Using online electronic newspapers in modern English-Language press corpora: Benefits and pitfalls

Tobias Rademann
Ruhr-University Bochum

1 Introduction

Newspapers have always provided a welcome basis for linguistic analysis. It is not only that considerable parts of major corpora such as the BNC, LOB or The Bank of English, for instance, draw heavily on newspaper articles; in recent years, the number of smaller projects where newspaper-based corpora are employed in order to investigate certain aspects of language (change) has risen considerably as well[1]. Although newspapers do obviously not present a reliable sample of an *entire* language, their high popularity with respect to linguistic corpora has different reasons: besides the fact that they are easily accessible, their most important advantage is that there is hardly any other domain which offers such a broad number of linguistically distinctive varieties (Crystal 1994: 388), because they contain, among numerous other text types, leaders, essays, reviews, columns and even advertisements and cartoons.

It is this significant variety within the genre itself that makes a collection of newspaper articles a much more representative sample of a given language than most others. In addition, newspapers can usually be classified according to social aspects (eg quality vs popular press) and regional aspects (eg UK vs US), thus enabling researchers to conduct synchronic studies investigating social and regional differences in language use. Furthermore, since at least the major newspapers tend to have considerably large target audiences, the language used in newspaper articles is often assumed to be characteristic of the respective period and society they are published in. Thus diachronic comparisons, ie studies investigating how far a given linguistic feature has changed over time, typically rely on historical newspaper corpora containing selected editions from different periods.[2]

This means, on the other hand, that *today's* newspapers present a valuable database for anyone interested in *current language usage.* And

although computer corpora containing current material can most advantageously be employed in lexicography, for instance, there is also a broad range of additional fields in which such corpora could be of great value. Released online or on CD-ROMs, they could be of assistance to the language student who does not know which context certain words may be used in or which connotations they have. They could also be employed by professional translators for highly specific text types, and they might even be taken as reference sources in machine translation.

In the past, one problem with constructing newspaper corpora has always been the way in which printed texts were transformed into electronic versions. As newspapers are typically printed on (recycled) paper or are only available in photocopied form, scanning has so far proved impossible as a result of the poor character quality, which means that articles have to be typed in manually. Even though this is only a somewhat minor technical problem, this transcription process tends to take up a considerable amount of precious time and manpower during the first stages of any project, and, in addition, it is always a source for potential mistakes, such as typing errors, missing words or lines, etc.

With the unequalled growth of the Internet, however, another medium has emerged, namely that of electronic online periodicals in general and electronic online newspapers (ENs) in particular. The following section will now present a brief overview over their most important characteristics, before their value for the construction of present-day English-language newspaper corpora will be discussed, laying special emphasis on the degree of conformity between printed and electronic editions of major English-language quality dailies. The third section will then be concerned with the various technical issues connected with employing online ENs in corpus construction, before the overall policies of some of the major publishing companies with respect to their electronic editions will briefly be outlined. Finally, the advantages and disadvantages of online electronic newspapers as compared to their offline counterparts (CD-ROMs) will be discussed. The paper closes with a critical summary of the benefits and pitfalls that arise when using online ENs in linguistic research and an outlook on likely future developments in this field.

2 Electronic Online Newspapers – An Introduction

Although electronic online newspapers can still be considered a comparatively unknown genre, there has been a steep increase in the number of newspapers that have begun to publish at least some of their material

on the Internet. Over the past four years, they have firmly established themselves as an essential source for retrieving news and information via the WWW, growing in number from 20 in 1993 and 100 by the end of 1994 to a current figure of more than 2,500 world-wide. Probably not least resulting from the decentralised structure of the Internet, there currently seems to be some disagreement as to the true number of ENs: while the E&P Directory of Newspapers (which has been heralded by the EU Newsletter *ISPO* and which claims to be 'the most comprehensive reference source of its kind') puts the number at 2,560 (cf Table 1)[3], its American counterpart, the American Journalism Review (AJR), lists more than 3,622 online electronic newspapers in its database.[4]

Table 1: Total Number of Online Newspapers in the E&P Interactive's Database[5]

Updated November 21, 1997

Total Number of online newspapers currently in the E&P Interactive's Online Newspaper Database: 2,560*

Number of Online Newspapers on the World Wide Web: 2,445

Online Newspapers by (parent company) Type	**Online Newspapers by Region** (These are WWW services only)
Dailies – 1484	Africa – 29
Weeklies – 687	Asia – 120
Business – 112	Canada – 139
Alternative – 71	Caribbean – 22
Publishing Groups – 119	Europe – 414
Specialty Papers – 80	Latin America – 149
News Magazines – 12	Middle East – 33
Online Commercial Services – 37	Oceania – 23
Dial-Ups (BBS) – 33	United States – 1591
Misc – 96	

It is of crucial importance, though, to realise that this overall number of online electronic newspapers is comparatively meaningless, since the quality of the individual Web sites tends to vary considerably with respect to both content and layout. Thus, especially for scientific analysis in general and corpus linguistics in particular, the number of potentially useful papers can be expected to be much smaller than the figures mentioned above might suggest at first sight. However, it must not be forgotten that, as this paper will attempt to demonstrate, even only a few dozen high-quality electronic newspapers will open up numerous new perspectives as far as corpus construction is concerned.

