Internet Linguistics

The Internet is now an integral part of contemporary life, and linguists are increasingly studying its influence on language. In this student-friendly guidebook, leading language authority Professor David Crystal follows on from his landmark bestseller *Language and the Internet* and presents the area as a new field: Internet linguistics.

In his engaging trademark style, Crystal addresses the online linguistic issues that affect us on a daily basis, incorporating real-life examples drawn from his own studies and personal involvement with Internet companies. He provides new linguistic analyses of Twitter, Internet security, and online advertising, explores the evolving multilingual character of the Internet, and offers illuminating observations about a wide range of online behaviour, from spam to exclamation marks.

Including many activities and suggestions for further research, this is the essential introduction to a critical new field for students of all levels of English language, linguistics and new media.

David Crystal is a freelance writer, lecturer and broadcaster, based in Holyhead, North Wales. He is author of numerous books including *Just a Phrase I’m Going Through* (Routledge 2009). The first Routledge David Crystal Lectures DVD, *The Future of Language*, was published in 2009.
‘Crystal draws on his wealth of expertise to shed light on the important issues related to language form and use online.’
   Mark Warschauer, University of California, Irvine, USA

‘David Crystal is a master linguist and master teacher. Given his expertise on language and the internet, he is the ideal author for this student text.’
   Naomi S. Baron, American University, USA

‘Crystal provides a unique overview of authentic applications for linguistics on the internet and the methodological issues raised in the case-studies will be relevant for a wide range of projects that readers may be working on. This will become essential reading for students in this area.’
   Charlotte Taylor, University of Portsmouth, UK
## Contents

**Preface**

1. Linguistic perspectives
   - Misconceptions
   - Terminological caution
   - Research challenges

2. The Internet as a medium
   - Speech vs writing
   - The Internet as a mixed medium
   - Differences with speech
   - Differences with writing
   - A new medium

3. A microexample: Twitter
   - Methodological issues
   - Content issues
   - Grammatical issues
   - Pragmatic issues
   - A variety in evolution
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Language change</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Orthography</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>Grammar</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Pragmatics</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Styles</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>A multilingual Internet</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Policy and technology</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Methodological issues</td>
<td>86</td>
</tr>
<tr>
<td>6</td>
<td>Applied Internet linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Problem areas</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>The focus on ambiguity</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>A lexicopedic approach</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>The centrality of semantics</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>An illustration</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>Other aspects</td>
<td>113</td>
</tr>
<tr>
<td>7</td>
<td>A forensic case study</td>
<td></td>
</tr>
<tr>
<td></td>
<td>An extract</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>A case study</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>Method</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>Results and discussion</td>
<td>127</td>
</tr>
<tr>
<td>8</td>
<td>Towards a theoretical Internet linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relevance and indexing</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>New directions</td>
<td>148</td>
</tr>
<tr>
<td>9</td>
<td>Research directions and activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 Debating roles (Chapter 1)</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>2 Audio issues (Chapter 2)</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>3 Distinctive forms (Chapter 2)</td>
<td>152</td>
</tr>
<tr>
<td></td>
<td>4 Testing hypotheses (Chapters 2 and 3)</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>5 Punctuation (Chapter 4)</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td>6 Spam (Chapter 4)</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td>7 Online translation (Chapter 5)</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>8 Localization (Chapter 6)</td>
<td>158</td>
</tr>
</tbody>
</table>
9  Taxonomy (Chapter 6) 159
10  Semantic targeting (Chapters 6 and 7) 161

Notes 163
Further reading 171
Index 172
How does one write a student guide to a subject that does not exist – or, at least, does not yet exist in such a recognized form that it appears routinely as a course in university syllabuses or as a chapter in anthologies of linguistics? Inevitably, it will be something of a personal account, informed by the various Internet projects with which I have been involved. The situation reminds me of the 1980s, when pragmatics was evolving as a field of study, and the various published introductions differed widely in their subject-matter. Internet linguistics is at that inchoate stage now. I can easily imagine other introductions to the subject – written perhaps by someone with a background in computational linguistics – which would look very different from this one. My background is in descriptive linguistics, and it shows. But it is an appropriate background to have, for the one thing Internet language needs, more than anything else, is good descriptions.

