



A Parser for News Downloads *Um Parser para o download de notícias*

Mike SCOTT

(Aston University - School of Languages & Social Sciences - Birmingham - UK)

ABSTRACT

This paper presents the Download Parser, a tool for handling text downloads from large online databases. Many universities have access to full-text databases which allow the user to search their holdings and then view and ideally download the full text of relevant articles, but there are important problems in practice in managing such downloads, because of factors such as duplication, unevenness of formatting standards, lack of documentation. The tool under discussion was devised to parse downloads, clean them up and standardise them, identify headlines and insert suitably marked-up headers for corpus analysis.

Key-words: *News corpus; Corpus clean-up; Duplicate texts; Building sub-corpora.*



This content is licensed under a Creative Commons Attribution License, which permits unrestricted use and distribution, provided the original author and source are credited.

RESUMO

Este artigo apresenta o Download Parser, uma ferramenta para gerenciar downloads de texto de grandes bancos de dados online. Muitas universidades têm acesso a bases de dados com textos completos que permitem ao usuário pesquisar e, em seguida, visualizar e, idealmente, baixar o texto completo de artigos relevantes. Todavia há problemas importantes na prática do gerenciamento de tais downloads, por causa de fatores como duplicação, falta de padrão de formatação e falta de documentação. A ferramenta em discussão foi concebida para analisar downloads, limpá-los e padronizá-los, identificar títulos e inserir cabeçalhos adequadamente etiquetados para análise de corpus.

Palavras-chave: *Corpus de notícias; Tratamento de Corpus; Textos Duplicados; Construção de sub-corpora.*

1. Introduction

The aim of this paper is to present a tool for handling text downloads from large online databases. Many universities have access to full-text databases held by organisations such as EBSCO host Research, Factiva, LexisNexis, ProQuest. These allow the subscribed user to search their holdings and then view and ideally download the full text of relevant articles. It is preferable if they allow a good number of relevant articles to be downloaded together in one text file. They may offer a variety of formats for the download such as plain text, PDF, Microsoft Word.

However, there are important problems in practice in managing such downloads, because of factors such as duplication, unevenness of formatting standards, lack of documentation. The tool under discussion was devised to parse downloads into their component articles, clean them up and standardise them, identify headlines and insert suitably marked-up headers for corpus analysis. With corpus software it is possible to study large amounts of text far exceeding what one research can read, seeking out not only patterns of collocation or of grammar, but also to trace references to themes of interest to researchers in sociology, history, law, literature, politics or medicine. Such corpora are theme- or topic-oriented specialised corpora and it is already evident that both

novices taking corpus linguistics courses and experts in various fields are finding them useful.

In this paper I will first explain the problems in more depth and then outline the procedures which the software adopts to attempt a solution. The evaluation will be complemented in the Conclusions with some important limitations and ideas for desirable further development.

The paper is presented in honour of my long-time friend and ex-colleague and fellow corpus enthusiast, Leila Barbara. She has encouraged generations of her colleagues and students at the Catholic University of São Paulo to use *WordSmith Tools* and I am hoping she will encourage them now to take an interest in the *Download Parser* software.

2. Problem

Holdings in the various databases depend on commercial agreements being signed between such online database agencies and relevant publishers such as the *Guardian*, *Daily Mail*, and their peers in many countries and in a wide range of languages. It is almost inevitable that these database agencies' acquisition policies will favour sources from the USA and its commercial allies. Naturally the number of years of access to the full text of such resources depends on each publisher having made their news accessible in electronic format and determined to entrust them to a large database company; while many newspapers have been organised with all their text produced and formatted in-house digitally since the 1980s or earlier, not all have passed on a full run of their holdings. Moreover, the formatting which suits one newspaper or news agency cannot be expected to correspond well with that of others.

In practice, using the Nexis holdings, it seems that a rather basic set of indicators can be found, usually indicating who wrote each article, the date (of which more below), which language it was written in, the name of the newspaper. Relevant articles appear one after the other up to the agency's limit such as 500 documents with this rather simple and incomplete format. The items one may search on in the agency's documentation are explained in help documentation (<http://bisinfo>.

lexisnexis.co.uk/hubfs/Resources/Nexis_User_Guides_Advanced_Search_Guide.pdf shows an example with two pages outlining the various choices one may make for an advanced search) but how the search operates internally in their database is not made clear. Function words get cut out so a search on “this” or “because” alone will fail, since the aim for most users will be to carry out historical or political research using content words. Accordingly such resources are not good for building a general corpus of news text. There are no options for requiring a search-term to occur more than a specified number of times in each article selected, and it seems that even if a search-term occurs once only in passing, the Nexis search-engine includes it. For example, a *Times* article of November 10th 2015 headed “Berlusconi bounces back in rightwing pact” was retrieved in a search on the term *austerity*, because it included a brief reference to Beppe Grillo, “a former comic whose anti-austerity Five-Star movement”...., although the text really concerns Berlusconi joining a coalition, hoping thereby to recover politically. The text is not about austerity, cuts, economic problems but about Italian politics.

