In Search of Oblivion? How the 'Right to be Forgotten' Could Undermine Web-based Corpora

Creese, S.

Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 95 – 102

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

In search of oblivion? How the 'Right to be Forgotten' could undermine web-based corpora

Sharon Creese

*Coventry University, Priory Street, Coventry, CV1 5FB, UK**

Abstract

Corpus linguists are now facing a new challenge to collecting accurate data for web-based corpora: the 'Right to be Forgotten'. This element of data protection legislation allows individuals to request that links to webpages be removed if the information contained there can now be considered inaccurate, irrelevant or excessive. The potential difficulties this poses for researchers are illustrated by my experience collecting data for a corpus of neologisms appearing in online versions of UK national newspapers.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: corpus linguistics; corpora; neologism; Google; newspaper; Right to be Forgotten;

1. Introduction

For decades, linguists have been using corpora as a means of collecting and examining language, from the broad-based corpora used in the creation of dictionaries (Atkins and Rundell, 2008:53-57) to highly specified corpora centered around texts belonging to a particular genre or discipline. John Sinclair, 'one of the pioneers of corpus lexicography' (Mugglestone, 2011: 39), defines a corpus as 'a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research' (Sinclair, 2004: 17).

By far the largest of all corpora has to be the World Wide Web itself, containing, as it does, 'hundreds of billions of words of text [which] can be used for all manner of language research' (Kilgariff and Grefenstette, 2003: 1). It is

* Corresponding author.
E-mail address: creeses@uni.coventry.ac.uk

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.424

an ideal data source both for corpus linguists wishing to present the broadest possible picture of language, and for those seeking to drill down into highly specific areas of linguistics; it costs very little to use (in the Western world at least) and is instantly accessible (Fletcher, 2013; Lüdeling, Evert and Baroni, 2007: 7-10) .

Until now, it has been fairly straightforward to access linguistic data on the web. Researchers have been able to employ either standard commercially available search engines like Google or Yahoo, or 'webcrawling' software which enables more guided web searches by crawling through the web following links based on pre-set criteria (Fletcher, 2013: 3-5). A recent change in privacy law, however, is complicating this process, by allowing members of the public to request that links to webpages about them be removed from search engine results. This new piece of legislation, called the 'Right to be Forgotten' (RTBF) is an element of EU law surrounding the processing of personal data. Guidelines for the reform of data protection rules were proposed by the EU in 2012, however the first test case was not brought until May of 2014, when Mario Costeja González won his case against Google Spain, forcing the search engine to remove links to notice of his home being repossessed in1998 (EU, 2012; Preston, 2014).

The new rules (EU, 2014: 1-2) mean that any individual can apply to a search engine to request that links to a webpage be removed, provided they can prove that the information on that page is:

- Now believed to be inaccurate
- No longer considered relevant
- Now deemed to be excessive.

Removal of potentially large swathes of linguistic data in this way means that builders of web-based corpora are faced with an increasingly patchy picture of language. Data may potentially be lost not only because it meets the RTBF criteria, but also, potentially, because of errors in the RTBF management process (Lee, 2014). This in turn could lead to growing inaccuracy in the conclusions drawn from corpus analysis, due to incomplete datasets, and a potential increase in the difficulties surrounding replication of corpus studies, due to 'disappearing' data.

This paper outlines the new 'Right to be Forgotten' rules, explores the implications of this for corpus linguists working with web-based data, and presents an illustration of the practicalities of working within these constraints, from my own research project.

2. Google and the 'Right to be Forgotten'

The new 'Right to be Forgotten' (RTBF) legislation applies to all search engines, however the media has mainly linked it to Google, due in part to the Costeja González test case, and the fact that Google is responsible for 90% of internet searches in Europe (*The Report*, 2014). In the three months following the Spanish test case, around 130,000 RTBF requests were made to Google, more than half of which involved cases of fraud, violent crime and child sex abuse. Around half of the requests were upheld without challenge, with some 30% requiring further investigation and approximately 20% being rejected out of hand (Ibid).

