5. Changing Language, Changing Technology

Michael Rundell, Dictionary and Corpus Consultant, Canterbury, UK

Abstract

New technology and mass communications are leading to new forms of English and to the rapid diffusion of new vocabulary. At the same time, computerization has made possible the collection of large databases of written and spoken English and the development of sophisticated software for automatic language analysis; this gives us some hope of keeping up with the pace of change in the language itself. On the one hand, therefore, the actual **content** of language courses is likely to undergo significant change (note for example the growing focus on phraseology and the lexical approach and the availability of teaching materials based on authentic **spoken** English), while new media and new approaches to teaching and learning are also beginning to emerge.

Introduction

One of the easiest ways to make a fool of oneself is to get involved in the business of making predictions, and with the pace of change as it is now this is even more hazardous than usual. If you look at predictions made just five or six years ago regarding the likely impact of new technology on language teaching (LT), you will find very little mention of CD-ROMs and no reference at all to the Internet - yet both technologies are now an important part of the LT (especially ELT) landscape. Indeed the CD-ROM is already such a "mature" product that it is now in the process of being replaced by something more up-to-date. Things are moving fast, as the following anecdote demonstrates. About three years ago, a spoof article in the *Guardian* newspaper on April Fools' Day described in great detail the Queen's new website: the humour of the piece derived from the supposedly incongruous juxtaposition of cutting-edge technology and Britain's hopelessly outmoded royal family. Yet within 18 months of this "report", Buckingham Palace did indeed have its own website up and running. It has become almost a cliché to talk about the rapidity of change, but reality constantly seems to outstrip our wildest forecasts.

What I propose to do here is, first, to look at some of the ways in which language itself - in this case, the English language - is changing under the impact of new technology; secondly, to review some of the technology that is helping us to keep up with these changes and provide a satisfactory, up-to-date description; and finally, to consider some of the implications - in terms of both resources and methodology - for the way people learn and teach English.

Changes in the language

It is well known that computer technology has spawned hundreds of neologisms, whether these be completely new words (such as **netiquette,** the rules of polite behaviour to be observed by users of the Internet) or - much more commonly - new uses of older words (such as **mouse, icon, surfing, virus,** and **unzip**). Less obviously, there is a trend whereby technical language that would formerly have been found only in rather specialized forms of discourse is now regularly used in quite mainstream text: words like **modem, baud, gigabyte** and **RAM** are by no means new but have now shifted in register to become (more or less) assimilated into the general language. (A similar trend can be observed in other contemporary "growth areas", such as the language of business and finance, or indeed the medical names for muscles: terms like **deltoids** and **triceps** are now routinely traded in conversation between people who have no medical background but spend a lot of time working out in the gym.)

Looking at the way that dictionaries themselves define terms like this throws an interesting light on this process of assimilation. Compare for example:

> **RAM** n [U] Random-Access Memory; a computer memory holding information that is needed by the computer for a limited period, and that can be searched for in any order one likes. *Longman Dictionary of Contemporary English* 1987 edition

> **RAM** n [U] Random-Access Memory; memory in a computer system, used as the main area in which current programs operate: *a PC with 8Mb of RAM. Longman Interactive Dictionary of American English 1997*

The earlier definition is not in any sense "wrong", but the focus on *explaining* the idea of random access ("can be searched in any order one likes") suggests that the writer is describing a concept that s/he does not fully understand. In the more up to

date definition, by contrast, the writer confidently selects the most relevant aspects of the word's meaning, and also throws in a typical example of its use, showing that RAM is typically measured in megabytes - though here again the current technology has overtaken the lexicographic record, because no computer on sale in 1998 would have as little as 8Mb of RAM.

Email: a new form of discourse?
The very rapid spread of email over the past two or three years is one of the more striking effects of new technology, and there are signs that a new form of discourse is emerging through this medium to challenge the old (and already rather shaky) speech vs writing dichotomy. The term electronic mail does not of course describe a stylistically uniform text-type, but there are nevertheless some recurrent characteristics and conventions that can now be regarded as very typical of this genre. Attitudes towards spelling and punctuation are typically fairly relaxed in email, and the grammar of email messages is often reminiscent of unedited, real-time communication, with the sort of false starts that are so characteristic of spoken text, for example:

> This idea of Netscape becoming, well, I don't exactly know what to call
> it - the main program for large organizations rather than a web browser
> for individual users - this was new to me...
> [from an email on the TESL-CA discussion list]

The register level in a given email is often fairly unstable, veering from sentence to sentence between, on the one hand, an informal conversational tone and, on the other, serious technical discussion with the high "lexical density" associated with academic discourse; with, thrown in for good measure, a number of abbreviations and other devices that are more or less specific to the medium. To quote an example from a corpus linguistics discussion group:

> Corpora clearly represent clustered samples of some sort and estimates
> of errors associated with estimates from those samples are presumably
> subject to the usual inflation factors. But what intrigues me more is the
> definition of the universe ... If the universe is defined as "all words in
> use in English at x time etc" then imho selecting corpi (?sp) is not
> analogous to selecting buckets from oceans. Just my $0.02, fwiw.

