Building and annotating a corpus for the study of journalistic text reuse

Paul Clough, Robert Gaizauskas, Scott Piao

Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield, S1 4DP, UK
{P.Clough,R.Gaizauskas,S.Piao}@dcs.shef.ac.uk

Abstract

In this paper we present the METER Corpus, a novel resource for the study and analysis of journalistic text reuse. The corpus consists of a set of news stories written by the Press Association (PA), the major UK news agency, and a set of stories about the same news events, as published in various British newspapers. In some cases the newspaper stories are rewritten from the PA source; in other cases they have been independently written by the newspapers' own journalists. We discuss the motivation for creating the corpus, its contents, the annotation of certain attributes for analysis of text reuse and finally the encoding of those annotations into a standardised corpus format: the Text Encoding Initiative (TEI).

1. Introduction

While the reuse of others' text without acknowledgement is, in academic life, a cardinal sin, there is one industry where this is not only accepted behavior, but is in fact standard business practice. In the newspaper industry most newspapers rely very heavily upon press agencies as their primary source of written news material. Upon payment of a subscription fee, the newspaper is free to reuse this material verbatim, or to edit it in whatever way it sees fit, often without requirement to acknowledge the source.

The process of gathering, editing and publishing newspaper stories is a complex and highly specialised task often operating within specific publishing constraints such as:

(1) short deadlines,

(2) prescriptive writing practice (see, e.g., Evans (2000)),

(3) limits of physical size during page layout,

(4) readability and audience comprehension,

(5) editorial bias, and

(6) a newspaper's house style.

Although news agency copy is reused in the creation of a news story, due to the above-mentioned publishing constraints it is unlikely that agency copy gets reused word-for-word, and almost invariably differences will arise. Previous research has identified the major rewriting operations used by journalists and editors as deletion, lexical substitution, changes in syntax (Bell, 1991) and summarisation (McKeown and Jing, 1999; Fries, 1997). More specifically, these include deletion of redundant information and deletion resulting from syntactic changes, substitution of synonymous words and phrases, changes in word order, conjunction and splitting of sentences, changes in tense, passive to active voice, use of abbreviation, verb/noun nominalisation, changes in the definite and indefinite article and changes in the use of name forms. For example, consider the following news agency source, from the UK Press Association (PA), and subsequent derived text published in subscribing British newspapers:

Original (PA) *A drink-driver who ran into the Queen Mother's official Daimler was fined £700 and banned from driving for two years.*

Rewrite (The Sun) *A DRUNK driver who ploughed into the Queen Mother's limo was fined £700 and banned for two years yesterday.*

Rewrite (The Mirror) *A BOOZY driver who smashed into the Queen Mums's chauffer-driven Daimler minutes after she had been dropped off was banned for two years and fined £700 yesterday.*

Rewrite (Daily Star) *A DRUNK driver who crashed into the back of the Queen Mum's limo was banned for two years yesterday.*

This simple example illustrates the types of rewrite that can occur even in a single very short sentence. The rewrites are all taken from the popular press whose style is markedly different from PA's. The use of slang and exaggeration helps to capture the reader's attention (e.g. "drunk", "boozy", "mum's", "limo", "smashed", "ploughed", "crashed"). The change of adverb "today" for "yesterday" is typical of all newspaper stories where writing takes place, or is published, the day after PA copy is produced.

We have been investigating this type of journalistic text reuse in the context of the the METER (MEasuring TExt Reuse) project [1], whose aim is to analyse the reuse of text between a news agency source and the corresponding newspaper articles, with the intent of producing an algorithm that can detect and measure "derived" newspaper text (with a given level of certainty). The commercial motivation behind the research is the potential for news agencies, such as the PA, to monitor and measure the amount of take-up from clients, such as national newspapers, for news which they publish. Academic interest lies in conceptual issues surrounding how to define reuse and in algorithmic issues of how best to detect and measure text reuse. The results of such research will be useful not just to those studying

[1] See http://www.dcs.shef.ac.uk/nlp/meter/

text reuse, but to the wider community of language engineering including computational stylistics, lexicon construction, natural language generation and automatic summarisation and paraphrasing.

In order to support this research we have created a corpus of relevant news materials. The METER Corpus is a collection of over 1,700 texts gathered specifically for analysis of journalistic text reuse. The corpus consists of news agency text produced by the main supplier of news in the UK, the Press Association (PA), and corresponding articles from national papers of the British Press who subscribe to the PA news service. These include *The Times*, *The Guardian*, *Independent*, *Telegraph*, *Daily Mail*, *Express*, *The Sun*, *Daily Star* and *The Mirror*. In the British media industry, it is common practice for newspapers to produce news articles based upon versions (called *copy*) released by the PA (regarded as "pre-fabricated" input by the journalist). Given that news agency discourse structure follows a similar format to that of a newspaper story, the chances that PA is reused over other documentary sources such as press releases, court reports and technical documents is high.