The new rules require search engine providers to take more responsibility, both for the content they link to, and for managing the RTBF process itself (Travis and Arthur, 2014). Already, numerous criticisms have been laid at the new system however including complaints that:

- There are too many inconsistencies in rulings
- Only search engine links are removed: webpages themselves can still be found by other means
- It is inappropriate to impose European law on US-based companies like Google
- It is unacceptable to allow search engines to police their own adherence to the new rules
- There is always a chance that articles have been removed in error (see section 3.1) (*The Report*, 2014)

In addition, some people may find themselves facing a great deal more publicity as a consequence of their RTBF request, when the media picks up their story. Daniela Holischeck, for instance, used the new legislation to seek

removal of links to stories about how she was assaulted by a group of teenage girls. Unfortunately, her request for removal was itself considered 'news' by the UK media, leading to even more publicity (*The Report*, 2014). Mario Costeja González, too, has found himself in the public eye simply for seeking out greater privacy online (Travis and Arthur, 2014).

3. Implications of RTBF for corpus linguists

Privacy concerns are not the only issues raised by the new RTBF rules; there are also significant implications for corpus linguists, in particular non-computational linguists, who focus their research not only on the volume of linguistic data available, but on the context or genre in which language occurs.

3.1 Impact of RTBF on accuracy and scope of corpus study

The Right to be Forgotten introduces a new level of uncertainty into the process of corpus building. Any search engine user is familiar with the way in which results lists are made more comprehensive through the addition of links to associated topics. A Google search for 'iPad' for example, brings up 1,220,000,000 results (22 March 2015), ranging from where to buy it, useful accessories, popular apps, reviews, help forums, and sites it can be used on, in particular gambling sites, from bingo to poker. Search engine providers intentionally offer these wide ranging results in order to encourage users to surf more widely.

For non-linguistic users this is a useful feature, and it has also been helpful for some computational linguists who create large web-based corpora (Fletcher, 2013: 3). However for corpus linguists working with highly specified corpora centered around texts belonging to a particular genre or discipline (for example discourse analysis), or texts written by a particular author or a type of author for a particular type of audience (Kilgariff and Grefenstette, 2003: 11-12) it is problematic, since it introduces enormous amounts of unwanted data which must somehow be weeded out before linguistic analysis can begin.

The Right to be Forgotten therefore complicates things for linguists collecting specific contextual information – for example production dates in a particular genre/domain, such as newspapers, since it potentially removes data which could contain vital clues to the pattern of language use/development the researcher is seeking to explain (as I myself discovered).

Since the majority of RTBF requests appear to relate to particular events (often criminal acts) or people (notably those mentioned in, or commenting on online articles) (Preston, 2014), texts associated with these individuals/events will be particularly affected by the RTBF, leading to a potential gap in the research and linguistic understanding of these fields.

This problem can be compounded by the possibility that links are being removed in error, as the BBC has already discovered through first-hand experience (Lee, 2014), and as I believe happened during the creation of my own corpus. This leads to the pool of potential corpus texts being even further reduced, and the potential skewing of results due to the researcher being forced to work from incomplete data.

3.2 Inconsistency in search results due to RTBF

The fact that RTBF requests can be made at any time also severely complicates the issue of replicating a corpus study. Links to texts which were available when the original study was conducted but may have been removed by the time a repeat study is attempted make it impossible to achieve the same results. Indeed, texts may 'disappear' part-way through any study, preventing the researcher from returning to the original data to collect additional contextual information, and hence rendering that data useless to the project. In addition, 'new' articles may appear on search results pages, due to the reinstatement of links which had previously been removed in error (Ibid).

3.3 Limitations of the RTBF sign-off line

Search results pages affected by RTBF requests carry a sign off line indicating that results (links) have been removed, however there is no way to be certain that all affected pages have been appropriately marked; it is possible that some could have been missed, leaving corpus builders unaware that there are additional texts which they are unable to access.

4. Implications for a specific study – '*Exploring New Words in Newspapers and in Wiktionary: Lexicographic and Linguistic Perspectives'*

4.1. An overview of the study

My own research fell afoul of the RTBF while I was collecting texts for a corpus of newspaper articles containing a series of neologisms whose behavior and development I am studying. My research involves investigating the relationship between lexicography and language change through the medium of neologisms in online versions of UK newspapers, and in the online collaborative dictionary *Wiktionary* (2015) I am specifically looking at a set of 35 neologisms over a 15-year period, to explore how the meanings used in the media differ between different newspapers aimed at different readerships, and how they compare with the definitions in traditional dictionaries and in *Wiktionary*. In particular, I am seeking to show how these have changed and developed over time, within journalistic language (governed by newspapers' own style sheets). To do this, I have created a corpus of newspaper articles – NTON, Neologism Tracking in Online Newspapers – containing instances of these 35 neologisms, from which I can conduct corpus analysis, and make comparisons with the different dictionaries.

