Using the Google N-Gram corpus to measure cultural complexity

Patrick Juola
Duquesne University, USA

Abstract

Empirical studies of broad-ranging aspects of culture, such as 'cultural complexities' are often extremely difficult. Following the model of Michel *et al.* (Michel, J.-B., Shen, Y. K., Aiden, A. P. *et al.* (2011). Quantitative analysis of culture using millions of digitized books. *Science*, **331**(6014): 176–82), and using a set of techniques originally developed to measure the complexity of language, we propose a text-based analysis of a large corpus of topic-uncontrolled text to determine how cultural complexity varies over time within a single culture. Using the Google Books American 2Gram corpus, we are able to show that (as predicted from the cumulative nature of culture), US culture has been steadily increasing in complexity, even when (for economic reasons) the amount of actual discourse as measured by publication volume decreases. We discuss several implication of this novel analysis technique as well as its implications for discussion of the meaning of 'culture.'

Correspondence:
Patrick Juola,
Evaluating Variations in Language Laboratory,
Duquesne University,
600 Forbes Avenue,
Pittsburgh, PA 15282, USA.
Email:
juola@mathcs.duq.edu

1 Introduction

It is common to hear assertions that culture is complex, that language is complex, and that while the complexity of language is a universal constant, the complexity of culture is increasing as technology and the increased pace of modern life creates new complexities. These statements are usually based on subjective assessments, often tinged by nostalgia for the 'good old days'. Can questions of cultural complexity be addressed quantitatively?

In this article, we apply information-theoretic techniques previously used (Juola, 1998, 2008) to study linguisitic complexity. Based on an admittedly simple intuitive meaning of 'culture', we confirm quantitatively and empirically a set of equally simple implications of this definition.

2 Background

2.1 Linguistic complexity

The study specifically of the complexity of language has been of interest to linguists for twenty years or more. Among laypeople, the question is usually expressed as some form of 'why is language X so hard?' or 'which is the most complex language in the world?' Sometimes it appears in a question about how complex all-languages-in-the-world, taken as a whole, are, for example, in the discussion of what sort of information would be needed to develop a truly 'culture-independent' world language. Sometimes it is the lament of a student trying to master the intricacies of a language far removed from her native tongue. Sometimes, more villainously, it is the underlying subtext to a statement about the inferiority of some racial or

Literary and Linguistic Computing, Vol. 28, No. 4, 2013.

doi:10.1093/llc/fqt017 Advance Access published on 29 June 2013

cultural group. For most of these questions, there are standard and accepted answers common to the community of linguists; for example, languages and structures that differ greatly from ones own are more difficult than ones one already knows, and thus native speakers of Japanese may find learning the definite/indefinite distinction in English to be difficult. Similarly, it is common knowledge that it is easier to learn a language by immersion than by studying it in school, and so forth.

Stripped of the learnability aspects, the question of overall structural complexity still admits of a meaningful overall answer. It is difficult to improve on Tony Wright's phrasing (on the LINGUIST-LIST, 14 August 1996):

> All languages are very complex, and probably equally complex, just in different ways. The lack of case marking on nouns and the lack of overt tense and agreement morphology with verbs makes English appear simple in comparison with Latin or Greek until you realize how complex and rigid English syntax is.

While useful, this answer is as much a doctrine as a scientific statement until and unless it can be confirmed via some sort of evidence. In practical terms, it was long considered unfalsifiable—it is difficult to determine exactly *when* a construct is more or less complex than another—for example, are PRO-drop languages such as Spanish more complex (than English) because one has an additional option to drop a pronoun or not, or are they less complex because one need not specify inferrable pronouns? Second, an analysis of linguistic complexity on this basis simply means that complexity is thrown over the fence; complexity and ambiguity on one tier (such as explicit morphological case-marking) can typically be resolved by examination of information at another tier, which can lead to the assumption that the unresolvable complexity is simply a function of a tier not yet examined. Finally, and most troubling, is the fact that complexity cannot often be compared. It would be a coincidence of the highest order if the complexity of English syntax were *exactly* the same as the complexity of Greek tense and agreement morphology. Absent any way of directly comparing the relative complexity, the apparent balance between syntactic and morphological complexity may be of little significance. And although studies of individual aspects of linguistic complexity, such as (Berlin and Kay, 1969) or (Perkins, 1992) are useful and interesting in their own right, they are difficult to unify into larger statements.