In general, then, the problems in using these downloads are that a rather mixed bag of incompletely formatted data is obtained containing a lot of relevant articles with some duplication and including articles where the search-term is peripheral.

3. Corpus Requirements

In recent years I have been involved in research using corpus linguistic methods on topics such as climate change, austerity, energy security. An example would be Grundmann and Scott (2014), where we used Nexis and searched news sources from the USA, UK, France and Germany looking for references to *global warming*, *climate change*, *greenhouse effect* and translations¹ in French and German. Berry (2016), Koteyko (2012) and Mercille (2014) likewise used Nexis

1. For French we used *changement climatique*(=*climate change*), *effet de serre* (*greenhouse effect*), *réchauffement de la planète* (*planet warming*) and *réchauffement climatique* (*climate warming*), and for German *Klimawandel*(*climate change*), *globale Erwärmung* (*global warming*), *Treibhauseffekt*(*greenhouse effect*), *Klimaschutz* (*climate protection*) and *Klimakatastrophe* (*climate catastrophe*).

sources for their searches and to obtain data and I believe there are many researchers who use similar resources.

A corpus building on news resources in that way should ideally

- ensure a standard format
- identify article features such as headlines, text boxes, sub-headings, captions of graphics, hypertext links
- filter out duplicate articles
- filter out articles where the main topic is not implicit in the search-term(s)
- store individual articles so that they can be retrieved easily
- build sub-corpora by grouping articles according to criteria such as date, publication-type, or within a set of specific publications

The Download Parser program looks like this when running:

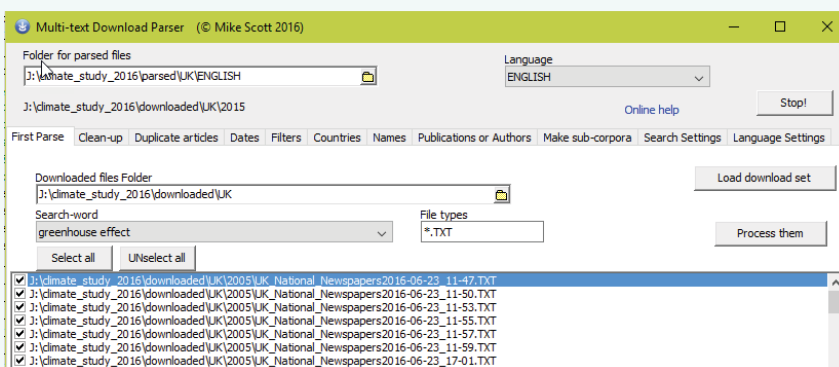
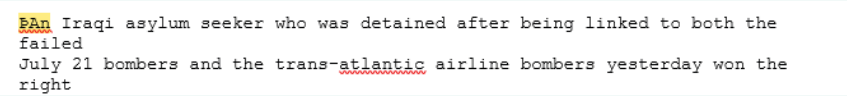


Fig. 1 – The Program.

A series of downloaded text files from the UK national press are listed, and the program user is ready to press the *Process them* button. Further tabs labelled *Clean-up*, *Duplicate articles* etc. can be seen. More detail is available at the Help file (Scott 2016).

4. Characteristics of the original download

Typically, downloads I have worked with have come in 500-document text files as downloaded from Nexis. They may have come through a variety of computers to the Nexis database. They can contain oddities such as additional strange characters. This is a view in Microsoft Word of a fragment of a download from the *Daily Telegraph*

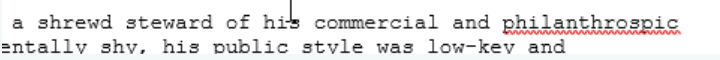


PAn Iraqi asylum seeker who was detained after being linked to both the failed July 21 bombers and the trans-atlantic airline bombers yesterday won the right

Fig. 2 – Thorn symbol text.

which shows a thorn² character P immediately preceding An. How this entered the original download is a mystery; some character-level contamination occurs regularly albeit in a minority of articles.

