

Six senses in the literature

The bleak sensory landscape of biomedical texts

Raul Rodriguez-Esteban & Andrey Rzhetsky

“It is beyond our power to fathom,
Which way the word we utter resonates,
Thus, like a sudden grace that comes upon us,
A gift of empathetic understanding emanates.”

Fyodor Tyutchev (1803–1873), Russian poet
(Translated by Mikhail N. Epstein
at Emory University, Atlanta, GA, USA).

When we read prose—whether technical or literary—our mind parses sentences to recover their meaning. Yet, the flow of the words themselves can invoke surprising or unexpected sensory responses, even for the writer. Even a very rational and technical text can typically affect the reader on multiple cognitive levels, in addition to its basic task of transmitting the author-intended meaning.

Prose in particular can modify an unsuspecting reader’s physiological and emotional states profoundly

Prose in particular can modify an unsuspecting reader’s physiological and emotional states profoundly. The semantic priming test in modern psychology exploits this phenomenon—for example, people start to feel and behave as though they have suddenly grown older after they have read a scrambled sequence of words enriched with ageing-related connotations (Srull & Wyer, 1979). The priming effect is largely independent of our conscious understanding of a text: autistic children, whose text comprehension is mildly impaired, respond to semantic priming the same way as non-autistic children (Saldana & Frith, 2006). Furthermore, our emotional response to a sequence of words depends in part on our genetic background. Children of parents with bipolar disorder, for

example, have been shown to react much more vividly to words that have undertones of social threat than children in a control group (Gotlib *et al*, 2005). Semantic priming can significantly affect the model of the outside world reported by our senses; merely naming an odour ‘cheddar cheese’ or ‘body odour’ can determine our perception of it as being pleasant or nauseating (de Araujo *et al*, 2005).

The selection of words in a composition also reveals the personality traits of the author to the reader. A person’s colour preferences and idiosyncrasies provide information relevant to their psychological evaluation (Lüscher, 1969). For example, Lüscher’s test links blue colour preference to a person’s “depth of feeling”, that is, the person tends to be concentric, passive, incorporative, heteronomous, sensitive, perceptive and unifying. Similarly, the colour yellow indicates spontaneity; yellow-loving people tend to be eccentric, active, projective, autonomous, expansive, aspiring and investigatory. Lüscher also attached special meaning to a person’s choice for combinations of colours. Schizophrenia patients—who are especially susceptible to semantic priming—have a characteristic utterance pattern: the patients’ own words generate diverse secondary associations in their minds. These self-inflicted associations surface in the patients’ utterances and disturb the clarity of their messages (Maher *et al*, 2005).

Of the vast swathes of published literature, scientific texts have a reputation for being factual, rational and ‘dry’ in contrast to other prose that is designed to evoke emotional responses. In this study, we therefore analysed the frequencies of use of sensory words and time-related terms in a large collection of biomedical texts, and compared the results with similar analyses of a collection

of news articles, a large encyclopaedia, and a body of literary prose and poetry. We found that, unlike literary compositions and news-wire articles, biomedical texts are extremely sensory poor, yet rich in overall vocabulary. It is likely that the sensory-deprived writing style that dominates the biomedical literature impedes text comprehension and numbs the reader’s senses and mind.

Increasingly sophisticated text-mining algorithms have made the task of the large-scale analysis of scientific publications more effective and efficient over the past years. The basis of such tools is typically the computational analysis of scientific language (Harris, 1988, 1989), which in itself provides a unique opportunity to look into the ‘collective unconsciousness’ of the scientific community. We used these analytical tools to compare the relative frequencies of various sensory terms—such as those related to the perception of colour, smell, taste, touch, sound and time—in order to infer the ‘collective sensory landscape’ of the biomedical literature and the hypothetical priming that biomedical texts exert on their readers.

Semantic priming can significantly affect the model of the outside world reported by our senses...