One influential attempt to quantify this notion is that of Nichols (1992), who defined linguistic complexity as the number of points in a typical sentence that are capable of receiving inflection. While obviously limited—aside from the difficulties inherent in defining a 'typical' sentence, this also focuses on morphology and does not address 'how complex and rigid English syntax is', and also misses some important distributional aspects of a more sophisticated analysis—this approach at least directly quantifies the idea of complexity in a way that allows Nichols to make comparative statements about different languages. A similar definition from McWhorter (2001, 2005) encompasses a number of similar ad-hoc measures (e.g. a language is more complex if it has more marked members in its phonemic inventory, or if it makes more extensive use of inflectional morphology), but he ties this, at least in theory, to a single numerical measure—the length of the grammar that a language requires. (In theory, of course, this 'grammar' describes not only syntactic rules but also morphological and phonological ones.) Despite the obvious practical difficulties (how do you compare two different inflectional paradigms, or how do you balance simple morphology with complex phonology), this provides a reasonable formulation for judging complexity.

In previous work (Juola, 1998, 2008), we have developed from first principles an information-theoretic measure of linguistic complexity that largely mirrors McWhorter's but extends and formalizes it. Using this definition, we have been able to confirm Nichols's analysis of morphological complexity and show an explicit relationship between morphological and syntactic complexity supporting Wright's intuitions.

2.2 Information theory

The discipline of information theory is largely based in Shannon's (Shannon, 1948, 1951; Brown *et al.*,

1992) work on communication channels, in which he defined the 'entropy' of a channel as the minimum capacity along which a set of (independent) messages could be sent and individually reconstructed or interpreted by the receiver. One of the key aspects of his formulation is that, like transmitting English messages via Morse code, not all characters [messages] are equally likely and that the unlikely ones can be transmitted using longer codes to allow a more efficient use of the channel by common messages. We omit the exact mathematical details for brevity but provide the following thought-experiment: if I asked you to play the yes/no game of twenty questions with me, given a set of a thousand objects that I might choose, you could always win in no more than ten questions, and if all objects were equally likely, it would almost always take you exactly ten questions. If, on the other hand, you knew I had a mania for a specific object and selected it half the time, you could win half the time by simply asking 'are you thinking of a buttered crumpet?' (or whatever the object were); the other half of the time would take ten additional questions to select from the universe of 999 objects, but you would average only 6 (the average of 1 and 11) instead of 10 questions, thus playing and winning the game with much greater efficiency. This kind of distributional imbalance reduces the 'entropy' of the game. Another common way of looking at this is that 'entropy' is a measure of uncertainty, and the more you know about my manias, the more certain you are of my fixation on crumpets.

This idea was extended by Kolmogorov (Kolmogorov, 1965; Chaitin, 1996; Li and Vitányi, 1997) who noted that while not all messages are independent, the degree of information provided in the message stream reduces the amount of uncertainty because some information can be recovered by examining the prior context at any point. 'Kolmogorov complexity' is thus defined as the size of a computer program that would be necessary to reconstruct a sample of messages from scratch. For example, the entropy of the string 'abababab . . .' is one bit per character (as the letters 'a' and 'b' occur with equal frequency), but the Kolmogorov complexity is nearly nil, as the recurring pattern quickly eliminates any uncertainty we have as to what the next letter is. This can be formalized by noting that a simple loop in any programming language of your choice (a program of a few characters) can generate an arbitrarily large sample of this string. By contrast, a sequence of a's and b's generated by a coin flip would have very high Kolmogorov complexity as the sequence of flips could not be reproduced by any computational process.

