The METER corpus: A corpus for analysing journalistic text reuse

Robert Gaizauskas†, Jonathan Foster‡, Yorick Wilks†, John Arundel‡, Paul Clough†, Scott Piao†

Departments of Computer Science† and Journalism‡
University of Sheffield, Sheffield, S1 4DP
(contact: R.Gaizauskas@dcs.shef.ac.uk; fax:(0114) 222 1810)

Abstract

As a part of the METER (MEasuring TExt Reuse) project we have built a new type of comparable corpus consisting of annotated examples of related newspaper texts. Texts in the corpus were manually collected from two main sources: the British Press Association (PA) and nine British national newspapers that subscribe to the PA newswire service. In addition to being structured to support efficient search for related PA and newspaper texts, the corpus is annotated at two levels. First, each of the newspaper texts is assigned one of three coarse, global classifications indicating its derivation relation to the PA: wholly derived, partially derived or non-derived. Second, about 400 wholly or partially derived newspaper articles are annotated down to the lexical level, indicating for each phrase, or even individual word, whether it appears verbatim, rewritten or as new material. We envisage that this corpus will be of use for a variety of studies, including detection and measurement of text reuse, analysis of paraphrase and journalistic styles, and information extraction/retrieval. To illustrate these potential uses we briefly describe some work we have done with the corpus to develop algorithms for detecting text reuse.

1. Introduction

The aim of the METER (MEasuring TExt Reuse) project[1] is to investigate how text is reused in the production of newspaper articles from newswire sources and to determine whether algorithms can be discovered to detect and quantify such reuse automatically. It is to be hoped that results will generalise beyond the newspaper-newswire scenario and provide broader insights into the nature of text derivation and paraphrase; but the newspaper-newswire scenario provides an ideal initial case study, and one with considerable potential practical application – see below.

To assist in this study it was necessary to create a comparable corpus[2] consisting of a selection of newswire texts and newspaper articles reporting the same stories, in some cases derived from the newswire texts and in some cases not. Because the Press Association, the major British domestic newswire service, is a collaborator in the METER project and have provided us with unrestricted access to their newswire service, we have used their archive as the source newswire for our corpus and texts from a variety of their subscribers in the British press as the candidate derived texts. Having assembled the corpus and annotated it to assist in our study of text reuse, we believed the corpus would be of wider interest to the corpus linguistics, natural language processing and language engineering communities, and hence decided to package and release the corpus on its own. This paper describes the design, structure and contents of the corpus, and illustrates its potential by briefly describing some experiments we have carried out using it. The METER Corpus is available free-of-charge for research purposes.

It should be stressed that the METER corpus is a pioneering corpus for the study of text reuse and that as such is no doubt flawed and limited in various ways. Resource limitations have meant limiting the size of the corpus and the amount of interannotator verification carried out on the annotations. Ideas as to how it should be annotated continued to evolve during the process of annotation, which means that complete consistency across annotations has probably not been achieved. Our hope is that despite these

[1] For further details of the METER project, see: http://www.dcs.shef.ac.uk/nlp/funded/meter.html.
[2] Johansson *et al*. (1996: 3) define comparable corpus as: "corpora consisting of parallel original and translated texts in the same languages".

limitations the corpus will still prove useful to others, even if only as a starting point for designing a better resource.

2. Text Reuse in the British Press

The Press Agency (PA) is the national news agency for the UK and Ireland. It provides regional, national and international news 24 hours a day, 365 days a year to its media customers throughout Britain. On a daily basis, the PA sources 1,500 news, sport and feature stories and one hundred news and sport photographs to the newspaper industry. The PA also supplies listings information in, for example, finance, arts and entertainment and television. In addition, the PA supplies text to specific news and sport internet websites, as well as providing news copy and information for other leading commercial and public sector organisations. Those using its services include national and international newspapers, regional morning and evening papers and terrestrial radio and television broadcasters. Thousands of weeklies, periodicals and magazines also receive various types of PA output.

Through its services, the PA quite clearly performs a critical function for the British media, whether they operate in print, audio or electronic forms. Because of its ongoing supply of both domestic and national news, it is clear that the PA has a critical role in setting the news agenda through its capacity to distribute salient information rapidly to news organisations. The PA, which has been operating since the nineteenth century, is in a unique position in British media industry and is widely regarded as a most credible, authoritative and trustworthy journalistic source for the newspaper and broadcast enterprises.