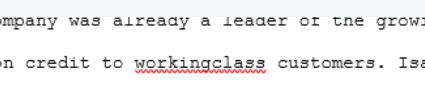
Second, we get some mis-spellings:



a shrewd steward of his commercial and philanthrospic entally shv, his public stvle was low-key and

Fig. 3 – Mis-spelling.

and words run together



mpany was already a leader of the growi
n credit to workingclass customers. Isa

Fig. 4 – words run together.

2. Microsoft's language handling resources in the case of German will correctly cause grüss, gruess and grüß to be treated as equal in terms of word order. Accordingly the combination P+AN in English was interpreted by Microsoft's resources in Windows 10 as THAN. This is automated mechanical matching, perhaps not unreasonable in theory, but weird in practice. It recently caused me and a student at Aston university to struggle to understand an anomaly we were getting in WordSmith's stop-list processing.

A Parser for News Downloads

Those come from the original journalist, sub-editor or editor, presumably. If a download process was carried out in Germany, the text may contain this:

```
Dokument 236 von 494

The Sunday Times (London)

September 2, 2007
```

Fig. 5 – Dokument.

which clearly derives from localisation software within Nexis. And in recent texts there seems to be evidence of mark-up:

```
2 of 500 DOCUMENTS

The Guardian

November 10, 2015 Tuesday 9:45 AM GMT

Moody's warns of global shockwaves from China slowdown - business live;
All the day's financial news, as rating agency warns that world economic
growth
will be hampered by emerging market problemsMoody's: Expect weak global
growth
until 2017Eurozone faces new headaches in Greece and Portugal

BYLINE: Graeme Wearden

SECTION: BUSINESS

LENGTH: 1555 words

block-time published-time 9.45am GMT

IEA: Oil price won't hit $80 until 2020

The world's top energy forecaster has predicted that oil prices will
remain
```

Fig. 6 – extraneous.

```
It is presumably tied into concerns that the ECB's bond-buying programme
violates the rules banning monetary financing within the eurozone.

enlrBREAKING: 3 lawsuits targeting the #ECB 's QE program filed with the
German
Constitutional Court. No hearing scheduled yet. #herewegoagain
```

Fig. 7 – oddities.

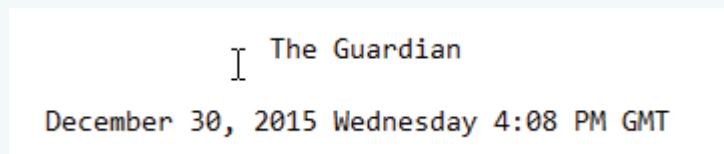
These show signs of the influence of the web, possibly Twitter hashtags but also of unmarked internal computer mark-up.

5. Parse Process

This section describes how the downloaded plain text files are parsed to identify where each article begins and ends and how they are cleaned up, marked-up and sorted for corpus use.

5.1. Interpreting weak mark-up

It is not possible or desirable in this paper to explain all of the technical aspects, but essentially the program operates first of all by converting all of each downloaded text to Unicode³. Then it goes through each download text looking for wording which marks the beginning of an article, such as “Document # of #”⁴. This wording and similar field markers can be edited in the program settings so that users may use the program with their own downloads. Once it finds where each text begins and ends it can keep a copy of each article and parse that copy for document fields. An example of the fields which state the publication and date is in Fig 8.



```

I The Guardian
December 30, 2015 Wednesday 4:08 PM GMT
```

Fig. 8 – Document fields A.

and fields showing the journalist’s name, the section of the newspaper and the length can be seen in Fig. 9.

3. Unicode allows the program to process text in most of the world’s languages because it allows a very great range of characters to be interpreted and displayed correctly.

4. Or *Dokument # von #* as the case may be.


```
BYLINE: Heather Stewart  
|  
SECTION: BUSINESS  
  
LENGTH: 926 words
```

Fig. 9 – Document fields B.

Note that the name of the publication and the date of the article are both not marked up by a word in capitals followed by a colon, unlike the byline, length and section, which use explicit fields in this article. The publication and date are typically not marked up by field names in my experience, in Nexis downloads. Likewise there is no explicit mark-up for the headline or any sub-headings. Back in 1997, Peter White described the ‘opening nucleus’ of hard news stories which typically starts articles off with what journalists call the headline and the lead (a story synthesis), after which the rest of the article goes in for what he calls ‘orbital textual development’ where the journalist elaborates, explains cause and effect, justifies, contextualises and appraises. The fact that there is no mark-up for the headline and the lead may be surprising.