This applies to language as well; to some extent, language is predictable (Shannon, 1951; Brown *et al.*, 1992). This can be seen intuitively (fill in the next word: *will you do me a . . .?*) and more formally by observing that a text file is fairly easy to compress effectively. In fact, file compression programs are good tools for approximating Kolmogorov complexity and formed the basis for our observation that all languages are equally complex. When one considers documents with similar contents, such as translations, the length of the texts themselves will vary considerably (Rybicki, 2010), but compressing those texts (Lempel and Ziv, 1976; Ziv and Lempel, 1977, 1978) will tend to produce much more similar document sizes (Juola, 1998, 2005, 2008; Juola *et al.*, 1998). In other words, there does not appear to be any systematic inter-language variation in complexity, In fact, we can identify the Baker's 'increased specificity' (Baker, 1993), claimed to be an aspect of translated text, as an increased information load and hence higher compressed size in translated texts as opposed to the original (Juola, 1997).

We thus have the basic idea is that a sample of language is complex if it contains a lot of information, defined formally as Shannon entropy or Kolmogorov complexity, which can be measured directly by methods as simple as compressing a text sample and looking at the size of the resulting file—the larger the resulting file, the more complex the original.

3 Cultural complexity

Key to this approach is the idea of discourse control; we are measuring how difficult it is to express a specific fixed concept in a given language and

comparing it to the same concept expressed in an another language. Culture, however, can be treated as the set of concepts that people choose to express. By eliminating the restriction of discourse control and instead investigating language chosen freely by the cultural participants, we may be able to tease apart the interaction between cultural and linguistic complexity. In particular, we can distinguish between linguistic and cultural complexity as follows: a language is complex if there is a lot of information contained in a topic-controlled discourse. A culture is complex if there is a large range of topics for discourse, or alternatively a lot of information contained in topical choice. Therefore, if we compare the complexity (however measured) of two language samples that are not topic-controlled, but instead are in some sense representative of the breadth of discourse present in a culture, we can calculate the differences attributable to discourse variation, and hence to cultural complexity.

As an illustrative example, we follow the approach of Spenser (Spencer, 1900; Denton, 2004) in that 'complex' means 'containing many different interdependent parts'. A complex political system has many parties and power groups, many different roles and offices, and many relationships among them. In a political discourse, many if not most of these parties and power groups would need to be explicitly named and distinguished from each other. By contrast, an autocratic monarchy is relatively simple: there is the monarch and then everyone else. A game is complex if it has many rules and strategies. A culture is complex if it contains many heterogeneous aspects such as technological specifications, social stratification, multilevel administrative hierarchies, or a large amount of object or object-types. Continuing this line of reasoning, a complex culture is one with lots of 'stuff' and where people do lots of things to or with 'stuff', where 'stuff' here refers not only to physical objects but also to people, groups, activities, abstractions, and so forth—anything that can be discussed among the group.

We therefore apply the previous methodology to a different sort of corpus; an uncontrolled corpus that represents the breadth of cultural experience. If the information contained in such a corpus is high, then we can say the culture is complex.

4 Corpus and Analysis

Several corpora may be suitable for this purpose; we have chosen to study the Google Books Ngram Corpus (Michel *et al.*, 2011). This contains all of the n-grams from the millions of books in the Google Books database, something like 20 million books, or approximately 4% of all books ever printed. While not strictly speaking representative (for example, 'publishing was a relatively rare event in the 16th and 17th centuries', and 'many more books are published in modern years'), and of course typically only the literate can write or publish books, this nevertheless gives us a time-stamped window into the scope of culture.

Based on the theory developed here, one could simply use vocabulary size as a proxy for culture. As with the mythical Eskimos (and real skiers) (Pullum, 1991) and their huge collection of words for snow, the more finely a culture divides the world, the more information is necessary to select among the divisions. Similarly, one of the primary features of cultural (especially technological) development is to create new stuff that must be named, whether this stuff be common nouns (airplane, transistor, iPad), verbs (to channel-surf, to google), newly created places (Disneyland, the Googleplex), or even characters (Sherlock Holmes, Sam Spade, Darth Vader).