Being a primary supplier, the news issued by the PA is widely reused, either directly or indirectly, in British newspapers. Even if not directly using the PA copy, journalists invariably refer to the agency's newswire service during report production. This ensures verification of the 'facts' in a story and also facilitates effective 'copy tasting' decisions (copy tasting is the process of assessing which of all the news available at a given time should be included in the current newspaper edition). Therefore, rich "real" examples of text reuse can be expected from a collection of PA copy and related newspaper articles.

The study of text reuse has, aside from its intrinsic academic interest, a number of potential applications. Like most newswire agencies, the PA does not monitor the uptake or dissemination of copy they release because tools, technologies, and even the appropriate conceptual framework for measuring reuse are unavailable. For the PA, potential applications of accurately measuring reuse of *their* text include: 1) monitoring of source take-up to identify unused or little used stories; 2) identifying the most reused stories within the British media; 3) determining customer dependencies on PA copy, and 4) new methods for charging customers based upon the amount of copy reused. This could create a fairer and more competitive pricing policy for the PA[3].

3. Construction of the METER corpus

The texts of the METER corpus were collected manually from the PA online service and the paper editions of nine British newspapers – *The Sun*, *Daily Mirror*, *Daily Star*, *Daily Mail*, *Daily Express*, *The Times*, *The Daily Telegraph*, *The Guardian* and *The Independent*[4].

Building a general newspaper corpus was beyond the time and resource limitations of the METER project, so we have limited the corpus to just two domains: British law court reporting and show business stories. Court stories were chosen because of the substantial amount of data available in both newspapers and PA and because of their regular recurrence in British news. Court stories also revolve around "facts" such as the name of the accused, the charge and the location of a trial or inquest, with limited scope for journalistic interpretation. This information is found in both newspaper and PA reports, even when newspapers do not use PA as a source. Courts generally sit from Monday to Friday; therefore PA copy was collected for these days and newspaper versions of the same story appearing on the succeeding day were also chosen. Court cases reported by newspapers for which PA did not

[3]For a more comprehensive summary of the PA and public access to a portion of their newswire, see the PA website: *http://www.ananova.com.*

[4] The first five of these papers are published in tabloid format and are viewed as the "popular" press; the latter four are referred to as "broadsheets" and viewed as the "quality" press.

produce copy were ignored. Stories were collected for cases that lasted just one day, as well as for cases that stretched over much longer periods.

The other domain included in the corpus is show business and entertainment news. This was chosen to contrast with the court domain. Show business press exhibits a more expansive style, with greater freedom of journalistic expression and interpretation. Show business stories tend to be reported in a more frivolous, light-hearted manner. Like court reporting, show business news is also a stable and recurring feature of contemporary British news. The show business news stories, however, form a secondary collection to that of court stories and less data from this domain is included in the corpus.

Just as practical restrictions on corpus construction limited the scope of news domains included in the corpus, so too did they limit the temporal extent of the material included. Too narrow a date range might have lead to a biased sample, but too wide a range would lead to too much material. In the end we settled on gathering material from a one year period. The text collection spans 24 days for the law court reporting and 13 days for show business stories from 12 July 1999 to 21 June 2000.

PA stories are classified under a number of discrete, identifiable news categories, such as *Courts*, *Showbiz* (show business), *Politics*, *Education*, etc. Under each of these categories, stories are further classified into sub-categories, such as *Courts (Axe)*, *Courts (Strangle)* and *Courts (Gamekeeper)*. Each of the sub-categories refers to an individual story, incident or event. Such a sub-category is called a *catchline*. For each catchline, the PA follows its development and keeps releasing updated reports throughout a day. Each report occupies a single web page – termed a *PA page*. Therefore, a catchline contains one or more PA pages.

On each day in the study, all PA catchlines and associated pages relating to courts or show business were identified through the PA categories, *Courts* and *Showbiz.* The PA pages under a selected catchline were downloaded into separate electronic files. For each selected PA catchline, the final southern editions of the nine British national newspapers from the next day were then examined. The newspaper reports about law and court stories were compared against the PA catchlines and all those that the PA had covered the day before identified. These newspaper articles were then manually scanned into separate electronic files (and later manually examined and spell-corrected). Original paper copies were used in constructing the METER Corpus because web-published versions of the stories were neither reliably available nor, even if available, reliably identical to the published paper copy, which is still viewed as the definitive form of publication for a newspaper.

Table 1 gives general statistical information about the METER corpus. As discussed, the corpus consists of two main parts, *law and court reports* and *show business reports*. These contain 1,430 texts (458,992 words) and 287 texts (76,158 words) respectively – here we use the term *text* to refer either to a PA page or a newspaper article. In terms of the PA-sourced and newspaper-sourced texts, 773 PA pages (239,679 words) versus 944 newspaper articles (295,471 words) were included. Of the PA-sourced texts, 611 courts texts and 112 show business texts are associated with 205 *Court* catchlines and 60 *Showbiz* catchlines respectively.

