

'O brave new world, that has such corpora in it!'
New trends and traditions on the Internet.

David Crystal

*Plenary paper given to ICAME 32,
Trends and Traditions in English Corpus Linguistics,
1 June 2011*

If there's one thing that unites all of us, in the field of corpus linguistics, it is that we assume we know a text when we see one. Make a random selection of texts recognized in the various corpus classifications: letter, interview, advertisement, radio sports commentary, news broadcast, shopping list, textbook, newspaper editorial, lecture, prayer, road sign, novel, poem... There are lots of issues to be debated, of course, such as how many examples to collect to make a representative sample, how large our samples should be, how to classify individual instances, and how to construct fruitful text typologies. But the bottom line is that, in all these cases, we are choosing units to study that are identifiable and determinate. They have definable physical boundaries, either spatial (eg letters and books) or temporal (eg broadcasts and interviews) or mixed medium (eg lecture powerpoint karaoke). They are created at a specific point in time; and once created, they are static and permanent. Each text has a single authorial or presenting voice (even in cases of multiple authorship of books and papers), and that authorship is either known or can easily be established (except in some historical contexts). It is a stable, familiar, comfortable world. And what the Internet has done is remove the stability, familiarity, and comfort. It is not good news for corpus linguists. We have to rethink everything.

Texts sans frontières

Written texts are defined by their physical boundaries: the edges of the page, the covers of the book, the border of the road sign... Spoken texts are defined by their temporal boundaries: the arrival and departure of participants in a conversation, the beginning and end of a broadcast, the opening and closing of a lecture... Internet texts are more problematic. Sometimes, as with a text message or an instant-message exchange, we can clearly identify the start and the finish. But with most Internet outputs there are decisions to be made, as the following examples show.

- Does a single email message constitute a text, or is the text everything available on a screen at a particular point in time, including previously exchanged messages that have not been deleted and any framed or intercalated responses sent by the recipient? If the latter, how does one deal, in any analysis, with the repeated linguistic features that appear in the various iterations? A typical sequence would begin with a single message:

To:	Hilary Crystal
Cc:	
Bcc:	
Subject:	an example
From:	Crystal David <davidcrystal1@...>
Signature:	Signature #3

This is an example of an email message, sent by me last week.

I need examples of a framed and an intercalated response for ICAME. But please keep them as short as possible, as they need to fit on a slide.

Many thanks.

David

It might then receive a response like this:

From: Hilary Crystal
Subject: **Re: an example**
Date: 23 May 2011 10:12:14 BST
To: Crystal David <davidcrystal1@gmail.com>

On 23 May 2011, at 10:09, Crystal David wrote:

| This is an example of an email message, sent by me last week.

Thank you.

| I need examples of a framed and an intercalated response for ICAME.

Will this do?

| But please keep them as short as possible, as they need to fit on a slide.

Okay.

| Many thanks.

| David

And this in turn might be given a further response, like this:

Further comments below.

D

On 23 May 2011, at 10:12, Hilary Crystal wrote:

On 23 May 2011, at 10:09, Crystal David wrote:

| This is an example of an email message, sent by me last week.

Thank you.

| I need examples of a framed and an intercalated response for ICAME.

| Will this do?

Perfect.

| But please keep them as short as possible, as they need to fit on a slide.

Okay.

| Many thanks.

| David

One could go on and on. But what is 'the text' in all of this? And does one include unchanging biodata, such as the sender's address, weblinks, and taglines, as here?



From: Crystal David <davidcrystal1@googlemail.com>

The previous example, for the sake of brevity, omitted the end-message data, shown below. If these are repeated in each exchange, it is easy to see how frequencies will be affected.