In the rest of this paper we first discuss conceptual issues underlying text reuse and the simple scheme we have used to classify reuse, then describe the METER corpus contents and its annotation based on our approach to classifying reuse, and finally present the TEI-conformant markup scheme we have used to realise the annotations in the corpus.

2. Conceptualising the problem

In thinking about how text reuse might be measured, it is important to realise that one is dealing with a continuous phenomenon that stretches from verbatim, or literal word-for-word reuse, through varying degrees of transformation involving substitutions, insertions, deletions and reorderings, to a situation where the text has been generated completely independently, but where the same events are being described by another member of the same linguistic and cultural community (and hence where one can anticipate overlap of various sorts). Bearing this in mind, we were very conscious of the difficulty, and possible futility, of trying to define reuse too precisely. We were also aware that a subjective element will remain in many judgements of whether reuse has occurred in specific cases.

2.1. The challenge of measuring reuse

Given two texts is it possible to determine, within acceptable levels of probability, whether one text is *derived* from the other? To clarify this, consider the scenario from Figure 1.

In the figure the dotted box denotes the "world of text"; outside it are "events" which occur in the non-linguistic world, and are described in texts. In this example, we consider three versions of the same story: the PA version (A), *The Times* version (B), and the *Daily Mail* version (C). Suppose the event being reported is a court case and assume the PA and *Times* reporters both attend the case; from this the stories A and B are written independently. While the stories are written independently, they will share much lexical content (e.g. the name of the accused, the charge, the verdict,

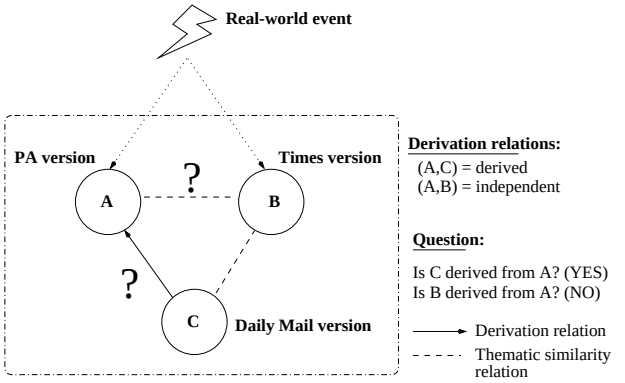

Figure 1: *Derivation* in journalistic text reuse

etc.) because they are reporting the same event, and structurally they will appear similar because they follow "standard" news discourse structuring. The *Daily Mail* reporter could not attend the press conference so instead rewrites the PA version to create text C. The following relationships exist: (A,B) are independent and (A,C) is a derived pair ((B,C) are also assumed to be independent). The question is whether we can correctly identify the dependent pair (A,C) and not falsely identify the independent pair (A,B), despite the inevitable similarities.

2.2. Classifying text reuse

As noted in the introduction to this section, text reuse is a continuous phenomenon. However, to make any headway in studying it, it is necessary to attempt some classification of *degree* of reuse. It is also necessary to specify the *level* of textual unit(s) at which one will identify reuse.

Driven partly by pragmatic concerns (potential utility for a news agency; feasibility of human annotators carrying out the task), we have opted to classify two levels of text unit and for each to distinguish three degrees of reuse.

At the highest level, the level of the whole document, one can make a single crude judgment as to whether one text has been derived from another. At this level we classify the degree to which a newspaper article depends upon PA copy as its source as:

(1) **Wholly-derived**: PA is the *only* source;

(2) **Partially-derived**: PA is a source, but *not* the only source; or,

(3) **Non-derived**: PA copy has not been used in the production of the newspaper text.

Document-level dependency measures the *extent* to which a newspaper has used PA copy ranging from the entire newspaper text, through parts reused, to none of the newspaper text derived from the PA, the source text.

The second level at which we identify reuse is at the word sequence, or lexical level. The classification of lexical-level information is a more ambitious scheme that attempts to capture reuse *within* the newspaper text itself down to the word or phrasal level. Again, a 3-category scheme has been devised to capture relationships between PA copy and the newspaper:

(1) **Verbatim**: text reused word-for-word to express the same information;

(2) **Rewrite**: text paraphrased to express the same information;

(3) **New**: text used to express information appearing in the newspaper, but not in the PA copy (could include word sequences in the PA, but not used in the same context).

This level of classification attempts to identify newspaper text which has been lifted from PA copy with no change, lifted but paraphrased and not lifted at all.