Source	**Domain**								**Total**	
	Law and Court				**Show Business**					
	Words	**Texts**			**Words**	**Texts**			**Words**	**Texts**
		WD	**PD**	**ND**		**WD**	**PD**	**ND**		
PA	206,354	661 (205 catchlines)			33,325	112 (60 catchlines)			239,679	773 (265 catchlines)
Other	1,269	0	3	2	0	0	0	0	1,269	5
Times	34,794	24	41	46	2,966	5	7	2	37,760	125
Star	14,021	15	28	27	7,590	10	19	7	21,611	106
Express	21,956	17	27	18	5,270	1	8	5	27,226	76
Mirror	17,359	22	32	28	4,211	7	11	6	21,570	106
Mail	31,686	21	29	7	6,414	0	7	7	38,100	71
Guardian	38,499	12	46	37	3,805	4	6	3	42,304	108
Telegraph	45,768	30	62	35	2,985	6	7	2	48,753	142
Sun	18,597	18	37	15	6,010	4	24	7	24,607	105
Independent	28,689	7	37	46	3,582	2	7	1	32,271	100
Total	458,992	1,430			76,158	287			535,150	1,717

Table 1: Statistics of the METER corpus

The primary aim of constructing the corpus was to provide a resource for studying reuse of text between the PA and various subscribing newspapers, and not to provide a resource to study how the same story is handled differentially across the British press. Nevertheless, there are a significant number of stories in the corpus which are carried across multiple newspapers. Figure 1 illustrates the distribution of shared catchlines across newspapers. Most catchlines have associated with them only a single newspaper article; but more than ten in the courts domain are to be found in all nine of the newspapers represented in the corpus.

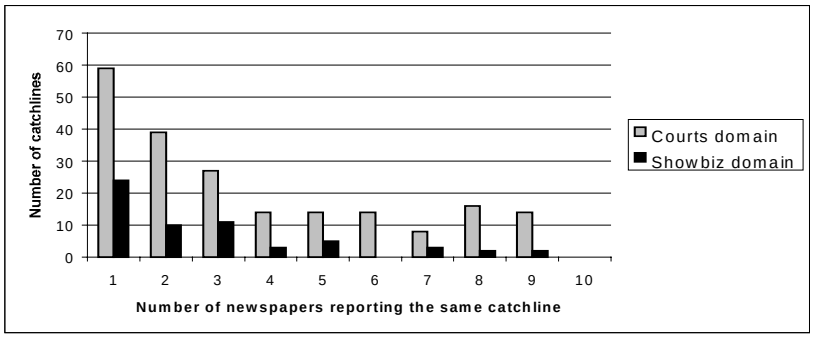

Figure 1: Distribution of shared catchlines across newspapers

The total effort expended in creating the corpus was about two person-years. Gathering stories, converting to electronic form and manual annotation took about one person-year. Subsequent OCR error checking and correction, organising the electronic texts into a structure, designing the mark-up scheme and electronic annotation took approximately another person-year.

4. Structure of the METER corpus

An important issue in corpus construction is deciding upon an appropriate structure in which to store the data, so as to facilitate human and machine access to it. In the current METER corpus, the texts (PA pages plus newspaper articles) are arranged in a tree structure. Texts are clustered according to their origin, topic and date of release. This structure provides a unique identifier for each text within the corpus. Figure 2 illustrates the overall structure of the METER corpus (lower level details are shown only for the *Courts* domain – the *Showbiz* part of the corpus is structured similarly).