As the feature analysis in the next section will show, what can be considered a (high-) quality online electronic newspaper does not necessarily have to correspond to what was traditionally regarded as a (high-)quality printed paper, which means that the two terms 'quality online EN' and a 'quality printed newspaper' are obviously not to be used interchangeably. Nonetheless, given the high costs involved in establishing a quality online Web site and the low returns this currently yields, in combination with the fact that the current potential readership on the Internet is most likely to come from the educated classes, there is a tendency for almost all quality online Web sites to have a well-known printed quality paper as their 'big brother'. However, while the statement 'Most quality online ENs also have a *quality* print edition' seems to be true in most cases, one cannot conclude from it that all of the low-quality online ENs will necessarily be electronic editions of a *popular* printed newspaper: There are dozens of electronic online newspapers that, though originally belonging to the group of so-called quality dailies (or weeklies) in the print sector, provide only very poor online editions (cf eg *The Guardian (Weekly)* or *The Observer*, to name but two).

The following section will be concerned with establishing the features that make up (high-)quality electronic online newspapers, since it is those which can most profitably be employed for linguistic research. For the purpose of doing so, special emphasis will be put on the differences between online and print editions.

3 Establishing the Quality of Online ENs

Given the highly divergent publishing environments, one might argue that the features determining the quality of online newspapers would have to differ considerably from those that were traditionally employed for establishing that of their printed counterparts. However, while this

section will demonstrate that there are indeed several new issues to be taken into consideration in an electronic publishing environment, it will equally prove that those are only *additional* criteria and that any online edition can only be regarded as a quality publication as long as the vast majority of traditional standards are satisfied as well. Some of these issues are, naturally, of primary importance with respect to the main concern of this paper, ie the construction of linguistic corpora, whereas others are only of minor importance in this context and will consequently be mentioned here only briefly.

As can be inferred from Figure 1, there are three major issues that

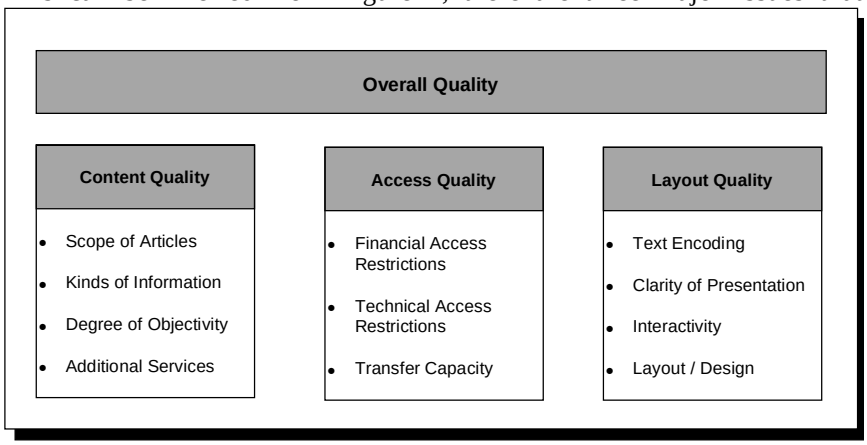

Figure 1: Criteria for determining the overall quality of online electronic newspapers

play a significant role in determining the overall quality of online electronic newspapers, namely the respective paper's content, layout, and access[6]. Of course, the various features subsumed under 'content' are of principal interest for linguistic research; but, as will be seen in the following paragraphs, at least from a technical perspective, the various access restrictions, for instance, also determine the overall use of a given EN to a considerable extent.

3.1 Content Quality

One of the main issues that affect a paper's content quality is indubitably the scope of its articles, which can be subdivided into the two features breadth and depth, where breadth refers to the number of subjects covered (eg politics, economics, sports, etc) and depth to the degree of detail the respective articles provide (ie whether they typically present in-depth analyses or superficial summaries). Quite in contrast to traditional printed newspapers, where the size and number of pages of a given issue heavily restrict the potential scope of articles, their electronic counterparts are subject to no such constraints at all. The fact that they

Figure 2: A typical newspaper article[7]

are published and distributed electronically implies that they can be as detailed as is regarded necessary for an adequate coverage of the respective topic, since the space on the publisher's hard disk (which would equal the number of pages in a print edition) is – at least theoretically – unlimited. Furthermore, the full-text version of any article in an electronic newspaper is usually displayed on its own and thus does not have to compete with others for the space on the screen (ie

the space on a page in print editions) (cf Figure 2).

