

- [OPINION](#)
- [WORLD](#)
- [BUSINESS](#)
- [FINANCE & ECONOMICS](#)
- [SCIENCE & TECHNOLOGY](#)
Technology Quarterly
- [PEOPLE](#)
- [BOOKS & ARTS](#)
- [MARKETS & DATA](#)
- [DIVERSIONS](#)





Economist Intelligence Unit
onlinestore

LIBRARY

- [Articles by subject](#)
- [Backgrounders](#)
- [Surveys](#)

Computers and language

-  [Printable page](#)
-  [E-mail this](#)

The elements of style

Feb 7th 2002
From The Economist print edition

Analysing compressed data leads to impressive results in linguistics

ZIPPING, as any computer buff knows, enables you to compress a file so that it can be stored efficiently, or sent quickly over the Internet. But Emanuele Caglioti and his colleagues at the University of Rome-La Sapienza have found a more esoteric use for it. Using zipped files, they can identify the authors of documents and reconstruct the family trees of languages.

The secret lies in the science of information theory, invented by Claude Shannon in the 1940s. Shannon pointed out that the length of the instructions used to encode a string of characters corresponds to the disorder, or "entropy", of that string. A repetitive sequence such as AAAAAA contains little entropy. It can be encoded with a brief command: "repeat A five times". On the other hand, a sequence such as QMTWZ can be encoded only with a set of instructions as long as the original sequence.

RELATED ITEMS

A new wave sampling
Jan 17th 200

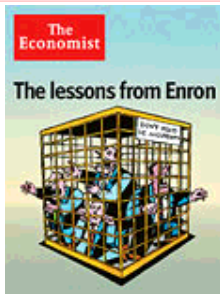
Websites

Dario Bened is published :
Review Lette
earlier article
Benedetto, al
language tree
zipping, is a
online.



ADVERTISEMENT

PRINT EDITION



Full contents
Subscriptions

CAREERS

Global Executive
with Whitehead Mann

Career guidance,
personalised advice, job
postings and more.
Click here

SHOP

Economist Shop

Books, diaries and more

CLASSIFIEDS

Business education,
recruitment, business
and personal: click here

ABOUT US

Economist.com
The Economist
Global Agenda
Contact us
Advertising info
Work for us

set of instructions as long as the original sequence.

In practice, the entropy of most writing lies somewhere between these extremes. Zipping programs work by replacing low-entropy data with instructions for reconstructing the replaced data. A good zipping program is able to work out the rules most applicable to a particular document as it goes along.

The length of a zipped file offers a rough-and-ready estimate of its entropy. Comparing the entropy of two texts, however, is slightly more complicated. One method is to feed the zipper some text in one language, then switch the input to a different language. The zipper suddenly finds that the tricks that it has picked up to encode the first language are not much help in encoding the second. In an English-to-French switch, for example, instances of "the" would abruptly become rare, whereas "le", "la" and "les" would crop up all over the place. The result is that the zipped file of such a hybrid document is longer than its monoglot equivalent. The less similar the languages, the more the extra length that is added to the hybrid zipped file. The same, to a lesser degree, is true of documents that have more than one author, and therefore more than one writing style.

Dr Caglioti and his colleagues have created a program that can categorise documents by language or authorship, based on these extra lengths. As he and his colleagues report in *Physical Review Letters*, they first tested it with ten texts apiece from ten official languages of the European Union. Using it, a snippet of text as short as 20 characters can be assigned unerringly to the language it was written in.

As a second test of the program's abilities, they used 52 versions of the document which, according to the "Guinness Book of Records", has been translated into more languages than any other in the world: the Universal Declaration of Human Rights. Forty-nine of these versions were in European languages or dialects. One was in Afrikaans, a South African language derived from Dutch. The other two were Uzbek and Turkish. The program calculated the relative entropies of all possible pairs of these 52 languages. It then used this information to construct a family tree that placed them into clusters.

When completed, this tree had sprouted branches representing all the main language groups in modern Europe: Romance, Celtic, Germanic, Slavic and so on. Moreover, the program was able to recognise the singularity of languages such as Basque and Maltese. It left these isolated, just as linguistic scholars do.

Measurements of relative entropy were enough to unmask anonymous authors, too. The program was fed a set of 89 texts written by nine Italians, including Dante, Machiavelli and Pirandello. Then came the test: by looking for the minimum amount of relative entropy, it tried to guess which of the nine was the author of a 90th text. More than 90% of the time the guess was accurate.

Of course, any self-respecting linguist could have performed these tasks as well as this, if not better. But the mathematicians' invention is still in its infancy, and will soon be set loose on languages that humans cannot easily learn-protein sequences, for example, or pieces of DNA. It would certainly be interesting if it managed to unmask an anonymous author behind these particular strings of text.

The best
deal from
The
Economist.

50% off the
newspaper



FREE
unlimited
access to
Economist.com

The
Economist



[OPINION](#) | [WORLD](#) | [BUSINESS](#) | [FINANCE & ECONOMICS](#) | [SCIENCE & TECHNOLOGY](#)
[PEOPLE](#) | [BOOKS & ARTS](#) | [MARKETS & DATA](#) | [DIVERSIONS](#) | [PRINT EDITION](#)

An Economist Group business

Copyright © The Economist Newspaper Limited 2002. All rights reserved.
[Legal disclaimer](#) | [Privacy Policy](#) | [Terms & Conditions](#) | [Help](#)
