Entropy and Divergence in a Modern Fiction Corpus

Craig, Hugh
hugh.craig@newcastle.edu.au
School of Humanities and Social Science,
University of Newcastle, Australia

The application of statistical methods to style is now well accepted in author attribution. It has found less favour in broader stylistic description. Louis Milic's pioneering quantitative work from the 1960s on the style of Jonathan Swift was vigorously contested by Stanley Fish, an attack which may well have had the effect of curbing enthusiasm for this kind of work. The other important exemplar is John Burrows' book on Jane Austen from 1987. I am not aware of any subsequent books of this kind.

In the proposed paper I aim to demonstrate the usefulness of two measures from Information Theory in the broad comparative analysis of text. One is entropy, which calculates the greatest possible compression of the information provided by a set of items considered as members of distinct classes (Rosso, Craig and Moscato). A large entropy value indicates that the items fall into a large number of classes, and thus must be represented by listing the counts of a large number of these classes. In an ecosystem, this would correspond to the presence of a large number of species each with relatively few members. The maximum entropy value occurs where each item represents a distinct class. Minimum entropy occurs where all items belong to a single class. In terms of language, word tokens are the items and word types the classes. A high-entropy text contains a large number of word types, many with a single token. A good example would be a technical manual for a complex machine which specifies numerous distinct small parts. A low-entropy text contains few word types, each with many occurrences, such as a legal document where terms are repeated in each clause to avoid ambiguity. Entropy is a measure of a sparse and diverse distribution versus a dense and concentrated one.

A second information-theory quantity which can serve for generalising about a set of texts is Jensen-Shannon Divergence (JSD). This gives a value to each set of items for the distance from a reference point, generally the mean for the whole grouping. This distance is calculated as the sum of divergences between the specimen and the mean for each of the classes represented in the set (Rosso, Craig and Moscato). In language terms the divergence value of a given text is the sum of the differences between the counts for the text for each word type used in a corpus and the corpus mean count for that word type. Some texts use language in a way that closely corresponds to the norm of a larger set, others use some words more heavily, and others more lightly, than the run of a comparable corpus. JSD is a measure of normality in this specialised sense.

There are important caveats for interpreting these two measures of the properties of a text. Both are sensitive to text length, if for different reasons. Given a finite number of word types available to a given user of a given language, as a text sample grows, more of the pool is exhausted, and there is a greater tendency to recur to already-used word types. Thus in a novel the word tokens of a single sentence may well be all different word types, and have maximum entropy, but this is unlikely to be true of a paragraph, and still less so of a chapter. In the case of divergence from a mean, the law of averages means that for longer texts local idiosyncrasies tend to be balanced out by a larger body of less unusual writing and indeed by contrasting idiosyncrasies.

It is also important to rule out the idea that entropy and divergence values relate directly to quality. Entropy is related to a simpler measure, type-token ratio, sometimes called 'vocabulary richness'. Yet 'richness' could scarcely be applied to a fighter plane manual, to revert to the example used above. One might associate divergence from the mean with originality or creativity, but it could just as well be the result of incompetence.

It is interesting that researchers have found genre to be a problem both with entropy work and with studies of intertextual distance when they are directed at authorship problems (Hoover, Labbé and Labbé). From a different point of view, this sensitivity to genre is part of

what makes the methods valuable for a more general assessment of the style of texts.

The corpus for the study in the paper consists of 377 fiction texts, being the first 25,000 words of all the texts with 25,000 words or more in the British National Corpus 'Imaginative Fiction' section. This amounts to 15,421,915 words in all. The texts are predominantly prose fiction published in the United Kingdom in the 1990s, taken from a wide variety of sources, short stories as well as novels, intended for young and young adult audiences as well as for a general readership. The usefulness of JSD results depend on the validity of the point of reference chosen. In the present study the mean of this large collection of texts of very varied authorship and genre, within the larger text type 'imaginative fiction', should be a good approximation of the mean for contemporary fiction in general.

At the time of writing this proposal work on this corpus with these methods is at an early stage, but there are some preliminary findings. The first is that entropy and divergence are positively correlated in this corpus. As density decreases and a wider range of word types are used for the same extent of text, samples diverge more from the mean. The individual exceptions to these broad tendencies are instructive and some individual examples will be discussed.

It is also possible to see at this early stage that within the universe of prose fiction these two quantities align with more impressionist views of style. High entropy fiction texts follow a traditional 'high style'. Their progression is linear, continuing to move on to new vocabulary, while low entropy texts retrace their steps and return to already used words. High entropy texts are demanding of the reader and dense in information. They constantly move to new mental territories; they are taxing and impressive. Low entropy texts are reassuring and familiar. They are implicit in their signification, assuming common knowledge, while high-entropy texts specify and create contexts for themselves. High-entropy texts contain more description and narrative, while low-entropy texts contain more dialogue.

The challenge for computational approaches to style is to use the power of statistics working on the abundant data available from texts to reveal tendencies which are important, yet would otherwise be invisible, or remain in the realm of the impressionistic. The argument of the proposed paper is that the entropy and divergence of words provide two useful ways of understanding fundamental properties of texts. Entropy and divergence are soundly based in statistical theory and informative on two fronts. They open the way to density and normality as fundamental ways of thinking about style; and they serve to place particular texts in relation to sets of comparison texts and thus to map them in a conceptual space. Short stories and novels may be virtual worlds, intensely personal meditations, and human dramas of love and conflict, but they are also sets of vocabulary items used with a given frequency, and it is surprising how much an analysis of that base level of their existence can reveal about them.

References

British National Corpus, version 2 (2001). *BNC World*. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen and an Experiment in Method*. Oxford: Clarendon.

Fish, Stanley (1980). 'What Is Stylistics and Why Are They Saying Such Terrible Things About It?'. *Is There a Text in This Class?*. Cambridge MA: Harvard University Press, pp. 68-96.

Hoover, David (2003). 'Another Perspective on Vocabulary Richness'. *Computers and the Humanities*. **37**: 151-78.

Labbé, Cyril, Labbé, Dominique (2006). 'A Tool for Literary Studies: Intertextual Distance And Tree Classification'. *Literary and Linguistic Computing*. **21.3**: 311-26.

Milic, Louis T. (1967). *A Quantitative Approach to the Style of Jonathan Swift*. Mouton: The Hague.

Rosso, Osvaldo, Craig, Hugh, Moscato, Pablo (2009). 'Shakespeare and Other English Renaissance Authors as Characterized by

Information Theory Complexity Quantifiers'. *Physica A*. **388**: 916-26.