Professor David Crystal
Akaroa
Gors Avenue
Holyhead LL65 1PB, UK

t 01407 762764

www.davidcrystal.com
<http://david-crystal.blogspot.com>
www.theshakespeareportal.com

Wee often see..a faire and beautifull corpes, but a foule vgly mind.
(Thomas Walkington, *The Optick Glasse of Humors*, 1607)

- A fortiori, does an entire website constitute a text, or are the texts the individual elements of the menu (Home, About, Contact, Help...), or the individual pages, or the functional elements seen on these pages (maintext, advertisements, comments...)? The distinction has commercial importance in online advertising, where an ad server is likely to serve a different range of ads to the top page of a site compared to its constituent pages. When I was writing this paper, Sky TV, for example, had a banking ad at the top of its home page, and a video games ad at the top of its sport page.
- What about translations? Many websites now are multilingual, with a list of language choices on the home page. What is the corpus here? How are these elements to be handled? Traditional corpus linguistics operates within single languages, but multilinguality is intrinsic to the Web.
- If an email, tweet, instant message, blog, or other output includes an obligatory hypertext link, is that link to be considered as part of the text? By obligatory I mean a link that forms part of the structure of a sentence or which provides information that is critical to the understanding of the page. Here is an example from the Web, which shows both of these types of dependence:

The next Passion in Practice workshop will be May 16th-20th 2011 in London.

Please head to www.passioninpractice.com for more details.

London Fringe Radio Live, Weds 16th @7.45pm

Posted on March 15, 2011 in: [Audio/Video](#), [Interviews](#), [Latest News](#) | [Comments Off](#)

Being interviewed on t'radio tomorrow about *Shakespeare on Toast...*

Details to be found here...

And here is an example from Twitter (the named individual here not being the author of this paper):

[MJMPR](#): **David Crystal** [@unikwaxcenters](#) at [#ewssmiami](#). Helping groom him for Mayor of North Miami Beach

3 days ago via [web](#) · [Reply](#) · [View Tweet](#)

Punctuation becomes critical, as we see in this next example, where grammatically we might wish to treat the final link differently when there is a period or a colon:

[axfelix](#): "Alcohol **language corpus**: the first public **corpus** of alcoholized German speech." I love linguistics very, very much. <http://bit.ly/ewEJgv> ([expand](#))

1 day ago via [Chromed Bird](#) · [Reply](#) · [View Tweet](#)

[Thuyrmo](#): Linking Up Contrastive and Learner **Corpus** Research (**Language & Computers**): <http://amzn.to/k2FEO0>

1 day ago via [twitterfeed](#) · [Reply](#) · [View Tweet](#)

- If security is an obligatory element (eg asking for user names, passwords, or other authentication), is this to be considered as part of the text? Are the glosses or images which appear when a mouse hovers over a string to be considered as part of the text? And do we include the keywords which identify the page, and which may not appear on the screen, but are only visible when one looks at the underlying code, as here?

```
<HEAD>
<TITLE>Stamp Collecting World</TITLE>
<META name="description" content="Everything you wanted to know
about stamps, from prices to history.">
<META name="keywords" content="stamps, stamp collecting,
stamp history, prices, stamps for sale">
</HEAD>
```

- How are we to define a text in an internet output which is continuously growing, as in a social networking site, a chatroom, a blog forum, or a bulletin board, which might last indefinitely? In these cases there is a dynamic archive, which in some cases goes back many years. Take this example from my blog, which was a game where one takes a well-known foreign phrase, changes a single letter, and then adds a humorous gloss. Here are some examples from the original post:


FELIX NAVIDAD
Our cat has a boat

HASTE CUISINE
Fast French food

E PLURIBUS ANUM
Out of any group, there's always one asshole

POSTED BY DC AT 11:35 89 COMMENTS

It received 89 comments, dating from 30 May to 11 January, such as this one:

 **Fran said...**

Je ne sais quoit - I'm rubbish at deck games.

30 May 2010 14:46



 **DC said...**

:))

30 May 2010 15:01

And of course the forum might be added to at any time. Question: are associated comments to be considered part of the text? As they are elicited by the maintext, and are semantically (and sometimes grammatically) dependent on it, as we see with this emoticon response, they cannot be taken as independent texts in their own right. There is an asymmetrical relationship. On the other hand, the maintext has autonomy: it does not need comments to survive. But comments could not exist without a maintext. And there is no theoretical limit to the number of comments a post might elicit.