2.3. Identifying derived texts

Given this scheme for classifying reuse, the question arises as to how this reuse is to be identified. Ideally one would observe a newspaper journalist at work, and see what use he or she makes of the agency source. Since this was impossible in our case, we have instead been forced to rely upon the judgments of a professional journalist who read the PA source and a candidate derived text, and applied the scheme *post hoc*.

Deciding whether a newspaper is derived from PA copy is similar to identifying characteristics that distinguish plagiarised texts from those written independently. Typically the decisions are subjective and hard to define, but the key factors in making judgments are:

(1) **Factual overlap** If all the information in the newspaper text is also found in the PA text, then it is a candidate wholly-derived text; if facts in the article are not in the PA text then clearly it is at best partially-derived.

(2) **Linguistic variation** The length, location, grouping and dispersion throughout PA and newspaper copy of verbatim matches and of potential paraphrases or rewritings signify potential reuse; certain sorts of rewriting are conventional and easily detected by a professional journalist.

(3) **Pragmatics of news production** A professional journalist uses his or her knowledge of how news is produced in the UK to assist in judgments about reuse.

- News agency copy is reused frequently, especially for particular sorts of stories, such as court cases, where the PA has a strong presence and wide coverage.
- The PA is the main UK news provider with a large customer base. Similarities between PA and a newspaper are likely to come from reuse of PA rather than other sources.
- Since customers pay up to £700,000 annually for PA copy, it is unlikely that they do not make use of PA as the source, except in cases where the newspaper has a strong interest in having its own reporter present – cases which can be predicted, to some extent, by the significance of the events being reported and the bias of the newspaper.

3. Constructing the corpus

To support the analysis and evaluation of algorithms for measuring text reuse, we have created the METER corpus. The corpus consists of 1,716 texts gathered manually over the duration of the METER project and collated into a suitable structure for both manual and computational analysis. A news story from a particular day is identified by the PA using a *catchline*: a short name indicating to journalists the topic of the story; e.g., the catchline *hamilton* identifies a story about the Tory MP Neil Hamilton and Al Fayed. For a given catchline, the PA output several pages of press reports presenting the facts of the story in news discourse. By this we mean that the structure of a PA story can be seen to follow typical structures used to convey written news to its reader (see, e.g., van Dijk (1988)). For each PA catchline, newspaper articles from the Brtish Press can be identified which report the same set of facts or event.

Bearing in mind that the PA release on average 1,500 news stories each day, determining the extent of news to cover formed an important part of the practical construction of the METER corpus. To make the amount of news collected for analysis manageable and given the limited effort available for corpus construction, we constrained the METER Corpus both in terms of the domains and date range covered. Given that PA copy is more likely to be used in stories which report daily news events (known as the "hard" news), we concentrated on two areas of news reporting: 1) law and courts, and 2) show business. Typically hard news is divided between tales of disaster, crime and incidents called "spot" news and those of politics or diplomacy. Soft news are stories that tend to deviate from the structure of hard news and include feature stories, editorials or commentaries. News agencies such as the PA are generally regarded as the main suppliers of hard news and custodians of its style. We chose law and court reporting from spot news and show business from soft news because:

- **Frequency** Law and court reporting and show business news stories are a recurring feature of British Press and occur daily.
- **Coverage** Stories from these domains are covered by all members of the British Press, where broadsheets typically cover more hard news and tabloids more soft news.
- **Suitability** These types of stories are suitable for reuse whereas editorials and commentaries tend to be written from primary sources such as interviews with those involved with the story.
- **Contrast** Law and court reports contrast in the style of reporting to stories from show business, with respect to freedom of expression, vocabulary and structure. For example, court stories tend to be written in a "standard" way and revolve around "hard" facts, which limit vocabulary usage and constrain the way in which text can be reused, making its identification easier; show business stories tend to be more informal and diverse. Including two contrasting domains enables the study of domain on the effects of reuse.

Date	Domain					
	Courts			Showbiz		
	Days	Catchlines	Articles	Days	Catchlines	Articles
July 1999	3	36	146	1	4	13
August	2	13	48	5	22	80
September	2	8	40	1	6	7
October	2	17	45	0	0	0
November	2	29	98	1	4	8
December	2	17	81	1	4	17
January 2000	3	27	79	2	11	32
February	2	17	52	0	0	0
March	3	15	66	0	0	0
April	2	20r	87	1	2	6
May	0	0	0	0	0	0
June	1	6	27	1	7	12
Total	**24**	**205**	**769**	**13**	**60**	**175**

Table 1: Number of days covered during July 1999 and June 2000

Temporal coverage was constrained to a one-year period between 12 July 1999 and 21 June 2000. An average of 3 days per month were selected for law and court reports resulting in 769 newspaper articles and 660 pages of PA copy covering 205 catchlines over 24 days. Court stories were made the principal focus of the METER corpus, therefore only 175 newspaper articles and 112 pages of PA copy were collected from the show business domain. This resulted in an average of 1-2 days per month over 13 days during the course of the year. Table 1 summarises the number of days covered and newspapers collected for those days and details about the resulting composition of the corpus can be found in Gaizauskas et al. (2001).

