Guide to the Diabetes News Corpus (DNC)

Please cite as:

Bednarek, M. & G. Carr 2019 [last updated April 2019]. *Guide to the Diabetes News Corpus (DNC).* Available at https://sydneycorpuslab.com/services-and-projects/projects/, ACCESS DATE.

1. PREFACE

The Diabetes News Corpus (DNC) is a dataset of newspaper articles on diabetes published in major Australian newspapers from 2013-2017. It is a small, specialised corpus (~250,000 words), representative of diabetes coverage in both national and metropolitan newspapers. The DNC was compiled in the Department of Linguistics at the University of Sydney in NSW, Australia in 2018. The project was conceived and overseen by Monika Bednarek, and the corpus was compiled by research assistant Georgia Carr. The design and building of the corpus was funded by the University of Sydney through a Multidisciplinary Arts and Social Sciences Inaugural Fellowship (MASSIF). For help with making the corpus available to other scholars through the CQPweb user interface, I am very grateful to Chao Sun, Andrew Hardie, and Andressa Rodrigues Gomide. For information on the CQPweb version, see section 2.5 below. Use of the DNC by other researchers is subject to the following conditions:

1. Access is only granted for the purposes of research or scholarship;
2. The corpus cannot be distributed to others;
3. No data from the corpus are permitted to be copied, duplicated or distributed, with the exception of 'fair use' in scholarly or educational texts or presentations;
4. Copyright for all material in the DNC remains with the original creators and the material can only be used for criticism, education, and scholarship;
5. The *Diabetes News Corpus (DNC)* must be acknowledged in any publication(s) and presentation(s) resulting from research on the corpus.

This manual reports on the construction of the corpus. It is structured as follows:

2. CORPUS BUILDING

2.1 Data collection

Using the online database Factiva, we searched for articles containing *diabet** in the headline or lead paragraph. The asterisk (*) acts as a wild card so that *diabet** captures any word beginning with *diabet-*, including the noun for the condition, *diabetes* (e.g. '**Diabetes** danger for families'), the noun for people living with the condition, *diabetic(s)* (e.g. 'Device takes pain out of daily glucose check for **diabetics**'), as well as the adjective *diabetic* (e.g. 'Exam relief for **diabetic** pupils', '**Diabetic** drug hits heart of problem').

Articles were collected from the following 12 Australian newspapers from 1 January 2013-31 December 2017:

- National: *The Australian Financial Review*; *The Australian*
- NSW: *The Sydney Morning Herald* (incl. *The Sun Herald)*; *The Daily Telegraph* (incl. *Sunday Telegraph*)
- Victoria: *The Age* (and *The Sunday Age)*; *Herald Sun* (incl. *Sunday Herald Sun*)
- ACT: *The Canberra Times*
- West Australia: *The West Australian*
- NT: *The Northern Territory News* (incl. *Sunday Territorian*)
- Queensland: *The Courier Mail* (incl. *Sunday Mail*)[1]
- Tasmania: *The Mercury*
- South Australia: *The Advertiser*

Factiva provides access to either print or online versions of these newspapers. The print version was chosen since the online versions only provided irregular, selected or interrupted coverage, and dates of first availability were not provided.

2.1.1 Refining search parameters

A pilot investigation was conducted searching for the term *diabetes* appearing anywhere in the article. This search was performed for two newspapers, *The Sydney Morning Herald* and *The Daily Telegraph*, for the period 1 January 2017 to 31 December 2017. This found 212 results (79 and 133 respectively). 196 of these 212 articles did not meet the criteria for inclusion (see **Exclusions**), leaving only 16 relevant results. As a result of this pilot investigation, the search parameters were revised. The search was limited to the headline and lead paragraph, and the search syntax was expanded to *diabet**. This resulted in a much more manageable number of articles, and a higher proportion of articles which met the criteria for inclusion. Searching with the initial parameters yielded 212 results, of which only 16 were relevant (7.5%). Searching with the revised parameters yielded 50 results, of which 16 were relevant (32%).

