

Johnson and the Internet

David Crystal

The Hilda Hulme Lecture, 21 April 2005

A linguist cannot help but be impressed by the Internet. It is an extraordinarily diverse medium, holding a mirror up to many sides of our linguistic nature. The World Wide Web, in particular, offers a home to virtually all the styles which have so far developed in the written language, and some of the spoken ones too – newspapers, scientific reports, bulletins, novels, poems, prayers – you name it, you'll find a page on it.

The Internet is not of course a single thing, but consists of several domains which use the technology — e-mails, the World Wide Web, real-time (or synchronous) chatrooms, asynchronous chatrooms where messages can be left for later reading, the world of fantasy games (of the 'dungeons and dragons' type), and, most recently, instant messaging. Each offers us novel possibilities of human communication which cumulatively I think can genuinely be called revolutionary.

I do not use the word 'revolutionary' lightly.¹¹ I think it is appropriate here because computer-mediated communication has allowed the evolution of a new medium of communication — a medium with different communicative properties from what was available before. The first medium, evolving some 50,000 years ago, was speech. The second, some 10,000 years ago, was writing. The third, the concept-based sign language of the deaf, is of uncertain origin, but we know of its systematic organization in the seventeenth century. And there has been no new medium until the emergence, in the 1970s, of the first electronic messages, which after 20 years of development led to the introduction of the World Wide Web in 1991 and the routine use of email from the mid-90s. We have to remember that, for most people, Internet communication, so familiar today, is still only around a decade old.

Internet communication — more precisely known as Computer-Mediated Communication (CMC) and informally as Netspeak² — has been called a technological revolution and a social revolution. Additionally, I call it a linguistic revolution. I do so because I believe that its properties, as a medium of communication, are unlike those found in traditional speech or writing. Indeed, we do not know quite what to call it. Is an

¹ For the development of this theme, see my *The Linguistic Revolution* (Cambridge: Polity Press, 2004).

² See my *Language and the Internet* (Cambridge: Cambridge University Press, 2001) and *A Glossary of Netspeak and Textspeak* (Edinburgh: Edinburgh University Press, 2004).

email exchange a conversation? People say they are 'talking' to each other when they are in a 'chatroom', despite the fact that they are typing. Even the Simpsons are unclear:³

Homer: What's an e-mail?

Lenny: It's a computer thing, like, er, an electric letter.

Carl: Or a quiet phone call.

Their confusion arises because CMC is not like speech nor is it like writing. It is something new.

Why is CMC not like speech? Primarily, because it lacks the fundamental property of conversation, without which a successful spoken interaction cannot take place: simultaneous feedback. When A is talking to B, B is not listening passively. There is a continuous stream of signals — some visual, such as head nods, some vocal, such as *mhm* — which let A know how the conversation is going. To withhold this feedback leads to an immediate breakdown in the conversation, with A unable to continue. Try withholding such signals in conversation (advisedly, with a friend) and see what happens!

The contrast with CMC should be obvious. There is no simultaneous feedback. If I send an email to you, you cannot give me feedback while I am writing it, because you do not know I am sending it. Only after it arrives on your screen can you react, and by then I have finished it. So computer messages are autonomous in a way that everyday conversation is not. In this respect they resemble the autonomy of most uses of the written language, where the messages have to be meaningful in the presence of the reader while tolerating the absence of the writer.

Computer users are only gradually beginning to realise the consequences of this autonomy. At the beginning of the CMC era there was a natural tendency to capitalise on the properties of the medium, which promoted informality of expression and the relaxation of the rules of standard written English. The first emails were notable for their erratic or missing capitalization and punctuation, for the presence of spelling errors arising out of fast or inadequate typing, and for the use of new symbols — chiefly, the smileys, or emoticons which expressed basic pragmatic notions, such as 'I'm joking' :) or 'I'm unhappy' :(. These features quickly became a characteristic of 'cool' computer communication, and they actually caused fewer problems of intelligibility than educationists feared — a language can survive very well without capital letters and punctuation, as the early history of English illustrates. But autonomy raises other issues of communicative efficiency.

³ Episode 12A6 of *The Simpsons* (Fox TV).