After the news articles have been copied from the original download files each article will have a set of headers gleaned from the weak mark-up supplied, looking like this:

```
<HEADER>  
|  
<Story 63 of 261 [THIS DOWNLOAD]>  
<FROM lines 8,704-8,800 of J:\climate_study_2016\downloaded\  
  UK\2015\UK_National_Newspapers2016-06-23_11-21.TXT>  
<SOURCE: INDEPENDENT.CO.UK>  
<DATE: DECEMBER 4, 2015 FRIDAY 3:42 PM GMT>  
<SECTION: UWE>  
<LENGTH: 580 words>  
<LOAD-DATE: December 4, 2015>  
<LANGUAGE: ENGLISH>  
<PUBLICATION-TYPE: Newspaper; Web Publication>  
<JOURNAL-CODE: WEBI>  
<COPYRIGHT: Copyright 2015 Independent Digital News and Media Limited>  
<COPYRIGHT: All Rights Reserved>  
</HEADER>
```

Fig. 10 – Header.

All the header lines starting with a word in capitals followed by a colon use the fields which were found in the article itself. All header lines are angle-bracketed so that corpus software can ignore them when handling the text alone.

A headline section follows:

```
<HEADLINE>  
    Climate change: Global carbon dioxide emissions  
    stall for second year in a row;  
</HEADLINE>
```

Fig. 11 – Headline.

Headlines are identified by seeking the end of the first complete sentence (ending in a full stop, a question mark or an exclamation mark), separating out as a chunk all the text up to the end of the first sentence and looking in that chunk for either a pair of line-breaks, or a line-break followed by several capital letters, or else the last line-break in the chunk. The headline is assumed to precede the place thus identified, the first sentence of the text (the lead or part of it) being the remainder.

5.2. Recoding

Cleaning up the text often requires some re-coding of the names of publications or of authors. For example the Guardian newspaper may have been labelled *The Guardian (London)*, or *Guardian Online* or *Guardian.co.uk* by the person who originally inputted the article into the Guardian or Nexis system. As stated above one requirement might be to construct a sub-corpus of one publication only or of broadsheets, and that requires one to merge the names appropriately. A further recoding is that the date of each article will usually be recognised by the computer operating system making sorting by date trivial. Finally, each file-name will encode the original search-word, the publication, author, date and whether the text is a duplicate or not.

5.3. Duplicate articles

Duplication occurs in four main ways. The same article can be retrieved in the database software if two similar search-terms are used, such as *global warming* and *climate change*. Such duplication is a simple artefact of the database search process. The other three causes of duplication can happen like this: an article gets re-published at a later date either in the same paper or elsewhere; it gets edited possibly for legal reasons; or it is re-issued for an online edition and there are editorial reasons for altering the style or length or accompanying images. A variant of the first of these happens also when two journalists are each using the same newswire source and echo wording in the newswire report. Where there is a date change, this can be marked up in the database, e.g. with a LOAD-DATE field.

It is easy to detect duplication where the text exactly matches or where the headers plus headlines match; computers are good at such simple matching. For the Download Parser to find cases where there has been minor change, however, is trickier. It first compares of the length of the two texts and if the difference is slight compares the vocabulary of the two texts taking into account each words' length and frequency. In practice duplication works reasonably well except in a growing number of cases where an online text gets added to with comments, because of the growth in text length.

The Download Parser offers further resources for checking that all or most of the dates in the download period got covered, a useful check that the sometimes repetitive download process was carried out thoroughly, and for filtering the articles, requiring one or more terms to be present and copying the articles thus found to suitable folders.

5.4. Corpus Output

The procedure for constructing sub-corpora can use a list of publications or of authors, and can create a sub-corpus organised so that all the texts belonging to the same day, month or year are glued together in one larger text. It is straightforward to generate any number of further sub-corpora using different criteria. Moreover each text will

be dated⁵ so that the operating system will recognise the date appropriately. The advantage of this is that time-line research is enhanced. In current work on both austerity and climate change my colleagues Reiner Grundmann, Kim-Sue Kreischer and I have usually found arranging data by month gives us the right degree of delicacy in research covering a decade, with 120 data points.

Also exported by the program automatically are various lists of authors, of publications, of headlines. Parsed output is organised by country and language, then by date.