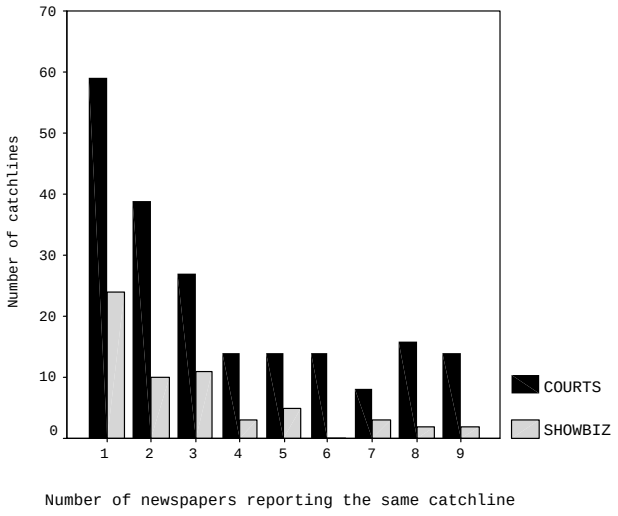

Figure 2: The number of catchlines covered simultaneously

The distribution of number of newspapers covering the same catchline is shown in Figure 2 and can be seen to range from between 1 and 9. Typically only those stories with high news values (these are determined by the editor and restrict which stories appear in the Press) appear across many newspapers. While multiple news stories on the same catchline are more likely to be useful to those who study the ways in which different authors express the same facts, e.g. the multi-document summarisation community, these tend to be harder to collect given the diversity in focus across the British Press.

To enable the study of the impact of journalistic *style* on text reuse, the number of newspapers from which stories were gathered was large enough to cover the wide diversity of reporting style in the British Press. Figure 3 shows the proportion of corpus content drawn from each newspaper in the corpus. This includes the 9 main newspaper sources and an additional *other* category which contains articles from additional newspapers initially considered for inclusion in the corpus, but subsequently not pursued.

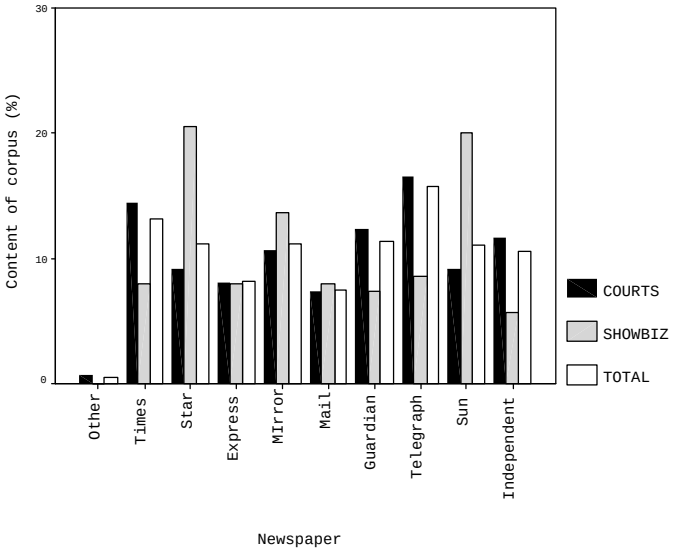

Figure 3: The proportion of content for each newspaper across domains

Careful selection of newspapers from different *registers* of press reporting was also made in order to allow analysis of the effects of register on press rewriting. In the British Press, newspapers are typically categorised according to their reporting style, language used and stories covered. Tunstall (1983) uses a three-way classification scheme. Those papers with a more "formal" style are known as the *quality* press (e.g. the *Guardian* and *The Times*), also referred to as *broadsheets*. Those newspapers whose style or writing is more relaxed are known as the *popular* press (e.g. *The Sun* and *Daily Star*) and are also called *tabloids*. A distinction is also made between those tabloids which used to be broadsheets (*Daily*

Mail and *The Express*), but would now class themselves as tabloids (called *middle-range tabloids*), and the so-called "national" tabloids or down-market tabloids (also called "red tops" because of their red coloured masthead/title on the front page), e.g. the *Sun*, *Mirror* and *Daily Star*. Figure 4 shows the proportion of newspapers of different registers by domain in the corpus. Notice that in the courts domain the majority of content is from broadsheets or the quality press rather than tabloids or the popular press; but the situation is reversed for show business content. This reflects the typical content of British Press, where broadsheets tend to contain a larger proportion of hard news and popular press more soft, advertising and feature stories.