There is a series of stages through which all new email users pass. People initially treat the medium as if it is conversation, and 'write as they speak', with informal syntax and punctuation; but without simultaneous feedback such messages can become ambiguous or unclear, or positively misleading. Users then encounter the consequences of unclear messaging: the 'what did you mean?' response, which requires them to rethink and resend their message, or worse, the phenomenon of 'flaming', where the recipient perceives an insult in a message and reacts angrily to it. Slowly, email and chatroom users realise that their messages are not like speech but need to be like writing — in other words, autonomous. They begin to take more care over their messaging, and develop new strategies, such as reading a message through before sending it, making more use of punctuation, and relying less on emoticons as a strategy for solving semantic inexplicitness.

There was in fact nothing revolutionary about such effects as typing inaccuracy, misspellings, and inconsistent capitalization and punctuation. These are rather minor effects, patently a special style arising out of the technological pressures operating on users of the medium, plus a natural desire (especially among younger — or younger-minded — users) to be idiosyncratic and daring. And that is how it is perceived. If I receive an e-mail from Professor Smith in which he mis-spells a word, I do not conclude from this that 'Smith can't spell'. I simply conclude that he was in a hurry. I know this because I do the same thing myself, when I am in a hurry. There is nothing truly revolutionary here.

What is revolutionary about e-mails is the way the medium permits what is called *framing*. You receive a message which contains, say, three different points in a single paragraph. You can, if you want, reply to each of these points by taking the paragraph, splitting it up into three parts, and then responding to each part separately, so that the message you send back then looks a bit like a play dialogue. Your sender can then do the same thing to your responses, and when you get the message back, you see his or her replies to your replies. You can then send the lot on to someone else for further comments, and when it comes back, there are now three voices framed on the screen. And so it can go on — replies within replies within replies — and all unified within the same screen typography. There's never been anything like this in the history of human communication. In this respect, therefore, CMC is unlike traditional mediums of expression, and supports my claim that we are dealing with something revolutionary here.

A second example is what we encounter when we see real-time Internet exchanges, as seen in chatrooms. You see on your screen messages coming in from all over the world. If there are 30 people in the room, then you could be seeing 30 different messages, all making various contributions to the theme, but often clustering into half a dozen or more sub-conversations. It's like being in a cocktail party where there are other conversations going on all around you. In the party, of course, you can't pay attention to them. In a chatroom you can't avoid them. It has never been possible before to 'listen' to 30 people at once. Now you can. Moreover, you can respond to as many of them as your mental powers and typing speed permit. This too is a revolutionary state of affairs, as far as speech is concerned.

The lack of simultaneous feedback and the promotion of simultaneous conversations are two ways in which CMC is fundamentally unlike speech. Here are two ways in which it is fundamentally unlike writing. Probably the most important distinctive feature of CMC is its hypertextuality. The *hypertext link* is the fundamental functional unit of the World Wide Web and other information-presenting electronic domains. This is the functionality whereby it is possible to click on an element on a screen page and be sent to another part of the same page, to a different page on the same site, or to a completely different site. It is essentially non-linear; the links can go in any direction. Traditional writing, by contrast, is essentially linear and unidirectional. The nearest we get to the hypertext link in writing is such a feature as the footnote or the cross-reference; but these are optional features (it is perfectly possible to have written text with no footnotes or cross-references), whereas without hypertext links there would be no World Wide Web.

A second example of a feature which differentiates CMC from writing is its dynamicity, which contrasts with the permanence of traditional written expression. You open a book at page 6, close the book, then open it at page 6 again. You expect to see the same thing. You would be more than a little surprised if the page had changed in the interim. But this kind of impermanence is perfectly normal on the Web — where indeed you can see the page changing in front of your eyes. Words appear and disappear, in varying colours. Sentences slide onto the screen and off again. Letters dance around. Pop-up advertisements irritate you. The Web is truly part of a new, animated linguistic channel — more dynamic than traditional writing, and more permanent than traditional speech. It is neither speech nor writing. It is part of a new medium.

And it is a medium which preserves its history. Technological catastrophes aside, nothing gets lost. You may think that, when you press delete on your email system, your

old messages are gone for ever. Likewise, you may think that, if you change a text on your Website, the earlier text is gone for ever. And you would probably be wrong. It is almost certainly out there somewhere. Occasionally you hear of an Internet investigation into fraud or suchlike, and the police go into the email records of a company. They are all there, on the host computer. Or take the way people add new data to the Internet — a firm advertises its 2005 range of Bermuda shorts. It hides from you its 2004 range, its 2003 range, and so on; but these old pages are still there, on the host computer. Indeed, more often than not they are still there in public view, because sites often fail to update their pages. Most of you, I imagine, will have had the experience of typing a search term into Google and finding that many of the pages you receive are well out of date.

