C. SAUNDERS

Clear as mud

It's not surprising that some academic papers seem to swim before our eyes — the scientific literature has become steadily less accessible over the past half-century. Can we stop this trend, asks Jonathan Knight.

"There is no form of prose more difficult to understand and more tedious to read than the average scientific paper," wrote Francis Crick in his 1994 book *The Astonishing Hypothesis*[1]. The observation is a caution to lay readers tempted to delve into the papers referenced in the book. But the co-discoverer of the structure of DNA was also acknowledging what everyone in science knows: research papers can be a nightmare to read.

It wasn't always so. Crick and others of his generation, who began writing scientific papers in the 1940s, have witnessed the transformation of scientific prose. A form that was as readable as the average newspaper has, in some fields, become a jungle of jargon that even those familiar with the territory struggle to understand.

The balkanization of science into subdisciplines, each with its own vocabulary, is largely to blame. Many journals are trying to tackle this, producing easy-to-read summaries of papers, and linking online papers to web-based glossaries. But these approaches tend to have a limited impact, whereas addressing other factors — notably writing style — could transform many papers. Writing takes practice, yet it is not part of standard scientific training. So could science become readable again if researchers went back to school and took writing lessons?

Readability itself is not easy to quantify. Microsoft's *Word* program features the Flesch Reading Ease scale, which measures the average length of words and sentences to calculate the number of years of education needed to comprehend a document. But such tools fail on several counts. For one, a long sentence that walks the reader down a path to its conclusion can be easier to follow than a muddled short sentence. And common words can be relatively long — technological or professor, for example — whereas many technical terms are short, such as meson, genome or glycan.

The common touch

Language experts generally agree that a better measure of accessibility is whether a piece of writing contains words in common usage — those that are at the front of the reader's mind, rather than tucked away in the recesses of memory. As a general principle, the greater the percentage of common words an article contains, the easier it is to comprehend.

Donald Hayes, an emeritus professor of sociology at Cornell University in Ithaca, New York, has used this principle for more than 20 years to analyse texts. He calls it lexical difficulty, and has developed a numerical scale, known as LEX, to quantify it. The scale is based on the *American Heritage Word Frequency Book*[2], which ranks 87,000 words by their frequency of use in textbooks, novels, magazines and encyclopaedias from US grammar schools in 1969.

Although the ranking is over 30 years old, it remains the primary word-frequency reference. 'The' is the most common word, with 'whooping' in 10,000th place. Among the scientific terms common enough to be included are 'bacteria' at 3,546, near 'pump' and 'fool'; and 'neuron', which ranks 23,595 — about as common as 'diddle'.

When calculating LEX scores, Hayes ignores the first 75 most common words as these contain little useful information. He then plots the 'cumulative proportion' of each word against the log of its rank. The cumulative proportion of, say, the 100th most common word — 'know' — is the percentage of the text made up of the words that lie between 75 and 100 in the frequency ranking. The graph (right) shows that the 1,000 most common words make up about 70% of all the words used by mothers when speaking to their children (orange line). In contrast, the same words make up only 20% of those used in the average *Nature* research article (blue line).

LEX values are generated by comparing the text's curve with the benchmark curve for newspapers, which have a LEX score of zero. The area under the text's curve is subtracted from the area under the newspaper curve to give the LEX value. Texts that use common words more frequently than newspapers have curves that rise rapidly, giving them a large area and a negative value; those that are skewed towards rare words end up with a positive value.

