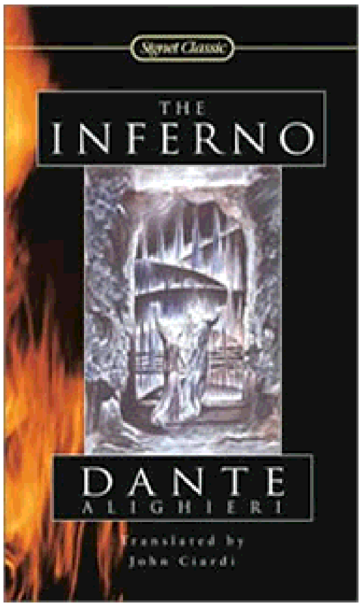
# Algorithm makes tongue tree

New computer programme could settle literary debates.
22 January 2002

PHILIP BALL




Compression helps a computer tell Dante from Machiavelli.

To date, unlike us, computers have struggled to differentiate a page of Jane Austen from one by Jackie Collins. Now researchers in Italy have developed a program that can spot enough subtle differences between two authors' works to attribute authorship[1].

The program can tell a text by Machiavelli from one by Pirandello, Dante or a host of other great Italian writers. It constructed a language tree of the degree of affinity between 50 different tongues. The tree identifies all the main linguistic groups, such as Romance, Celtic, Slavic and so forth and highlights Maltese (an Afro-Asiatic language) and Basque as anomalies.

As well as settling a few literary arguments, the

relatedstories

*Language gene found
4 October 2001

*Physicists decode the Bard
21 September 2001

*Speaking in syntactical tongues
30 March 2000

*Making sense of sentences
12 November 1999

morenews

*Car work for quantum mechanics
30 January 2002

*Air traffic tussles up annoyance
30 January 2002

*Jellies roll back time
30 January 2002

*Moon mirrors life's colours
29 January 2002

*Cylinders make circuits spontaneously
29 January 2002

technique might be useful for comparing other information-rich sequences of data. These might include genetic sequences, medical-monitoring measurements and stock-market fluctuations.

**Style over substance**

Identifying the language of a particular text is generally not hard in itself: one need simply look for the greatest overlap between the words used and those in a reference list for each language. Classifying linguistic styles is altogether more tricky.

One obvious approach is to compare the range and frequency of words in the sample text against reference texts from various candidate authors. That might work for markedly different styles: it would quickly distinguish Shakespeare from Tom Clancy.

But literary scholars often argue furiously about attributions for old texts. The task can become immensely difficult even for those with a great deal of knowledge about the candidate authors' writing styles.

**Clash of symbols**

So Dario Benedetto and colleagues at the Universita 'La Sapienza' in Rome try a different approach. They start from the premise that written language is in the end no more than a string of symbols. It might look rather random, but it is not.

Some groups of characters recur commonly (such as 'the' in English), and particular authors favour certain constructions and turns of phrase. These can be measured, rather than being reliant on subjective impressions or anecdotal comparisons.

Entropy is a measure of how much redundancy a message contains

The team begin from the classic insight of telecommunications engineer Claude Shannon in the 1940s that the information content of a message is related to its entropy. Roughly speaking, entropy is a measure of how much redundancy a message contains. It can be defined as the smallest program that will produce the original message as the output.

For a random string of characters, this program would simply specify every character - it would be the same size as the original message. For a string of just A's, the program could be very concise: 'repeat A'. Most real messages lie somewhere in-between: they can usually be compressed a little without losing significant information. This is the basis of data-compression computer algorithms, used to make 'zip' files, for instance.

Benedetto and his colleagues borrow the principles of data-compression algorithms to calculate a kind of relative entropy for two different character strings: a measure of how much they differ. This distance between two texts is smaller for two works by the same author than for two works by different authors.

A similar 'information-based' method has been used previously to reconstruct phylogenetic evolutionary trees by looking at the similarities between DNA sequences of different organisms[2].

References

1. Benedetto, D., Caglioti, E. & Loreto, V. Language trees and zipping. *Physical Review Letters*, 88, 048702, (2002).
2. Li, M. et al. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17, 149 - 3154, (2001).