A growing number of linguistics students, at undergraduate and postgraduate levels, are now beginning to study the subject, and I have written this book primarily for them. It will I hope also be of interest to those who are taking a language course as part of a degree in media or communication studies. I have assumed that
readers have completed an introductory course in linguistics, or at least read an introduction to the subject, and are familiar with the various domains that constitute the Internet, including the most recent developments. They will not find here an exposition of syntax or sociolinguistics, or of blogging or social networking. It is an account written for people who are comfortable with the basic tenets and methods of linguistics, well versed in Internet activities, and curious about the relationship between the two. It is also for those, within this population, who are fascinated by the way Internet language is evolving, and want to research it. I have therefore given as many pointers as I could to topics where research is needed. My aim is not just to inform but to inspire more linguists to work in this field, for, as will become apparent—and surprising as it may seem—the subject is urgently in need of them. In particular, I have illustrated my points almost entirely from English, and this limitation needs to be overcome if the conclusions are to be robust.

This book is very different from my Language and the Internet. The emphasis in that work was on the stylistic diversity of the medium, so there was a focus on the linguistic features which identify language varieties. In the present book, general issues of characterization and methodology take centre stage. The descriptive chapter on Twitter would not have been out of place in the earlier book, but in other respects Internet linguistics tries to live up to its title and provide a wider perspective which Language and the Internet lacked. A certain amount of overlap has been inevitable, but I hope it is not intrusive.

My thanks are due to those who reviewed this text on behalf of the publisher, and also to Sacha Carton, Ian Saunders, and others in the companies (AND, Crystal Semantics, Adpepper Media) with whom I have had the opportunity to develop the approaches described in Chapter 6. Above all, I owe an enormous debt of gratitude to my wife and business partner Hilary, who has shared my close encounter with the Internet, professionally and privately, over the past 20 years.

David Crystal

July 2010
LINGUISTIC PERSPECTIVES

Wherever we find language, we find linguists. That is what linguists are for: to seek out, describe, and analyse manifestations of language everywhere. So when we encounter the largest database of language the world has ever seen, we would expect to find linguists exploring it, to see what is going on. It has begun to happen. And a new field is emerging as a consequence: Internet linguistics.

The name is not yet in universal use, partly because other terms have been proposed to focus on the communicative function of the Internet. In the 1990s, computer-mediated communication (CMC) became widely known, a usage which was much reinforced when it appeared in the title of an influential online publication, the Journal of Computer-Mediated Communication. However, from a linguistic point of view, this term presented a problem: it was too broad. It included all forms of communication, such as music, photographs, line-drawings, and video, as well as language in the strict sense of the word. It is this ‘strict sense’ that forms the foundation of any course on linguistics, where linguists point out the important difference between spoken, written, and signed language, on the one hand, and such figurative notions as ‘the language of painting’ and ‘the language
The terms language and communication are not synonymous.

The name computer-mediated communication is still widely used, though, as are two other terms which have an even broader remit. The emergence of mobile technology placed a certain strain on the notion of ‘mediation by computer’. People do not really feel they are holding a computer up to their ear when they talk on their cellphone, notwithstanding the fact that a great deal of computational processing is involved in making the arrangement work. And the unease was increased by the proliferation of interactive speech devices. Whether a machine is talking to us (as with satellite-navigation car instructions or airport tannoy announcements) or we are talking to a machine (as with a telephone-booking service or a voice-activated washing-machine) or reading an e-book, we do not primarily think of the devices as ‘computers’. Or, at least, they are very different ‘computers’ from the kind we are used to seeing on our desks or carrying in our briefcases. Many people have thus begun to use the more inclusive names electronically mediated communication (EMC) or digitally mediated communication (DMC). It is too soon to say which of these will become standard – or, indeed, whether some other name will emerge from cyberspace. Either way, from a linguistic point of view they are still too broad, blurring the distinction between language and other forms of communication.