However, we can use the information in n-grams to develop a more finely grained measure of culture that includes not only information about the types of stuff in the culture, but also about its distribution and relationships. By focusing on frequency distribution, we can learn how common words are, knowing from Shannon that distribution plays a key roll in complexity. Furthermore, by examining 2-grams (word pairs), we can first observe not only the distribution of 'stuff' but also some of the relationships between 'stuff'—for example, the number and range of word pairs beginning with 'expensive' will inform us about changing opinions regarding money and the types of goods considered luxurious and pricey. The words associated with 'cellular' will describe the rise of cellular electronics as mobile phones become part of culture and leave biology behind; for an even more pronounced example,

consider 'spam', and whether anyone in 1950 would want to 'filter' or 'block' a meat product.

We therefore used the Google Books American 2-Gram Corpus to measure changes in the complexity of American culture at ten-year intervals between 1900 and 2000. This corpus simply contains a frequency list of all two word phrases used in American-published books in any given year. For example, the phrase 'hamburgers with' appeared only eight times in print in 1940, compared to forty-five in the year 2000. Focusing strictly on the US during the 20th century avoids many of the problems with mass culture, as publishing was a well-established industry and literacy was widespread. However, the number of books published in this time of course tended to increase.

Table 1 Number of 2-gram types by year

Year	# types
1900	17,769,755
1910	22,834,741
1920	22,409,426
1930	19,745,549
1940	20,369,679
1950	23,632,749
1960	27,379,411
1970	34,218,686
1980	34,458,083
1990	37,796,626
2000	41,654,264

Table 2 Entropy of 2-grams by year

Year	Entropy (bits)
1900	17.942357
1910	18.072880
1920	18.072325
1930	18.133058
1940	18.241048
1950	18.336162
1960	18.391872
1970	18.473447
1980	18.692278
1990	18.729807
2000	18.742085

5 Results

5.1 Predicted results

The increased number of books is of course a measure of increased cultural complexity; more books are more stuff to read. Beyond this, more books become more fodder for discussion. Culture is cumulative as a book published in one year may be influencing cultural content for years or decades to come; while a new invention like the car can reshape society, people still remember (and publish about) the horse and buggy era. Our first observation, then, is that culture may be increasing simply from the number of different things to talk about, which would be reflected in the vocabulary.

However, we also expect an increase in cultural complexity beyond mere vocabulary. Measuring the number of word pair types shows more directly the increase in things and their relationships. The number of different word pair types per year increased dramatically, more than doubling from 1900 to 2000, as given in Table 1.

This alone indicates an increase in the complexity of written culture, although this process is not continuous and some years during the Depression show a loss. Intuitively, we do not feel that culture became less complex in 1930 simply because people are no longer able to afford to publish/sell as many books as in 1920. Measuring the Shannon-entropy of the n-gram distribution provides an alternative measure that incorporates distributional information. Our expectation is that this distributional information will also show an increase in complexity. The results of this analysis are attached as Table 2.

This further analysis illustrates that a more sophisticated measure of complexity shows a continuous process of increasing complexity, even in times when (for example due to economic downturn) the actual volume of words published decreases. In particular, note that in 1920, the Google sample shows a complexity of 18.072 bits based on a sample of 22.4 million word pair types. With only 19.7 million word pair types, 1930 showed a complexity of 18.13 bits, and 1940 showed 18.24 bits with 20.4 million word pair types. In other words, even when the quantity of writing dropped, the measured complexity of the writing continued to increase. Even when people are writing less, they

still have more 'stuff' about which to write, showing the cumulative nature of culture (todays current events are tomorrows history, but still suitable material for discussion and analysis—part of culture).