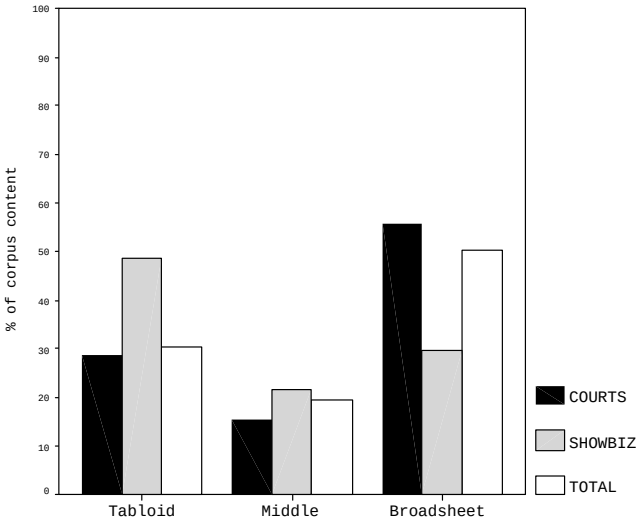

Figure 4: The proportion of content for each newspaper *register* across domains

To ensure continuity between newspaper articles, only *Southern* editions were used and stories were chosen varying between the *1-off* story and the *running* story, i.e., which appears on more than one day, such as a court trial. Other resources could have been considered such as magazines and weekend editions of newspapers, but previous informal analysis had shown daily newspapers to exhibit the most text reuse and more likely to be paying customers of the PA. Paper editions were used rather than on-line versions because these are considered the archival versions of the stories and because the process of creating on-line versions was not clear to us; further, it would have not been possible to gather all articles required from on-line newspapers in the selected date range and across required sources. To create the electronic corpus, articles from paper editions were scanned and manually corrected for OCR errors before saving as plaintext ASCII.

We also attempted to collect newspaper stories of varying lengths to provide a representative sample of stories ranging from one-sentence summaries, called the *News in brief* stories, or NiBs, to longer court and show business reports of around 500-1000 words in length. The mean file length for court stories is 328 words (1,555 words of PA copy) and 244 for show business stories (785 words for PA copy). Different "types" of PA copy were also collected for each catchline to reflect the days output from PA on each story. This included nightleads, snaps, subs, fact-files and corrections. A "feature" of PA copy is that some catchlines contain duplicated copy where later PA copy includes previously released text with additional information or correction.

4. Annotation – enriching the corpus

As is well known, a collection of texts is greatly enriched by annotation – interpretative information added to the base text (Leech, 1997). In the case of the METER Corpus we have added annotations capturing a professional journalist's view of the derivational relationship between PA copy and a corresponding newspaper text, according to the two-level, three-degree scheme for classification of reuse outlined in section 2.2..

All the texts in the newspaper portion of the corpus have been annotated at the document level. Analysis of the corpus reveals that 78.4% of court stories were judged to be derived from PA copy, with 34% wholly-derived and 44.4% partially-derived. 77.2% of show business stories were also judged to derive PA copy, with 22.2% being wholly-derived and 54.9% partially-derived.

At the lexical level, 445 newspaper articles across both domains were manually analysed and classified – resource restrictions did not permit more. In these texts every word sequence has been classified as *verbatim*, *rewritten* or *new*. This annotation was carried out using a sweep-and-click annotation tool with annotations being exported as SGML, as described in the next section. Due to manual resource and software constraints, we have not provided links between verbatim and rewritten text in the newspaper text and the most likely source text in the PA copy. This is valuable information that could be included in a future release of the METER corpus.

All of the annotation was carried out by a single annotator, a professional journalist. A small subset of the annotations was reviewed by another journalist who concurred with the judgments of the initial annotator, but this was an informal check and not an independent annotation exercise. Resources did not permit the sort of multiple independent annotation that is necessary to establish convincingly the objectivity of the annotation process.

5. Standardising the resource

Two versions of the METER Corpus have been created. In the first version documents were stored in a hierarchical directory structure reflecting their domain, date, topic, and (newspaper) source, and marked up according to an SGML DTD. In the second, more recent version, the texts are stored in a flatter structure based upon date, and marked up to comply with the XML version of the TEI (Goldfarb and Prescod, 2001).

5.1. The SGML version

In the SGML version of the corpus documents are stored both in plain ASCII and in SGML using a DTD to define document annotations. The directory structure of the SGML corpus is shown in Figure 5. The corpus follows a

parallel structure, divided between PA and newspaper stories. Each of these subtrees is then subdivided into rawtexts (plain ASCII text) and those annotated according to the original SGML schema. The corpus is further divided by domain (courts and show business), date of publication *for the PA catchline* (post-dated newspapers are stored according to release by PA to make the structure parallel), catchline with the texts themselves – either newspaper articles, or pages of PA copy – occurring as the leaves of the tree structure. More information about this structure can be found in Gaizauskas et al. (2001).