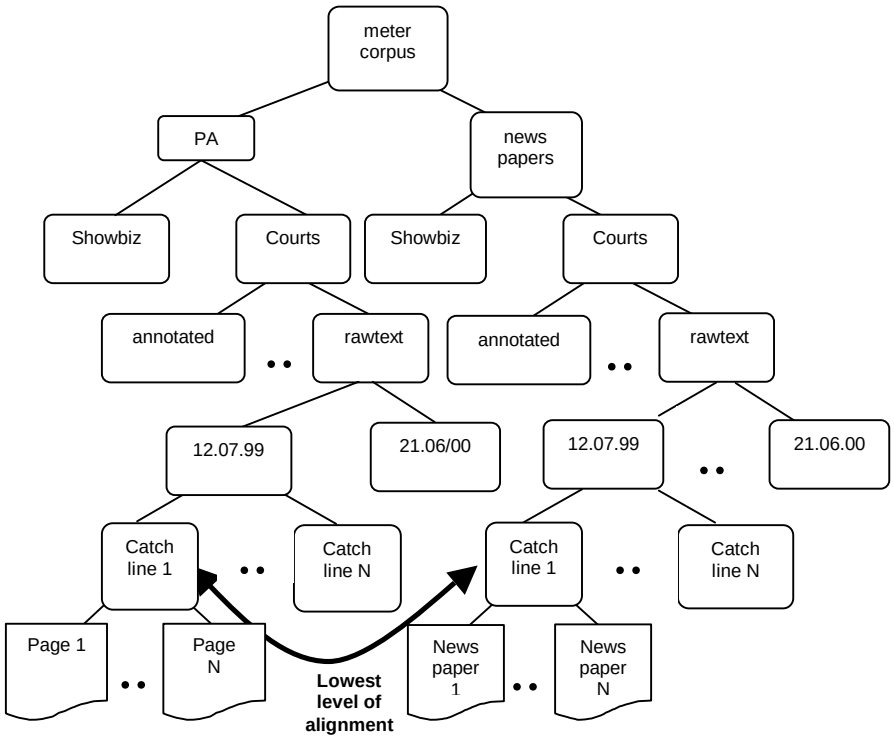

Figure 2: Topology of the METER corpus

As shown in Figure 2, the PA and newspaper stories are stored in a six-level tree structure. At the topmost level, the corpus is divided according to the *source* of materials, PA texts under one branch and related newspaper texts under the other. At the second level, both branches split into two according to the *domain* of the materials, *Courts* stories under one branch and *Showbiz* stories under the other. At the third level, all branches bifurcate again, into *raw* and *annotated* sub-branches. The raw sub-branch contains exact copies of all texts as downloaded from the PA website or scanned in and OCR error-corrected from the newspapers; the annotated sub-branch contains marked-up versions of the texts as described in more detail in section 5 below.

The fourth level of the tree classifies the texts by the *date of release*. All texts issued on the same day, both in the PA and newspaper divisions, are grouped together under the same directory, which takes the date as its name, in the form *day.month.year*, e.g. *12.04.99*. Note, however, that the newspaper texts in the METER corpus are released one day later than the corresponding PA texts. However, for the sake of retaining a parallel data structure between PA and newspapers, the PA dates are used as directories for both PA and newspaper sections. Thus newspaper texts are actually released a day later than their directories indicate.

At the fifth level, i.e. within date, in both the PA and newspaper divisions the texts are grouped into *catchlines*. However, it should be noted that the texts under the same *catchline* are related in different ways within the PA and newspaper halves. In the PA division, all the PA pages reporting a given story on the same date are stored under the *catchline* under which the PA released them. In the newspaper division, all the newspaper versions of a given story are stored beneath the PA *catchline* for that story, for example, the *Times*', *Sun*'s and *Telegraph*'s versions of reports about a killing.

Example filename	Code	File type description
meter_corpus/PA/courts /21.06.00/sergeant/ sergeant1.txt		A page of a story released by the PA (number indicates page).
meter_corpus/PA/courts /21.06.00/sergeant/ sergeant1lead.txt	*lead*	A page of PA copy that summarises the major aspects of a story.
meter_corpus/PA/courts /21.06.00/sergeant/sergeant1nl.txt, meter_corpus/PA/courts /22.11.99/inqheart/inqheart1ld.txt	*nl* or *ld*	As lead - compiled in the afternoon or evening. Especially useful for the next day's daily newspapers.
meter_corpus/PA/courts /21.06.00/sergeant/ sergeant1snap.txt	*snap*	A single sentence giving urgent, breaking news.
meter_corpus/PA/courts /21.06.00/sergeant/ sergeant1sub.txt	*sub*	A page of copy that develops previous material sent by the PA.
meter_corpus/PA/courts /07.10.99/wife/wifecorr.txt	*corr*	Amendment to earlier copy provided.
meter_corpus/PA/courts /16.07.99/care/care1nlcorr.txt	*nlcorr*	Nightlead correction.
meter_corpus/PA/showbiz /14.12.99/mccartney/mccartney1ff.txt	*ff*	Fact file - A series of bullet points to accompany an existing story.