It is especially the depth of articles that also comprises the amount of extra-textual information. And again, this is one area where the possibilities of the World Wide Web can brilliantly be employed to offer a considerably higher quality of service than the print sector ever could. The Web enables publishers to embed any kinds of multimedia elements, which comprise not only simple graphics or photographs, but also such sophisticated means as (three-dimensional) animations or even audio- and video-sequences, guaranteeing that the news consumer will get a multi-faceted account of the subject. Although these extra-textual items will usually be neglected in traditional linguistic research (especially in text-based corpus construction), their future importance in online newspapers may be a very interesting issue to watch.

Furthermore, an electronic publishing environment adds yet another dimension of in-depth reporting to online newspaper texts, since articles may also include hypertext references to related articles or even to the Web sites of third parties. Although this presents a valuable means of obtaining a fuller, ie often even more objective view on a given topic (since both the ongoing development of the case at hand and the opinions of the other parties can now be accessed as well), this service is only of minor importance for our interests.

As far as the breadth of articles published in online electronic newspapers is concerned, one can typically find news, comments, stories, columns, letters to the editor, science and technology, sports news, entertainment, education, book reviews, etc in any average-quality electronic newspaper. In addition, there are even some papers which already feature such sections as classifieds, advertisements, real estate, horoscopes, cartoons, etc. Although having largely been neglected in newspaper corpora so far, such features as advertisements and cartoons, for instance, present linguistically quite interesting categories. This results not least from the fact that they possess a considerable degree of importance with respect to various communicative issues (among others, both are typically used to convey information in a very condensed manner, which is often coupled with additional features such as irony, or requires a considerable degree of culture-specific background information on the part of the reader). However, it must be said that the sections just mentioned will most likely remain exceptions in online ENs until there is a sufficiently large target audience to justify the expenses necessary for offering comparable services (cf below for more details).

Although the scope of articles published in electronic newspapers

certainly plays an important role in constituting the framework for the content quality, it obviously does not say anything about the journalistic value of the articles as such (which is obviously of primary importance). When investigating this issue, one must bear in mind that today's potential online target audience still differs considerably from one which could – theoretically – buy a printed newspaper. The vast majority of today's Internet users consists of the educated upper and middle classes, with students and international business executives probably being the two largest groups at present (at least as far as the European market is concerned). Apart from these sociological aspects, one other important issue has to be kept in mind when talking about the (potential) target audience of online ENs: resulting from the fact that the Internet is a global network, there are no constraints as far as the regional availability of an electronic newspaper is concerned. Every user connected to the Internet – regardless of his geographic location – may now access the Web site of any electronic newspaper at all times, thus making the entire Internet community world-wide its potential target audience. For determining the target audience of online ENs, one also needs to bear in mind that, since the dominant language in international communication is English, the newspapers which one can expect to find on the Internet (at present!) will most likely be comparable to what in the print sector has been known as (high) quality English-language international dailies. Papers that are published in any other language do not yet have a sufficiently large target audience, which means that they will not play an important role until a much larger share in population has Internet access (see right column in Table 1). Finally, resulting from the typically highly culture-specific content of most popular press newspapers, these cannot be expected to go online in large numbers in the near future.

As far as the construction of English-language press corpora is concerned, it thus follows that, while any linguistic research project based on quality papers will greatly benefit from the existence of online electronic newspapers, it will probably not be possible in the next few years to use the Internet for corpora which require articles taken from many different popular press papers or locals. Furthermore, it should be mentioned that, although there are no limits with respect to the regional availability of online ENs, constraints are, of course, put on the content as a result of a paper's regional, national, or ethical focus. As paradoxical as this might seem for the Global Village at first sight, on the Internet, too, newspapers can still be categorised in international, national, regional, or local papers – mainly resulting from the fact that the vast majority

of them have (well-known) printed counterparts.[8]

Although most of what can be summarised as 'additional services' is of hardly any value for the purpose of this study, there is one feature of online electronic newspapers that can be of great benefit for constructing linguistic corpora, namely their electronic archives. Being one of the most powerful tools in an electronic publishing environment, online archives enable the scientific researcher to search a paper's database for articles matching a certain keyword-based query (eg *USA Today*) or even to browse entire back issues of a given paper (eg *The London Times*, *The San Francisco Chronicle*). These archives typically cover between two weeks to one month (eg *The Chicago Tribune* or *The Washington Post*), but – given the current technological developments – it is most likely that archives for entire years can soon be accessed online (the electronic archive of *The London Times*, for example, already comprises every issue of the paper from January 1st, 1996, and the back issues of *The San Francisco Chronicle / Examiner* can be retrieved to as early as January 1st, 1995)[9]. Obviously, these archives are of importance for corpus construction in that they grant the researcher instant access to either certain issues (eg all Thursdays) or to articles on a given topic (eg elections, stock exchange reports, etc). At least as far as the major archives are concerned, this service means that the researcher does not have to wait for weeks once he knows which kinds of articles he wants to focus on, and does not have to check the Web site every single day in order not to miss out on a certain report.