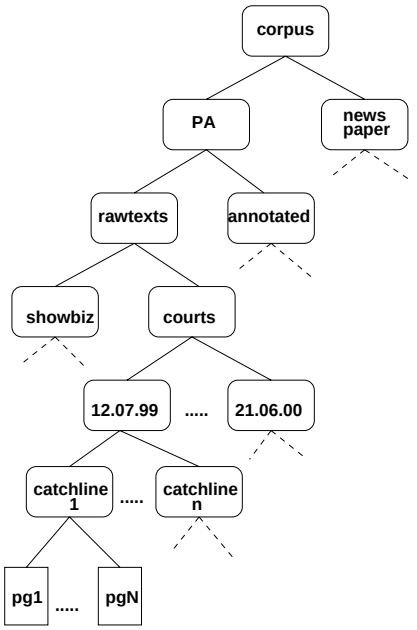

Figure 5: Directory structure of the SGML version

Subsequent to an initial beta release of the SGML version, feedback from some users suggested that the deep hierarchical file structure was difficult to work with. Furthermore, our SGML markup scheme did not conform to international corpus encoding formats such as the Text Encoding Initiative, TEI, which aim to promote standardisation and exchangeability. To address these concerns we transformed the SGML version into a TEI-conformant structure which includes a physical re-structuring and re-annotation of the corpus data into TEI format (see Sperberg-McQueen and Burnard (1999) for more information on TEI).

5.2. The TEI version

In the TEI version, the main body of the corpus is divided into 27 files based upon the publication date of PA catchlines. Information about the corpus as a whole including publication information and the definition of attributes specific to the METER corpus (e.g. the document and lexical-level text reuse annotation scheme) are defined in a global header file. Attributes are associated with *elements* (or tags) which encapsulate the corpus in TEI. Files for each day contain a local header defined by the <teiHeader> tag that includes the XML filename, publication date and domain identifier: “courts”, “showbiz” or “courts showbiz” (some dates contain texts from both domains). PA copy and newspaper texts are grouped by catchline.

```
<TEI.2>
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>The Meter Corpus data for 01’03’2000</title>
  </titleStmt>
  <publicationStmt>
    <p>Section 22</p>
  </publicationStmt>
  <sourceDesc>
    <p>File path: "meter01_03_2000.xml"</p>
  </sourceDesc>
 </fileDesc>
 <profileDesc>
  <textClass>
   <catRef target="courts"/>
  </textClass>
 </profileDesc>
</teiHeader>

<text id="M01032000">
<group>

<text id="M01032000-1" n="burstein">
<body>

<div id="A622" n="pa-01032000-1" type="courts"
ana="src">
<pb n="1"/>
<head type="patext">COMPOSER CHALLENGED...</head>
<p>
<s n="1">Composer Keith Burstein, seeking....</s>
</p>
</div>

<div id="M747" n="times-01032000-1" type="courts"
ana="pd">
<pb n="3"/>
<head type="news">Musician stunned...</head>
<p>
<s n="1">A COMPOSER of classical music was....</s>
</p>
</div>

<div id="M748" n="guardian-01032000-1" type="courts"
ana="pd">
<pb n="10"/>
<head type="news">Composer’s ‘incredulity’...</head>
<p>
<s n="1">
<seg ana="rewrite">A</seg>
<seg ana="verbatim">composer</seg>
<seg ana="rewrite">yesterday</seg>
<seg ana="verbatim">told the high court</seg>
.
.
.
</s>
</div>
</body>
</text>
</group>
</text>
</TEI.2>
```

Figure 6: An example METER corpus day file

All catchlines for the date are encapsulated within a <group> tag and individual catchlines are demarcated using the <body> tag.

Each page of PA copy and newspaper article is demarcated using the tag and within each text, paragraphs and sentences are marked using the <p> and <s> tags respectively. Sentences are numbered from 1 to n for each text and all texts within the corpus are given unique identifiers. A number of newspapers are further annotated at the lexical-level and text is marked as either verbatim, rewrite or new using the <seg> tag where the “ana” attribute is used to indicate the category of lexical-level reuse.