Table 2: Different PA file types

The sixth and final level of the corpus contains the leaf nodes of the tree structure, the actual text files which make up the corpus content. Each text is given a filename that conveys basic information about it. The general naming convention for PA texts is: *catchline+PA-page-number+{story-type}.txt* ("+" here indicates concatenation, "{}" indicate optionality). *catchline* and *PA page* have been previously explained. *story-type* indicates whether a PA text is a "non-standard" press release, such as nightlead, snap, sub, etc. For example, the filename *sergeant1snap.txt* refers to a text which is a single sentence giving urgent, breaking news about an incident or event involving a sergeant. Table 2 lists PA text types included in the METER corpus.

On the other hand, the filenames given to newspaper texts consist of three parts: *catchline+filecode_newspaper-name*.txt. The second component *filecode* is a unique number assigned to each newspaper article for identification. For example, a newspaper filename

thomas353_telegraph.txt refers to a newspaper article from the Telegraph newspaper under the *catchline* of *thomas* with the file code 353.

The parallel structure of the PA and newspaper divisions in the METER corpus facilitates manual and automatic search for related PA and newspaper texts. Furthermore, the indicative directory and file names make it possible to retrieve basic information about text(s) automatically, e.g. data source, text domain, date of release and topic, etc. This information may be obtained directly from the corpus structure and text filenames, without the requirement to look inside any of the texts either at text content or embedded markup. Such information provides one straightforward route to exploit the corpus; however, all this information, and more, is also available from metadata embedded in the annotated portion of the corpus.

5. Annotation of the METER corpus

Leech (1997: 4) suggests that "corpora are useful only if we can extract knowledge or information from them". While raw texts contain useful data, except basic information such as word frequency and collocations, it is difficult to extract more complex information automatically from them. Such information needs to be encoded explicitly in the corpus.

In order to increase the utility of the METER corpus, we have encoded information pertaining to text reuse in the newspaper section of the corpus. Each of the newspaper articles has been manually classified into one of three general categories, indicating degree of derivation from the PA. In addition, approximately 400 newspaper articles have been subject to detailed annotation down to the sentence, phrase, or even word level. The annotation of the METER corpus is described in the following sub-sections.

Due to limitations of project resource and time, all annotations were carried out by one person, a professional journalist. However, we are in the process of getting second judgements from another expert to verify the choices made by the original journalist in the general classification task, for 5% of texts in each category. Resources permitting we will validate the more detailed annotations later.

5.1. General classification at the document level

As mentioned earlier, every newspaper text in the METER corpus shares a topic with one of the PA catchlines. However, the newspaper articles are related to their PA counterparts in a variety of ways. Some of them use whole pieces of PA text without any change; some of them modify PA texts to fit their specific requirements; some of them supplement PA materials with other content that is not to be found in the PA materials; and some do not appear to have consulted the PA at all, even though the PA provided relevant materials for the story.

In order to capture general information about the reliance of newspapers on the PA, each newspaper text is classified by an expert journalist into one of the following three categories:

a) **Wholly derived** (WD) – all content of the target text is derived only from the PA.
b) **Partially derived** (PD) – some content of the target text is derived from the source text. Other sources have also been used.
c) **Non-derived** (ND) – no content of the target text is derived from the source text. Although verbatim and rewritten text may appear in the target text, the context, overlap of entities or use of source text is not indicative of reuse.

Note that this classification is based upon judgements concerning the source of the *content* in the newspaper article, not simply upon surface criteria, such as presence of a certain number or length of shared tokens. Since we are interested in studying the mechanisms of text reuse, we cannot begin by presuming what we hope to discover; i.e. we cannot begin by *defining* text reuse in terms of surface linguistic criteria. Otherwise, we can never hope to discover more than our initial definition. Instead we rely on the judgement of expert journalists who bring years of experience to bear, comprising specialised linguistic and world knowledge.

In a wholly derived newspaper text, all of the facts in it can be mapped, with varying degrees of directness, to PA text(s) under the shared *catchline*. In the most direct cases, the whole or part of a PA

text is copied verbatim to form the newspaper text. In the other cases, a PA text is modified in various ways before being deployed in a newspaper article, including change of word order, substitution of synonyms, and paraphrase. In such cases, the relation of the newspaper text and its counterpart PA text(s) is often not so clear, sometimes even difficult for a human to infer.

In a partially derived newspaper text, part of the text can be mapped to the corresponding PA text(s), but for other parts no related materials can be found in the PA texts. In other words, newspaper articles in this category contain new facts not found in the PA. This category represents an intermediate degree of dependency of newspaper texts on the PA. Within this category, the level of dependency varies considerably, from the majority of text being derived to only one or two sentences being derived from the PA.

The last category covers those newspaper articles that are written independently from PA. Note that we are considering newspaper and PA texts covering the same event. This means that the PA provides coverage of the event, but that it has not been utilised by the newspaper. Instead, the newspaper has used other journalistic sources. This category represents the null dependency of the newspaper on the PA. The number of texts in each category can be obtained from Table 1.