- Similarly, how are we to define a text in an internet output which is continually changing - where there is permanently scrolling data, regularly updated, such as stock-market reports and news headlines? Here there may be no archive: old information is deleted as it is replaced. The content comes from an inventory which is fixed at any one point in time, but frequently refreshed. Some sequences that appear on-screen are cyclical (such as the recurring headlines we see in a news ticker service or a retail store); others are randomly generated (such as the pop-up ads or banner ads taken from a large inventory, which may change in front of your eyes every few seconds. This is not the first time we have seen dynamic text, of course - neon images in Times Square and Piccadilly Circus come to mind. But these don't solve the problem, only reinforce the question of how to handle it (for I do not recall neon signs being a recognized category in corpus classifications hitherto).
- What do we do with a message sequence (as in emails or a bulletin board) where the subject-line identifies a semantic thread? Is the text the set of messages that relate to that thread? They may be separated by other messages, as in this example from a Shakespeare forum.

4	Arden3 The Merchant of Venice
5	Thoughts on Double Falsehood
6	Arden3 Sir Thomas More
7	2011 Blackfriars Conference Announcement
8	From New York to Santa Fe
9	Arden3 "The Merchant of Venice"
10	Hamlet's Enemy

Do we follow the header? If so, what do we do with cases where (a) the discussion continues but someone changes the header in the subject-line, or (b) the header in the subject-line remains the same, but the discussion veers off-topic? Which takes priority?

- Are we to include in the text elements automatically inserted by cookies, such as site preferences, shopping cart contents, and visitor tracking, or the features which are available to users, such as helplines and analytics reports?
- How do we view texts rendered incomplete by the technology, as when a tweet exceeds the 140-character limit and is truncated by the software? This is shown by ellipsis dots on screen.



[dailynews24org](#): Childcare benefit 'cuts' warning: **David Cameron's** flagship promise to make work pay may be "shattered" by cuts t... <http://bit.ly/m7Jlop>

(expand)

3 minutes ago via *twitterfeed* · [Reply](#) · [View Tweet](#)

I do not know how to answer these questions using the traditional notion of 'text'. And this makes me think that we are asking the wrong question. It looks as if some broader, more inclusive notion is going to be needed. Clearly, what we see in all these examples are aggregates of functional elements, which interact in various ways in different Internet outputs (as I call them, Crystal 2011). We need terms for both the elements and the aggregates and the . Dürscheid and Jucker (2011), for example, call the elements 'communicative acts', and the aggregates 'communicative act sequences'. Doubtless other proposals will be forthcoming, as corpus linguists explore these phenomena in more detail. In the meantime, some general observations may be in order.

Panchronicity

The above examples are not a complete list of the boundary decisions which have to be made when we are trying to identify Internet texts, but they are representative of what is 'out there'. And they raise quite fundamental questions. In particular, Ferdinand de Saussure's classical distinction between synchronic and diachronic does not adapt well to these kinds of communication, where everything is diachronic, time-stampable to a micro-level. Texts are classically treated as synchronic entities, by which we mean we disregard the changes that were made during the process of composition and treat the finished product as if time did not exist. But with many electronically mediated texts there is no finished product. And in many cases, time ceases to be chronological. For example, I can in 2011 post a message to a forum discussion about a page which was created in 2004. From a linguistic point of view, we cannot say that we now have a new synchronic iteration of that page, because the language has changed in the interim. I might use in my comment vocabulary that has entered the language since 2004, or show the influence of an ongoing grammatical change. Content is inevitably affected. I might refer to Twitter - something which would not have been possible in 2004, for that network did not appear until 2006. I might even - as is possible with Wiki pages - insert information into the maintext of a page which could not have been available at the time of the page's creation. In the case of my blog, I might go back to a post I wrote in 2004 and edit it to include material from 2011.

Here's a specific example of the kind of complication which can arise. A little while ago I wrote a blog post in which I gave a certain statistic. A reader send a comment to the forum pointing out an error. I published the comment, but - to avoid

other readers being misled by my error I called up my original post and made the correction. The main text is now correct; but the comment referring to a now nonexistent error still remains. One type of confusion has been replaced by another. I tried to reduce the confusion by sending a comment explaining what I had done:

 **mollymooly said...**

Slight overfractionation: $2405 / 144461 = 1.66\%$

Oddly, "amongst" and "amidst" are British whereas "unbeknownst" is American.