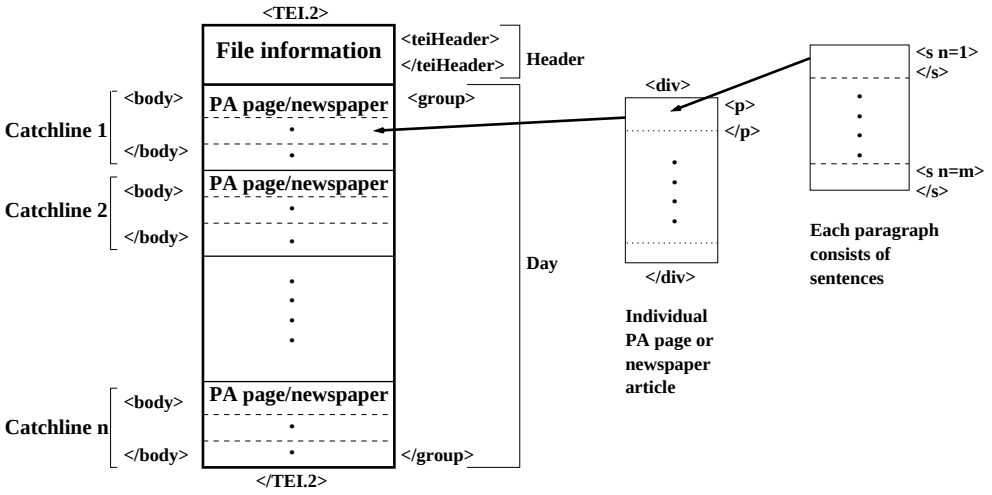

Figure 7: Markup of individual files which represent days

Document-level reuse is captured in the tag by the "ana" attribute. Each text also has a <head> tag which captures the title (also known as a headline) and an optional tag indicating the author of the news story if present.

Figure 6 is an example of a METER corpus file in TEI for the day `01'03'2000`. Before <group> and <body> tags are <text> elements with the 'id" attribute specifying the date and date/catchline respectively. The example contains a page of PA copy and two newspaper articles from the Times and Guardian respectively. Only the second newspaper text contains lexical-level annotation. All texts in the corpus are sequentially numbered as specified by the "id" attribute for the tag which also contains the following attributes: 1) "n" - to identify each text within the day file, 2) "type" - indicates the domain and 3) "ana" - indicates the document-level reuse category for a newspaper (*wd*, *pd* or *nd*), or *src* for PA. The "n" attribute of the <pd> tag contains the page number of PA copy or the page number of the article within the original newspaper. Catchlines within each day are alphabetically ordered and PA pages ordered by the time of their release. The order of newspaper texts is arbitrary. Figure 7 illustrates the composition of a TEI-encoded file for any given day in the corpus, and more detailed information can be found in Piao et al. (2002).

The whole Meter corpus is encapsulated by a single TEI tag: <teiCorpus> contained in the driver file. However, physically the corpus is separated into 27 files. We split the corpus to allow individual blocks to be processed rather than the entire corpus which collectively amounts to roughly half a million words and could prove too large to process conveniently as a single block. To enable parsing of the entire corpus, however, a *driver* file is used which allows physical separation of data but facilitates collecting them together when needed. The driver file brings together the global header file for the METER corpus, the 27 files (called *sections*) which are defined as entities in the *textlist.ent* file and other auxiliary documents required for parsing the corpus. The TEI version of the corpus consists of four main parts (see also Figure 8):

1. a global <teiCorpus> header file;
2. corpus data stored in twenty-seven files reflecting catchline dates;
3. the TEI2 DTD and other auxiliary documents such as an entity file defining character reference names for non-ASCII characters;
4. a driver file linking all of the components.

The TEI DTD file can either be installed locally or be accessed via a URL[2], allowing access from anywhere in the world. In Figure 8, the dotted line encloses the whole Meter corpus and a parser accesses each component of the corpus through the driver file. In order to facilitate easier distribution and utilisation, we transformed the corpus into a structure conforming to the XML version of the TEI (see, e.g., Goldfarb and Prescod (2001)). This version of the METER corpus will be published by ELRA[3], the European Languages Resources Association, and we envisage that it will provide an ease to use and highly useful language resource.

6. Conclusions

We have presented the METER corpus, a novel corpus primarily built for the analysis of text reuse in journalism and for the evaluation of automatic approaches to measure text reuse. Although limited in terms of the domains, dates and stories selected the METER corpus is the first corpus of its kind and its creation has illuminated a set of issues surrounding the construction of such a resource.

Construction of the METER corpus has benefited from journalistic and computational linguistic expertise, the former to determine and gather its content, the latter to make the resource accessible through computational means to a wide audience. The data has been carefully selected and the composition of domains, newspapers, styles of newspaper, lengths of stories and variety of catchlines makes this a representative sample of contemporary British Press. The

[2]For example, the Meter driver file accesses the TEI DTD file via the Web at: http://www.tei-c.org/P4X/DTD/tei2.dtd

[3]Release is scheduled for Spring/Summer 2002 pending some final copyright clearances.

