Text analysis

ONE of the main contributors to a text's difficulty is its pattern of word choice. In English, this choice is from an estimated 600,000 word-types (terms having unique orthography). A log-normal model[1] of word choice predicts that when the words from a large representative sample of texts are arrayed by the log of their frequency of use, the resulting cumulative distribution will be linear. British and US newspapers have closely followed this pattern of word choice since at least 1730. Because of this stability, the pattern's simplicity and their wide readership, newspapers were adopted as the standard for comparison.

Taking social interactions into account, however, leads to a more complex and general model. Speakers and authors normally tailor texts to their intended audience's interests and knowledge. Compared to newspaper word choice, texts of spontaneous speech underuse the more common grammatical words, overuse the more common substantive words and underuse the rarer substantive words, producing an S-shaped cumulative distribution. Difficult technical texts have the opposite biases, producing the reverse S-shaped distribution. Lexical difficulty is represented by this spectrum of lexical patterns.

The software used in the work described here calculates the discrepancy between a specific text's pattern of word choice and the linear pattern of newspapers. First, each text of 1,500+ words is derived by multistage stratified simple random sampling and edited to a common standard. Second, a cumulative curve is generated from the words in that text beginning with the proportion of the most common English word, 'the'; to which is added the proportion of the second ('of'); the third ('and'); and so on through the 10,000th most common word. (Reliable estimates for word frequencies beyond 10,000 are not available.) Third, the 75 most common words in English, accounting for about half the words in texts, are deleted as they contain little information.

Finally, the area beneath that text's cumulative curve is integrated and subtracted from the corresponding area beneath the cumulative curve for newspapers. Texts with negative lexical difficulty scores are skewed towards common words; those with positive scores are skewed towards rare words. The values quoted in this article represent the extent to which word choice is skewed relative to that of newspapers. **D.P.H.**

1. Herdan, G. *Language as Choice and Chance* (Noordhoff, Groningen, 1956).
2. Carroll, J. B., Davies, P. & Richman, B. *Word Frequency Book* (Houghton Mifflin, Boston, 1971).
3. Hayes, D. P. *J. Mem. Lang.* **27**, 572–585 (1988).

on a reef, science magazines must compete for essential resources: important authors and papers, subscribers and, for some, advertisers. They may have to compete for or exploit lexical niches as well.

For example, in the late 1970s it must have become apparent to other publishers that *Scientific American* had left its old niche at 0.0 and was not going to return. In the United States, four general science magazines were created to fill the gap. *Science Digest* transformed itself from a *Readers' Digest* format into a *Scientific American* lookalike (−2.6 in 1986). The American Chemical Society changed the name of the publication *Chemistry* to *SciQuest*, and broadened its message, coverage and appeal (2.2 in 1986). The American Association for the Advancement of Science (publisher of *Science*) developed *Science-80* (−1.0 in 1986) to fill a void in part created when the research articles in *Science* had risen in difficulty from 7 in 1960 to 17 in 1980. Only *Discover* survives (−0.4 in 1986, but −3.6 in 1992). For a brief period, all four magazines occupied the 0.0 niche.

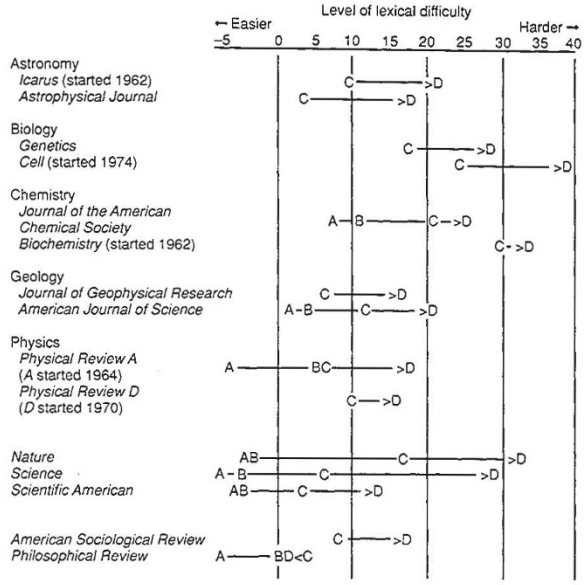

FIG. 2 Change in lexical difficulty in ten basic science journals, the three general journals and two journals dealing with other disciplines. A, 1900; B, 1925; C, 1950 or the first year of publication; D, 1990.

The growth of science has greatly enlarged the audience for general and technical science publications. As their technical articles became more difficult, the general science journals and magazines vacated their former lexical niches. These were soon filled (coincidentally at the vacated levels) by new publications or by ones which moved there from some other niche. Such publications now fill most niches between −22.6 and 38. In particular, professional societies and science publishers have produced several single-science magazines tailored to specific audiences (for example *Physics Today* 13.3, *BioScience* 16.8, *Geology Today* 11.2, and *Chemistry in Britain* 12.6). There is even a chemistry newspaper, *Reaction Times* (7.8). A final adaptation to this trend has been for journals to differentiate parts of each issue, setting each section to a different lexical level, so all readers will find something they can read.

What, though, are the consequences of the drift towards inaccessibility? Specialization in science has produced unprecedented levels of knowledge, but the unwelcome side-effects are clear. These days, more expertise than ever is required to understand published research and theory in other fields and to referee papers and proposals in one's own discipline. The broad consequences are that ideas flow less freely across and within the sciences, and the public's access to (and maybe trust in) science is diminished.

To scientists this trend represents a narrowing of their range of expertise, even while the depth of their knowledge grows. So they may change specialities less often as the costs of becoming expert in another area grow. One response, I suspect, has been an increase in collaboration with scientists in other specialities. Another has been to develop still more complex research teams whose members have complementary skills and knowledge. Complicated sociological structures such as this can be productive but they introduce new kinds of tension, for instance disputes over the order in which names appear on a paper.

Projecting the trend summarized in Fig. 2, there will soon be basic science journals whose average article difficulty will exceed 40, and before long some journal may consistently exceed 50 (indeed, many articles in *Cell* currently exceed 40, and a few now exceed 50). No mainstream science journal was as high as 10 in 1900. And of the nine journals examined and published as recently as 1950, only one was above 20. This erection of higher and higher barriers to the comprehension of scientific affairs must surely diminish science itself. Above all, it is a threat to an essential characteristic of the endeavour — its openness to outside examination and appraisal. □

Donald P. Hayes is in the Department of Sociology, Cornell University, Ithaca, New York 14853, USA.