28 February 2011 18:14

 **DC said...**

Oops. Thanks. Now changed in the post.

28 February 2011 18:21

But I doubt whether this kind of confession is commonplace. In theory there should be no problem, because all contributions are time-stamped, so in principle it is possible to find out when a contribution or change was made. But time codes are often not visible on the screen; if they are, they are usually not very prominent, and even if they are easy to read, most readers would pay them no attention.

We need a new term for this curious conflation of language from different time periods. We are very familiar with texts which include language from earlier periods (*archaisms*). We need a way of describing features of texts which include language from later periods. The traditional term for a chronological mismatch is *anachronism* - when something from a particular point in time is introduced into an earlier period (before it existed) or a later period (after it ceased to exist). Anachronisms can be isolated instances - as when Shakespeare introduces striking clocks into ancient Rome (in *Julius Caesar*) - or a whole text can be anachronistic, as when a modern author writes a play about the 17th century and has everyone speak in a 21st-century way. But these cases don't quite capture the Internet situation, where a chronological anomaly has been introduced into an original text. This is a new take on the grammatical notion of 'future in the past' - or perhaps better, 'back to the future'. And I think we need a new term to capture what is happening. A text which contains such futurisms cannot be described as synchronic for it cannot be seen as a single *état de langue*: it is a conflation of language from two or more *états de langue*. Nor can it be described as diachronic, for the aim is not to show language change between these different *états*. Such texts, whose identity is dependent on features from different time-frames, I propose to call *panchronic*.

Wiki pages, such as those seen on Wikipedia, are typically panchronic. They are the result of an indefinite number of interventions by an indefinite number of

individuals over an indefinite number of periods of time (which become increasingly present as time goes by). We are only twenty years into the Web, so the effect so far is limited; but think ahead 50 or 100 years, and it is obvious that panchronicity will become a dominant element of Internet presence. From a linguistic point of view, the result is pages that are temporally and stylistically heterogeneous. Already we find huge differences, such as standard and nonstandard language coexisting on the same page, often because some of the contributors are communicating in a second language in which they are nonfluent. Tenses go all over the place, as this example illustrates:

Following his resignation, Mubarak did not make any media appearances. With the exception of family and a close circle of aides, he reportedly refused to talk to anyone, even his supporters. His health was speculated to be rapidly deteriorating with some reports even alleging him to be in a coma. Most sources claim that he is no longer interested in performing any duties and wants to "die in Sharm El-Sheikh."^{[59][60]}

On 28 February 2011, the General Prosecutor of Egypt issued an order prohibiting Mubarak and his family from leaving Egypt. It was reported that the former president was in contact with his lawyer in case of possible criminal charges against him.^[61] As a result, Mubarak and his family had been under house arrest at a presidential palace in the Red Sea resort of Sharm el-Sheikh.^[62] On Wednesday 13 April 2011 Egyptian prosecutors said they had detained former president Hosni Mubarak for 15 days, facing questioning about corruption and abuse of power, few hours after he was hospitalized in the resort of Sharm el Sheik.^[63]

Note the way for example, we move from past tense to present tense in paragraph 1, and from *was* to *had* in paragraph 2. Note also the way *former president Hosni Mubarak* is introduced in the last sentence, as if this were a new topic in the discourse. And how are we to interpret such usages as *was speculated*, *in case of*, and *few hours*?

In pages like this, traditional notions of stylistic coherence, with respect to level of formality, technicality, and individuality, no longer apply, though a certain amount of accommodation is apparent, either because contributors sense the properties of each other's style, or a piece of software alters contributions (eg removing obscenities), or a moderator introduces a degree of levelling. The pages are also semantically and pragmatically heterogeneous, as the intentions behind the various contributions vary greatly. Wiki articles on sensitive topics illustrate this most clearly, with judicious observations competing with contributions that range from mild through moderate to severe in the subjectivity of their opinions. And one never knows whether a change introduced in a wiki context is factual or fictitious, innocent or malicious.