4.2. The importance of context

Crucial to this project, is the contextual information surrounding the use of these words in a selection of British newspapers. Articles were selected for inclusion in NTON based upon a number of contextual factors:

- Date (only articles post-2000).
- Newspaper.
- Article author (for example, press agency articles were excluded, since the same article could appear across all the newspapers).
- Type of article (for example letters and reader comments were excluded as they are not written by journalists. Also articles which would ONLY have appeared in the online version of newspaper – for example minute-by-minute blog-style coverage of sports events – were excluded, as the study seeks to examine journalistic use of these words as a whole, not just internet use where different linguistics conventions can apply.
- Paid-for articles were excluded, as these are often not written by journalists, and are not required to adhere to the newspapers' style guidelines, meaning the data could be misleading.
- Duplicate articles were excluded (this often occurred where a new reader comment had been added to the page; which would cause search engines to view the page as a new entry). This also meant that archive pages were excluded, since the original articles had already been collected.

Using this contextual information, I am able to track use of selected neologisms across time – 15 years, beginning in 2000 – across newspapers and their readerships, and across dictionaries.

4.2 Google Advanced Search and The Sun

Use of the 35 neologisms under study was to be assessed through their appearance in five UK national newspapers: *The Guardian* (Guardian News and Media, 2015), *Independent* (Independent.co.uk, 2015), *Daily Mail* (Associated Newspapers, 2015), *The Sun* (Newsgroup Newspapers, n.d.) and the *Express* (Northern and Shell Media Productions, 2015). A key objective was to demonstrate the development and behavior of these new words not only

over a 15-year time period, but also across the newspapers and their readerships, particularly reader demographics. Thus it might be that new words are more widely used in tabloid newspapers aimed at lower class, less-well educated readers, than in broadsheets. The newspapers used were therefore chosen based on the National Readership Survey's (NRS) social grade categories covering issues such as employment and social class (see Table 1) (businessballs.com, 2015). According to Ipsos Mori data, *The Guardian* and *Independent* are largely read by those in groups A and B, the *Daily Mail* and the *Express* by those in C1 and C2, and *The Sun* by C1/2, D and E (Duffy and Rowden, 2005: 24).

Table 1. NRS social grade definitions (businessballs.com, 2015)

Social grade	Social status	Occupation
A	Upper middle class	Higher managerial, administrative or professional
B	Middle class	Intermediate managerial, administrative or professional
C1	Lower middle class	Supervisory or clerical, junior managerial, administrative or professional
C2	Skilled working class	Skill manual workers
D	Working class	Semi and unskilled manual workers
E	Those at lowest level of subsistence	State pensioners or widows (no other earner), casual or lowest grade workers

The use of neologisms in these newspapers was originally assessed through a media scoping study conducted in April 2014, using the newspapers' own internal search engines, however there was a significant lack of reliability in the results, including:

- False positives
- Inconsistencies in search engine functionality, for example:
  - The way in which dates were represented
  - The search parameters allowed
  - The presentation of results
- Unexpected changes in search engine functionality, for example implementing a paywall.

The biggest problem was the false positives, which resulted in some 41,000 potential texts being identified for inclusion in the corpus, yet when several of these were randomly sampled to ensure the neologism was present, it became clear that most of these texts contained false positives, for example a search for 'iPadable' returned thousands of results for 'iPad'.

Testing 'iPadable' on the Google search engine showed that these false positives did not appear. It was therefore clear that an external search engine would be required to conduct the main study, and use of the same one across all of the newspapers would eliminate the inconsistencies found during the scoping study. A number of different external search engines were trialed, including Google Advanced Search (GAS), Yahoo Advanced Search and Yahoo News Search. Searches were limited to the neologisms within the newspapers' domain names and, where necessary, excluding similar or oft-repeated phrases. GAS was found to be the most reliable under these circumstances, and to best address the difficulties mentioned above

The medial scoping study was repeated in September 2014, using GAS instead of internal search engines, and this generated much more accurate results (judged by further random sampling of potential corpus texts). One unexpected problem, however, was that several texts from *The Sun*, which its own internal search engine

consistently located and which were clearly in place and did indeed contain the neologism in question, did *not* appear in any of the external search results lists. Thus although *The Sun*'s internal search engine returned 36 articles for 'frenemy' (someone who seems a friend, but really is not), 29 of these were missing from the GAS results list. At the same time, both of the positive results for 'floordrobe' (a pile of clothes on the floor) were missing from GAS, as was the single entry for 'diabesity' (diabetes caused by obesity). In addition, there were also a number of 'disappearing' articles – those that appeared on the GAS results page one day, yet were gone the next, even though they remained on the internal search results page throughout.