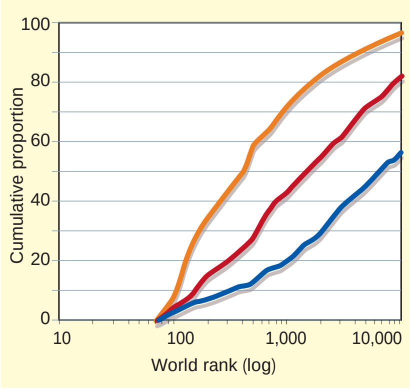
In a 1992 analysis[3], Hayes found that fiction for nine-year-olds scored about −32, and a transcript of farm workers talking to dairy cows — "Let's go. Over here. You dummy, over here." — had a value of −59. Scientific papers in *Nature* and *Science* scored about 30. When *Nature* asked Hayes to repeat his analysis last year, papers in both magazines had risen to the mid-30s. This trend is not new: in the early 1900s, papers in *Science* and *Nature* had accessibility scores of close to zero (see graph, above right), similar to those of newspapers such as *The Daily Telegraph* and *The New York Times*.

Alphabet soup

What happened, says *Science*'s editor-in-chief Donald Kennedy, is that sometime after the Second World War, the number of people active in science increased dramatically, creating new subdisciplines. As they entered ever more specialized fields, new vocabularies arose. The subdisciplines of biology are among the worst for jargon. In the past 20 years, immunologists have uncovered a new world of proteins and processes, each requiring a new name or acronym. Cell-signalling research is also packed with unfamiliar terms. The average paper in *Cell*, for example, has a LEX score of about 40.

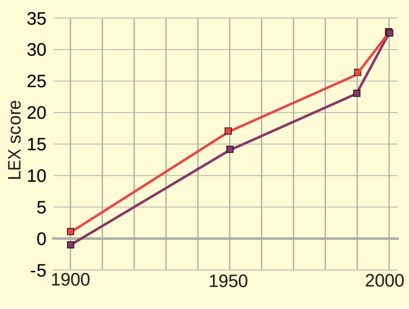
The physical sciences do a bit better. Earth scientists often use relatively common words to describe what they study, such as 'ice sheet' or 'volcano'. Specialized vocabulary exists, but there is less of it. And according to a recent unpublished study by Hayes, average papers in *Physical Review D* and the astronomy journal *Icarus* have LEX values of about 22.

D. HAYES



Typical LEX curves. The red line represents a newspaper, orange is a mother talking to her baby, blue is an average research paper in *Nature*.

D. HAYES



Sign of the times: the LEX scores for *Nature* (red) and *Science* have risen steeply since 1900.

The effects of an increasingly opaque literature are easy to imagine, if difficult to quantify. If opening paragraphs or abstracts are difficult to understand, researchers may miss opportunities for collaboration between disciplines. If whole papers are unclear, students get diverted to other interests and the public's fear and mistrust of science, which in part arises from difficulties in understanding new research, may increase.

Some journals are taking small steps to tackle the problem. Earlier this year, *Science* began adding one-line explanations of its papers to its table of contents. *Development* and the *Journal of Cell Science*, together with other journals published by the Company of Biologists in Cambridge, UK, have added a section to highlight half-a-dozen papers in each issue in language that is accessible to all biologists. *Nature* and *Science* have similar sections, which are complemented by longer pieces written by other academics discussing the newly published papers.

The Internet is also playing an important part in the solution. Each week, one of *Science*'s 'Perspectives' — a commentary on a published paper — in its online edition appears with links from technical terms in the text to web glossaries or sites with further information, a practice also followed by the review journals of the Nature Publishing Group. Articles in the forthcoming online journals of the Public Library of Science, a San Francisco-based organization that promotes free access to scientific literature, will be paired with lay-language summaries. And Cell Press journals now include general-interest summaries of articles in tables of contents sent out by e-mail.

But these are not perfect solutions. Scientists can be suspicious of lay summaries, fearing that they are oversimplified or inaccurate. And the Internet could exacerbate, instead of lessen, the balkanization within science. Scientists reading online are less likely to scan eye-catching figures or cogent abstracts that might entice them out of their field. And although summaries and web links lower the barrier to understanding caused by jargon and acronyms, they can't eliminate it. "These are Band-Aids," says Kennedy. "It will be tremendously difficult to solve the problem."

It is also easy to forget why jargon is there in the first place. Technical terms are problematic for outsiders, but they are indispensable for specialists. They allow accurate shorthand for substances and processes that would take paragraphs to define. Apart from adding brief notes of explanation where space permits, the editors of top journals contacted for this article all agreed that there is little that can be done about jargon in research papers — it is here to stay.