2.1.2 Exclusions

Personal announcements, obituaries, calendars, captions, letters, weather news, food items, routine traffic reports, sports and recreation stories were excluded

automatically through Factiva's optional news filter. Remaining results were surveyed manually, and the following items were excluded: short articles (fewer than 150 words), articles that discuss diabetes in non-humans, and articles that only mention diabetes in passing. These exclusions follow Gounder & Ameer's (2018) study. It must be noted that the question of what counts as a 'mention in passing' is a subjective one. We decided to include articles that were not exclusively about diabetes but that did include longer stretches of content about diabetes rather than just mentioning the condition in passing. For example, while most articles are predominantly about diabetes, we also included items such as the following:

- An article about the danger of energy drinks, including that they spike insulin and could predispose people to diabetes.
- An article about several diseases/conditions that are predicted to be eradicated in 50 years, including type 1 diabetes. The article discusses several of these diseases/conditions, including about 220 words on type 1 diabetes.
- An article about an imprisoned drug dealer who received a payout because his diabetes worsened while in prison.
- An article about the stock price of a drug for treating diabetic eye disease.
- An article about diabetic people and their consumption of caffeine.
- An article about Diabetes Research WA wanting to change perceptions of obesity, because it is a key factor for type 2 diabetes. The article discusses both diabetes and obesity, although overall more attention is paid to obesity.

The disadvantage with this kind of approach lies in introducing researcher decisions at the stage of data construction (see Bednarek & Caple 2017: 168). The disadvantage of the alternative (no exclusions) is that the corpus would contain every item that includes the string *diabet** regardless of the subject matter of the item and would contain many 'mentions in passing'. As noted below (see 2.1.4), only about 37% of the search results from Factiva were relevant. A corpus that includes *any* mention of *diabetes, diabetic, diabetics,* etc could be used to analyse the use of diabetes-related words in general rather than focussing on items that are predominantly or at least partially *about* diabetes. Any problematic issues regarding matters of exclusion/inclusion were identified by the research assistant and resolved by the lead researcher.

Identical items within a given newspaper were automatically excluded by Factiva. However, each newspaper was searched separately and so identical items in different newspapers (rather than within the same newspaper) were not excluded by this filter. Identical items frequently appear across different newspapers, since they are often part of the same parent company (e.g. NewsCorp operates *The Australian, The Daily Telegraph, The Herald Sun, The Courier Mail, The Advertiser, The Mercury* and *The NT News*). In these cases, it would be arbitrary to include the article for one of the newspapers it appeared in and not all (e.g. this would affect the number of articles per newspaper). In addition, it is part of the Australian media landscape that particular articles are published across different newspapers, and such information is important. As such, all copies were retained. Cases of identical duplicates, as noticed during the data collection process, were recorded in a spreadsheet. This was the case for 26 items. Partial duplication (e.g. using the same quotes) also occurs across non-identical articles, which is again typical of news media coverage more generally.

2.1.3 Number of results

After all exclusions, 694 articles were included in the corpus. Table 1 shows the items by year, while Table 2 shows them by newspaper:

Year	Texts
2013	151
2014	119
2015	152
2016	140
2017	132
Total	694

Table 1 Number of articles per year

Newspaper	Items
The Australian Financial Review (F)	9
NT News (T)	21
The Age (G) incl. *Sunday Age* (U)	28
The Australian (A)	41
The Canberra Times (C)	42
The Sydney Morning Herald (S) incl. *Sun Herald* (N)	47
The Daily Telegraph (D)	54
Mercury (O)	56
The Courier Mail (R)	84
Herald Sun (H)	87
The Advertiser (V)	110
The West Australian (W)	115
Total	694

Table 2 Number of articles per newspaper, with newspaper abbreviations as used for file IDs

Several factors may explain the different number of results for different newspapers. Fewer items may appear in *The Daily Telegraph, The Sydney Morning Herald, The Northern Territory News* and *The Hobart Mercury* as they had a number of exclusions for length (i.e. items with fewer than 150 words). This includes items which were clearly focussed on diabetes, and some items which discussed diabetes in Aboriginal or Torres Strait Islander peoples. More items may appear in *The Advertiser* as this paper has a regular writer who is a 'diabetes educator, dietitian and pharmacist'. Further, more items may appear in 2013 for *The Sydney Morning Herald* which ran a series of articles in July 2013 as part of the 75th anniversary of the Australian Diabetes Council, and for *The West Australian* which ran a large number of articles (10 articles) on July 1-2 2015, corresponding with the opening of the new Telethon Type 1 Diabetes Family Centre on July 2 2015. This correlation of frequency data with events (spikes/peaks, troughs) is well-known in corpus linguistic research on newspaper data (Gabrielatos & Baker 2008, Gabrielatos et al 2012).