3.2 Access Quality

There can be no doubt that the overall content quality of a given EN will be the crucial factor in determining its suitability for linguistic research in general and corpus construction in particular. However, as has been hinted at in the introductory passage of this section, there are also two other areas that have to be taken into consideration, namely the 'access quality' and the 'layout quality'. Even though it might be argued that issues concerning the access quality of an online electronic newspaper are no relevant criteria for establishing whether or not a given online EN is to be included in a corpus, it is the aim of this study to discuss the overall suitability of this new genre for its use in corpus linguistics. And since it would not make sense to recommend the use of online ENs in the construction of press corpora if these ENs were not readily accessible for the vast majority of researchers, these

issues will have to be investigated more closely in a paper like this.

All things considered, the 'access quality' comprises three major issues: financial access restrictions, technical access restrictions, and finally the site's transfer capacity (bandwidth). A fact which is quite encouraging, as far as the employment of online ENs for academic research is concerned, is that most of today's editions are available free-of-charge. Nevertheless, there are already companies that charge users for accessing their sites (eg *The New York Times* or *The Wall Street Journal Interactive Edition* [*WSJIE*]). Subscription rates for electronic newspapers tend to lie well below those of their printed counterparts (*The WSJIE*, eg, currently charges US$ 49.— per annum for those who do not subscribe to the paper's printed edition and US$ 29.— per annum for those who do); a clear exception is *The New York Times*, though, the fee of which is extremely high, namely US$ 35.– per month.[10] However, when comparing the services offered by fee-based quality-online ENs [eg *The WSJIE*, *The NYT*] to those that are offered free of charge [eg *FT.com*, *The London Times*, *USA Today*], hardly any difference in service and content quality is observable.

There is no widespread agreement as to how this scenario will change in the future, as the Internet develops into a genuine info-communications mass medium. On the one hand, it could be argued that, by granting interested readers free access to their sites in today's developing stage, publishers can gather valuable experience for later projects to be implemented when the general public is ready for the true onset of the Information Age – a point at which most publishers would be likely to bill their customers for accessing their pages. However, it could also be maintained that by continuing to allow free access to their sites, the number of visitors will remain much higher, which, in turn, will enable publishers to charge more for on-site advertising.[11] Obviously, it is especially in an academic environment, where funds are getting smaller every day, that no – or at least comparatively low – access fees present an interesting option.[12] If one bears in mind the costs (and manpower in some cases) involved in either buying international editions of the major quality dailies or even of flying to London in order to photocopy newspaper articles from the national archive, the potential savings are enormous.

3.3 Layout Quality

As an intersection between the access quality and the last point to be discussed in this chapter, namely the layout quality, the way in which articles are encoded plays another important role with respect to the suitability of online ENs for corpus linguistics. Although there are numerous ways in which articles may be made available in an electronic publishing environment, the three most common are (a) as plain ASCII text (eg in an eMail), (b) as HTML (Hypertext Mark-Up Language)-encoded text (on the WWW), and finally the most recent alternative (c) as Adobe's PDF (Portable Document Format)-files (also on the WWW and special retrieval programs). Let alone the fact that the various publishing formats are usually closely associated with a certain distribution method (see brackets above), there are various other issues that are of interest when it comes to working with them in corpus linguistics. While ASCII-based articles can only contain plain text made up of the 256 ASCII symbols, it is a big advantage of HTML that not only text, but all kinds of multimedia elements can be embedded in one file. Nevertheless, despite its numerous advantages, there are some considerable disadvantages connected to the use of HTML as well, the two most important of which are the lack of author control with respect to the article's layout (every browser will display a given article differently, among other things depending on the screen size or the preferences the user has entered for features such as font and character size) and the fact that any HTML-file can easily be copied, altered in form and content, and re-distributed by anyone possessing only some basic understanding of HTML programming. Although PDF-files are only beginning to emerge as a publishing format on the Web, their most important advantage consists in an increased author control over the respective documents with regard to layout and content. It is extremely difficult at present to modify both form and content of files published in PDF, which makes it easy to include copyright notices within the individual documents and ensures that the content cannot be redistributed in a modified way.[13]

All things considered, the quality analysis of online electronic newspapers in the passages above has demonstrated that it is important to realise that, in order to determine the overall quality of online ENs, it is necessary to investigate several additional criteria which mainly result from the electronic publishing environment. However, as has been shown, a considerable number of (high-) quality electronic newspapers have been made available on the Internet over the past few years, which indubitably present valuable alternatives for constructing English-language

press corpora; there are even some ways, in which this new genre opens up whole new perspectives for the research in this field, eg by means of global availability at hardly any costs, electronic archives, and so on.

4 Corpus Construction and Electronic Newspapers – Technical Issues

While the preceding section has demonstrated that the employment of online electronic newspapers can be of great benefit for the construction of press sections of modern English-language corpora, it is the task of this chapter to give some insight into the various technical issues that have to be taken into account when working with electronic newspaper articles. Among others, these mainly concern three major steps involved in preparing texts for their inclusion in a corpus, namely (a) finding suitable texts, (b) saving them on a local HDD, and finally (c) editing them.