Jargon busters

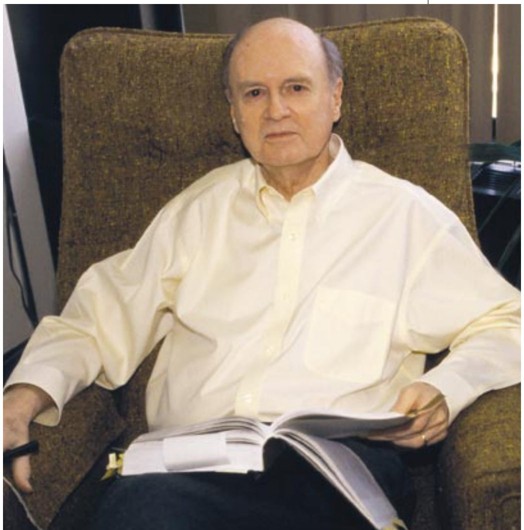
But there are other ways to improve readability. "Jargon is less pernicious if you can understand what is going on," says writing instructor Judith Swan of Princeton University in New Jersey, a former biochemist who now runs workshops to help scientists to improve their papers. Swan's courses stem from a collaboration with George Gopen, a lecturer in English at Duke University in Durham, North Carolina, in which the pair developed principles of clear scientific writing by analysing published papers[4].

Researchers who attend the workshops expect to be told never to use the passive voice or split an infinitive, but Swan takes a different approach. "The one rule I subscribe to is that there are no rules," she says. "One doesn't follow rules, one exercises judgement."

Take passive voice. Active sentences do pack more punch, says Swan, but passive ones are sometimes clearer. For example, there is no need to begin every sentence with "We". Scientific papers tell stories about experiments and data, not scientists. Rather than use 'We found the value to be *x*', it is fine to say 'The value was found to be *x*', suggests Swan. "The passive is a marvellous way to hide agency when agency is not important," she says.

In general, Swan and Gopen recommend

C. HARRINGTON/CORNELL UNIV.



Brought to book: Donald Hayes has devised a scale for measuring the accessibility of an article.

focusing attention on the expectations of the readers. Linguists know that information is easier to interpret if it is placed where readers expect it to be. So, for example, when a subject is introduced in a sentence, readers expect to find a verb soon after it. Everything that comes between the subject and the verb gets little attention.

A place for everything

Take this sentence from a paper in a recent issue of *Science*: "The emergence of virulent *Plasmodium falciparum* in Africa within the past 6,000 years as a result of a cascade of changes in human behaviour and mosquito transmission has recently been hypothesized." After the subject — "emergence" — the reader must wade through 25 words before reaching the verb — "has been hypothesized". Readers will focus too much attention on the anticipated verb to notice the importance of the intervening material.

Swan suggests the following rewrite: "According to a recent hypothesis, virulent *Plasmodium falciparum* emerged in Africa within the past 6,000 years as a result of a cascade of changes in human behaviour and mosquito transmission." Not only are the subject and verb snugly together — "*Plasmodium falciparum* emerged" — but now the important information occupies a key position in the sentence: the end.

The last part of a sentence is what linguists call the stress position. Readers naturally emphasize the information at the end of a unit of discourse, such as a sentence or paragraph, making it the logical spot for new information. Old information does better near the beginning of a sentence, where it grounds the reader in preparation for the mental leap to come. And the more closely the structure matches the reader's expectations, the more likely the reader is to comprehend what the author is trying to say.

Another mistake often occurs right at the start of the sentence, in the 'topic position'. Readers expect to find some sort of bridge between sentences here. If a completely new word or phrase occupies this spot, Swan says, the reader is momentarily confused. For example, a recent paper in *Cell* begins: "We demonstrate that the tendons associated with the axial skeleton derive from a heretofore unappreciated, fourth compartment of the somites. Scleraxis (Scx), a bHLH transcription factor, marks this somitic tendon progenitor population at its inception, and is continuously expressed through differentiation into the mature tendons."

