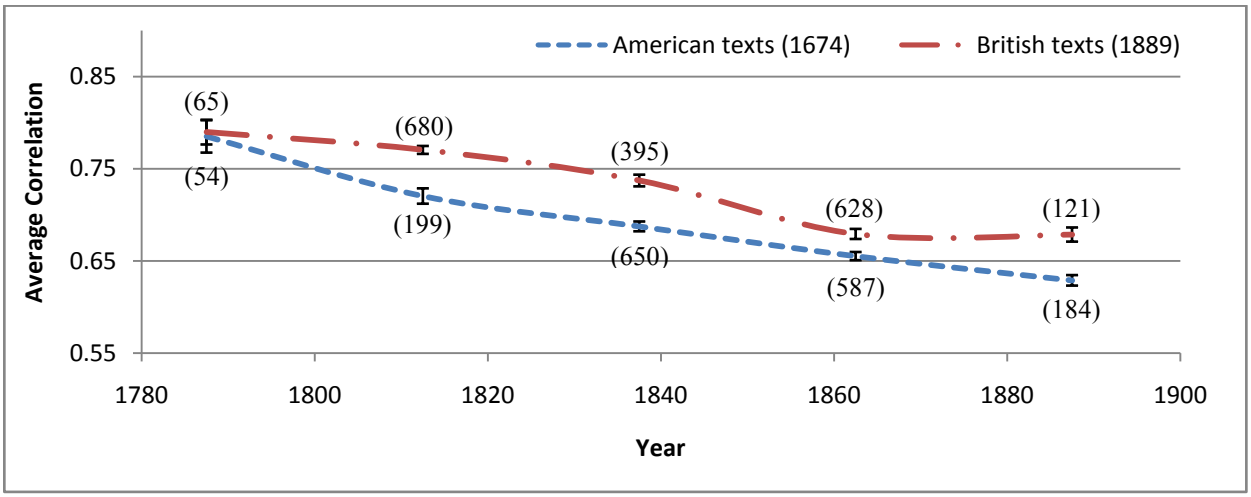

Figure 1: Average correlations of British and American texts with the average document vector for British texts from 1600 to 1775 (error bars represent standard error of the mean, number of texts is given in parenthesis)

we hypothesize that: (1) Over time the correlation of documents with the baseline vector should decrease; (2) The document vectors for texts originating in the United States should have a lower correlation with the baseline vector than texts originating in the British Isles; (3) There will be an interaction such that the decrease in correlation observed in (1) is accelerated in the case of texts originating in the US compared to those originating in the British Isles.

The Corpus

In order to perform this comparison, we generated a semantic space from a corpus containing 4034 works from Project Gutenberg dating as far back as the 14th century. We determined the cultural association of each document based on the nationality of its author. Furthermore, we divided the range of time between 1775 and 1900 into 5 25-year periods. However, because precise authoring dates are unavailable for many of the works involved, we decided to use the date of birth of the author[4] when deciding the period to which a work belongs. As our cultural baseline we used the average document vector based for British texts whose authors were born between 1600 and 1775 (226 texts).

Results and Discussion

We calculated the average document vector for each culture and period and correlated these vectors with the average document vector of our baseline period. The results of this analysis are summarized in Figure 1.

As predicted, the overall correlation between document vectors and the baseline vector decreased over time ($F(4, 3553) = 42.141$, $p < .001$). In addition, American texts generally demonstrated a lower correlation with the baseline vector than British texts ($F(1, 3553) = 28.43$, $p < .001$). This effect was statistically significant within each individual time period except for 1775-1800. Finally, there was a significant interaction between the culture and period variables ($F(4, 3553) = 9.261$, $p < .001$), indicating that the two cultures indeed grew apart between 1775 and 1900.

However, in contrast with our prediction, the majority of this cultural divergence seems to have occurred between 1775 and 1825 as the slope of cultural decline following that period is highly similar (the difference between the two correlation scores is .05 for the periods 1800-1825, 1825-1850, and 1875-1900). Interestingly, the overall linear trend observed for the rate of cultural change is broken by British texts written by authors born between 1850 and 1875. One possible interpretation of this abnormality is that historical events around that time period temporarily brought the two cultures together.

Following the results presented in this section, we hypothesize that the rate of cultural change tends to be constant across time. However, major historical events, such as the Revolutionary War, might cause an increase in the rate of change. Consequently, such events might result in cultures drifting apart. Nevertheless, this cultural drift seems to be limited to a relatively short period of time. One possible mechanism that could account for this pattern is that certain historical circumstances (e.g., wars) cause a reduction in the cultural exchange between societies. In turn, this reduction leads to a divergence between the cultures. However, once the specific event has run its course, cultural exchange often returns to its normal patterns and this divergence is halted (although not necessarily reversed).