It has been illustrated that there are at least three common ways of publishing electronic newspaper texts online at present: plain ASCII, HTML, and Adobe's PDF. Obviously, all of these require different steps as far as the their preparation is concerned. However, as will be explained below, only HTML-files have to be discussed at some greater length: ASCII files are typically sent via eMail, which means that no 'finding' will be required here. In addition, ASCII source code is precisely what is required as input for most corpus projects, which also implies that ASCII files do not have to be edited. As the text of PDF files cannot be extracted, these newspaper versions are useless for our purposes, and thus do not have to be discussed here either.[14] Furthermore, these two publishing methods only account for a minimal fraction of online electronic newspapers at present, and it can be expected that this will remain to be the case in the future. Consequently, the following passages will centre on the different steps involved in preparing articles published in HTML for their inclusion in corpora.

Primarily depending on the underlying interests, there are two major ways in which texts in electronic newspapers can be 'found'. Obviously, the first is to browse the respective newspaper's Web site, a process which roughly corresponds to browsing through a printed copy by skimming the different articles page-by-page. However, the layout of Web-based newspapers and their printed counterparts differs considerably from one another: A printed newspaper might have between 50 and 100

pages per copy, with several articles appearing on the same page, whereas an EN might feature links to several thousand articles which will all be displayed on separate screens (in the Web edition of the *WSJIE*, for instance, an average of 44,000 different articles are published every new day). Thus, reading an online edition by browsing through it is usually only of interest for the very first time a researcher visits the site in order to get a basic impression of the respective paper's structure and (likely) content. Should he be interested in retrieving a certain kind of text, such as that day's leader(s) or, for instance, a column, he will usually not have to browse through the whole paper but can simply bookmark the respective section page (eg 'columnists') which will include a direct hyperlink to the required article. A major advantage of electronic newspapers can be exploited, if the task is to retrieve all texts on a given topic. Here, it will be most useful to refer to the on-site search engine which is included in every (major) online newspaper: the researcher can conduct a keyword search on the topic of his interest, searching either the current or even back issues, depending on the size of the individual archive.

Independently of the method employed to retrieve the text via the Web, the result will always be a HTML-based file displayed on a user's screen. In order to be able to use it for future work, it is now obviously necessary to save and edit it on the local hard disk (HDD). Although the saving of articles appears to be a comparatively simple task at first sight, there are a few things that have to be kept in mind:

First of all, it is necessary to distinguish between frame-based and non-frame-based Web sites. For frame-based Web sites (such as that of *The London Times*), it is important that the user clicks in the frame that includes the article before he saves it, otherwise the menu frame, for instance, might be saved instead. Furthermore, the user has to decide whether he wants to save the file as a HTML- or a TXT-file, an issue which is of particular relevance for the editing process. Although one would intuitively assume that TXT-files are the better choice for our purposes, both alternatives have their advantages and disadvantages. While the latter only contain ASCII characters without any of the additional material mentioned above, the original paragraph formatting will be lost, since the conversion process typically results in paragraph breaks (¶) being automatically inserted at the end of every line displayed on the screen.[15] Should the former be chosen, on the other hand, the output file will include all of the HTML-tags. Since a typical HTML-file will look like the following extract of the text in Figure 2:

<h2>Japan: GDP falls 2.9% after slump in consumption</h2>FRIDAY SEPTEMBER 12 1997<hr>By Michiyo Nakamoto in Tokyo<hr><img src="12w404.gif" alt="Japanese GDP" align=right border= 0> Japan's gross [...]

it is especially for the purpose of integrating texts taken from online electronic newspapers in linguistic corpora that their 'text-only' versions are required; ie they will typically have to be converted to plain ASCII format, eliminating (stripping) all 'unnecessary' information such as graphics, formatting, HTML tags, and so on. Fortunately, quite a number of converters have been introduced lately that will assist the user in removing these tags from the source and save the file as a plain ASCII text file.[16] Once it has been saved on the local HDD, the output can be modified to fit the individual requirements of the corpus, which means that the file could be tagged or parsed, for instance.

While such plain ASCII text files are precisely what is required for further linguistic analysis, every researcher will want to keep a (backup) copy of the original files in his records. With printed newspapers, this is frequently a photocopy of the respective article. Of course, this can be done for electronic articles as well. As long as one has a high-quality laser printer at hand, every article can be printed out and archived as such. Nevertheless, given the rising quantity of multimedia elements included in online reports, a considerable amount of (extra-textual) information will be lost if articles are archived in printed form only. If the researcher wants to keep copies in his archive that contain all of the original information (including such features as animations, sound and video sequences, and so on), the texts would have to be saved the way they were published in the first place – electronically. Nevertheless, the process of obtaining fully compatible electronic copies will require quite a lot of additional post editing. It is not sufficient to save the HTML source only, since it merely contains links (ie references to additional items such as graphics, photographs or sound and video sequences) and not the objects themselves. Thus, saving just the HTML source on the local HDD will result in the text being displayed with dummies for every embedded object (see Figure 3). Consequently, the user will have to retrieve and save all of the embedded multimedia elements of such files manually. For the purpose of doing so, the user needs a good understanding of HTML because he has to open the article's source code and look for the directories on the host computer

(ie the paper's Web site) in which the various multimedia elements have been stored. Finally, he has to change all these directories in the source code of the local copy to match those on his HDD – a task which can be rather time consuming.