To ensure that the problem did not lie with *The Sun*'s searching for individual words, I conducted internal searches for the neologisms by date, by author of the article and of course by the neologism itself; in all cases, the article(s) appeared on the results page and were available via the website. I widened my GAS searches to include these additional parameters – article author and article date – but received the same results: the articles were either missing or appeared intermittently. The other external search engines also returned the same results. Whatever the search method, these articles appeared only to be available via the newspapers' own search engine.

5 Conclusion

5.1 Impact of RTBF on NTON corpus. Outcomes and updates

The central issue, then, was that articles containing the neologisms 'floordrobe', 'diabesity' and 'frenemy' were posted on *The Sun*'s website, but they were ONLY accessible by its internal search engine – external search engines simply could not find them. As other articles on the website were available through external search engines, it did not seem likely that the problem was caused by the way in which *The Sun*'s articles are coded online, since this would presumably have affected all webpages within the site. Thus something was preventing external search engines from displaying links to certain articles; the only reason that seemed to fit all of the evidence was that the articles in question had been subject to an RTBF removal request, and hence their links had been removed, or they had been taken down in error during an RTBF culling.

However it was noticeable that none of the GAS search results pages carried the RTBF sign-off line (see section 3.3), suggesting that the removal-in-error explanation was the most likely, however this could not be proven – it could just be a case of human error, with someone simply forgetting to apply the sign-off line as they were supposed to.

As previously mentioned, for my research, it was crucial to collect all the articles containing the neologisms under study so that I could track their behavior and development across time, newspaper and readership; any missing article could contain a vital clue, such as the very first use of a word in a particular newspaper, or a first use of a slightly altered meaning of that word. Any such missing data could drastically impact upon the results of the study.

The only solution, as things stood in September 2014, was to remove *The Sun* from the list of newspapers under study, meaning that the project would be limited to neologism usage in just four newspapers: *The Guardian*, *Independent*, *Daily Mail* and *Express*. This resulted in some 9,500 potential corpus texts, a significantly more accurate – and manageable – number than the original 41,000. The decision to exclude *The Sun* was made possible by the fact that newspapers had been chosen to provide a cross-over of readerships' demographic groupings: as *The Sun* largely corresponds with that of the *Daily Mail* and the *Express*, removing it would be unlikely to significantly skew the results achieved from the project.

Returning to look at *The Sun* in March 2015, and comparing results with those achieved in September, there appears to have been some slight change in the situation. Now, the articles believed to have been affected by the RTBF appear to be accessible via external search engines, but only when searching for keywords that are included in the article's headline. Thus where 'frenemy' appears in the headline, GAS is now able to locate the article, where

previously it could not, yet articles in which 'frenemy' only appears in the body copy still return a null result in both Google and the other external search engines previously used. Interestingly, a similar phenomenon was experienced by BBC blogger Robert Preston, who received notification from Google that links to one of his posts would henceforth be removed from search results pages, yet some time later the article could still be found if a search was carried out for a particular name from the story (Stan Neal) (Preston, 2014). This anomaly when searching for my neologisms was not present during the media scoping phase of data collection, as this was one of the many variations of search parameter which I employed in trying to locate the 'missing' articles in Google et al.

The implications of this change are varied: it could indicate that the RTBF was not the cause of the problems encountered in September, or it may simply be the result of a change in the coding of *The Sun*'s website. Recent changes to the coding of *The Guardian* site have presented a number of new challenges to data collection, all of which have been resolved through slight methodological tweaks, proving that this kind of research is vulnerable to such outside influences. However the changes to *The Sun* have remained intractable and the unreliability of data from this source from the very start of the project means that the situation must be approached with caution. I have therefore taken the decision not to return *The Sun* to my dataset.