The buckets and oceans analogy invokes shared contextual knowledge (referring to an earlier email on this subject), while the expressions **imho** ("in my humble opinion") and **my $0.02, fwiw** ("my 2 cents' worth, for what it's worth") come from a growing list of abbreviated discourse markers (that also include **btw,** by the way, and **tia,** thanks in advance) which have become a standard feature of email, along with the well-known "emoticons" like =) which are used to indicate a smile.

Another point worth noting here is that a very high percentage of the communication carried on through special-subject, global online discussion groups is taking place in English. Many, perhaps most, of the participants in these exchanges are not, in the old-fashioned sense, "native-speakers" of English, but much of the time it is impossible to know and usually it is of no relevance. For various historical and economic reasons, English has become the automatic means of communication for, say, a Peruvian agronomist who wishes to exchange ideas with a colleague in Finland or Indonesia, and all this is contributing to the growth of what is variously described as "Offshore English" or "English as an International Language". Rather than seeing this as a manifestation of linguistic imperialism, one could equally interpret it as showing that the ownership of English is becoming more widely dispersed.

The ever-more-widespread diffusion of English is one clear consequence of the electronic revolution, and as learners of English become participants in, and consumers of, these new forms of communication, they will increasingly find themselves exposed to less carefully edited, perhaps less "correct" forms of English. The debate over the relative merits of "authentic" and specially-written, pedagogically motivated teaching material still continues (see the recent contributions from Carter 1998 and Cook 1998), but in the future, students may not have so much choice about what they encounter.

Before we move on to look at the technology available for monitoring developments in language, a few more salutary comments about the pace of change. As recently as 1993, the term **information superhighway** first saw the light of day and, as always, dictionary publishers scrambled to include it in their books and blazon it on their covers as proof of their absolute up-to-date-ness. Yet by 1996, the term had already become so hackneyed, and was so evidently misleading - as one commentator pointed out, the Internet in its present form would more accurately be described as an "information goat-track" (Eastment, 1996) - that anyone who drew attention to it as a "new word" would simply be proving how out of touch they were. An *Oxford Dictionary of New Words* (ODNW)

published in 1992 includes among its neologisms the term **daisy wheel printer,** a device now so antediluvian that most people cannot even remember what it looked like. At the same time, some expressions originally used in computing contexts have now become so well assimilated that they are used almost exclusively for more general purposes. **User-friendly,** the most obvious example, is rarely used these days to discuss software interfaces, but has evolved new uses in many other contexts, sometimes even being used to describe people. And the following corpus lines for the phrase **what you see is what you get** - defined in ODNW as "a slogan applied to computer systems in which what appears on the screen exactly mirrors the eventual output" - show that, although it has become almost redundant in the field of software, it has made a new life for itself as a way of describing people who are open and without affectation:

...make moral judgements.	**What you see is what you get.**	I am a defeated ...
She has few pretensions	**What you see is what you get:**	honest opinions,
...have to live with that.	**What you see is what you get**	with him. You either...
With her,	**what you see is what you get.**	Yet if ...

Keeping track of language change
If things are changing this fast, how can those of us who describe language keep up, and provide those who teach and learn the language with a reliable description of what is to be taught and learned? The same technology that is contributing to language change also supplies powerful tools to help us monitor these developments. The new discipline of "corpus linguistics" began in the early 1960s with the arrival of the first electronic corpora of English - that is, large collections of text taken from a variety of sources (such as newspapers, novels, and academic writing), stored in machine-readable form and capable of being automatically analyzed using special corpus-inquiry software. Early text corpora were relatively small, typically consisting of 1 million words of text. But from the 1980s corpora grew rapidly, in parallel with the increases in processing power and storage space of ordinary desktop computers. By the mid 1990s, corpora of both written and spoken text were typically measured in hundreds of millions of words, providing huge amounts of raw material for linguists and lexicographers, and this now forms the basis for most serious descriptions of the language. (For an overview of these developments, see Sinclair 1991, Rundell 1996, Aston & Burnard 1998.)