It can thus be summarised that, while there are indeed some issues that have to be kept in mind when preparing electronic articles (especially HTML-files) for their use in corpora, this process does not require too much work and experience. However, it has been shown as well that if

Figure 3: Figure 2 as saved on a local HDD without its multimedia elements

one chooses to keep the electronic file *including* all its multimedia elements, this can become quite a difficult task, requiring both a considerable degree of experience in handling HTML and a lot of time.

5 Policies of Online Electronic Newspapers

Apart from all these technical issues, one of the more important questions that will probably be of interest for those researchers thinking about using online electronic newspapers for their press corpora is how far the content of online editions matches that of their printed counterparts.

In a recent eMail survey answered by the editors of about a dozen quality English-language Internet dailies, the following results were obtained. The subsequent quote from the Managing Director of News International Publishing (*The London Times / The Sunday Times*) sums up quite nicely the tenor of the answers received in response to the survey:

> Our policy is to reproduce all of the content of the published versions of *The Times* and *The Sunday Times* in our Internet editions. [...] I would estimate that we publish something of the order of 97 per cent of the text and 65 per cent of the photographs of our papers. Because our text is crucially important, we feel it vital that the reader in any country should be able to see the 'same' paper, presenting the particular view of the world from the offices of *The Times*, as the reader who actually bought the printed version. Certainly that will continue to be our policy for the immediate future, and we hope both to complete the content we cannot use at the moment, and to augment it with more text outside the papers themselves.[17]

As can easily be inferred from this quote (which is representative of most others that I received), many of the quality papers already publish (almost) the full content of their print editions online, though there is, of course, also a considerable number that post only 'selected highlights' (*The Age* [Australia], for instance). As far as the content of online editions is concerned, it appears that the vast majority of Internet editors does not intend to modify them (eg *The London Times*, *USA Today*, *The Wall Street Journal Interactive Edition*), except for such minor features as embedding additional links to older articles or third-party Web-sites (cf eg *The Washington Post*).

What is striking, however, is that many editors have either already enhanced the content of their Web editions or are determined to do so in future. *USA Today*, for instance, clearly states that 'our print version is a subset of what can be reached online', and *The WSJIE* 'contains everything that is in *The Wall Street Journal*, along with additional articles and features unavailable in the print Journal.' These additional features typically comprise such issues as 'live coverage [...], interactive reactions [...], and news updates' (*The Age*). However, at least as far as the continuous updates are concerned, there are also papers that pursue a different approach, as for example *The Washington Post*, which

takes care that their 'on-line publishing is timed not to preempt the print paper.'

6 Online Newspapers vs CD-ROMs

In recent years, some of the major English-language quality dailies, such as *The Times*, *The Guardian*, *The Independent*, and *The Financial Times*, have begun publishing CD-ROMs containing the complete editions of the respective paper for a whole year in electronic form.[18] As useful as some of these editions have turned out to be for the purpose of constructing press corpora, there are also considerable disadvantages connected to their employment in this field, especially with respect to the following issues. One of the most important disadvantages to almost any newspaper CD-ROM – particularly in today's academic environment where adequate funding is increasingly difficult to receive – is the fact that they are usually quite expensive: prices for CD-ROM editions range from £395 for the 1997 subscription of *The Guardian* to as much as £950 for the 1997 subscription of *The Financial Times*. In addition, one has to take into account that these editions contain 'the text of the newspaper with certain exclusions' only, which means that there will be no copyrighted material, letters or advertisements, for instance. As far as technical issues are concerned, most search interfaces are extremely poor, not allowing for multi-word searches or for queries containing so-called stop words, which presents a major disadvantage as far as linguistic analysis is concerned. It is another essential drawback of this medium that the number of quality dailies which go through the effort of publishing such versions is confined to about a dozen of the major papers. Thus, no CD-ROMs have so far been released of popular newspapers, for instance. Furthermore, it has to be kept in mind that CD-ROM versions of newspapers are a medium which is only available in industrialised nations; with respect to other countries, such as India, where English-language newspapers are quite common (even on the Web) and could present valuable additions for some corpora, no CD-ROM editions exist of the papers that are published in these countries. Finally, two further issues have to be taken into consideration when working with CD-ROMs, though these are only of minor importance for any linguistic analysis, namely the fact that the original article layout is usually not maintained and that there is (obviously) a considerable temporal delay between the date of issue of the printed version and the release of the CD-ROM (ie in the worst case, the paper of January 1st

of a given year will not be available on CD-ROM before some time the next year).