6 Discussion

One of the issues with cultural studies is the difficulty in gathering wide-ranging empirical data. (This issue is not confined purely to cultural studies, but arises in linguistics as well; see (Juola, 2000, 2003, 2012)). It is difficult to define 'culture' tractably and equally if not more difficult to define 'change' in a clear and useful manner. Michel's work shows how specific aspects of culture can be addressed lexicographically by searching for specific words—for example, analyzing Nazi-era censorship by looking at how frequencies of names and concepts differ between German and English-language books over an extended period of time. For broader, overarching concepts, such as the rate of change or the degree of overall cultural similarity, or notions of complexity, however, these small-scale focused analyses may not be sufficient. However, the same large corpora that permit empirical studies on a focused small scale can also allow larger scale overarching studies as demonstrated here.

We acknowledge that this is a preliminary study only. Google Books offers cultural snapshots at much greater frequency than ten-year intervals. Google Books also offers corpora in other languages (including German, French, Spanish, Russian, and Hebrew) as well as another English-speaking culture. Use of a more balanced corpus (such as the Google Books English Million corpus, a corpus balanced at about 1 million words/year to offset increased publication), or the BYU Corpus of Historical American English might help clarify the effects of publication volume. Simply re-running this type of experiment on other corpora reflecting other cultures would be useful. Comparing complexity changes in different cultures or at different times might go further and illuminate some of the causes of cultural change. As a simple example, it has been shown (Juola, 2003) that there is a relationship between language change and war. Is there a relationship between war and cultural change, or for that matter between the economy and cultural change? Such a relationship has been shown for language and war. Certainly, having a booming economy may create a market for new 'stuff' for people to buy, but economic trouble and uncertainty may create a market and incentive for new ways to use existing stuff and new applications of what we already have. If intuition permits both arguments in both directions to be credible, an objective analysis can help.

Analysis of n-grams at sizes other than 2 would illustrate other types of complexity—in particular, 1-grams (words) would show changes in lexical but not syntactic complexity and hence an analysis of 'stuff' but not relationships among 'stuff'. Trigrams (3-grams) are likely to show many instances of 'something of something-else', such as 'parts of speech', 'deck of cards' or 'gaggle of geese'. These 3-grams may capture object-object relationships (such restrictions on collective nouns) that, for grammatical reasons, would not be captured in English 2-grams. (Of course, other languages with different grammar might capture these relationships using a direct noun-noun compound, or even a single compound word.)

A better and broader question, however, is what aspects of 'culture' this simple information-theoretic approach ignores, and what instruments or measurement techniques might bring them out. (Is there, for example, a way of comparing the complexity of musical offerings without going through text first, perhaps by way of concert reviews?) Similarly, there are many aspects of culture that the proposed technique captures implicitly, but that can be made explicit. One possibility, for example, is to look at the different types of n-grams and their changing distribution. Words in English (and many other languages) can be broadly divided into 'content' and 'function' words; function words are typically common words like prepositions that have grammatical function but little meaningful content ('of' is a classic example). Of our 2-grams, we expect (but have not confirmed) that many of them are pairs of function words ('of the', 'instead of', 'by a'), many of them are pairs of content words ('air compressor'), and many are one of each ('the house', 'onions with'). We equally

expect but have not confirmed that most of the function-word-only pairs will not change much in frequency or complexity over time, and that most of the action, so to speak, will be in the content word pairs. But is it possible to perform an intermediate-scale analysis, larger than Michel's but smaller than the present work, that will focus on broad-scale aspects of cultural complexity and change?

As a proposed experiment, consider bigrams beginning with the word 'lady' or 'woman'. These bigrams would include, inter alia, pairs such as 'lady novelist'. As cultural expectations of women's roles change, would we expect to see an increase or a decrease in the complexity of such gender-marked word pairs? Certainly this author has not seen 'lady novelist' in any context outside of Gilbert and Sullivan (or George Elliot) in years.