This three-way classification is encoded in the METER corpus as described in section 5.3 below. With such information explicitly available, the METER corpus can be used for training/evaluating algorithms for detecting text reuse in journalistic domain, studying journalistic styles, etc.

5.2. Detailed classification at the lexical or phrasal level

In addition to the document level classification of the whole of the newspaper portion of the METER corpus, detailed annotation of about 400 of the wholly or partially derived newspaper articles was also carried out. In this annotation, individual words, phrases or sentences in the newspaper texts were tagged with information about their derivation from the PA.

The detailed classification parallels the text-level classification discussed in the previous section. Three categories were used:

1) **Verbatim**: text that is reused from PA word-for-word in the same context;
2) **Rewrite**: text that is reused from PA, but paraphrased to create a different surface appearance. The context is still the same;
3) **New**: text not appearing in PA or apparently verbatim or rewritten, but used in a different context.

Of these three categories, the rewrite appears in various forms, such as change of word order or paraphrase. Such modification of the PA texts occurs due to four main reasons: a) the PA text may be re-written to comply with the house style of the newspaper; b) it may be re-written to fit the space available in the newspaper; c) information dispersed throughout the PA texts may be re-arranged into a single, coherent sequence; d) a PA text may be dramatised by replacing neutral words with more dramatic words. Accordingly, a PA text can be modified in the following ways:

1) Rearrangement of word/phrase/sentence order or position;
2) Substitution of original terms with synonyms or other context dependent substitutable terms;
3) Deletion of original materials;
4) Insertion of minor new materials (e.g. addition of words like *by* in passivisation),

In the annotation, newspaper materials falling into any one of the above four categories were tagged as rewrite. There are some controversial cases in which subjective judgement had to be made. On the other hand, the tagging of the verbatim and new materials was generally straightforward, although minor mistakes are inevitable for a manual annotation. All the information about the dependency of newspaper materials on the PA is encoded in an SGML annotation scheme, as described in the following sub-section.

5.3. Annotation of the METER corpus

The METER corpus is annotated in SGML (Standard Generalised Markup Language) to comply with the international mark-up standard (Goldfarb, 1990) and a customised SGML DTD (Document Type

Definition) was developed for it. The DTD provides a framework for recording general information about the texts (such as date, catchline, etc.) as well as for recording information specifically pertaining to reuse. Figure 3 illustrates the structure of the METER SGML DTD document[5].