I find Internet linguistics the most convenient name for the scientific study of all manifestations of language in the electronic medium. It provides the required focus, compared with human communication as a whole (for which the name Internet semiotics might be more appropriate). And it is certainly a more satisfactory label than some of those which were proposed in the early days of the Internet. Cyberspeak, Netspeak, and other -speak coinages were often used in accounts aimed at a general public, but their weakness was that they placed undue emphasis on the potential linguistic idiosyncrasy of the medium and suggested that the medium was more homogeneous than it actually is. The predominance of English on the Internet led to such names as Netlish and Weblish, but -lish terms are far too restricting today, given the increased e-presence of Chinese and other languages.
Electronic discourse and computer-mediated discourse also had some use, and their focus on interaction and dialogue have kept them alive in a social networking era. The e- prefix generated e-language and e-linguistics, though neither seems to have caught on; nor has cyberlinguistics. Sometimes it was the kind of activity that generated a new label, as in the case of searchlinguistics. Internet linguistics, as I am using the term, includes them all, as does netlinguistics. It is the study of language on the Internet – or language@internet, as the title of an online journal has it.4

As a domain of academic enquiry, Internet linguistics is in its infancy, but we can see how it is likely to develop. All the recognized branches of linguistics are in principle available. We can anticipate studies of Internet syntax, morphology, means of transmission (phonological, graphological, multimedia), semantics, discourse, pragmatics, sociolinguistics, psycholinguistics, and so on. A balance needs to be maintained between the study of the formal properties of Internet language and the study of its communicative purposes and effects. As descriptive and theoretical findings accumulate, we can expect a fruitful domain of applied Internet linguistics to emerge, providing solutions to problems of language encountered by the various users of the Internet, such as in search, e-advertising, and online security. Indeed, as we shall see, a great deal of research into Internet language has already been motivated by applied considerations.

MISCONCEPTIONS

As has happened repeatedly in the history of language study, an important part of the linguist’s job is to eliminate popular misconceptions, and the Internet has certainly provided plenty of these. The prophets of doom have been out in force, attributing every contemporary linguistic worry to the new technology, and predicting the disappearance of languages and a decline in spoken and written standards. When we investigate the worries, we invariably find they are based on myths. The moral panic that accompanied the arrival of text-messaging (or SMS, the ‘short-messaging service’) provides an illustration.
When text-messaging became popular in the UK, around the year 2000, many people saw it as a linguistic disaster. Five years later, when it began to be popular in the USA, the same reaction appeared there. There was a widespread belief that texting had evolved as a modern phenomenon, full of abbreviations that were being used in homework and exams by a young generation that had lost its sense of standards. A typical comment appeared in the *Daily Mail* in 2007 from the broadcaster John Humphrys. In an article headed ‘I h8 txt msgs: How texting is wrecking our language’ he says that texters are ‘vandals who are doing to our language what Genghis Khan did to his neighbours eight hundred years ago. They are destroying it: pillaging our punctuation; savaging our sentences; raping our vocabulary. And they must be stopped.’ He was not alone. Other disparaging comments have labelled the genre as ‘textese’, ‘slanguage’, and a ‘digital virus’.

It was difficult to counter these views in the absence of relevant linguistic research. But several studies have now shown that the hysteria about the linguistic novelty (and thus the dangers) of text-messaging is misplaced. All the popular beliefs about texting are wrong. To summarize the results of a growing literature: only a small part of text-messaging uses distinctive abbreviations (textisms); these abbreviations are not a modern phenomenon; they are not restricted to the young generation; young people do not pour them into their homework and exams; and texting helps rather than hinders literacy standards.

Text-messages are not ‘full of abbreviations’. In one American study, less than 20 per cent of the text-messages showed abbreviated forms of any kind – about three per message. In a Norwegian study, the proportion was even lower, with just 6 per cent using abbreviations. In a collection I made myself, the figure was about 10 per cent. People evidently swallowed whole the stories that appear from time to time asserting that youngsters use nothing else but abbreviations when they text. The most famous case was a story widely reported in 2003 claiming that a teenager had written an essay so full of textisms that her teacher was totally unable to understand it. An extract was posted online, and quoted incessantly. The whole thing was a hoax – which everyone believed.
Nor are text-message abbreviations ‘a modern phenomenon’. Many of them were being used in chatroom interactions that predated the arrival of mobile phones. Several can be found in pre-computer informal writing, dating back a hundred years or more. The most noticeable feature is the use of single letters, numerals, and symbols to represent words or parts of words, as with \( b \) ‘be’ and \( 2 \) ‘to’. They are called rebuses, and they go back centuries. Adults who condemn a ‘c u’ in a young person’s texting have forgotten that they once did the same thing themselves when they played word games. Similarly, the use of initial letters for whole words (\( n \) for ‘no’, \( gf \) for ‘girlfriend’, \( cmb \) ‘call me back’) is not at all new. People have been initializing common phrases for ages. \( IOU \) is recorded from 1618. There is no difference, apart from the medium of communication, between a modern kid’s \( lol \) (‘laughing out loud’) and an earlier generation’s \( SWALK \) (‘sealed with a loving kiss’).

Nor is the omission of letters – as in \( msg \) (‘message’) and \( xlnt \) (‘excellent’) – a new phenomenon. Eric Partridge published his \textit{Dictionary of Abbreviations} in 1942. It contains dozens of SMS-looking examples, such as \( agn \) ‘again’, \( mth \) ‘month’, and \( gd \) ‘good’. Texters also use deviant spellings, such as \( wot \) ‘what’ and \( cos \) ‘because’. But they are by no means the first to use such nonstandard forms. Several of these are so much part of English literary tradition that they have been given entries in the \textit{Oxford English Dictionary}. \( Cos \) is there from 1828 and \( wot \) from 1829.

The most important finding of the research studies is that texting does not erode children’s ability to read and write. On the contrary, literacy improves. Strong positive links have been found between the use of textisms and the skills underlying success in standard English in pre-teenage children. Interestingly, the more they used abbreviations, the higher they scored on tests of reading and vocabulary. The children who were better at spelling and writing used the most textisms. And the younger they received their first phone, the higher their scores. Sample sizes are small, but the results all point in the same direction.

These results surprise some people. But why should we be surprised? Children could not be good at texting if they had not already developed considerable literacy awareness. Before you
can write and play with abbreviated forms, you need to have a sense of how the sounds of your language relate to the letters. You need to know that there are such things as alternative spellings. You need to have a good visual memory and good motor skills. If you are aware that your texting behaviour is different, you must have already intuited that there is such a thing as a standard. If you are using such abbreviations as *lol* and *brb* (‘be right back’), you must have developed a sensitivity to the communicative needs of your textees, because these forms show you are responding to them.

It will be a while before the moral panic surrounding the language of text-messaging dies down. It does not take long for a myth to be established in the mind of the general public, but it can take a lifetime to eradicate it. That is one of the chief responsibilities of linguists – to demythologize. They need to build up databases using larger samples, patiently publicize findings, and try to establish a more positive climate. They can also contribute to educational projects, suggesting ways in which the Internet in general (and text-messaging in particular) can be introduced into the classroom so as to facilitate learning about language. A fruitful exercise is the ‘translation’ of text-messages into a more formal kind of standard language, and vice versa, in order to develop the student’s sense of the appropriateness of styles of language in particular situations. Several schools also engage in creative projects, such as the writing of text-messaging poetry.

What linguists cannot do is contribute professionally to the debates which take place about the social, psychological, legal, and other dangers associated with the Internet. Should a teacher confiscate a mobile phone being used by a student in class? Should parents control the amount of time their children spend on their computer? Should employers monitor the use of computers for work-unrelated activity? Should the Internet be censored? Should advertising be controlled? How can we prevent excessive keyboard or keypad use causing muscular damage? There are many such questions, about which I (as a human being) have my opinions; but these opinions do not relate to my expertise as a linguist. Rather, they fall under the remit of sociologists, psychologists, physiologists, educationalists, lawyers, and
others. They are not part of an Internet linguistics, though applied linguistic collaborations with these other domains are likely to prove illuminating.

What I, as a linguist, see on the Internet is a remarkable expansion of the expressive options available in a language – far exceeding the kinds of stylistic expansion that took place with the arrival of printing and broadcasting. These earlier media introduced many new varieties of language, such as news articles, advertisements, sports commentaries, and weather forecasts. The same sort of thing has happened on the Internet, illustrated by such new varieties as email, chat, texting, blogging, tweeting, instant messaging, and social networking. The difference is that the Internet is so much larger than the earlier media – it is capable of subsuming the worlds of print and broadcasting – and changes more rapidly. We therefore need to learn to manage it, and this point applies not only to Internet content but also to the language in which the content is expressed.

It is not always easy to use language clearly and effectively on the Internet. The interaction between sender and receiver is different from traditional conversation. The anonymity of participants alters familiar communicative expectations. Written language on a screen does not behave in the same way as writing on a traditional page. We write it differently and we read it differently. It is easy to be ambiguous, misleading, or offensive, as is shown by the proliferation of netiquette guides which offer advice about how people should behave online. In short, we need to take care. But we cannot take care if we do not understand the strengths and weaknesses of the various linguistic options that are available to us. We need to understand how electronically mediated language works, how to exploit the strengths and avoid the dangers, and this is where the developing branch of Internet linguistics can make a significant contribution.

TERMINOLOGICAL CAUTION

Students of Internet linguistics need also to be aware that some of the terminology they associate with the subject of linguistic science appears on the Internet in a different guise. This is not
the first time this has happened. Linguistics has often proved to be useful to other intellectual disciplines, which borrow its terms and then change their meaning. The Internet has done the same, notably with the words *semantic* and *semantics*.

Semantics began as a branch of linguistic science. Indeed, the word *science* is used in its original definition: the French philologist Michel Bréal, who introduced the term in the 1890s, defined it as ‘la science des significations’ – the science of meaning in language. It came to be seen as a level of linguistic investigation, alongside phonetics, phonology, morphology, and syntax, in such seminal works as Leonard Bloomfield’s *Language*; but the abstract and indeterminate nature of ‘meaning’ meant that it remained a neglected branch of linguistics for many decades. The first full-scale linguistic treatment was John Lyons’ two-volume *Semantics* in 1977, now regarded as a classic statement of the ‘state of the art’ within linguistics and linguistic philosophy. In the meantime, in the absence of a linguistic characterization, other fields found the notion of semantics useful and began to employ it in individual ways.

The philosopher Charles Morris gave semantics a more general interpretation in 1946, defining it as the interpretation of signs in general – *signs* here being used in an abstract sense to include everything that conveys information. It therefore included facial expressions, bodily gestures, road signs, railway signals, and other non-linguistic systems. Also in the 1940s, the term achieved a certain notoriety in popular usage, where ‘it’s just semantics’ began to refer to an irritating or pointless quibble. Psychologist Charles Osgood took the term in a different direction in 1953, referring to the judgements people make about words, and devising a system of rating scales which he called a ‘semantic differential’ – whether words are judged as strong/weak, good/bad, active/passive, and so on. Sometimes the term was narrowed, as when it began to appear in medicine with reference to a clinical syndrome – ‘semantic aphasia’, where people lose the ability to use words after brain damage. Sometimes it was broadened, as when Alfred Korzybski developed ‘general semantics’ in the 1930s as a method of enabling people to avoid the ideological traps built into language. But the term has achieved one of its widest
extensions in the notion of the ‘Semantic Web’, where it includes all concepts and relationships within human knowledge.

‘The vision I have for the Web is about anything being potentially connected with anything’, says the web’s inventor, Tim Berners-Lee, on the first page of his biographical account, Weaving the Web. The Semantic Web will evolve ‘without relying on English or any natural language for understanding’, he says a little later. There could be no broader definition of semantics than that, and no definition that is further away from the original linguistic intention. The Semantic Web is seen to be an evolution of the web: the existing web is human readable, whereas the Semantic Web will be machine readable. Faced with the web in its current form, it is the human user who has to specify, find, and implement the links between one page or site and another; in the Semantic Web, the links will be processed by computers without human intervention. Both a linguistic and an encyclopedic dimension will be involved. For example, to achieve a presence for automobile on the Semantic Web, the linguistic definition (as found in a dictionary) would include such features as ‘vehicle’, ‘wheels’, ‘drive’, and ‘road’; the encyclopedic account would include such elements as the different makes of car, their cost, and their safety record.

Semantics has achieved a buzz word status on the Internet these days, with many companies and approaches to knowledge management calling themselves ‘semantic’ (see further, Chapter 6). It must not be assumed that they are all talking about the same thing, or focusing on the same aspects of language. And this cautionary note applies in principle to any use of a linguistic term when found in the context of the Internet.

A rather different terminological question is what to call the various entities which form Internet discourse, such as email, blogs, chats, and tweets. A main aim of Internet linguistics is to establish their linguistic character. They are often described as genres, but that suggests a homogeneity which has not yet been established. The same question-begging would arise if they were called varieties or dialects or registers or any of the other terms for situationally related uses of language provided by sociolinguistics and stylistics. Linguists have to demonstrate linguistic
coherence, not assume it. We need a term that is theoretically neutral, from the linguistic point of view, and for the present book I propose to use outputs. I shall talk about email, for example, as being one of the outputs of Internet technology. The term implies nothing about its linguistic character, or how it relates to other outputs.

RESEARCH CHALLENGES

There are several properties of Internet language which constitute a challenge to linguists wanting to explore this medium. The amount of data it contains, first of all. There has never been a language corpus as large as this one. It now contains more written language than all the libraries in the world combined, and its informational content is rapidly increasing as more parts of the world come online, video storage grows (via such networks as YouTube), and voice-over-Internet becomes routine.

Secondly, there is the diversity of the language encountered on the Internet. The stylistic range has to recognize not only web pages, but also the vast amount of material found in email, chatrooms, virtual worlds, blogging, instant messaging, texting, tweeting, and other outputs, as well as the increasing amount of linguistic communication in social networking forums (over 170 in 2011) such as Facebook, MySpace, Hi5, and Bebo. Each of these outputs presents different communicative perspectives, properties, strategies, and expectations. It is difficult to find linguistic generalizations that apply comfortably to Internet language as a whole.

Part of the reason for this is another linguistically challenging property: the speed of change. It is not easy to keep pace with the communicative opportunities offered by new technologies, let alone to explore them in the required linguistic detail. By way of anecdotal illustration, the first edition of my Language and the Internet appeared in 2001: it made no reference to blogging and instant messaging, which had achieved little public presence at that time. A new edition of the book was therefore quickly needed, and that appeared in 2006. It included sections on the language of blogs and of instant
messages, but it made no reference to the social networking sites, which had achieved little prominence, and certainly no mention of Twitter, which arrived in the same year. Linguistic studies of the Internet always run the risk of being out of date as soon as they are written.

Even within a single output, it is difficult to keep pace. How can we generalize about the linguistic style of emails? When email first became prevalent, in the mid-1990s, the average age of emailers was in the 20s. Today, it is in the late 30s: the average in the UK rose from 35.7 to 37.9 in the year October 2006 to October 2007, according to Nielsen Online.8 Doubtless similar increases are to be found in other countries. This means that many emailers, for example, are now senior citizens – ‘silver surfers’, as they are sometimes called. The consequence is that the original colloquial and radical style of emails (with their deviant spelling, punctuation, and capitalization) has been supplemented by more conservative and formal styles, as older people introduce their norms derived from the standard language.

Another example of rapid change comes from Twitter, which uses a prompt to elicit a user response. In November 2009 the nature of the prompt changed from ‘What are you doing?’ to ‘What’s happening?’ As the Twitter blog explained:

The fundamentally open model of Twitter created a new kind of information network and it has long outgrown the concept of personal status updates. Twitter helps you share and discover what’s happening now among all the things, people, and events you care about. ‘What are you doing?’ isn’t the right question anymore – starting today, we’ve shortened it by two characters. Twitter now asks, ‘What’s happening?’9

The blogger added: ‘We don’t expect this to change how anyone uses Twitter’. But in fact a change from an inward-looking question to an outward-looking one could not fail to alter the content of the site. Twitter now has far fewer isolated postings and far more semantic threads (see further, Chapter 3). In the terminology of classical linguistics, we are faced with a new language state (Saussure’s état de langue), which raises the question of how we investigate the old ones.
For most people, the Internet became a reality following the arrival of the web in 1991, and a searchable reality after the arrival of Google in 1999. In that time, it went through several changes, reflecting the technological developments of the time. Each of these changes will have had linguistic consequences. For example, the kinds of constraint which gave a particular linguistic character to online games (MUDs, MOOs) in the 1990s have long been superseded. This means that the language of those games (1990s era) is in some ways like a period in the history of a language, needing to be studied in its own terms. But defining the boundaries of that period proves to be extremely difficult. The start-point of a new language output is relatively easy to establish, as it is linked to the innovative technology: people conversant with the history of the science can say with some precision when the language we associate with text-messaging, blogging, and tweeting began. What is more difficult is to identify endpoints, when a technology becomes outmoded or evolves into something different. And even when one has a sense of start- and end-points, tracking down the relevant data can be surprisingly difficult.

The Internet is sometimes wonderfully specific about its temporal identity, and at other times frustratingly inspecific. Beneath every page there is information about when the page was created; but only in a proportion of instances does that date appear on screen. This can cause great confusion, when (for example) a search for the population of a country yields several conflicting figures, and it remains unclear whether these reflect a synchronic or a diachronic perspective. When dates do appear, they are sometimes incomplete: many news sites, for example, give the day and the month, but not the year. There are techniques for finding the creation date of a page, or the date when the page was first spidered by a search engine or later updated, but they are cumbersome to nonspecialists.

When the dates are available, linguists find themselves faced with a different kind of problem: how to handle the unprecedented specificity? Linguists are used to being vague when it comes to describing language change: a word is said to have entered the language ‘in the early sixteenth century’ or in ‘the
Indeed, with rare exceptions, it has been impossible to identify the precise moment at which a new word or sense arrives in a language. But the time-stamping of web pages, and the ability to track changes, opens up a whole new set of opportunities. If I introduce a new word such as *digitextualization* on my website tomorrow at 09.42, it will be possible for lexicographers to say that the first recorded use of this word was at 09.42 on that day. This sort of chronological specificity has hitherto been of professional interest only to forensic linguists, concerned to identify patterns of criminal interaction, but it will in future be of much broader relevance. It is not yet clear how Internet linguistics will handle this level of descriptive detail.

Finally, leaving aside questions of dating, some kinds of Internet language present a rather different kind of challenge: inaccessibility. There is of course no problem in finding and downloading data from the pages of the web, within the various legal and commercial constraints imposed by website-owners. But it is a different matter when dealing with such outputs as email, chat, and text-messages. People are notoriously reluctant to allow their private e-communications to be accessed by passing linguists. There are now some excellent corpora of emails and chatroom interaction, but issues of reliability and representativeness have yet to be fully explored, and some domains, such as text-messaging, remain elusive, especially in languages other than English. The research literature is characterized by a great deal of theoretical speculation but relatively few empirical studies.

Another research issue arises out of the practice of anonymity. Normally, linguists take great pains to establish the situational factors which motivate or condition a use of language. Factors such as age, gender, class, and ethnicity are critical. But in a medium where a large number of participants hide their identity, or where we cannot trust the self-disclosed information about themselves which they place online, it is difficult to know how to interpret observed usage. Even fundamental distinctions, such as whether a netizen is male or female, or a native or non-native speaker, can be obscured. The Internet is not the first medium to allow interaction between individuals who wish to remain anonymous, of course, as we know from the history of telephone and
amateur radio; but it is certainly unprecedented in the scale and range of situations in which people can hide their identity, especially in chatgroups, blogging, and social networking. The effect of anonymity on linguistic behaviour also needs to be explored. Operating behind a false persona seems to make people less inhibited: they may feel emboldened to talk more and in different ways from their real-world linguistic repertoire. They must also expect to receive messages from others who are likewise less inhibited, and be prepared for negative outcomes. There are obviously inherent risks in talking to someone we do not know, and instances of harassment, insulting or aggressive language, and subterfuge are commonplace.

Ethical considerations also need to be taken into account: what kinds of permission are needed to use Internet data? The same questions that linguists had to address in the 1960s, in the early days of corpus construction – such as the distinction between public and private language – have risen again in electronic form. If I send a message to the Internet, I have presumably let it go into the public domain: do I then have any right to object if a linguist includes it in a corpus? Who owns the text-messages in my mobile phone archive: are they mine or the senders’? In an increasingly litigious world, linguists need to take care that their data-collection procedures are robust with respect to the question of ownership.

As the old saying goes: turn a challenge over and you see an opportunity. The Internet offers linguists unprecedented opportunities for original research. Because we are dealing with an electronic medium, we need to not only investigate the new kinds of language introduced by the technology (blogging, tweeting, etc.), but also reinterpret everything we already know about language as realized through the older mediums of speech, writing, and sign. Whatever facts were established about, say, the differences between spoken and written vocabulary and grammar, these now have to be revisited, because the way we use language on the Internet is different in salient respects from the way we use it in traditional speech and writing. Which Internet styles of writing promote the use of abbreviations and emoticons? How does column width affect discourse structure? Do hypertext links influence the way a written text is organized? How does speech lag affect the
character of a spoken conversation on Skype or iChat? Every use of language on the Internet will display features that do not correspond to the features identifying that use in traditional speech or writing. Written language has to be graphically translated\textsuperscript{10} so that its content appears clearly on screen and can be easily accessed and navigated. Spoken language too needs to be processed so that its content can be indexed and navigated, with the possibilities here dependent on progress in automatic speech recognition.

Even when the electronic medium simply scans texts for viewing on screen, it presents those texts in new ways, allowing us to do new things with them. We can zoom in on an ancient manuscript and see detail that was not easily visible before, or carry out linguistic searches which were not practicable before. Well-studied uses of speech and writing appear in fresh guises. News journalism, for example, can look very different on screen compared with how it would appear in the traditional medium of print – paragraph size, for example, is often shorter. A poem on a screen is a very different reading experience from one in a printed book, especially when, as in text-messaging poetry, the small screen allows only a small part of the poem to be seen at any one time. The novelty is most apparent in the written language, for the Internet to date has been a predominantly graphic medium; but spoken language is also affected. Even the ‘listen again’ feature in a broadcasting station offers new possibilities: the programme is the same as it was on the radio, but the listener now has the opportunity to stop it at will, to listen to something a second time, to skip sections, and to move forwards and backwards along the timeline. The management of the auditory experience has transferred from the producer to the receiver.

The first step, then, in an Internet linguistics, is to establish the properties of the medium which condition the language experience and behaviour of its users. The most illuminating way of doing this, in my view, is to start by distinguishing it from the familiar worlds of spoken and written language.

See also ‘Research directions and activities’, p. 151.