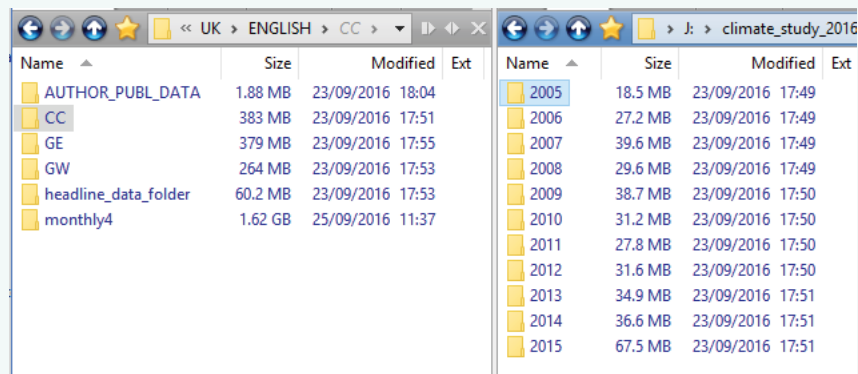


Fig. 12 – folder output.

Figure 12 shows UK data on climate change (CC) with one sub-folder per year. The year sub-folders each contain 12 month folders and the individual texts are dated appropriately both in their file-names and in terms of Windows.

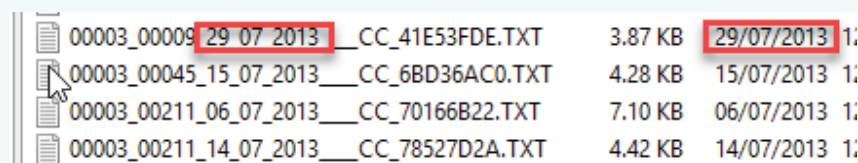


Fig. 13 – article dates.

The first article in Fig. 13 was from publication 3 and author 9, its search term was *climate change*.

5. If covering a date from January 1st 1970.

6. Conclusion

This paper has described a program freely available for Windows computers which takes as its input a set of downloaded plain text files as exported by a large online database such as Nexis, and which generates a cleaned up and formatted set of articles together with the power to create further sub-corpora at will.

6.1. Limitations

There are limitations which still need to be addressed. This program has grown over the past five years or so; it will continue to develop as necessary.

In the first section I said the procedure should ideally:

1. ensure a standard format
2. identify article features such as headlines, text boxes, sub-headings, captions of graphics, hypertext links
3. filter out duplicate articles
4. filter out articles where the main topic is not implicit in the search-term(s)
5. store individual articles so that they can be retrieved easily
6. build sub-corpora by grouping articles according to criteria such as date, publication-type, or within a set of specific publications

Objectives 1, 2, 5 and 6 seem to be met satisfactorily at present. The filter for duplicate articles (objective 3) needs to be improved chiefly because of the growing amount of online publication where comments may be included in the database output. Comments not only are supplied by new voices other than the established journalist, sub-editors, editor, that is voices who are not on the payroll and for which the publication is only partly responsible. They are also clearly variable in quality, language and tone. This is not to reject a corpus

which contains them but I feel it is important to identify and mark up such sections. Objective 4 could be met perhaps using the WordSmith key words procedure but not straightforwardly. Determining that the Berlusconi article refers to politics and sex scandals might be obvious to a human reader, and to the same person if merely shown the key words generated from that text using WordSmith, but automating which lexis belongs to austerity versus politics and sex scandals is not straightforward. A human reader who knows about late 20th and early 21st century Italy is not surprised to see politics and sex scandal vocabulary in the same article but that knowledge is not easily codified for automatic corpus procedures. Another idea would be requiring the headline text to contain certain words. That is problematic especially in journalism where there is often punning or other word play in headlines. A less binary solution might involve scoring all articles with an estimate of relevance but the Download Parser at the time of writing (October 2016) does not yet attempt that.

6.2. Named Entity Recognition (NER)

A further need is to identify Named Entities, such as names of politicians and scientists, organisations, places. The Download Parser as offered on the Lexical Analysis Software site cannot do that. At home I am able to run a program which accesses the freely available for private use Stanford Named Entity Recognition software. This produces detailed output on the names,

```
CATEGORY : PERSON
  Obama (42217 mentions)
    2005 (Dec, 3 mentions)
    2006 (Nov,Dec, 12 mentions)
    2007 (Jan-Dec, 405 mentions)
    2008 (Jan-Dec, 6478 mentions)
    2009 (Jan-Dec, 11103 mentions)
    2010 (Jan-Dec, 4783 mentions)
    2011 (Jan-Dec, 2109 mentions)
    2012 (Jan-Dec, 4855 mentions)
    2013 (Jan-Dec, 4745 mentions)
    2014 (Jan-Dec, 3193 mentions)
    2015 (Jan-Nov, 4531 mentions)
```

and organisations.