As can be inferred from the passages above, Internet versions of newspapers (as long as they exist, of course) have significant advantages over their offline counterparts, CD-ROMs. First of all, they will almost certainly be much cheaper; secondly, there is also a greater diversity, both from a regional / national and a content-related point of view; thirdly, there is no temporal delay as far as their release is concerned – they are typically available in the early morning of the day that the printed paper is published. In addition, online editions often provide extra services and material, though these are not of primary interest for linguistic analysis. However, CD-ROM editions, too, do have certain advantages as compared to their online counterparts. First of all, there is not much time involved in retrieving articles, though in most cases some post-editing has to be done in order to prepare them for inclusion in a corpus. In addition, the researcher does not depend on the temporal scope of the electronic archive; it has been mentioned above that most online archives currently contain up to a month of back issues, which might not be sufficient for someone preparing a larger corpus. If a researcher buys a CD-ROM edition, he has got the full past 12 months at his immediate disposal. Finally, CD-ROM editions might be more comprehensive than most Internet editions.

7 Conclusion

This paper has demonstrated that the new genre of online electronic newspapers that has emerged on the World Wide Web over the past four years can already be regarded as a valuable new medium for constructing press sections of modern English-language corpora. The preceding sections have illustrated that quality online ENs do not only present interesting alternatives to their printed counterparts, but also that they open up whole new perspectives as to what can be done in corpus linguistics. While lexicographers now have a vast amount of easily accessible up-to-date material from a broad number of linguistically distinctive varieties at their disposal, online ENs also enable researchers to conduct projects which would have been rather difficult or troublesome so far. They can compare variations in linguistic features across English-language quality dailies from different nations of the world (eg the US, UK, Australia, India) or between national and international editions of a given paper without having to go through much trouble for obtaining

the required copies. In addition, language students and even professionals can use these corpora to improve their understanding of the collocations and connotations of English vocabulary, for example, if they use up-to-date corpora in connection with such tools as OUP's WordSmith. Finally, tailor-made special-purpose corpora can be built much faster – and far cheaper. Corpora consisting of language used in certain fields, such as in stock-exchange or sports reports, can be constructed, and even thematic corpora (eg elections or such emotional events like the death of Princess Diana) including material from all over the world.

Although it cannot be expected that the genre of online ENs will continue to grow as fast as it has over the past year(s), the increasing number of companies and private individuals that do not only have access to the Internet but will also get more and more used to employing its resources for their own purposes, should guarantee a continuously rising target audience, which will, in turn, secure at least the major publishing houses sufficiently large advertising returns to keep their Web services running. In my opinion, it will be one of the most interesting issues to watch over the next few years, though, how far the fact that the WWW as a medium allowing a simultaneous transmission of information which has so far had to be sent via two different channels (newspapers and television / radio) will influence the use of language and certain linguistic features, and how far there will be a convergence between what we now still refer to as newspapers and such Web sites as, for instance, that of CNN.

Notes

1 For some interesting projects see, among others, Diller (1993), who investigates the personal styles of selected British columnists, Schneider (1997), whose study is concerned with the construction of a historical newspaper corpus to analyse the changes of several linguistic features over time, or Reynolds & Cascio (1997), who have used a newspaper corpus for the purpose of establishing how far the use of contractions in printed English has become more wide-spread in recent years.

2 See Kytö (1993) for a summary of various projects or Schneider (1997) and Hundt (1997).

3 E&P Directory of Newspapers (1997). See also European Commission (1997).

4 See AJR News Link (1997a) at http://www.newslink.org/news.html.

5 Table taken from E&P Directory of Newspapers (1997), http://www.mediainfo.com/ephome/npaper/nphtm/stats.htm.

6 While the first two features can be applied to printed newspapers in a similar fashion, 'access' here refers to the various issues concerned with accessing the Web sites of the respective EN, such as bandwidth and technical / financial access restrictions (see 3.2 for some more details).

7 Figure taken from the Web site of The Financial Times, FT.Com.

8 It will be one of the most interesting issues to observe over the next few years if the possibilities offered by electronic publishing will lead to the traditional classification scheme of newspapers (ie the regional approach) being abandoned in favour of one that is more suitable for an electronic environment (ie one where the primary distinctions between newspapers are based on the interests of their target audience rather than their regional availability).

9 Unfortunately, it can be expected that accessing such databases will probably not be free of charge as their maintenance tends to be comparatively expensive. There are already quite a few highly sophisticated commercial databases, containing thousands of recently published newspaper articles as well as valuable past files; some outstanding examples are NewsNet, Dow Jones, Lexis-Nexis, and Dialog (for more information see Notess 1996). In addition, *USA* Today's archive has just been transformed into a pay-per-article database as well. While it now enables researchers to access articles from as early as 1987, users have to pay US$ 1.- for every article they download from the paper's archive.

10 For a more detailed value-for-money analysis of several individual newspapers see Rademann (1996: 26–41). However, given the costs involved in connecting to the paper's Web site in addition to the restricted readability (only in front of the PC), rates such as that charged by The New York Times do by no means seem justified – even though the NYT has won many awards, among others the two top honours in the 'World's Best Online Newspaper Awards' (see E&P 1997 for more details) and in the American Journalism Review's 'Top 50 news sites' (cf AJR 1997b and Meyer 1997b).