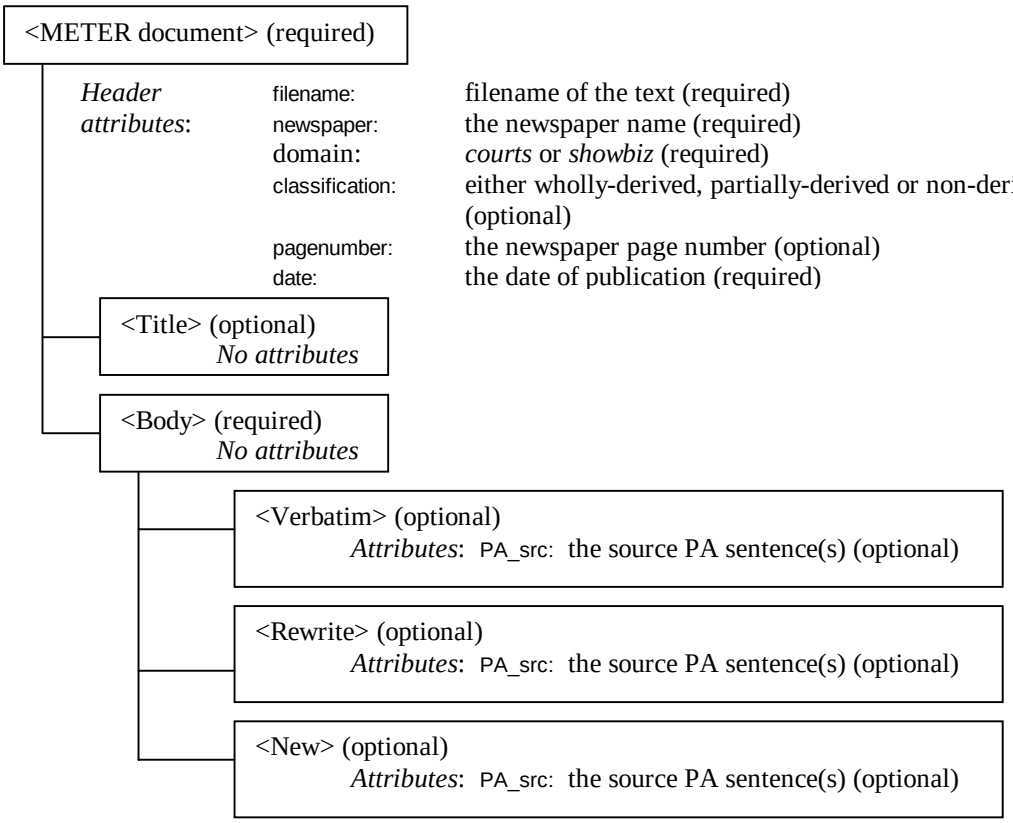

Figure 3: Structure of the METER DTD

As shown in Figure 3, the METER mark-up scheme consists of three levels of tags. The topmost is the *header* which keeps general information about files, including file name, source newspaper, newspaper page number on which the report appeared, date of release and *catchline*. For newspaper articles it also includes the text level classification of the article as wholly derived, partially derived, or non-derived. This header mark-up is found in every text in the whole corpus.

The second level consists of two elements: *title* and *body* of the document. The *title* is optional.

The third layer consists of three elements: verbatim, rewrite and new (for their definitions, see section 5.2). They are sub-elements of *body*, and apply to any individual tokens or token sequences in the newspaper article, including single words, phrases, sentences and punctuation marks. Generally the annotation focuses on words rather than punctuation marks. But the latter were included in the annotation, for they may have significant role in some cases, e.g. in the recognition of quotations.

The detailed annotation was carried out in two stages. First, a journalist analysed and classified the contents of about 400 newspaper articles on paper. Later, the classification was transcribed into tags in the electronic version of the corpus using an annotation tool. Figure 4 displays a sample of annotated newspaper text.

In this example the first line informs an SGML parser of the document type and the location of the DTD file. The *header* indicates that the sample is a newspaper report about a court story related to the PA catchline "banker" on the 4th page of the *Telegraph* released on 16 July 1999 and is wholly derived from the PA. The text body is broken into segments each of which is tagged with one of the three

[5] In SGML, the DTD document defines the constituent parts of a document in terms of objects, known as *elements*. Each element may have parameters called *attributes*.

categories: *verbatim*, *rewrite* and *new*. The value of the attribute *PA_src*, which indicates location of the PA source sentence(s), is now blank, but may be completed in a subsequent release of the corpus.

Original PA version

BANKER'S BITTERNESS LED TO SYSTEMATIC THEFTS<
By Lyndsay Moss, PA News<
A middle-aged banker who stole more than £270,000 from his bosses because he resented younger staff being promoted over his head, was jailed for four years today.<
Trusted Derek Boe, 48, used some of the money to splash out on holidays, buy a car and a caravan, and pay for expensive home improvements.<

Telegraph version:

A BANKER who stole more than £270,000 from his bosses because he resented younger staff being promoted over his head, was jailed for four years yesterday.

Derek Boe, 48, used some of the money for holidays, to buy a car and a caravan, and to pay for home improvements.

Annotated Telegraph version:

```
<!DOCTYPE meterdocument SYSTEM "meter_corpus/dtds/meter.dtd">

<meterdocument filename="meter_corpus/newspapers/annotated/courts/16.07.99/banker/banker125_telegraph.sgml",
newspaper="telegraph", domain = "courts", classification="wholly-derived", pagenumber="4", date="16.07.99",
catchline="banker">
        <body>
                <verbatim PA_src="">A </verbatim>
                <verbatim PA_src="">BANKER who stole more than </verbatim>
                <rewrite PA_src="">£270,000 </rewrite>
                <verbatim PA_src="">from his bosses because he resented younger staff being promoted over his head,
                was jailed for four years </verbatim>
                <rewrite PA_src="">yesterday. </rewrite>
                <verbatim PA_src="">Derek Boe, 48, used some of the money </verbatim>
                <rewrite PA_src="">for </rewrite>
                <verbatim PA_src="">holidays, </verbatim>
                <rewrite PA_src="">to </rewrite>
                <verbatim PA_src="">buy a car and a caravan, and </verbatim>
                <rewrite PA_src="">to </rewrite>
                <verbatim PA_src="">pay for </verbatim>
                <verbatim PA_src="">home improvements. </verbatim>
        </body>
```

Figure 4: A sample annotated newspaper article from the METER Corpus

6. Preliminary experiments with the METER corpus

Currently, the corpus is being used in the METER project for training and evaluating algorithms to detect text reuse. Our initial investigations have utilised the document-level annotations only and have addressed the following task: given a PA text and a candidate derived text, determine whether the text is wholly derived, partially derived or non-derived. We have tried three approaches to this task: the dotplot, n-gram overlap and text alignment. Details can be found in Clough *et al.* (2001); here we just give a brief overview of the work to illustrate how the corpus can be useful.

