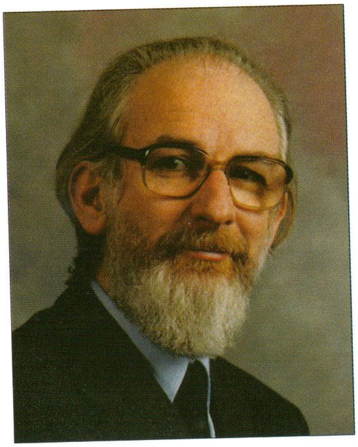
David Crystal: Research Profile

Honorary Fellow



One theme has dominated my research over the past twenty years: the evolution of language in electronically-mediated communication (EMC) – a term which includes the many domains encountered through personal computers (such as the Web, chat rooms, and email), the use of spoken and written language on mobile phones, and the linguistic content of communication devices such as satnav. It is a field characterized by rapid change. I wrote an initial account of it in *Language and the Internet*, which appeared in 2001. A mere four years later the book needed significant revision, for it made no mention of instant messaging and blogging – two developments which were virtually unknown in 2001 but which had become fast-growing areas of internet activity by 2003. The second edition of my book came out in 2006. Already it needs significant revision, for it makes no mention of such interactive domains as YouTube, MySpace, and FaceBook, which again were virtually unknown in 2005. Text messaging provides another illustration. It seems to have been with us forever, and yet for almost all users it is less than ten years old.

EMC presents linguistics researchers with some unusual problems. Getting hold of the data, for a start. It proves to be extremely difficult to build a corpus of emails, chat room conversations, or text messages. People are remarkably reluctant to share their e-exchanges. Would you let me see yours? And even when people do agree to provide messages, a certain amount of sanitization takes place. People send me only what they want me to see. Knowing I am a linguist, someone once told me 'Yes, I'll send you some, but I'm cleaning up my grammar first!' – thereby, of course, misunderstanding what linguistics is all about.

The lack of uncontaminated data is one of the reasons why linguistic research in EMC has been slow to develop – and why so many urban myths abound about its character. For example, virtually everyone believes that text-messaging is full of novel abbreviations (such as *C U L8r*): in fact, typically less than ten percent of the words used in texts are abbreviated in this way, and almost all the common abbreviations can be traced back to a period long before mobile phones were invented (code-puzzles such as *Y Y U R, Y Y U B*... were popular in Victorian England). A related myth is that texting harms children's language growth – something that research studies are now demonstrating to be false. On the contrary, the more children text, the better their literacy scores.

This is the descriptive and experimental side to research into EMC: establishing the linguistic facts. How is language actually used? How much variation is there? How fast does e-language change? Are there differences of age, social background, gender ...? One tiny observation to illustrate: women texters use far more exclamation marks than men. A small point, which by itself is of little significance, but when seen in association with other points of gender difference allows us to make some interesting deductions about how people vary the emotional content of texts and what the functions of text-messaging are.

The other side of EMC research is applied in character. Three problems illustrate the need.

- You type the word *depression* into Google, wanting results in economics, and you are annoyed to get thousands of hits from psychiatry.
- You type *mobile phones* into an online retail site, and the site says it has no mobile phones (but you know it must have them).
- A news report about a street stabbing has ads down the side of the screen which say, appallingly, 'Buy your knives here'.

The research goals are clear. Search engines need assistance to improve the relevance of results (by devising lexical filters which exclude pages irrelevant to your search interest). E-commerce needs to improve the accuracy of online enquiries (by anticipating all variables – in the above case, only the search-term *cellular phone* was being accepted by the software). And advertising agencies need to improve the appropriateness of ad placement on web pages by not relying on oversimple word frequency counts (which highlighted only *knife/knives*, in the above example).

These solutions depend on a single methodology. The task is to anticipate the words that users employ when interacting with websites. Which words will you be likely to use when talking about *depression* in the meteorological sense? Which, if it is the psychiatric sense? Which, if it is the economic sense? To ensure comprehensiveness, the initial research task was to work through an English dictionary, assigning all content words and their meanings to appropriate knowledge categories, and to build a device (which I call a *sense engine*) that would take web pages and classify them accurately. This has now been done, and the technology is being used initially in the advertising domain. Further applications include automatic document classification, to facilitate the retrieval of information in large electronic databases; and internet security, to monitor sensitive or dangerous online content. It is a long-term programme, for it needs to be applied to all languages which have a significant Web presence. So far a translation/localization has begun for just four languages. Internet linguistics will keep a lot of linguists happily employed for quite some time.