Overall, we analysed a large collection of scientific texts, including almost 250,000 full-text articles from 78 biomedical journals (Fig 1; *Journals*). We compared the properties of these biomedical texts with those of news reports (*Reuters*), the open-access encyclopaedia Wikipedia (*Wiki*), and the complete collected works of Edgar

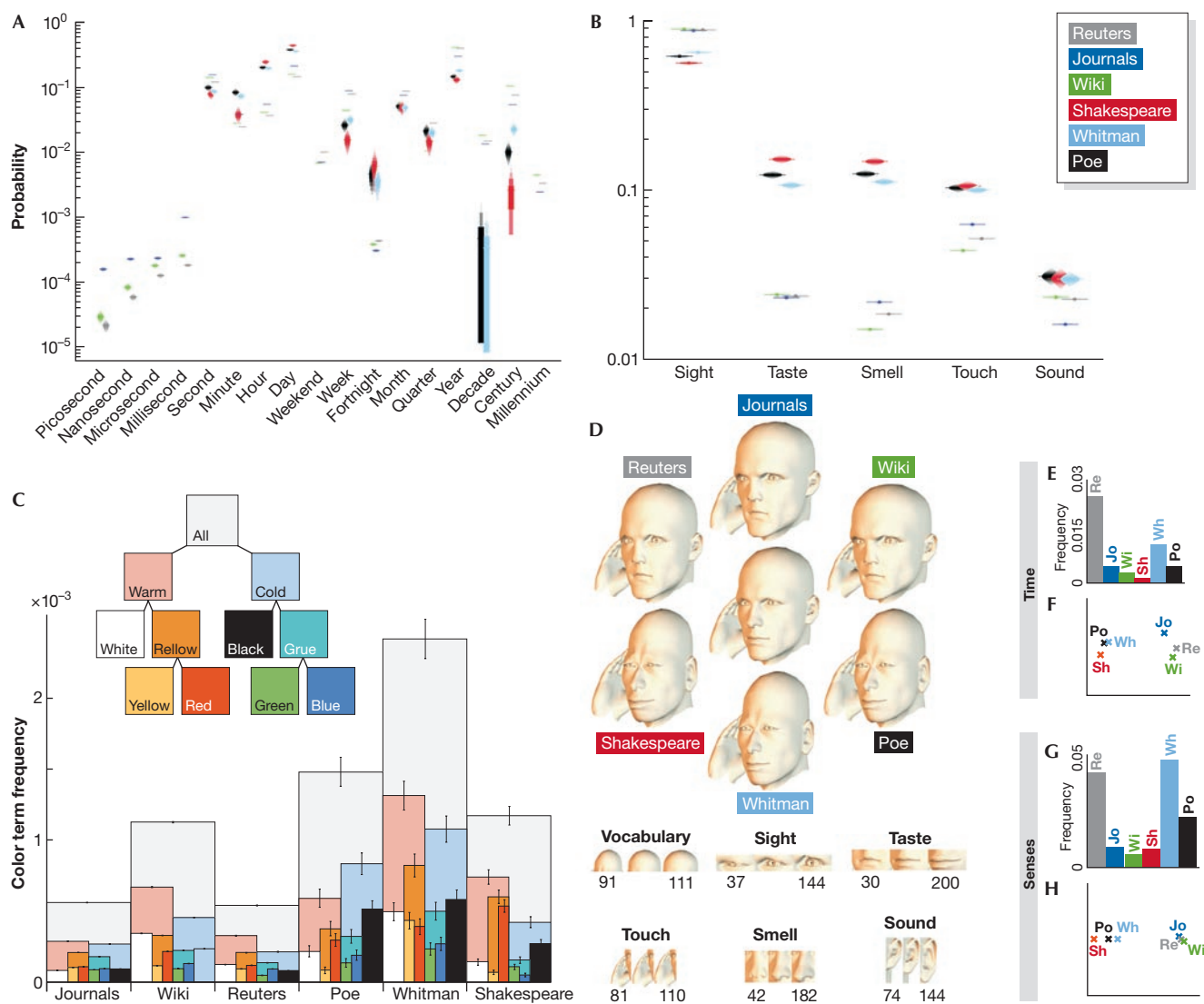


Fig 1 | Analysis of the frequencies of sensory words in six large corpora: *Journals*, *Wiki*, *Reuters*, *Shakespeare*, *Whitman* and *Poe*. (A) Frequencies of time-related terms. (B) Share of sensory terms divided into five sense groups. (C) Frequencies of colour-related terms grouped into a Berlin–Kay-like taxonomy (inset) computed for six large text corpora. (D) Balance of sensory terms in different corpora compared with the average (face in the centre). (E) Combined frequencies of time-related terms. (F) Multidimensional scaling of time-related word frequencies. (G) Combined frequencies of five-sense-related terms. (H) Multidimensional scaling of five-sense-related word share. *Journals* (GeneWays 6.0) is a collection of nearly 250,000 full-text articles from 78 research journals (Cokol *et al.*, 2005). *Wiki* (Wikipedia) is an open-access encyclopaedia that rivals Encyclopaedia Britannica in accuracy and far surpasses Britannica in breadth and coverage. *Reuters* (Reuters Newswire 2000) is a corpus of news articles in multiple languages (we used only the articles in English). *Shakespeare* (William Shakespeare, 1564–1616) is a complete collection of the works of probably the best known English playwright and poet. *Whitman* (Walt Whitman, 1819–1892) is a corpus of compositions of probably the most famous American poet. *Poe* (Edgar Allan Poe, 1809–1849) is a complete collection of works by the prolific and influential writer and poet, whose life was short and tragic.

Allan Poe (1809–1849, *Poe*), William Shakespeare (1564–1616, *Shakespeare*) and Walt Whitman (1819–1892, *Whitman*). We grouped these corpora into collective works (*Journals*, *Reuters* and *Wiki*) and individual works (*Poe*, *Shakespeare* and *Whitman*).

When describing ‘time’ (Fig 1A), all six corpora most frequently mention ‘days’

and ‘years’. In individual corpora, ‘days’ predominate over all other time terms, followed by ‘hours’ and ‘years’. Collective corpora most often mention ‘years’, followed by ‘days’ and ‘seconds’. Whereas individual corpora remain exclusively within the second-to-century range, collective corpora reach into picoseconds on

the short-term timescale, and into millennia—and even millions of years—on the long-term timescale. Within individual corpora, *Whitman* was the most concerned with centuries, and *Shakespeare* the least. *Reuters* is almost twice as obsessed with time as *Whitman*; all the other corpora are several-fold poorer in time-conscious