11 See Meyer (1997a), for instance.

12 Although one could differentiate between access and copyright fees, the latter only arise when one retrieves articles from third-party archives, such as Document Delivery Services. At present, there is no sign that most online newspapers would plan to charge any

copyright fees for texts retrieved from their own Web sites, as long as they are used for 'personal, noncommercial use'.

13 Since the three features 'clarity of presentation', 'Interactivity', and 'Layout/Design' are of hardly any interest for the purpose of this study, they will not be discussed here. The interested reader is referred to Rademann (1997) for a more detailed analysis.

14 It should be added, that PDF is most likely to be used for scientific publications rather than news reporting, since such issues as copyright protection and layout control are much more important there.

15 However, these additional line breaks can easily be removed by executing three 'edit/replace' commands (eg '¶¶' => '*'; '¶' => ''; '*' => '¶').

16 All major HTML editors (such as eg the HTML Assistant Pro) have an option that will allow the user to strip the HTML tags off a given source. It has to be mentioned, though, that it is especially the more recent developments like Java-Script code, for instance, that sometimes cause some confusion among those converters, which means that the output for selected files should be checked manually in order to make sure that the converter does work for the respective task. Depending on the task the corpus is used for, the HTML tags might even remain in the files, because some of the more sophisticated tools used in linguistic analysis of texts, as for example Oxford University Press's WordSmith Tools, can be configured to ignore them.

17 Murphy (1997).

18 Actually, in the UK it is not so much the individual papers but Chadwyck-Healey Ltd who pub-lishes most of these CD-ROM editions.

References

AJR News Link. 1997a. Newspapers. November 4th – 9th. http://www.newslink.org/news.html

AJR News Link. 1997b. The year's Top 50 news sites. *AJR News Link online edition*. May 6th. [http://www.newslink.org/bestresults.html]

Crystal, David. 1994. *The Cambridge Encyclopaedia of Language*. Cambridge: Cambridge University Press.

Diller, Hans-Jürgen. 1993. British Columnists – why study them, and

how? *Anglistik and Englischunterricht (A&E)*. Vol. 51, 7 – 25.
E&P (The Editor & Publisher Co.) Directory of Newspapers. 1997. *Online Newspaper Statistics*. http://www.mediainfo.com/ephome/npaper/nphtm/stats.htm
European Commission. 1997. More than 1,600 newspapers online. *ISPO – Information Society News*. No. 11. March 1997, p 4.
Hundt, Marianne. 1997. The Press Sections of Standard One-Million-Word Corpora. Paper held at ESSE/4 – The European Society for the Study of English. 4th Conference, Debrecen, Hungary September 5–9. To be published in *Anglistik und Englischunterricht (A&E)*. Fall 1998.
Kytö, Merja. 1993. *Corpora Across The Centuries – Proceedings of the First International Colloquium on English Diachronic Corpora*. M. Kytö, M. Rissanen and S. Wright (eds). Amsterdam: Rodopi.
Meyer, Eric K. 1997a. An Unexpected Wider Web for the World's Newspapers. *AJR News Link online edition*. http://www.newslink.org/emcol10.html
Meyer, Eric K. 1997b. Best of the Web? It Depends on Your Perspective. *AJR News Link online edition*. May 6th. http://www.newslink.org/best.html
Murphy, Mike. 1997. eMail response to a survey on the Internet policies of major English-language quality dailies.
Notess, Greg. R. 1996. News Resources on the World Wide Web. *DATABASE Magazine*. Feb / March 12–20. http://www.onlineinc.com/database/FebDB/notess2.html
Rademann, Tobias. 1997. Electronic Newspapers on the Internet – An Introduction. *TEXT Technology – The Journal of Computer Text Processing*. Vol. 7, No. 2 (Summer 1997), 118 – 139.
Rademann, Tobias. 1996. *Cultural Studies Online. Information Resources in English-Language Electronic Periodicals*. Ruhr-University Bochum. Department for English & American Studies. http://www.ruhr-uni-bochum.de/www-public/rademtbu/private/pubs/ep.pdf
Reynolds, Mike and Giovanna Cascio. 1997. It's Short and It's Spreading: The Use of Contracted Forms in British Newspapers – A Change Under Way. Paper held at ESSE/4 – The European Society for the Study of English. 4th Conference, Debrecen, Hungary September 5–9. To be published in *Anglistik und Englischunterricht* (A&E). Fall 1998.
Schneider, Kristina. 1997. Exploring the Roots of Popular English Jour-

nalistic Style: A Preliminary Report on a Corpus-based Project. Paper held at ESSE/4 – The European Society for the Study of English. 4th Conference, Debrecen, Hungary September 5–9. To be published in *Anglistik und Englischunterricht* (A&E). Fall 1998.