The problem exists even when the person introducing the various changes is the same. The author of the original text may change it—refreshing a Web page, or revising a blog posting. How are we to view the relationship between the various versions? This is not the first time we have encountered this problem. It is a familiar problem for medievalists faced with varying versions of a text. It is a routine question in the case of, say, Shakespeare: did he (or someone else) go back and revise an earlier manuscript? It is something we see all the time in the notion of a 'second edition', where the two layers of text may be separated by many years. But what is happening on the Internet is hugely different from the traditional process of revision, because it is something that authors can do with unprecedented frequency and in

unprecedented ways. A website page can be refreshed, either automatically or manually. The issue is particularly relevant now that print-on-demand texts are becoming common. It is possible for me to publish a book very quickly and cheaply, printing only a handful of copies. Having produced my first print-run, I then decide to print another, but make a few changes to the file before I send it to the POD company. In theory (and increasingly common in practice), I can print just one copy, make some changes, then print another copy, make some more changes, and so on. The situation is beginning to resemble medieval scribal practice, where no two manuscripts were identical, or the typesetting variations between copies of Shakespeare's First Folio. The traditional terminology of 'first edition', 'second edition', 'first edition with corrections', ISBN numbering, and so on, seems totally inadequate to account for the variability we now encounter. The same problem is also present in archiving. The British Library, for example, has recently launched its Web Archiving Consortium. My website is included. But how do we define the relationship between the various time-stamped iterations of this site, as they accumulate in the archive?

Anonymity

I mentioned five criteria above: texts have definable physical boundaries; they are created at a specific point in time; they are static and permanent; they have a single authorial or presenting voice; and - apart from in some historical contexts - authorship is either known or can easily be established. None of these criteria are necessarily present on the Internet. And in the case of the last of these, its absence presents corpus linguists with a particularly difficult situation. When we classify texts into types we rely greatly on extralinguistic information. This is something we have learned from sociolinguistics and stylistics: the notion of a language variety (or register, or genre, or whatever) arises from a correlation of linguistic features with extralinguistic features of the situation in which it occurs, such as its formality or occupational identity. In principle we know the speaker or writer - whether male or female, old or young, upper-class or lower-class, scientist or journalist, and so on. And when we do research we try to take these variables into account in order to make our corpus comparable to others or distinguishable from others in controlled ways. In short, we know who we are dealing with.

But on the Internet, a lot of the time, we don't. The writer is anonymous. In a wide range of Internet situations, people hide their identity, especially in chatgroups, blogging, spam emails, avatar-based interactions (such as virtual reality games and Second Life), and social networking. These situations routinely contain individuals who are talking to each other under nicknames (*nicks*), which may be an assumed first-name, a fantasy description (*topdude*, *sexstar*), or a mythical character or role (*rockman*, *elfslayer*). 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 legion. Terminology has evolved to identify them, such as flaming, spoofing, trolling, and lurking. New conventions have evolved, such as the use of CAPITALS to express 'shouting'.

While all of these phenomena have a history in traditional mediums, the Internet makes them present in the public domain to an extent that was not encountered before. But we do not yet have detailed linguistic accounts of the

consequences of anonymity. All that is clear is that traditional theories don't account for it. Try using Gricean maxims of conversation to the Internet (Grice 1975): our speech acts should be truthful (maxim of quality), brief (maxim of quantity), relevant (maxim of relation), and clear (maxim of manner). Take quality: Do not say what you believe to be false; Do not say anything for which you lack evidence. Which world was Grice living in? A pre-Internet world, evidently. Pragmatics people traditionally assume that human beings are nice. The Internet has shown that they are not. Is a paedophile going to be truthful, brief, relevant and clear? Are the people sending us tempting offers from Nigeria - beautifully pilloried in Neil Forsyth's recent book, *Delete This at your Peril* (2010)? Are extreme-views sites (such as hate racist sites) going to follow Geoffrey Leech's (1983) maxims of politeness (tact, generosity, approbation, modesty, agreement, sympathy)? And if brevity was the soul of the Internet, we would not have such coinages as *blogorrhea* and *twitterrhea*.


Electronically mediated communication 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, and exploit their anonymity in ways that would be difficult to replicate offline. And the linguist is faced with a growing corpus of data which is uninterpretable in sociolinguistic or stylistic terms. A different orientation needs to be devised, in which intention and effect become primary, and identity becomes secondary. In short, our text typologies need to become increasingly pragmatic.