For the linguist, of course, such historical archiving, whether conscious or inadvertent, is a goldmine. Putting it into linguistic terms, the Internet in general, and the World Wide Web in particular, is the largest historical corpus there has ever been. We are still working out the best way of investigating it. But note that it is historical in two senses. It is, most obviously, providing an ongoing record of CMC communication in our own age, year by year, hour by hour even — primarily of writing and increasingly (with the advent of audio files and computer-mediated telephony) of speech. Ferdinand De Saussure would have been amazed to see his synchronic 'axis of simultaneities'⁴ spelled out in such an explicit way, with each text file date-stamped down to the level of the minute and second. Historical linguists have never had it so good — or, depending on your point of view, bad, for there is only one thing worse than too little data and that is too much data.

Not only is the Internet providing us with this detailed kind of ongoing record, it is also filling in the gaps in our linguistic knowledge of earlier (pre-computational) historical periods. This is a less obvious but very important point. There are now a number of projects around the world which are providing electronic text resources of works which previously would have been available only in specialist libraries.⁵ And this means that the procedural limitations or biases of past philological projects are slowly being overcome. One of the most famous is the way the researchers in the early days of the *Oxford English Dictionary* (OED) selected their authors. Shakespeare, obviously, was a candidate for thorough examination, as were certain other leading Elizabethan

⁴ Ferdinand de Saussure, *Course in General Linguistics* (1916), translated by Wade Baskin (New York: Philosophical Library, 1959).

⁵ For example, the Electronic Text Center at the University of Virginia.

dramatists. But if you were a minor Elizabethan dramatist, you would have no future as an exemplar of usage in English lexicography.

The result has been that, when people look for evidence of words coming into English, using the OED's first recorded usage as evidence, then Shakespeare is enormously over-represented (29,305 quotations, to be precise). For a recent book, I used the electronic edition of the OED to find all instances of these usages.⁶

Excluding 54 cases of malapropisms and nonsense words (e.g. *gratality*, *allicolby*) there are 2035 of them. These are said to be the words that Shakespeare invented, and when we look at such striking instances as *anthropophaginian* and *unshout*, we are probably right to assert his individual creativity. But it is always wise to treat grand totals with scepticism. Shakespeare is also the earliest recorded user of *clack-dish* (that beggars used) and *'sblood*, and this tells us nothing about his creativity. No-one would seriously suggest that he was the first to use the word *'sblood*.

Now, with the electronic availability of other texts from the period, we have evidence whether anyone had used the word before him. *Lonely* is an example. The *OED* gives it first to *Coriolanus*, when he tells his mother 'I go alone, / Like to a lonely dragon' (IV.i.30). But in *The Tragedie of Antonie*, a translation of Robert Garnier's *Antoine* by Mary Sidney, the Countess of Pembroke, we find 'By fields whereon the lonely Ghosts do treade'. This was first published in 1592, some 15 years before *Coriolanus* was written. Probably the Countess wasn't the first to use the word either. But whatever the Shakespeare total was before, after learning this fact it is now one less. How many such revisions we are likely to see I do not know, but I suspect it will be a very large number, which will probably reduce our impression of Shakespeare's lexical contribution to the English language by half. None of that is to belittle his linguistic achievement. Any of us would be delighted to introduce just one new word into the English language, let alone a few hundred.

In the last few minutes you will have noticed a sea-change in the terminology of this lecture. I have begun to leave behind such words as *email* and *emoticons* and *chatroom* and to use such words as *Shakespeare* and *corpus* and *lexicography*. We are, in short, moving steadfastly in the direction of Johnson. Now the reason I am doing this is quite simple. In this week, of all weeks, I have no alternative, if one has any respect at all for anniversaries. Last Friday, 15 April, saw the 250th anniversary of the publication of Johnson's *Dictionary*. And just as 23 April each year is traditionally the date on which the

⁶ The data are reported in *The Stories of English* (London: Penguin, 2004), Chapter 13.

publishers of the world compete to publish books on Shakespeare — notwithstanding the fact that the differences between the Julian and Gregorian calendar mean that what was 23 April in 1564 is equivalent to 3 May today — so 15 April 2005 has come to be a date which has motivated publishers to publish books on Johnson, and on his *Dictionary* in particular.⁷ One selection from the *Dictionary* has already appeared, at the beginning of the year.⁸ Another is to appear at the end. That one is mine, so you will understand why Johnson is on my mind.