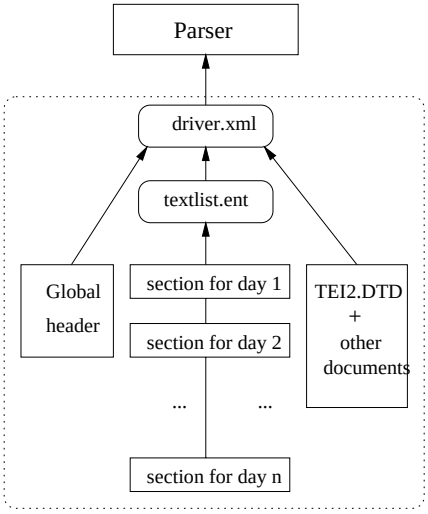

Figure 8: Outline of the Meter Corpus structure

fact that all data in the corpus were manually collected and tuned guarantees high reliability and quality, making the METER corpus a unique resource for computational linguists.

We foresee the following applications:

- **Multi-document summarisation** Because the corpus provides different versions of the same events, it could be used to train and test summarisation algorithms.
- **Automatic headline generation** Given newspaper and PA stories with headlines, it would be possible to analyse headlines and derive automatic method for generation.
- **Text classification** Given articles from newspapers writing in different registers, it should be possible to determine lexical and syntactic cues that could be used to route documents according to their register, e.g. tabloid or broadsheet. This could be used to re-rank a list of relevant documents returned from an information retrieval system.
- **Discourse interpretation** Newspaper and PA texts could be used to analyse the discourse structure of news stories in the British Press and evaluate existing methods for discourse interpretation.
- **Text reuse** Primarily built for this purpose, the corpus can be used to analyse the reuse of text by journalists from PA copy and to test algorithms for automatically measuring text reuse.
- **Plagiarism detection** The manner of journalistic rewriting is similar to that used by plagiarists. Given the difficulty of obtaining examples of plagiarism, the METER corpus could be used as a common resource for testing and evaluating methods for free-text plagiarism detection.

We are aware that the METER corpus has two significant weaknesses:

1. lack of multiple independent annotation to verify the objectivity of the document and lexical-level classifications of reuse;
2. no explicit links between the lexical-level annotations in the derived (newspaper) texts and source (PA) texts

Both of these issues could be addressed in future versions of the corpus. Other future work on the METER corpus includes increasing the size of the corpus and increasing the number of domains to include perhaps politics, sports and editorials.

7. Acknowledgements

We would like to thank Jonathan Foster and John Arundel, members of Department of Journalism Studies, University of Sheffield, who helped conceive the METER corpus and without whose help the corpus would not have been built. We would also like to thank Lou Burnard for his help on the encoding of the METER corpus in TEI. The Press Association has been extremely helpful in making texts available and enthusiastically supporting this research. Finally, we thank the UK Engineering and Physical Sciences Research Council (EPRSC) for funding the METER project via a ROPA award (GR/M34041/01).

8. References

A. Bell. 1991. *The Language of News Media*. Blackwell, Oxford, UK.

H. Evans. 2000. *Essential English for Journalists, Editors and Writers (Revised Edition)*. Pimlico, London, UK.

U. Fries, 1997. *Summaries in Newspapers: A Textlinguistic Investigation*, pages 47–63. In The Structure of Texts, U.Fries (editor), Swiss Papers in English Language and Literature, Tübingen, Narr.

R. Gaizauskas, J. Foster, Y. Wilks, J. Arundel, P. Clough, and S. Piao. 2001. The meter corpus: a corpus for analysing journalistic text reuse. In *Proceedings of Corpus Linguistics 2001, Lancaster, UK*, pages 214–22.

Charles F. Goldfarb and Paul Prescod. 2001. *The XML Handbook*. Prentice Hall PTR, Prentice Hall PTR, Upper Saddle River, NJ 07458, USA, third edition.

K. McKeown and H. Jing. 1999. The decomposition of human-written summary sentences. In *22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 129–136.

S. Piao, P. D. Clough, R. Gaizauskas, and J. Arundel. 2002. Meter corpus documentation (version 2.0). Technical report, Department of Computer Science, University of Sheffield.

C.M. Sperberg-McQueen and L. Burnard, editors. 1999. *Guidelines for Electronic Text Encoding and Interchange*. TEI P3 Text Encoding Initiative, Oxford, UK, revised reprint edition.

J. Tunstall. 1983. *The Media in Britain*. Columbia University Press.

T. van Dijk. 1988. *News analysis: Case Studies of International and National News in the Press*. Hillsdale, NJ.: Lawrence Erlbaum.