words (Fig 1E). *Journals* are among the poorest in time-related terms, although *Wiki* and *Shakespeare* are even poorer.

It is likely that the sensory-deprived writing style that dominates the biomedical literature impedes text comprehension and numbs the reader's senses and mind

When we consider words related to the five basic human senses (Fig 1B,G), sight-related terms are most frequent in all six corpora (Fig 1B). The collective prose is significantly more visual than the individual, but the trend is reversed for taste-, smell-, touch- and sound-related terms. Among the individual corpora, *Shakespeare* is the least visual, but the richest in taste-, smell- and touch-related terms, when sensory word frequencies are normalized to one within each corpus. However, if we look at the absolute frequencies of sensory terms, the differences between corpora are staggering (Fig 1G): *Whitman* is the richest in sensory terms, closely followed by *Reuters*. *Poe*, the next in ranking, reaches barely half the frequency of sensory-related terms found in *Reuters* and *Whitman*. Compared with *Whitman* and *Reuters*, sensory terms are almost absent from *Journals*, *Shakespeare* and *Wiki*, with *Wiki* being the most sensory deprived. The balance between different sensory terms—combined with the overall dictionary size—is visually highlighted in Fig 1D: individual corpora have, understandably, more limited vocabularies and are significantly richer in non-visual sensory terms, but poorer in visual sensory terms, than collective corpora. Note that these differences in vocabulary richness, as well as all other properties in Fig 1D, are logarithmic rather than linear in scale.

To highlight the similarities and dissimilarities in the frequencies of numerous sensory terms among the six corpora, we used a multidimensional scaling technique (Cox & Cox, 2001; Fig 1F,H). Multidimensional scaling in its 'true' form involves reconstructing a geographical map from a set of known distances between cities. In our case, we used it to try to arrange points corresponding to our six corpora on a plane, so that the resulting distances are as close as possible to the Euclidean distances between

corpus-specific vectors of frequencies of sensory terms: the closer two corpora are on the map, the more similar they are in terms of the frequency of cognitive terms. Both in the case of the time-related (Fig 1F) and the five-sense-related (Fig 1H) terms, the collective and individual corpora form two distinct groups. *Shakespeare* appears to be an outlier in both cases, whereas *Poe* and *Whitman* are rather similar among the individual corpora. Among the collective corpora, *Journals* and *Wiki* are the most dissimilar, with *Reuters* occupying an intermediate position.