7 Conclusions

Despite these weaknesses and the huge amount of work still to be done, this article illustrates that culture-wide analysis of abstractions like 'increasing complexity' is both practical and fruitful. This article also shows that large text collections such as the Google N-gram Corpus can be effectively used for this kind of large-scale analysis, and indeed even less well-curated corpora would work well, possibly even better.

Our results are largely a confirmation of intuition—culture is in large part cumulative, so it will continue to become more complex, even when the economics of book publishing dictates less media discourse, the discourse itself will still be informed both by the past and by recent/present developments to create increased complexity. On the other hand, this experiment represents a new way for our intuitions about complexity to be informed by rigorous and objective quantitative analysis.

Acknowledgements

This material is based on work supported by the National Science Foundation under Grant No. OCI-1032683 and by DARPA under BAA-12-06. 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 or DARPA.

References

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In Baker, M., Francis, G., and Tognini-Bonnelli, E. (eds), *Text and Technology: In Honour of John Sinclair*. Philadelphia: John Benjamins Publishing, pp. 233–50.

Berlin, B. and Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA: University of California Press.

Brown, P. F., Della Pietra, V. J., Mercer, R. L., Della Pietra, S. A., and Lai, J. C. (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics*, **18**(1): 31–40.

Chaitin, G. J. (1996). A new verson of algorithmic information theory. *Complexity*, **1**(4): 55–9.

Denton, T. (2004). Cultural complexity revisited. *Cross-Cultural Research*, **38**(1): 3–26.

Juola, P. (1997). *A numerical analysis of cultural context in translation, Proceedings of the Second European Conference on Cognitive Science*. Manchester, UK, pp. 207–10.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, **5**(3): 206–13.

Juola, P. (2000). The rate of language change. *Proceedings of Qualico-00*. Prague: Czech Republic.

Juola, P. (2003). The time course of language change. *Computers and the Humanities*, **37**(1): 77–96.

Juola, P. (2005). Compression-based analysis of language complexity. Presented at *Approaches to Complexity in Language*. Finland, Helsinki.

Juola, P. (2008). Assessing linguistic complexity. In Miestamo, M., Sinnemaki, K., and Karlsson, F. (eds), *Language Complexity: Typology, Contact, Change*. Amsterdam: John Benjamins Press.

Juola, P. (2012). *Using the Google Ngram Corpus to measure dialectical and cultural differences, Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*. Chicago.

Juola, P., Bailey, T. M., and Pothos, E. M. (1998). *Theory-neutral system regularity measurements,*

Proceedings of the Twentieth Annual Conference of the Cognitive Science Society (CogSci-98). Madison, WI.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission*, **1**: 1–7.

Lempel, A. and Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, **IT-22**(1): 75–81.

Li, M. and Vitányi, P. (1997). An introduction to Kolmogorov Complexity and its applications. *Graduate Texts in Computer Science,* 2nd edn. New York: Springer.

McWhorter, J. H. (2001). The world's simplest grammars are creole grammars. *Linguistic Typology*, **6**: 125–166.

McWhorter, J. H. (2005). *Defining Creole*. Oxford: Oxford University Press.

Michel, J.-B., Shen, Y. K., Aiden, A. P. *et al.* (2011). Quantitative analysis of culture using millions of digitized books. *Science*, **331**(6014): 176–82.

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. Chicago, IL: University of Chicago Press.

Perkins, R. D. (1992). Deixis, Grammar and Culture. *Typological Studies in Language*. Amsterdam: John Benjamins.

Pullum, G. K. (1991). *The Great Eskimo Vocabulary Hoax*. University of Chicago Press.

Rybicki, J. (2010). *Original, translation, inflation. Are all translations longer than their originals?, Proceedings of Digital Humanities 2010*. London, June 2010, pp. 363–4.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**(4): 379–423.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, **30**(1): 50–64.

Spencer, H. (1900). *First Principles.* 6th edn. Akron, OH: Werner.

Ziv, J. and Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, **IT-23**(3): 337–43.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable rate coding. *IEEE Transactions on Information Theory*, **IT-24**(5): 530–6.