All items were downloaded in Rich Text Format (RTF). These were then cleaned, split and converted to plain text (see **Data sanitisation and cleaning-Data splitting**).

2.1.4 Corpus representativeness

In order to create a recent, representative, and comprehensive sample, we collected items that appeared in 12 Australian newspapers in the last five years. Where available, newspapers were chosen from both of Australia's main news organisations Fairfax and NewsCorp, and Sunday editions were included. However, the corpus is not representative of local/regional newspapers and other types of news. As the corpus does not contain any photographs, etc, it is not suitable for multimodal analysis.

Regarding the revised search parameters (see Refining **search parameters**), it is possible that articles which did focus on diabetes mentioned relevant search terms (e.g. diabetes, diabetic, diabetics) in the body of the article but not in the headline or lead paragraph. These articles would have been missed by these search parameters. However, this method resulted in a more manageable number of articles and a much higher proportion of results which fit the criteria for inclusion: searching 12 newspapers with the revised parameters found 1857 articles over a five-year period, of which 694 met the criteria for inclusion (37.4%). While we cannot be 100 per cent certain that the search parameters retrieved every article on diabetes, the selected items are clearly representative of newspaper items about diabetes in metropolitan/national newspapers in Australia.

2.2 Data sanitisation and cleaning

Boilerplate information (e.g. newspaper title, date, author) was removed and recorded as metadata (see **Metadata**). Other details were retained, e.g. if the article was labelled as an 'exclusive', or if the location was mentioned before the main body of the article ('Paris', 'London'). Minor passages which were deemed as not part of the article were excluded, e.g. links to the newspaper's website ('theage.com.au read more', 'tell us what you think heraldsun.com.au'), instructions on how to write to the newspaper's advice column ('What's your problem?', 'Where to write'), and mentions of other newspapers as sources (e.g. 'New York Times'). Minor formatting changes were also made (e.g. removing bullet points) to allow for easier processing in corpus analysis software. Photo captions were deleted to be consistent with the Factiva search criteria (i.e. some captions which should have been excluded automatically by Factiva were removed manually). After the data cleaning process, 4 articles were less than 150 words, presumably from removing boilerplate information. These articles were retained as they met the original 150+ words criterion according to the Factiva search.

2.3 Data splitting

All articles were split and saved as plain text (.txt) files and given a unique file ID (see **File IDs**). Each file contains the file ID, headline, and article main body, as follows:

File ID
Headline
Main body

```
<text_id="R14N019">
Gen Y at great risk of being #gendiabetes
OVERWEIGHT and obese young people have been urged to prioritise their health in the new
year, with estimates that up to a third of their diet is made up of "treat" foods.
Diabetes Queensland has launched the Turn It Around campaign, urging Gen Ys to cut one
junk food from their diet each day, which could see them save just over 10kg of weight
gain and $1100 a year.
Nearly 40 per cent of 18 to 24-year-olds in Queensland are overweight or obese but this
jumps to almost 50 per cent for 24 to 34-year-olds.
"Over consumption can start innocently enough at the beginning of the week with a couple
of sweet biscuits with your morning cuppa but by the end of the week you're drinking too
much and snacking on chips and burgers," Diabetes Queensland chief executive Michelle
Trute said.
"In almost 60 per cent of cases a person can prevent or delay the onset of type 2 diabetes
by cutting back on these types of foods." She said Gen Y was at risk of turning into
"Generation Diabetes" with a 63 per cent increase in diabetes in young adults in the past
10 years.
Dietitian Michael Peterman said "small sustainable changes" each day add up and it was
important for people to "retrain" their bodies.
"I see a lot of people who think they might be addicted to sugar. That 3pm sugar pick-me-
up is very common but that's a good time to take advantage of a healthier snack." Plumber
Dave Dower, 25, of Carina Heights, indulged in a treat every day but has taken up the
challenge. "Soft drink is probably my weakness, followed by beer," Mr Dower (pictured)
said.
```

In two instances, an article was split into two parts because it contained sufficiently distinct sections (e.g. a fact sheet on diabetes and a biography on someone with diabetes).

2.4 Data coding

The following information was recorded in a spreadsheet for each item (see below for an explanation of each):

- File ID
- Metadata
- Whether the item was ