David Crystal

Internet Linguistics

A Student Guide

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*

Internet Linguistics: A Student Guide

David Crystal

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2011
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
270 Madison Avenue, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group, an informa business

Typeset in Sabon and Scala by
Swales & Willis Ltd, Exeter, Devon
Printed and bound in Great Britain by
TJ International Ltd, Padstow, Cornwall

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Crystal, David, 1941–
Internet linguistics : a student guide / David Crystal.
p. cm.
Includes index.
1. Computational linguistics. 2. Internet. 3. Internet—Social aspects. I. Title.
P98.5.I57C75 2011
004.601'4—dc22
2010034571

ISBN 13: 978-0-415-60268-6 (hbk)
ISBN 13: 978-0-415-60271-6 (pbk)
ISBN 13: 978-0-203-83090-1 (ebk)

Contents

Preface

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

1

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*.[1] 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

of the face', on the other.[2] 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,[3] 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:[5] 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 *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 *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.[6] 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*.[7] 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