The scope of Internet linguistics

David Crystal

Abstract
The paper outlines a frame of reference for a putative 'Internet Linguistics'. The Internet, as a medium of computer-mediated communication, needs to be characterized in terms of its formal character, which differs in fundamental respects from traditional conversational speech and from writing. A sociolinguistic perspective is also required, examining the way in which a language evolves new varieties, and especially increasing its expressive range at the informal end of the stylistic spectrum. There are immediate educational consequences, because children need guidance in handling and exploiting the Internet's linguistic potential. Of particular interest are the ways in which the technology is fostering new kinds of creativity through language, as illustrated by blogging. The communicative capabilities of the Internet can also be exploited in ways that present many social problems, suggesting that we are at the beginning of an era of Applied Internet Linguistics.

Introduction

The emergence of a new branch of an academic discipline does not take place very often, but the arrival of the Internet has had such an impact on language that I believe the time is right to recognise and explore the scope of a putative 'Internet linguistics'. I would define this as the synchronic analysis of language in all areas of Internet activity, including email, the various kinds of chatroom and games interaction, instant messaging, and Web pages, and including associated areas of computer-mediated communication (CMC), such as SMS messaging (texting). The speed of change in the past 15 years has been such that it is already possible to see a diachronic as well as a synchronic dimension to this subject - a historical Internet linguistics, studying language change - but like no other study of language change in linguistic history, for the Internet allows us to follow, like never before, the rate and reach of language change in vocabulary, grammar, spelling, and (increasingly) pronunciation. It is also possible to see the rapid evolution of a comparative Internet linguistics, as the medium becomes increasingly multilingual.

The scope of Internet linguistics is evidently very wide, and all I can do in the time available is attempt to identify major areas which I think require investigation. I will focus only on the synchronic issues, and illustrate[j] only from English. I will look briefly at three points to do with the medium: its formal character, its use, and its exploitation.

The formal character of the medium

The Internet has permitted language to evolve a new medium of communication, different in fundamental respects from traditional conversational speech and from writing. The properties which differentiate CMC from speech include its lack of simultaneous feedback (critical to successful conversation), the absence of a nonsegmental phonology (or tone of voice, which emoticons attempted, but failed, to express), and its ability to carry on multiple interactions simultaneously (in chatrooms). The properties which differentiate CMC from writing include its dynamic dimension (through such effects as animation and page refreshing), its ability to frame messages (as in email cutting and pasting), and its hypertextuality (only hinted at in traditional writing through such notions as the footnote).

These features are, to my mind, much more important than the rather minor effect CMC is having on the 'surface' properties of languages, introducing novel features of grammar,

vocabulary, and spelling. Even in vocabulary, which is always the primary manifestation of language change, the number of new expressions is not very large. No systematic survey has been done, but we are talking about only a few hundred new words and idioms coming into the written language - a tiny number, compared with the thousands of words which come into English each year from all sources globally, and certainly tiny compared with the size of the English lexicon as a whole - well over a million lexemes. Moreover only a small fraction of this fraction has as yet made any kind of impact on general spoken usage. I collected as many as I could find for my little *Glossary of Textspeak and Netspeak,* last year, and ended up with only a few dozen spoken language examples, and in many of these cases it is unclear just how widespread the usages are. Words like *geek* and *glitch* have achieved a general status, but if I say to someone 'Mike's 404' - meaning he's not around - I wonder how many people would understand it?

Using the medium: the sociolinguistic perspective
The Internet has given language new stylistic varieties, in particular increasing a language's expressive range at the informal end of the spectrum. This is the area which has attracted most public attention, because people notice informality in language use and worry about it, thinking that it causes deterioration in a language. The prophets of doom emerge every time a new technology influences language, of course - they gathered when printing was introduced, in the 15th century, as well as when the telephone was introduced in the 19th, and when broadcasting came along in the 20th; and they gathered again when it was noticed that Internet writing broke several of the rules of formal standard English - in such areas as punctuation, capitalization, and spelling. All that has happened, in fact, is that the language's resources for the expression of informality in writing have hugely increased - something which has not been seen in English since the Middle Ages, and which was largely lost when Standard English came to be established in the 18th century. Rather than condemning it, therefore, we should be exulting in the fact that the Internet is allowing us to once more explore the power of the written language in a creative way.