Last year I was commissioned by Penguin to compile an anthology for their Penguin Classics⁹ — surprisingly, no edition of the *Dictionary* has ever been published in that very wide-ranging series. And there is a second reason: in the same month, I was invited by the Johnson Society to be their president this year, presumably also for anniversarial reasons. I accepted with curiosity, if not alacrity, having heard excellent reports of the Johnson anniversary celebrations in Lichfield each September, which apparently rival those at Stratford every April. And in the same month, I was invited to give this lecture, and to suggest a topic. Most of my research for the past year or so, as the first part of this lecture suggests, has been into the language of the huge Internet corpus, so I knew that would have to be my theme. On the other hand, knowing that my Johnson labours would be finishing this month — in a nice coincidence, I approved the cover copy for the anthology on April 15th — I knew his presence would still be with me. And indeed, each time I dive into the Internet corpus, I am reminded of Johnson's dictum: 'A large work is difficult because it is large' (Preface). I felt sure that there were interesting parallels between his task and mine which might usefully be explored. I chose my title, 'Johnson and the Internet', and hoped — in the naive way that title-choosers do, six months ahead of their lecture — that a bridge between the two topics would emerge in due course.

In the event, it was Hilda Hulme who provided my bridge. When I was an undergraduate at University College, in the early 1960s, Hilda Hulme taught me Shakespeare. Though I didn't know it at the time, she was writing her book *Explorations in Shakespeare's Language*, which was published in 1962, and I would like to think that my own fascination with the language of Shakespeare owes not a little to her tutorials.

⁷ See, for example, Henry Hitchings, *Dr Johnson's Dictionary: the Extraordinary Story of the Book that Defined the World* (London: John Murray).

⁸ Jack Lynch, *Samuel Johnson's Dictionary* (New York: Atlantic Books, 2005).

⁹ David Crystal, *Dr Johnson's Dictionary* (London: Penguin, 2005).

Certainly I have referred to it often since, and I did so again, when thinking about this talk. At the very beginning of her book, she writes:

Of the language of art ... two things, apparently contradictory, are plainly true: first, that there is no single way of responding to its meaning; what one finds depends on what one brings. And equally, what one finds is there already; the meaning is there in the language.¹⁰

She illustrates her point, a few lines later, by referring to Johnson, who was typical of his age in admiring Shakespeare — as J R Sutherland had put it — 'rather in spite of his language', so that Johnson praises only the 'ease and simplicity' of Shakespeare's dialogue, finding his 'ruggedness or difficulty' a fault, his conceits 'idle', and his equivocations 'contemptible'.

Two things struck me, as I reread that quotation. First, it could just as well apply to the Internet — substituting the word *Internet* for the word *art* — and second, it applied very clearly to my feelings about Johnson and the rather superficial way he has been treated by the media this year, where commentators have looked at the *Dictionary* and seen only what they expected to see. In fact, one of the reasons I took on the Johnson job was not because I relished the prospect of reading through the entire *Dictionary* from beginning to end — that is not something one normally does to dictionaries, except when one is writing them — but because I wanted to get behind some of the mythology which surrounds that work. Even if you have never studied Johnson, as such, you will probably have encountered some of his definitions, because they have entered most popular books of quotations. He is the most quoted figure after Shakespeare. So most people are aware that he defined lexicographers as harmless drudges, and that he was apparently rude about excisemen and the Scots. Just a fortnight ago, indeed, the *Independent* published a double-page spread celebrating the anniversary¹¹ — all praise to them for that — but the myths abound in it.

Take, for example, the view that Johnson's definitions were eccentric. This is what the newspaper article authors say: 'Though generally admired, Johnson's idiosyncratic definitions were criticised'. And they say, of his definition of *network*: 'One of today's most fashionable buzzwords famously confounded Johnson when he attempted a definition: "Anything reticulated or decussated, at equal distances, with insterstices between the intersections".' Let us look at this criticism in more detail.