Towards a new typology




Do existing text typologies apply to web pages? This is the question asked by Marina Santini in a paper she wrote for the ICAME journal (2006). Only with great caution, is her conclusion. I would go further. My answer is: No. There are three reasons.

Because of the sheer size of the Internet, only automated analyses will be able to cope with the diversity and rate of change of language within the medium, and at the moment we are still a long way from a means of natural language processing which will suit the needs of corpus linguists. There are several issues, some of which are more easily solvable than others. Most methods assume linearity of text, and run into difficulty with nonlinear text, such as tables, lists, headings and subheadings, buttons, links, and the location of blocks of information in a page design. There are problems with integrating video, sound, and images, as well as text. Even the basic terminology for describing page elements is in its infancy. Progress has been made in standardizing metadata elements, thanks to such projects as the Dublin Core, but the basic set of elements is, from a linguistic point of view, very general: title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. The good news is that all the information that makes up an Internet page is in principle available. A page is the result of a design, and all the elements of the design are specified in the underlying code. We need to get at that, and filter it in a way that is linguistically useful. This isn't easy, as the scrape has to distinguish between linguistically relevant and irrelevant data. For example, when I was scraping pages as part of a web classification, I compiled lexical sets which would identify the content of a category. However, if the lexemes in the set happened to contain items that were used in the underlying code (eg *bin*, *clear*, *align*, *type*...), then there was a real risk of a misclassification.

A second problem is the trustworthiness of the data. Pages contain all kinds of errors introduced by typos, processing errors (especially through automatic text recognition procedures), lag, and suchlike. Translating between software systems can introduce all kinds of distortion - old formatting lost, new formatting inserted, and so on - such as sending a page from Word to iPages. The language is also often more informal than what is typically found in written texts, with punctuation, capitalization, and spelling nonstandard, as in this example:

 Finally - Obama calls for Israel's return to pre-1967 borders

this is a start al tho a pre 1948 borders would be the best outcome. Lol and BBYahooSoFuckinNutz is supposed to come to the whitehouse in next few days , I think Obama just Bitch slapped him for BB previous actions towards obama.

Israel change is gonna come , one way or the other   

The contrast is striking. All non-Internet public written texts have been edited and proofread. The graphical and grammatical homogeneity often seen there, which has fuelled goodness knows how many statistical studies, is partly an artefact of copy-editorial interference, which has clothed our writing in the forms favoured by institutional house-styles. Only now, on the Internet, are we seeing this sort of writing in its most naked form. When I write my blog, there is no copy-editor looking over my shoulder and telling me how I should be punctuating, spelling, and phrasing my sentences. The down side of this is that all kinds of graphical idiosyncrasies are found in Internet outputs - variations in hyphenation, for example, which complicate word identification, or variations in upper and lower case (which are usually ignored in the search engines). Deviant spellings are not just the result of error, of course, but can be deliberate, as in the proper names of products and organizations (*Kwik*, etc). It should be remembered that on the Internet the traditional distinction between dictionary and encyclopedia disappears. And data goes well beyond the linguistic, including for example model numbers of products.

A third problem is developing a system of unit classification which is appropriate to the Internet, and here linguists face their greatest challenge. The kind of classifications which are currently used in corpus linguistics are inappropriate for Internet analysis. Take the classification which informs the Survey of English Usage:

Speech

To be heard

Now (the norm)

Later, e.g. telephone answering messages

To be written down

As if spoken, e.g. police statement, magazine interview

As if written, e.g. letters, dictation

Writing

To be read (the norm)

To be read aloud

As if spoken, e.g. radio/TV drama

As if written, e.g. radio/TV newsreading

To be partly read aloud, e.g. broadcasting continuity summaries

Mixed medium

- To self, e.g. memoranda, shopping list
- To single other, e.g. co-authorship sessions
- To many others, e.g. spoken commentary on a handout or blackboard

This kind of approach can handle only a small proportion of Internet texts. For the Internet, we need to adopt an additional pragmatic perspective, recognizing the factors which govern the selection of texts, such as the intention users have in mind or the effect they wish to convey. This will produce a classification along the following lines:

Electronically mediated communication

As an end in itself

- To be read (the norm)
 - In the displayed language
 - In another language (if available), e.g. translate, click icon
- To be read
 - Statically (the norm)
 - Dynamically, e.g. news feeds, incoming results, market reports
- To be added to, e.g. chatroom, forum, post a comment, Facebook wall
- To be acted upon
 - To obtain information, e.g. contact us, help
 - To review or evaluate, e.g. consumer reviews
 - To persuade, e.g. ads, wish lists, more like this
 - To purchase, e.g. payment methods, buying procedures
 - To spread news, e.g. retweets

As a means to an end

- On the same page
 - To be searched, e.g. advanced search, archive, track order
- On a different page
 - Hyperlinks, e.g. quick links, permalinks, tweet expansions
- Using another medium, e.g. podcast, video link, YouTube link

In the present state of research, a list of this kind can only be indicative, not comprehensive. One reason is that the pace of change is very rapid. What is the pragmatic purpose of Twitter, for example? There are many types of tweet, but at the most general level the pragmatic purpose is easy to see from the prompt used by the network, which defines the intention behind it. This would seem to be a gift for linguists. But we have to be on our guard, for that prompt can change - as it did in November 2009, when the original 'What are you doing?' became 'What's happening?' The result was a shift in the character of the tweets, which took on more of the features of a news service, as well as attracting more advertising content. And who knows how many more changes in pragmatic function there might be?

The stylistic analysis of the texts relating to each of the above categories is in its infancy. Plainly, there is a scale of online adaptability. At one extreme, we find outputs where no adaptation to electronically mediated communication has been made—a pdf of an article on screen, for example, with no search or other facilities—in which case, any linguistic analysis would be identical with that of the

corresponding offline text. At the other extreme, we find outputs which have no counterpart in the offline world. For example, there are texts whose aim is to defeat spam filters, and which produce such strings as: *supr vi-agra online now znwygghsxp*. There are outputs designed to ensure that a webpage appears in the first few hits in a web search, by manipulating the language to increase the frequency of certain words. Text-messaging and tweeting are examples of outputs whose linguistic characteristics have evolved partly as a response to technological constraints.

The biggest question mark relates to the way the Internet is developing - always difficult to predict. Although the Internet is supposedly a medium where freedom of speech is axiomatic, controls and constraints are commonplace to avoid abuses. These range from the excising of obscene and aggressive language to the editing of pages or posts to ensure that they stay focused on a particular topic. A good example of content moderation is in the online advertising industry, where there is a great deal of current concern to ensure that ads on a particular web page are both relevant and sensitive to the content of that page. Irrelevance or insensitivity leads to lost commercial opportunities and can generate extremely bad PR. Irrelevance can be illustrated by an online report about apple crop yields, where the ads down the side of the screen carried ads for Apple Macs—the software being evidently unaware that the fruit sense of 'apple' in the news report did not match the computer sense of 'apple' in the ad inventory. Insensitivity can be illustrated by a page which was reporting a ferocious civil war in a certain country, and the adjacent ad was for a tourist agency advertising holidays to that country, much to the embarrassment of all concerned (the ad network, the client company, the reader). Putting this in the terminology of pragmatics, the perlocutionary effect was not what was intended. The solution known as 'semantic targetting', as used in the iSense and Sitescreen products I developed for Ad Pepper Media, carries out a complete lexical analysis of web pages and ad inventories so that subject-matter is matched and ad misplacements avoided. Ad misplacements are everywhere: no car manufacturer wants their ad to appear against a story about a horrific car accident, for example; but this sort of thing happens all the time, as we see here:

The screenshot shows a news website interface. At the top, there is a navigation bar with links like 'SHAKSPER: T... Conference', 'Hopkins, Ge...', '118. Poems', 'Bullokars Bo...', 'Molcaster', 'Type IPA ph...', 'languages', 'Home: Oxfar', and 'Dictionary'. Below this is a banner advertisement for Toyota with the text '#1 FOR A REASON' and 'See local offers'. The main content area features a large headline 'ON DEADLINE' with the subtext 'Breaking news and must-read stories'. Below this is a search bar and navigation links for 'Home', 'Archives', 'Forum', and 'About'. The main article is titled 'Randy 'Macho Man' Savage dies in car accident' and is dated '02:08 PM'. It is attributed to 'By Douglas Stanglin, USA TODAY'. To the right of the article is an advertisement for Edward Jones, featuring a photo of a man and a child. Below the Edward Jones ad is a section titled 'About Doug Stanglin' with a small photo of the author. At the bottom of the page, there is a loading message: 'Loading "http://content.usatoday.com/communities/ondeadline/post/2011/05/andy-macho-man-savage-dies-in-car-accident/1", completed 62 of 54 items'.

And, to show that the Internet issues I discuss in this paper are not restricted to English, here is an example of a German ad misplacement: I do not think this car manufacturer would have wanted its triumphant ad to be placed next to a story about an accident involving a burnt-out car on the motorway:

News Lokales Sport Videos Dossier Fotos Handy Anzeigen Leserservice AZ Webservice

AZ-WEB.DE Aachener Zeitung

Lokales / Aachen

Google-Anzeigen

Lippenmodellierung
Wir beraten Sie gerne in Düsseldorf Hyaluronsäure und BTX auf der KO
www.koe-aesthetic.de/Falt

4% Tagesgeld-Zinsen
Tagesgeld-Konten mit Top-Zinsen im aktuellsten Online-Vergleich!
Vergleich.de/Tagesgeld

Düren Ladenlokal-SB-Markt
Direkt vom Eigentümer zu mieten Maklerfrei - Ruf 02242-88111
www.vvstbau.de

Aachen Immobilien
Wohnungen, Häuser, Grundstücke. kaufen/mieten, verkaufen/vermieten.
www.immowelt.de/aachen

Hotel Paseo Aachen
Spanisches Themenhotel in Aachen in direkter Nähe zum Klinikum RWTH
hotel-paseo.de

Auto ausgebrannt: Stau auf der A4
(EVA) 25.10.2009, 17:10

INNE MEINUNG ARTIKEL DRUCKEN EMAIL AN REDAKTION ARTIKEL VERSENDEN A A SCHRIFT GRÖSSE

Aachen. Auf der Autobahn 4 in Fahrtrichtung von Holland in Richtung Aachener Kreuz ist es am Sonntagmittag zu massiven Rückstaus gekommen.



Grund waren Lösch- und Sicherungsarbeiten auf einem direkt vor dem Kreuz gelegenen Parkplatz.
Dort war am Vormittag ein Auto aus bislang unbekannten Gründen in

In extreme cases, such as a firm which does not want its ad to appear on a particular page (e.g. a child clothing manufacturer on an adult porn site), ads can be blocked from appearing. As a result, from a content point of view, the text that appears on a page appears more semantically coherent and pragmatically acceptable than would otherwise be the case.

This paper has largely focused on the written language of the Internet. The main issue for the future will be how to deal with the increased presence of spoken outputs, as a result of growth in Voice over the Internet and mobile communication. There are several new kinds of speech situation here, such as the modifications which are introduced into conversation to compensate for the inevitable lag between

participants, automatic speech-to-text translation (as when voicemail is turned into text messages), text-to-speech translation (as when a web page is read aloud), voice recognition interaction (as when we tell the washing machine what to do), and voice synthesis (as when we listen to GPS driving instructions). Each of these domains is going to introduce us to new kinds of output over the next twenty years. Evidently, we ain't seen nothin' yet. Whatever the trends and traditions were in corpus linguistics during its first half century, they will be very different in its next.

References

Crystal, David. 2011. *Internet linguistics: a student guide*. London: Routledge.

Dürscheid, Christa and Andreas Jucker. 2011. Text as utterance: communication in the electronic media. Paper given to the conference 'Language as a social and cultural practice: advances in linguistics', University of Basel.

Forsyth, Neil (aka Bob Servant). 2010. *Delete This at your Peril*. Edinburgh: Birlinn.

Grice, H.P. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds), *Syntax and semantics 3: speech acts*. New York: Academic Press, 41-58.

Leech, Geoffrey. 1983. *Principles of pragmatics*. London: Longman.

Santini, Marina. 2006. Web pages, text types, and linguistic features: Some issues. *ICAME Journal* 30, 67-86.