General Discussion

In this paper we demonstrated how corpus analysis might be used to explore the rate and scope of cultural change. While the results reported here are limited in scope and interpretation, future enhancements would hopefully allow us to identify specific cultural difference as well as focus on how individual cultural concepts differ between cultures and

[4] While we report results based on the author's date of birth, the analysis shows the same patterns when the author's date of death is used instead.

change over time. Nevertheless, the results presented suggest that written corpora can provide insights regarding the culture from which the texts originate. Such texts can be used for both historical analysis and as a tool for a synchronous comparison of similar cultures.

Furthermore, similarly to the Cultural Consensus Model, the method presented in this paper provides a means of quantifying aspects of the underlying concept of culture. Such quantification can be helpful in transforming culture from an abstract notion into an empirically-defined variable. For instance, one interesting extension of this method might be to apply a clustering analysis to a set of document vectors. In cases where a culture might be comprised of several distinct subcultures, such an analysis is likely to group the documents into clusters that correspond with these subcultures. Such a correspondence can be used as an empirical test of the hypothesized cultural composition.

This method can also be applied in a more experimentally rigorous setting by recruiting informants or participants and asking them to generate specific types of texts. A statistical analysis of these texts can then reveal whether one group of participants might differ from another in its cultural knowledge, whether due to differences in background or some experimental manipulation.

Finally, while the analysis we presented here focused on vectors that represent entire texts, researchers interested in studying culture might also be interested in examining how the meaning of specific words and the concepts they represent varies by culture or changes across time. It is possible to examine such changes by focusing on the contexts in which words occur. Vectors representing such contexts have been successfully applied to semantic disambiguation (e.g., Schütze, 1998) and comparison (e.g., Sagi, Kaufmann, and Clark, 2009). A similar application might prove useful as a tool for cultural investigations.

References

Abu-Lughod L. (1991), Writing against Culture. In Fox, R. G. (ed) *Recapturing Anthropology: Working in the Present.* Santa Fe, NM: School of American Research Press.

Abu-Lughod L. (1999), Comment on Brumann (1999). *Current Anthropology*, 40, S13-S15.

Atran, S., Medin D., & Ross N. (2005) The Cultural Mind: Environmental Decision Making and Cultural Modeling Within and Across Populations. *Psychological Review, 112,* 744-776.

Berry, M. W. (1992) *SVDPACK: A Fortran-77 software library for the sparse singular value decomposition.* Tech. Rep. CS-92-159, Knoxville, TN: University of Tennessee.

Berry, M. W., Do, T., O'Brien, G. Vijay, K., & Varadh S. (1993) *SVDPACKC (Version 1.0) User's Guide*, Tech. Rep. UT-CS-93-194, Knoxville, TN: University of Tennessee.

Bierlien, J. F. (1994) *Parallel Myths*. NY: Ballentine Wellspring.

Brumann C. (1999) Writing for Culture: Why a Successful Concept Should not be Discarded. *Current Anthropology*, 40, S1-S27.

Dawkins R. (1976) *The Selfish Gene*. NY: Oxford University Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990) Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41, 391-407.

Firth, J. (1957) *Papers in Linguistics, 1934-1951*, Oxford University Press.

Halliday, M. A. K., & Hasan, R. (1976) *Cohesion in English*. London: Longman.

Henrich, J. & Boyd, R. (2002). On modeling cognition and culture: Why cultural evolution does not require replication of representations. *Journal of Culture and Cognition*, 2(2), 87-112.

Hoey, M. (1991) *Patterns of Lexis in Text*. London: Oxford University Press.

Infomap [Computer Software]. (2007). http://infomap-nlp.sourceforge.net/ Stanford, CA.

Kaufmann, S. (2000) Second-order cohesion. *Computational Intelligence.* 16, 511-524.

Manning, C.D. & Schütze H. (2002) *Foundations of Statistical Natural Language Processing*. MA: MIT Press.

Project Gutenberg (n.d.) http://gutenberg.org

Richerson, P. & Boyd, R. (2002). Culture is part of human biology: Why the superorganic concept serves the human sciences badly. *Probing Human Origins*, American Academy of Arts and Sciences, 59-86.

Romney, A. K., Weller, S., & Batchelder, W. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88, 313-338.

Ross, N. (2004) *Culture and Cognition: Implications for Theory and Method*. Thousand Oaks, Sage Publication.

Sagi, E., Kaufmann, S., and Clark, B. (2009). Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In Basili R., and Pennacchiotti M. (eds.), *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics.* Athens, Greece. 104-111.

Schütze, H. (1996) *Ambiguity in language learning: computational and cognitive models*. CA: Stanford.

Schütze, H. (1998) Automatic word sense discrimination. *Computational Linguistics* 24(1):97-124.

Takayama, Y., Flournoy, R., Kaufmann, S. & Peters, S. (1999). Information retrieval based on domain-specific word associations. In Cercone, N. and Naruedomkul K. (eds.), *Proceedings of the Pacific Association for Computational Linguistics (PACLING'99)*, Waterloo, Canada. 155-161.

The American Heritage® Dictionary of the English Language, 4th Ed. (2000). Houghton Mifflin Company.

Wierzbicka, A. (1992) *Semantics, culture, and cognition: Universal human concepts in culture-specific configurations.* Oxford: Oxford University Press.