There are surprising and significant differences in the usage of colour-related terms among our six corpora—as Shakespeare put it: “our wits are so diversely coloured” (*The Tragedy of Coriolanus*, Shakespeare, 1623). We grouped colour terms according to the taxonomy proposed by the anthropologists Brent Berlin and Paul Kay (Berlin & Kay, 1969; Kay & Berlin, 1997), which describes a hypothetical historical origin and diversification of colour terms summarized over 19 distinct cultures (inset in Fig 1C). The authors suggest that colour description is rather universal across cultures, owing to the universality of the anatomy and physiology of human vision. Very briefly, according to their theory, as language develops in a typical culture, the description of colour goes through several stages of complexity. The first stage involves just two colour terms, such as ‘warm’ and ‘cool’, followed by the isolation of pure white and pure black. At the later stages of colour term differentiation, so-called ‘yellow’ colour splits into red and yellow, while so-called ‘grue’ splits into green and blue.

We suggest that apparently cognitively bleak biomedical texts can and should be transformed into perceptually richer prose

Journals and *Reuters* are almost tied in the contest for the title of the visually bleakest corpus. *Reuters*, the bleakest corpus colour-wise, but not in all sensory terms, is significantly biased towards warm colours, while in *Journals* the frequencies of various colour-related terms are nearly uniform.

In all corpora but *Poe*, warm colours dominate over cold colours. In *Poe*, not only do the cold colours prevail, but also black dominates over all other colours at an extremely high level of significance. *Poe*'s prose and

poetry is literally dark. In Shakespeare's writing the significantly dominant colour is red—Shakespeare's prose is probably tinted by action in which blood is spilled.

Unlike *Poe* and *Shakespeare*, *Whitman* produced texts with colour term frequencies almost perfectly evenly distributed among the six major categories: white, black, red, yellow, green and blue. This is particularly curious in the light of the observation that *Whitman*'s writing overall is twice as rich in colour terms as that of *Poe* and *Shakespeare* and almost five times as rich as *Journals*.

What then is the likely priming effect of biomedical texts on readers? Our conjecture—which can be tested rigorously by experimental psychologists—is that reading such a text is similar to the effect of a long journey through a colourless flat terrain devoid of prominent features: a numbing of the senses. We suggest that apparently cognitively bleak biomedical texts can and should be transformed into perceptually richer prose, though we are not implying that it is an easy task. It is, nevertheless, important, because the mapping of abstract concepts onto objects with meaningful sensory properties acts as a stepping-stone to the solution of complex problems. Consider the following quote from Richard Feynman: “I had a scheme, which I still use today when somebody is explaining something that I'm trying to understand: I keep making up examples. For instance, the mathematicians would come in with a terrific theorem, and they're all excited. As they're telling me the conditions of the theorem, I construct something which fits all the conditions. You know, you have a set (one ball)—disjoint (two balls). Then the ball turns colours, grows hairs, or whatever, in my head as they put more conditions on. Finally they state the theorem, which is some dumb thing about the ball which isn't true for my hairy green ball thing, so I say 'False!' If it's true, they get all excited, and I let them go on for a while. Then I point out my counterexample. 'Oh, I forgot to tell you that it's my Class 2 Hausdorff homomorphic.' 'Well, then,' I say, 'It's trivial! It's trivial!' By that time I know which way it goes, even though I don't know what Hausdorff homomorphic means” (Feynman *et al*, 1985; Dennett, 1991).

The human brain was shaped by evolutionary adaptations, each of which was invoked by an acute necessity to ensure our survival in a constantly changing

environment. Our neural system is therefore an eclectic ensemble of disparate pieces of hardware, which are perfected for solving specialized problems, such as the detection of potentially threatening bilateral vertical symmetry—a lurking predator—in a chaotic environment, or the prompt recognition of the faces of the numerous members of our own tribe. To make more efficient use of our neural machinery, we need to translate abstract problems into concrete sensory-grounded symbols that can be efficiently processed by our brains. This is like trying to perform a general computation using graphics-oriented hardware: to make the computation efficient, we have to translate our task into spatial translations of three-dimensional primitives.