Word up: Polly Matzinger wants to see prizes awarded for the best-written scientific papers.

The authors unintentionally trip the reader by starting the second sentence with a brand new term — the *Scleraxis* gene. Although readers can manage this jump, it forces them to divert some of their attention from the science to the writing, particularly if the pattern recurs. To avoid this, Swan suggests sliding one of the familiar items from later in the sentence to the front to prepare the readers the new information. "This somitic tendon progenitor population is marked at its inception by the gene *Scleraxis* (*Scx*), a bHLH transcription factor." Although passive, the revised sentence smoothes the prose, so readers can focus more intently on the scientific content.

Swan's advice is distributed in a variety of ways. Some institutions, such as JILA, a physics laboratory at the University of Colorado, Boulder, have offices that are dedicated to editing papers and helping scientists with their writing, and which draw on materials developed by Swan and Gopen. Swan also gives about eight scientific-writing workshops a year in the United States. The Earth sciences division of the Lawrence Berkeley National Laboratory in California hosted one workshop last year. Divisional director Bo Bodvarsson says such training is essential to a successful scientific career, particularly for students whose first language is not English, because clear communication opens doors.

Do the write thing

Despite such enthusiasm, most scientists receive no such training. Part of the reason may be that a significant minority of researchers believe that good writing cannot be taught. Among them is Christopher Miller, a biochemist at Brandeis University in Waltham, Massachusetts, known among editors for submitting clearly written papers and reviews. He says that he doesn't follow a set of writing rules, but writes instinctively and only when he is in a "writing mood".

This instinct isn't something that Miller feels he can convey to his students, so he takes a different approach. Anyone in his lab has two chances to write a paper. "You give me a draft and it will stink, I will write a few things on it and you get another chance," he says. If that advice doesn't result in an acceptable paper, Miller writes it himself. Some students and postdocs aren't interested, but those who are often produce good second drafts nearly ready for submission, he says. "It has turned out to be productive and fun."

Pens down: Christopher Miller believes that the instinct for good writing cannot be taught.

For those who do not have access to advisers such as Miller or Swan, professional manuscript editing services can help. Brian Leonard runs Exact Science Communications, an editing service based in Surprise, Arizona. He says that most of his clients are non-native speakers of English — others want to polish their manuscript to improve the chances of publication. In some cases the suggestion to seek professional help comes from journal editors or referees, he says.

Those editors could also do a lot more to draw better writing from contributors, says Polly Matzinger, an immunologist at the US National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, and a scientific adviser to the Council for the Advancement of Science Writing, which aims to improve the quality of science journalism. Many editors already spend a great deal of time whipping manuscripts into shape, with particular emphasis on the abstract and first paragraph. But Matzinger thinks that journals should push harder and expect good writing in all submissions, possibly even rejecting papers on that basis alone. "Play it up, talk about it, insist on it," she says. One idea would be for journals to announce an annual award for the best-written paper of the year.

With a lack of big ideas for addressing the jargon problem, bit-part solutions, such as prizes and the range of other 'Band-Aid' measures, are currently the best hope for promoting accessibility. Together with the techniques promoted by Swan, they should help to attract scientists to new kinds of abstracts, and keep them hooked until the end of the article. Thanks for sticking with this one. It has a LEX score of −4.1, so you don't have to be smart enough to read *The New York Times* to understand it — but it's probably too complex for cows. ■

Jonathan Knight writes for *Nature* from San Francisco.

1. Crick F. *The Astonishing Hypothesis* (Scribner, New York, 1994).
2. Carroll, J. B., Davies, P. & Richman, B. *American Heritage Word Frequency Book* (Houghton-Mifflin, Boston, 1971).
3. Hayes, D. P. *Nature* **356,** 739–740 (1992).
4. Gopen, G. D. & Swan, J. A. *Am. Sci.* **78,** 550–558 (1990).