The most typical and familiar form of corpus inquiry is to ask the computer for a "concordance" of a given word or phrase: the software then generates a list of every instance of the specified form together with its immediate context. The example below shows a concordance, from the British National Corpus (BNC), for the string

"**taste** (verb, in all its forms) + **like** (preposition):

Dot wondered what melon	tasted	like. It looked so green a
it correctly. This Pepsi	tastes	like Brasso with Canderel
Brits expect chocolate to	taste	like Cadbury's Dairy Milk
know it today. Most of it	tasted	like "something between
and their soured milk	tasted	like a cross between yoghurt
poured them so that they	tasted	like a horse's kick
red cherry beer, Kriek	tasted	like alcoholic cough mixture
-based fluid that smells and	tastes	like a regular bar, but with
seeds and rind. Celeriac	tastes	like a sweet and nutty
no surprise that the mixture	tastes	like a very solid fudge.
days beyond recall when bread	tasted	like bread and lawns were
like purple cauliflower, yet	tastes	like broccoli. Cut into

The concordance is a highly efficient way of revealing regular features of language behaviour, and its use has revolutionized the process of dictionary-making in the last 15 years. Increasingly now, this method is supplemented by various types of statistical analysis, such as the "mutual information" (MI) test, a measure of significance which automatically identifies - and lists in order of importance - all those words that are most likely to occur in conjunction with another word, which makes it an invaluable way of charting collocational behaviour (see Barnbrook 1996 for more details). An MI test for words most likely to follow the verb **cause,** for example, shows that the ten most significant collocates - again, based on information in the BNC - are: **grievous, bodily, havoc, harm, disruption, uproar, obstruction, inconvenience, damage,** and **distress.** Now, English dictionaries usually define **cause** as something like "to make something happen". Yet the evidence here shows quite graphically that **cause** is overwhelmingly associated with bad things happening. On the basis of this, recent learner's dictionaries have been able to improve their account of this word - for example the *Longman*

Essential Activator (1997) now defines cause as "to make something happen, especially something bad", and the *Cambridge International Dictionary of English (1995)* does something very similar. An obvious enough point, one might think, yet one that was missed by English dictionaries for hundreds of years.

The value of such corpus-derived evidence depends to a high degree on the quality of the data being searched: a corpus can only give a reliable picture of the way a language is used if it is itself a "good" sample of that language. Precisely what constitutes a good corpus is a matter of ongoing debate, but most people in the field would agree that a general-language corpus should aim to include text taken from the widest possible variety of sources. A simple model for corpus design can be based on the following five main headings:

mode	medium	genre	variety	topic
spoken	book	novel	America E	football
written	newspaper	short story	British E	philosophy
	magazine	play	S African E	cooking
	email	acad. writing	Singapore E	linguistics
	unpublished	lecture	International E	cosmology
	ephemera	conversation	etc.	politics
	spoken: direct	seminar		romance
	spoken: radio	etc.		etc.
	etc.			

This is by no means an exhaustive list, but the large corpus collections used by major dictionary publishers will typically aim to include most of the text-types listed here (and many others besides). Once each text in a corpus is "tagged" to show its own specific attributes (for example, to show that a particular text is (1) written, (2) in the form of a book, (3) a piece of academic writing, etc.), the software can analyze and compare the linguistic features of written and spoken discourse, of specific genres or regional varieties, and so on.

Learner corpora - that is, collections of text produced by language students in their exams, class essays, and homework - can help us to identify the lexis and structures that are especially problematical for learners, and dictionaries and other teaching materials can then target these areas for special treatment. Longman's learner corpus, for example, provides strong evidence of problems around the use

of the verb **remind,** with many learners omitting the preposition from the pattern VERB+OBJECT+PREP (**of**), thus: * *That woman reminded me an old friend.* This can then form the basis for explicit warnings in the relevant dictionary entry. (For a more detailed discussion, see Gadsby & Gillard forthcoming.)

Much of what has been said here has focussed on improvements in dictionaries, and it is true that many other types of teaching material are still based more on writers' own intuitions than on empirical evidence of language in use. But this is beginning to change. Because of the technology now available, corpus building has ceased to be a major, capital-intensive enterprise, and is therefore no longer the preserve of large well-funded institutions. The practicalities of assembling a large collection of electronic text (at least, of *written* text) have become much less daunting, and it is now perfectly feasible for individual writers or researchers to produce their own corpora for their own specific purposes. Thus, alongside very large general-purpose text collections such as the BNC and Birmingham University's Bank of English, we now see hundreds of smaller, more narrowly-focussed corpora. At the same time, powerful software tools for interrogating a corpus are also now available at low cost for use on ordinary PCs (see for example Barlow and Scott in the bibliography), so the world of corpus linguistics is effectively opened up for a far wider range of users.