The dotplot is a tool adapted from the biological domain to visualise the similarities and differences between input streams of either text in electronic format, or software code (Helfman, 1993). This technique was tested on METER texts from the three categories: wholly derived, partially derived and non-derived. It was found that when compared with candidate PA source texts, newspaper texts from the three categories could generally be identified by distinct dotplot patterns. However, a disadvantage of the dotplot is that without more complex processing, no quantitative similarity value is produced above and beyond the visual image. Building the dotplot for large input streams is also computationally expensive.

An alternative approach is to measure similarity between texts by simply measuring the number of word n-grams shared between them. From initial experiments, we found that derived texts appeared to share more n-grams of lengths 3 words and above. This follows the intuition that derived texts share longer matching strings than non-derived texts. However, we also found some instances of non-derived

texts which contained shared 10-grams (due to shared directly quoted text). Using the METER corpus we trained the classifier by selecting optimal threshold values for various parameters. Results in testing ranged from 50-70% correct in classifying documents as wholly derived, partially derived or non-derived.

The final approach to this classification task that we have explored so far is text alignment (see, e.g., Manning and Schütze (1999) for a review of statistical alignment techniques). Given a candidate derived newspaper text, we first carry out best-match alignment at the sentence level. Then we estimate the dependency of the whole candidate derived text on the source, using parameters derived from the METER corpus. Preliminary results show 80-90% correct classification, depending on the setup. The METER corpus has provided invaluable data for the development and evaluation of this algorithm.

These experiments are but a few examples of potential applications of the METER corpus. When it becomes widely available to the research community, we envisage that a much wider range of applications will be found for it.

7. Conclusion

In this paper, we have described a new corpus – the METER corpus – built as part of a project to investigate text reuse in the world of newspaper journalism. It contains texts from the domains of law courts and show business reporting. The texts were collected from the Press Association and nine British national newspapers. The data were manually collected and classified by a professional journalist. All of the newspaper articles are classified at the document level based on their dependency on the PA. Some of the newspaper articles are also annotated at the phrasal or even lexical level to indicate material that is verbatim, rewritten or new.

This is an innovative corpus resource for the communities of corpus linguistics and natural language processing/engineering. The critical role of the PA in the contemporary British media industry and the breadth of the newspaper sources from which the data were collected all warrant that the METER corpus is highly representative of contemporary media in the given domains. The fact that all of the data in the corpus were manually selected and tuned guarantees high quality and reliability. Until now, no corpora marked up with information about text reuse have been reported. The METER corpus fills this gap. We envisage that the corpus will be of use for a range of research, including the study of text reuse, plagiarism, journalistic style, lexicon building, document clustering and language generation.

Acknowledgements

The authors would like to acknowledge the UK Engineering and Physical Sciences Research Council for financial support for the METER project (GR/M34041). We would also like to thank the Press Association for supplying us with access to their newswire archive and for discussions regarding text reuse in the British press. Finally we would like to thank Andrea Setzer for supplying and helping us to customise the annotation tool we used to annotate the METER corpus.

References:

Clough P, Gaizauskas R, Piao S, Wilks Y 2001 *METER: MEasuring TExt Reuse*. Department of Computer Science Reseach Memorandum CS-01-03, University of Sheffield.

Goldfarb C 1990 *The SGML Handbook*. Oxford, Oxford University Press.

Helfman J 1993 Dotplot: A Program for Exploring Self-Similarity in Millions of Lines of Text and Code. *Journal of Computational and Graphical Studies* 2(2): 153-174.

Johansson Stig, Ebeling Jarle 1996 Exploring the English-Norwegian parallel corpus. In Percy Carol E., Meyer Charles F., Lancashire Ian (eds), *Synchronic corpus linguistics*. Amsterdam-Atlanta, GA, Rodopi B. V., pp 3-15.

Leech Geoffrey 1997 Introducing corpus annotation. In Garside Roger, Leech Geoffrey, McEnery Anthony (eds), *Corpus annotation – linguistic information from computer text corpora*. London & New York, Longman, pp 1-18.

Manning C, Schütze H 1999 *Foundations of Statistical Natural Language Processing*. Cambridge, MA, MIT Press.