Using the medium: the educational perspective
Of course, such change has to be managed. The Internet is altering our sense of responsibility towards language use, forcing people to rethink their role as communicators. If new informal varieties of language are emerging, then a fresh relationship between nonstandard English and Standard English is likely, and this has immediate educational implications. Children have to be taught about their language. They have to learn to read and write, how to improve their abilities to listen and speak, how to use language appropriately, and how to respond intelligently to other people's language use. They have to learn about the importance of Standard English, as a medium of educated communication. All this is in the hands of the teaching profession. And it is the teachers who have to take on responsibility for helping children to handle and exploit the linguistic potential of the Internet. In particular, it is up to the teachers to spot any signs of communicative inefficiency - such as children being stylistically inconsistent, or conflating features that belong to different styles. For example, if SMS abbreviations turn up in a formal school essay, then the inappropriateness should be pointed out. SMS abbreviations were designed to suit a medium where there is a technological limit on what can be communicated; they have no place in a medium where such limitations do not exist (allowing, as ever, for the controlled use of stylistic features in a literary context). This is simply another manifestation of the principle of stylistic appropriateness, which has governed intuitions in language teaching for centuries.

Using the medium: the stylistic perspective
Of particular interest are the ways in which the Internet and its associated technology is fostering new kinds of creativity through language, especially in literature. I have been very impressed with the expressive potential of the cellphone, for instance - not, obviously, in its daily communicative manifestations, where the creative possibilities are presumably quite limited in messages whose sole purpose is to inform someone that the train is late or that

everyone is meeting at Macdonalds - but in such cases as the *Guardian* text-messaging poetry competition, where within the 160-character limit some very effective pieces have been written. The formal constraint is no more artificial than, say, the syllable and line limitations of the haiku. The question is: what can be done, creatively, when imposing a discipline of this kind? And the answer seems to be: quite a lot.

Even more noticeable is the creative energy which is going into blogging, itself a multifaceted phenomenon, but already providing evidence of a new genre 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. It is the beginning of a new stage in the evolution of the written language, and a new motivation for child and adult literacy. It is exciting to be in at the outset of it. But it will be even more exciting to see what happens when the medium evolves a more routine spoken dimension, and when the results of streaming different modalities foster fresh forms of expression, such as in interactive television.

Exploiting the medium: the applied perspective
It is a commonplace these days to acknowledge the way the Internet can be exploited for good or evil, and this tension applies equally to its communicative capabilities. As an example of its linguistic exploitation for good I would cite the way it offers the world a first-hand encounter with multilingualism - an environment that will become increasingly diverse as more language communities come online. In particular, it offers minority and endangered languages a new opportunity for progress, in two respects. Firstly, it aids documentation, because digital techniques allow the easy recording of ethnological material - and, of prime importance, doing so using a technology which the teenage generation think of as 'cool'. The teenagers, after all, are the critical generation for any endangered language, as they are the parents of the next generation of speakers, and if their interest goes, the language is lost. Secondly, the Internet aids language revitalisation, by enabling speakers separated by space to maintain a virtual contact through email, chat, and instant messaging environments, and there are some very lively interactions out there now. For the moment, these are restricted to literate individuals, but it will not be long before the availability of telephone broadband communication, using such means as Skype, are put to use in the service of language revitalisation.

On the negative side, the exploitation of the Internet as a medium which can foster such activities as terrorism, fraud, and paedophilia carries with it corresponding concerns for security and protection. This is a domain which is full of potential, though it has received very little research from a forensic linguistic point of view. For example, I spent some time last year developing a chatroom child protection procedure based on semantic filtering, but found very little linguistically-oriented literature to help in the task.

We are at the beginning of an era of linguistically based online applications - an era which will see the emergence of an Applied Internet Linguistics. The above examples are illustrative, but a linguistic perspective is also turning out to be of value in improving the relevance and coherence of results in several online areas - in document classification, search, contextual advertising, and e-commerce. How quickly the field develops will depend on collaborations between linguists and major online players. The task is a large one, in terms of both time and money. Each of these activities requires a detailed investigation of the aims, methods, and expectations of the target domain, and requires a linguistics of the old-fashioned, time-consuming kind, involving a great deal of linguistic description and language analysis. For example, a lexical tool for improving the relevance of results in online search and classification took my team of lexicographers four years to develop, in the late 1990s, and an investment of some 4 million dollars. This is small beer, compared with the investments

that have taken place in Internet circles in recent years, but they are large sums compared with the kinds of research proposals typically encountered in academic linguistics. So I remain unsure just how much Applied Internet Linguistics we will see in the next year or so. But in the longer term, I am quite certain: we ain't seen nothin' yet.