¹⁰ Hilda M. Hulme, *Explorations in Shakespeare's Language* (London: Longmans, 1962).

¹¹ Christopher Hirst and Geneviève Roberts, 'The A-Z of Dr Johnson's Dictionary', *The Independent*, 31 March 2005, pp. 14-15.

Yes, there are a number of definitions which have achieved a certain degree of notoriety due to the personal opinions they express. Boswell was the first to point them out in his *Life* of Johnson. Characterizing them as instances of 'capricious and humorous indulgence', he lists *Tory*, *Whig*, *pension*, *oats*, *excise*, 'and a few more' — by which he means such entries as *lexicographer*, *patron*, *leader* (sense 4), *reformation* and *reformer*, *aleconner*, *palmistry*, and *stockjobber*. As a characteristic of Johnson's lexicography, their fame far exceeds their significance. Although there are judgemental nuances scattered throughout, in my view there are *less than twenty* really idiosyncratic definitions in the whole work — out of 42,773 entries (in the first edition) and 140,871 definitions. The most famous definition of all — *oats* defined as 'grain, which in England is generally given to horses, but in Scotland supports the people' — was almost certainly one of those in-jokes that lexicographers love to bury in their books. It would have been no more than a friendly dig at his amanuenses, five of whom, as Boswell points out, were from Scotland, and whose influence is reflected in dozens of allusions to Scottish English throughout the *Dictionary*. A similar sympathy pervades his famous definition of *lexicographer*. I have never met one of these individuals who did not delight in the characterization of their profession as 'harmless drudgery'.

We must not dismiss that characterisation of 'drudgery'. Every lexicographer knows what this is — the need to handle with precision the grammatical words of the language (such as *what*, *as*, *of*, *but*), the everyday words (*one*, *two*, *three*, *January*, *December*), the remarkable number of words beginning with such prefixes as *un-* and *self-*, or those 'light verbs' (as modern linguists call them) — verbs of 'vague and indeterminate' use, as Johnson puts it in his Preface — which play an important part in English idiom, such as *make* and *do*. In Johnson's case, the longest entry is for *take*, whose 134 uses (including phrasal verbs) take up 11 full columns of print; but a special mention should be made of the verbs *set* (88 uses), *put* (80), *stand* (69), *go*, and *run* (both 68). Such mammoth entries were unprecedented in English dictionaries, and they are remarkable in their attention to semantic nuance.

Of the two major dimensions in any dictionary — coverage (which items to include) and treatment (how to deal with them) — Johnson is in no doubt that treatment is the greater problem. As he says in his Plan, after talking about issues to do with selection and identification:

The great labour is yet to come, the labour of interpreting these words and phrases with brevity, fulness and perspicuity.

It was indeed a huge labour, and when we look at a sequence of Johnsonian definitions today, it is obvious how much thought must have gone into them. They are the dictionary's primary strength, and its chief claim to fame. Anyone can get a sense of the problem by trying to formulate for themselves appropriate definitions for such words as *effect*, *nature*, *relation*, and *sign*, and comparing their attempt with Johnson's entries. The plural, 'definitions', is important: most words in a language have more than one sense. Some, as we have seen, have dozens. Abstract words pose particular problems, but all words require definitions that are clear, succinct, well-sequenced, and contrastive (with words of related meaning), and Johnson's achievement can be seen on virtually any page. For clarity and succinctness, take *acquiescence*:

A silent appearance of content, distinguished on one side from avowed consent, on the other from opposition.

or *message*

An errand; any thing committed to another to be told to a third.

His definitions are often elegant (*history*: 'A narration of events and facts delivered with dignity'), thoughtful (such as his additional note to *sorrow*: 'Sorrow is not commonly understood as the effect of present evil, but of lost good'), and perceptive, such as his definition of *sorry*:

Grieved for something past. It is generally used of slight or casual miscarriages or vexations, but sometimes of greater things. It does not imply any long continuance of grief.

They can also be humorous, such as his cheeky alliteration in *heresiarch*:

A leader in heresy; the head of a herd of hereticks.