...we need to translate abstract problems into concrete sensory-grounded symbols that can be efficiently processed by our brains

When we read and compose sensory-deprived prose, we probably leave a large part of our nervous system uninvolved—different words and meanings are processed by distinct brain areas (Pulvermüller, 2001). We conjecture that a piece of sensory-poor prose does, on average, a poorer job of engaging the reader's imagination than a sensory-rich one, although the former can be much more precise and concise than the latter. Within a narrow scientific subfield, an expert would undoubtedly prefer to read a concise technical text rather than a longer one replete with metaphors and analogies. However, the situation is different for a scientist trying to read a paper

from a neighbouring subfield: a dry technical description might require a prohibitive amount of a non-expert's time to read and grasp its content. It is in the writer's best interest to ensure that his or her work is as widely accessible as possible.

In short, we believe that scientific prose should be enriched with sensory words, provided that they clarify the meaning rather than obscure it, in much the same way as a good statistical data visualization involves the mapping of abstract data into colours and three-dimensional shapes, to help the reader or viewer discover meaningful patterns.

ACKNOWLEDGEMENTS

We thank Murat Çokol for suggesting the design of Fig 1C, Marc Hadfield for programming support, and Murat Çokol, Ivan Iossifov and Rita Rzhetsky for comments on the earlier version of the manuscript. This work was supported by the National Institutes of Health (GM61372 to A.R.).

REFERENCES

- Berlin B, Kay P (1969) *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA, USA: University of California Press
- Çokol M, Iossifov I, Weinreb C, Rzhetsky A (2005) Emergent behavior of growing knowledge about molecular interactions. *Nat Biotechnol* **23**: 1243–1247
- Cox TF, Cox MAA (2001) *Multidimensional Scaling* 2nd edn. Boca Raton, FL, USA: Chapman & Hall/CRC
- de Araujo IE, Rolls ET, Velazco MI, Margot C, Cayeux I (2005) Cognitive modulation of olfactory processing. *Neuron* **46**: 671–679
- Dennett DC (1991) *Consciousness Explained*. Boston, MA, USA: Little, Brown & Co.
- Feynman RP, Leighton R, Hutchings E (1985) "Surely You're Joking, Mr. Feynman!": *Adventures of a Curious Character*, pp. 85–86. New York, NY, USA: W.W. Norton
- Gotlib IH, Traill SK, Montoya RL, Joormann J, Chang K (2005) Attention and memory biases in the offspring of parents with bipolar disorder: indications from a pilot study. *J Child Psychol Psychiatry* **46**: 84–93

- Lüscher M (1969) *The Lüscher Color Test*. New York, NY, USA: Random House
- Harris ZS (1988) *Language and Information*. New York, NY, USA: Columbia University Press
- Harris ZS (1989) *The Form of Information in Science: Analysis of an Immunology Sublanguage*. Dordrecht, The Netherlands: Kluwer
- Kay P, Berlin B (1997) There are non-trivial constraints on color categorization. *Behav Brain Sci* **20**: 196–202
- Maher BA, Manschreck TC, Linnet J, Candela S (2005) Quantitative assessment of the frequency of normal associations in the utterances of schizophrenia patients and healthy controls. *Schizophr Res* **78**: 219–224
- Pulvermüller F (2001) Brain reflections of words and their meaning. *Trends Cogn Sci* **5**: 517–524
- Saldana D, Frith U (2006) Do readers with autism make bridging inferences from world knowledge? *J Exp Child Psychol* **96**: 310–319
- Srull TK, Wyer RS (1979) The role of category accessibility in the interpretation of information about persons: some determinants and implications. *J Pers Soc Psychol* **37**: 1660–1672



Raul Rodriguez-Esteban is at the Department of Electrical Engineering, Columbia University, New York, NY, USA. Andrey Rzhetsky is at the Departments of Medicine and Human Genetics, Institute for Genomics & Systems Biology, and Computation Institute, University of Chicago, Chicago, IL, USA.
E-mail: arzhetsky@uchicago.edu

doi:10.1038/embor.2008.15