In addition to all this, we now have available the phenomenal riches of the Internet, which can be used in a variety of ways to monitor language use and language change. Suppose, for example, you are teaching (or learning about) the language of the stock market, and you come across the expression **dead cat bounce.** What you really need to know here is, first, how the phrase is used, and secondly whether it is a common usage or merely a whimsical, one-off coinage. A search on the Internet, using the AltaVista search engine, returns over 700 instances of **dead cat bounce** (suggesting that it is a frequent enough term to be worth looking at further), and provides useful information about its meaning and use, as in the following extract from an online stock market enthusiasts' group in the US:

> Spectral Diagnostics has had quite a slide, dropping steadily from around $30 in 1993 to its recent low of around $5... When a stock with this sort of performance gains a few points suddenly, those ever colorful traders call it a "dead cat bounce".

Every possible subject area, from the most serious to the most frivolous, is well catered for on the Intemet, so there is an enormous amount of evidence for language in use in a wide variety of contexts and registers.

Resources and methodologies

Clearly, then, technology has already had a major impact on *what* we teach - that is, on our capacity for producing a description of English that corresponds much more closely to the way the language is actually used. The impact of technology on how we teach, on the resources and methodologies used for enabling learners to exploit this knowledge, is perhaps less clear at present. As we have seen already, there is something on the Internet for everyone: for teachers, an endless source of authentic material from online newspapers, tourist offices, museums, football clubs and the like and for students, any amount of data for research and project work. In addition to these general materials, there are a growing number of internet sites devoted specifically to the teaching and learning of English, and offering discussion forums for both students and teachers, teaching materials, stories, quizzes, help with grammar or idioms, and much more. (For some pointers, see Appendices in Eastment 1996, and p144-150 of *The EFL Directory,* 1988, Summersdale Publishers.) Inexpensive corpus inquiry software of the type mentioned above can also be used by students themselves to develop language awareness (see Tribble & Jones 1997), while rapidly-improving voice-recognition software is now finding its way into interactive programs geared to testing and improving pronunciation. The use of email for pen-pal (or rather, "keypal") projects is already well-established, and standard presentation software can enable students to create professional-looking projects and present them to the class.

For many students, there is much here that is highly motivating. Perhaps there are elements of novelty and of instant gratification, but there is too the real possibility of students feeling more in control of their own learning agenda. Having worked with students using corpus data, I have found that (while some are admittedly overawed by it) the ones who enjoy it and benefit from it do so because of the sense of independence it gives them in formulating their own questions and then testing them against live data. In this area and many others, new technology can help to make possible a more "democratic" learning paradigm in which students do a lot of discovering for themselves, rather than having "knowledge" transmitted to them by the teacher.

The reality, of course, is not always as rosy as this. Slow data-transfer speeds and endless blind alleys are still a familiar experience to anyone who uses the Internet, and there is certainly a need, as Eastment points out (1996) for some sort of quality control (in terms both of content and presentation) to apply to ELT-specific websites. The new modes of teaching and learning that are emerging now may not appeal to everyone, but the one safe prediction we can make is that the pace of change both in language and in technology is unlikely to slacken in the coming years, and this is sure to affect the way that teachers teach and students learn.

Biodata
Michael Rundell - Professional lexicographer since 1980, working on a wide range of pedagogical English dictionaries. Managing Editor at Longman Dictionaries (1984-94), and editor of (among others) *Longman Language Activator* (1993) and *Longman Dictionary of Contempory English* (1987 and 1995 editions). Leading role in the design of lexicographic corpora (including Longman Lancaster Corpus and British National Corpus). Currently freelance dictionary consultant and occasional teacher-trainer/lecturer. Interests include the phraseology of spoken English and the value and use of corpora as a resource for English language teaching. **email: ae345@dial.pipex.com**

Bibliography
Aston, Guy & Lou Bumard *The BNC Handbook*
(Edinburgh: Edinburgh University Press 1998)

Barnbrook, Geoff *Language and Computers*
(Edinburgh: Edinburgh University Press 1996)

Carter, Ronald "Orders of Reality: CANCODE, Communication, and Culture",
ELT Journal 52/1 1998, 43-56

Cook, Guy "The Uses of Reality: a Reply to Ronald Carter", *ELTJournal* 52/1 1998, 57-63

Eastment, David *The Internet and ELT* (English 2000)
(Manchester: The British Council 1996)

Gadsby, Adam & Patrick Gillard "Using a Learner's Corpus in Compiling a Dictionary" in Sylviane Granger (Ed.) *Learner English on Computer*
(Harlow: Addison Wesley Longman 1998)

Rundell, Michael "Computer Corpora and their Impact on Lexicography and Language Teaching" in Christine Mullings et al, (Eds) *New Technologies for the Humanities*
(London: Bowker Saur 1996), 198-216

Sinclair, John *Corpus, Concordance, Collocation* (Oxford: OLTP 1991)

Tribble, Chris & Glyn Jones *Concordances in the Classroom* (2nd edition)
(Houston: Athelstan Press 1997)