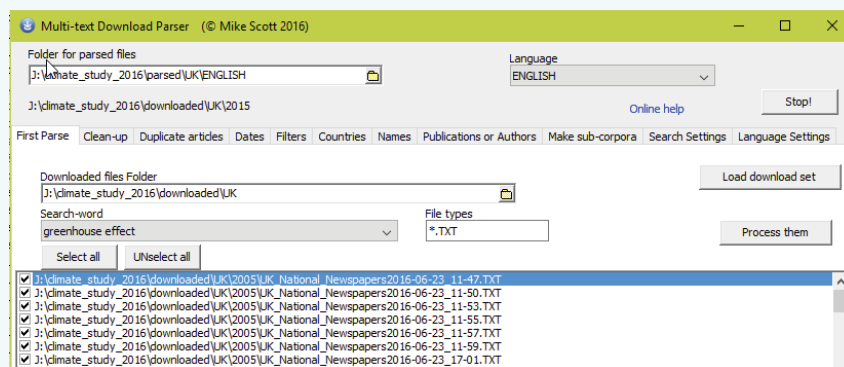


Fig. 15 – NER Organisations.

And we use this information to plot how central figures rise and fall in time-lines (Grundmann & Scott 2014).

The Download Parser has already proved a useful resource but in the future I hope it will meet needs more completely. If the Download Parser is but an infant, Corpus Linguistics and online text databases are also still developing fast and I think it is inevitable that we will see an enormous growth in the next years. It is likely that generations of students taking Corpus Linguistics modules in many countries will be creating their own specialised content corpora from online databases. To them and to their teachers, such as Leila Barbara, this paper is dedicated.

Recebido em: 05/01/2017
Aprovado em: 15/03/2017
E-mail: mike@lexically.net

Bibliography references

- BERRY, Mike. 2016. The UK Press and the Deficit Debate. *Sociology*, **50/3**: 542-559.
- EBSCO host Research. Available at <https://www.ebscohost.com/public/newspaper-source>.
- FACTIVA. Available at <https://global.factiva.com>.

- FINKEL, Jenny Rose, Trond Grenager & Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*: 363-370. Available at <http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf>.
- GRUNDMANN, Reiner & Mike Scott. 2014. Disputed climate science in the media: Do countries matter? *Public Understanding of Science*, Vol. **23/2**: 220-235.
- KOTEYKO, Nelya. 2012. Managing carbon emissions: A discursive presentation of 'market-driven sustainability' in the British media. *Language and Communication*. Vol. 32: 24-35.
- LEXISNEXIS. Available at <https://www.nexis.com/>.
- MERCILLE, Julien. 2014. The role of the media in fiscal consolidation programmes: the case of Ireland. *Cambridge Journal of Economics*. **38**: 281-300.
- National Library of Australia. Available at <http://trove.nla.gov.au/newspaper/>.
- ProQuest. Available at <http://www.proquest.com/>.
- SCOTT, Mike. 2016. *Download Parser*. Stroud: Lexical Analysis Software. Available at <http://lexically.net/DownloadParser/> with help at <http://lexically.net/DownloadParser/HTML/index.html?introduction.html>.
- STANFORD NER. Available at <http://nlp.stanford.edu/software/CRF-NER.shtml>.
- UNITED NATIONS DOCUMENT SYSTEM. Available at <https://documents.un.org/prod/ods.nsf/home.xsp>.
- WHITE, Peter R.R. 1997. 'Death, Disruption and the Moral Order: the Narrative Impulse in Mass-Media Hard News Reporting.', in *Genres and Institutions: Social Processes in the Workplace and School*, Christie, F. Martin, J.R., London, Cassell: 101-133. Available at [http://www.prrwhite.info/prrwhite,1997,Death,DisruptionandtheMoralOrder\(innews\).pdf](http://www.prrwhite.info/prrwhite,1997,Death,DisruptionandtheMoralOrder(innews).pdf)
- WORLDPRESS. Available at <http://www.worldpress.org/>.