5.2 Implications for corpus builders going forward

Whilst the solution to the problems caused by RTBF was relatively straightforward for my research project – simply discard the offending newspaper – for other corpus builders the situation could be significantly more complicated. RTBF removal requests could affect many text sources which are not so easily disentangled from the corpus at large, leading to large swathes of data being inadvertently excluded from future corpus studies. This could lead to a reduction in the accuracy of the results of corpus analyses, since researchers are unknowingly working with incomplete data, and it could make replicability of studies – one of the cornerstones of validity in any research project (Lüdeling, Evert and Baroni, 2007: 10) – increasingly difficult to achieve.

The result, then, is that corpus builders must be increasingly vigilant and aware of the problems presented by the RTBF, in order to continue to produce the most accurate and reliable linguistic data possible.

Acknowledgements

I would like to thank Professor Hilary Nesi of Coventry University for her help and guidance in both navigating the issues raised by the new RTBF legislation, and in producing this paper and the associated conference presentation.

References

Associated Newspapers (2014) *Mail Online* [online]. Available from http://www.dailymail.co.uk/home/index.html [17 December 2014]

Atkins, BTS., and Rundell, M. (2008) *The Oxford guide to practical lexicography* (1st ed). Oxford: Oxford University Press

Businessballs.com (2015) *Demographics classifications* [online]. Available from http://www.businessballs.com/demographicsclassifications.htm#nrs-social-grade-definitions-uk [30 March 2015]

Duffy, B. and Rowden, L. (2005) *You are what you read? How newspaper readership is related to views [online]*. London/Edinburgh: Ipsos Mori. Available from https://www.ipsos-mori.com/researchpublications/publications/240/You-are-what-you-read.aspx [30 *March 2015]*,

European Commission (2012) *Commission proposes a comprehensive reform of the data protection rules to increase users' control of their data and to cut costs for business* [online]. Available from: http://europa.eu/rapid/press-release_IP-12-46_en.htm [30 March 2015]

European Commission (2014) *Factsheet on the 'Right to be Forgotten' ruling* [online]. Available from: http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_en.pdf, [1 March 2015]

Fletcher, W.H. (2013) Corpus analysis of the world wide web. In CA. Chapelle (ed.), *The encyclopedia of applied linguistics* [online] Blackwell Publishing. Available from http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0254/full, [12 February 2013]

Guardian News and Media (2014). *The Guardian* [online]. Available from http://www.guardian.co.uk/ [17 December 2014]

Independent.co.uk (2014). *The Independent* [online]. Available from http://www.independent.co.uk/ [17 December 2014]

Kilgariff, A. and Grefenstette, G. (2003) Web as corpus, *Computational Linguistics*, 29:3

Lee, D. (2014) 'Google reinstates 'forgotten' links after pressure', *BBC News Technology* [online]. Available from:

http://www.bbc.co.uk/news/technology-28157607, 28 [February 2015]

Lüdeling, A., Evert, S., and Baroni, M. (2007) 'Using web data for linguistic purposes', *Language and computers – studies in practical linguistics*, 59, 7-24

Mugglestone, L. (2011) *Dictionaries: a very short introduction* (1st ed). Oxford: Oxford University Press

News Group Newspapers (n.d.) *The Sun* [online]. Available from http://www.thesun.co.uk/sol/homepage/ [18 December 2014]

Northern and Shell Media Productions (2014). Express [online]. Available from http://www.express.co.uk/ [16 December 2014]

Preston, R. (2014) 'Why has Google cast me into oblivion?' *BBC Business News*, 2 July 2014 [online]. Available from: http://www.bbc.co.uk/news/business-28130581 [28 February 2015]

Sinclair, J. (2004) Corpus and text – basic principles' in *developing linguistic corpora: a guide to good practice*. In M. Wynne (ed.) [online] Oxford: Oxbow Books. Available from <www.ahds.ac.uk/linguistic-corpora/> [21 February 2013]

The Report (2014) BBC Radio 4 *The 'Right to be Forgotten'*, [18 September 2014, 8pm.]

Travis, A. & Arthur, C. (2014) 'EU court backs Right to be Forgotten: Google must amend results on request' *The Guardian*, 13 May 2014 [online]. Available from: http://www.theguardian.com/technology/2014/may/13/right-to-be-forgotten-eu-court-google-search-results, [21 November 2014]

Wiktionary (2015) *Wiktionary – a wiki-based open content dictionary* [online]. Available from http://en.wiktionary.org/wiki/English [30 March 2015]