There are many illustrations of the care he takes to sequence his definitions in a semantically related way, and to provide a balance between definition and associated quotation (illustrative are *fierceness*, *flesh*, *knowledge*, *ring* (noun), *shade*, and *taste*). His concern to relate words to other words can be seen in his synonym lists, as at *careless*, *chafe*, and *flatter*. Most lexicographers would be satisfied with just two or three synonyms: Johnson's *careless*, for example, gives twelve. And the way in which he draws attention to contrasts in meaning can be seen in such entries as *tempest* (vs. *breeze*, *gale*, *gust*, *storm*), and *sore* (vs. *wound*, *tumour*, *bruise*) — a feature which is particularly noticeable in the second half of the *Dictionary*. 'It is necessary likewise to explain many words by their opposition to others; for contraries are best seen when they stand together', he commented in his Plan. In this respect he anticipates twentieth-century structural semantics.

Then take the supposed 'difficulty' of his definitions, in such cases as *network* above or *cough* ('A convulsion of the lungs, vellicated by some sharp serosity'). Here too their role has been exaggerated, for there are only a couple of dozen of them. But here, as in so many other ways, he anticipated his critics:

sometimes easier words are changed into harder, as *burial* into *sepulture* or *interment*, *drier* into *desiccative*, *dryness* into *siccity* or *aridity*, *fit* into *paroxysm*; for the easiest word, whatever it be, can never be translated into one more easy. But easiness and difficulty are merely relative, and if the present prevalence of our language should invite foreigners to this dictionary, many will be assisted by those words which now seem only to increase or produce obscurity. (Preface)

'... easiness and difficulty are merely relative.' To modern eyes, such definitions do often seem lexically abstruse, but they have to be seen in the context of the time, which was a period when 'hard words' were much more routine than today. There had already been several dictionaries of 'hard words', dating from Robert Cawdrey's in 1604. The definitions would have been challenging, but not obscure, to Johnson's contemporaries. And the frequency with which some of the hard words were used makes them more palatable, even to the modern reader: *reticulated* is one of several words in the dictionary beginning with *reti-*; *interstice* turns up in a number of entries (*dense, imporous, mesh, net*), both in definitions and quotations, and also has a entry of its own. We must not assume that the 18th-century sense of lexical difficulty is the same as ours today. This is the danger identified in my quotation from Hilda Hulme: 'what one finds depends on what one brings'.

The mythology about Johnson has had all the press attention, as it were, and hidden some of the properties of the dictionary which deserve much more widespread recognition and which, in several respects, anticipate the way in which the Internet corpus can be exploited. Let us remind ourselves, firstly, of that great moment when Johnson's mindset moved from purist to linguist. In his Plan, he had been unequivocal:

one great end of this undertaking is to fix the English language.

In his Preface he realises how absurd this notion had been:

Those who have been persuaded to think well of my design, require that it should fix our language, and put a stop to those alterations which time and chance have hitherto been suffered to make in it without opposition. With this consequence I will confess that I flattered myself for a while; but now begin to fear that I have indulged expectation which neither reason nor experience can justify.

One of the consequences of this change of mind can be seen throughout the *Dictionary*, in the detailed attention he pays to etymology and in his recognition of the importance of regional and social variation. The entries which contain information about regional dialects are often ignored, in accounts of Johnson's lexicography, but they are an important innovation. There are not many of them, but they fall into three main types: words from his home-town Lichfield and Staffordshire (*gnarled, goldfinch, moreland, orrery, shaw*), occasional observations about other English dialects (*amper, atter, haver, onset*), and above all usages from Scottish English (*mon, scamblor, sponk*), which are common enough to suggest that his amanuenses were being used for far more than their copy-writing skills.

Similarly, the dictionary contains a great deal of information about social and stylistic variation — observations about eighteenth-century usage or, at least, Johnson's opinion about contemporary usage. The stylistic range of the *Dictionary* is in fact very wide. At one extreme we find highly formal words of classical origin (*adumbrate, prognostication, sagacity*); at the other we find colloquial interjections (*ay, fob, hist, look, right, tush, tut*). The latter never attract the attention of the journalist. Nor do his inclusion of social locutions (*howd'ye*), terms of address (*servant*), and gender differences ('women's words', such as *frightfully* and *horrid*). At the same time, being part of the spirit of his age, he routinely draws attention to words he considers improper, using such terms as 'bad', 'low', 'vulgar', 'cant', 'barbarous', 'ludicrous', and 'corrupt' to describe such words as *alamode, budge, cajole, coax, desperate, nowise, plaguy*, and *sconce*. We can sense his concern to warn his readers about words which it might be dangerous to use in eighteenth-century 'polite' society. However, we should not exaggerate his attitudes: terms such as 'low' and 'vulgar' may have been intended to convey no more than the labels used by modern lexicographers, such as 'informal'.

Johnson gave regional, social, and stylistic variation a presence in his dictionary that had not been seen before. We should perhaps not be surprised. Johnson was in no two minds about it. In Chapter 20 of Boswell's *Life*, we find the following report. Johnson said: 'By collecting those [words] of your country, you will do a useful thing towards the history of the language.' He bade me also go on with collections which I was making upon the antiquities of Scotland. 'Make a large book — a folio.'

BOSWELL: But of what use will it be, sir?

JOHNSON: Never mind the use; do it.

And it is this which makes me feel that Johnson would have felt very much in sympathy with the linguistic dimension of the Internet, for the Internet is giving a home to variation in English — and to languages in general — in an unprecedentedly wide-ranging and detailed way. Let me look firstly at the way it is giving a new home to regional and social variation within languages. Think back a decade: if you wanted to find out about any regional dialect, or to hear examples of a regional accent, how would you have done it, apart from ringing up a local phonetician and asking for some articulation over the phone? Now you can find dozens of sites for all the major dialects, most of which allow you to download examples of the local speech. And by 'major dialects' here, I mean two things: major intranational dialects, such as Scots, Yorkshire, and Geordie; and major international dialects, such as Australian, Indian, and Singaporean English. We must not forget that, of the 1.5 billion users of English in the world, three-quarters are not native speakers, and the distinctive features of their emerging dialects is just as much a part of the 'English language mix' as are the older distinctions such as British vs American. All of course are now easily accessible via the Web, and colloquial forms of these dialects — including the mixed-language scenarios such as Singlish (which mixes English and Chinese) can be encountered in a multitude of chatrooms.

And secondly, the Internet is giving a home to all languages — and especially the minority and endangered languages of the world — in a way that was not possible before. This point sometimes surprises people, who think of the Internet as a predominantly English-language medium. Indeed it was exclusively English when it began, back in the 1970s, but with the Internet's globalization, the presence of other languages has steadily risen. By the mid-1990s, about 80% of the Net was in English — a figure derived from the first major study of language distribution on the Internet, carried out in 1997 by Babel, a joint initiative of the Internet Society and Alis Technologies. This showed English well ahead, but with several other languages entering the ring — notably German, Japanese, French, and Spanish.

Since then, the estimates for English have been steadily falling, and it will not be long before the Web (and the Internet as a whole) will be predominantly *non*-English, as communications infrastructure develops in Europe, Asia, Africa, and South America. A Global Reach survey just after the turn of the century estimated that people with Internet access in non-English-speaking countries increased between 1995 and 2000 from 7 million to 136 million. In 1998, there was another surprise: the number of newly created Web sites *not* in English passed the total for newly created sites that *were* in English. And

in 2003, a magic figure was passed, with less than 50% of Web hosts now operating in English.

Spend an hour hunting for languages on the World Wide Web and you'll find hundreds. In 2001 I spent a few days tracking down as many examples as I could find, for my book *Language and the Internet*. It's not difficult to find evidence of a Net presence for all the more frequently used languages in the world, and for a large number of minority languages too. My estimate is that about a quarter of the world's languages – that's about 1500 — have some sort of cyber existence now. And I am talking about language presence in a real sense. These aren't sites which only analyse or talk about languages, from the point of view of linguistics or some other academic subject. They're sites which allow us to see languages as they are. In many cases, the total Web presence, in terms of number of pages, is quite small. The crucial point is that the languages *are* out there, even if they're represented by only a sprinkling of sites.

The Internet is the ideal medium for minority and endangered languages. It is still a not sufficiently widely known fact that at least half the world's languages are so seriously endangered that they will disappear in the course of the present century; by the end of the next century that will probably fall to 80 percent. If you are a speaker or supporter of such a language – an aboriginal language, say, or one of the Celtic languages – you're keen to give the language some publicity, to draw its plight to the attention of the world. Previously, this was very difficult to do. It was hard to attract a newspaper article on the subject, and the cost of a newspaper advertisement was prohibitive. It was virtually impossible to get a radio or television programme devoted to it. But now, with Web pages and e-mail waiting to be used, you can get your message out in next to no time, in your own language – with a translation as well, if you want — and in front of a global audience whose potential size makes traditional media audiences look minuscule by comparison. Chat rooms, moreover, are a boon to speakers living in isolation from each other, as now there can be a virtual speech community to which they can belong.

So the Internet is linguistically interesting for a whole series of reasons, and all of them, I like to think, would have intrigued Johnson. I cite five points in particular.

- First, CMC has provided us with a new medium of communication, in addition to 'traditional' speech or writing. It will lead to a whole new domain of linguistic enquiry, what we can call *Internet linguistics*. And applications of this new subject will be many and various. Examples on which I have worked myself over the past couple of years include the development of a chatroom child protection procedure based on semantic filtering,

and improving the relevance and coherence of results in such activities as online document classification, search, contextual advertising, and e-commerce. These kinds of activity illustrate the putative domain of *applied Internet linguistics*. They are the tip of an iceberg of applications.

- Second, CMC has extended the stylistic range of the language, especially at the informal end of the written formality spectrum, where the spontaneous options available in emails, chatting, and the like have taken the written language in fresh directions. Whatever the 'most informal' writing was before — perhaps informal letter-writing between intimates — the language pales in comparison with the kind of informality we can read in chatrooms and instant messaging exchanges and in the abbreviations which are ubiquitous in mobile phone text-messaging. And, I might add, the cheeks pale too. I have often heard (what we might call) creative taboo language spoken on the streets, but I have never encountered it so extensively in written form in a public domain as one finds on the Internet.

- Third, CMC has speeded up the process of language change. Nobody knows quite how long it takes for a new coinage to spread around the world and to appear in dictionaries. I recall *OED* editor Robert Burchfield saying that one should always be prepared to add an earlier notional 25 years to the estimate of a first recorded usage date in the *OED*. I am sure this was right. But today a new coinage can be around the world in a matter of hours only, and usages can turn up in online dictionaries within days. The big dictionary projects have yet to work out how best to cope with this situation.

- Fourth, CMC is developing new genres which are permitting more and more variation. My chief example here is the creative energy which is going into blogging. Thanks to easy-access Web software, anyone may now write their Web log, or *blog*, on their topic of interest, whether this be a simple account of their daily life or a focus group on a subject of obscure interest. Blogging is the fastest growing area of Web activity: there are already over 3 billion blogs, and they are increasing at the rate of over a million a month. It is as yet a little-studied phenomenon, but it is already providing evidence of a new kind of diary writing, which was a genre that a few years ago was thought to be dying out as a literary domain. From a linguistic point of view, what we see in blogs is written language in its most 'naked' form — without the interference of proofreaders, copy-editors, sub-editors, and all the others who take our written expression and standardise it, often to the point of blandness and boredom. It is a kind of English which has not been seen since the Middle Ages, before the rise of standard English. I suspect it is the beginning of a

new stage in the evolution of the written language, and a new motivation for children's literacy. And, as they say, 'we ain't seen nothin' yet' — or perhaps I should say 'heard nothin' yet', for CMC as a medium is now beginning to have a spoken dimension, and the results of streaming different modalities is fostering fresh forms of expression, such as in interactive television.

- And lastly, the issue of endangered languages. The Internet arrived at the right time, for these languages. The World Wide Web was introduced in 1991, and the first widely used chatrooms became a reality a couple of years later. In 1992, at The Quebec Linguistics Congress, we learned for the first time of the crisis affecting the world's languages. These languages are doomed to extinction unless something happens to give them a new lease of life. That something could be the Internet.

And it is this last point, perhaps, which would have most appealed to Johnson. Anyone who is for multidialectism, as Johnson was, has to be for multilingualism too. And so he was. Recall his famous remark about endangered languages: 'I am always sorry when any language is lost because languages are the pedigree of nations'.¹² It is difficult to think of anyone who has contributed more to the institutionalisation of that pedigree, in the case of English. And it is difficult to think of any medium that could record the history of that pedigree more efficiently than the Internet.

¹² James Boswell, *Tour to the Hebrides*, Sept 1773.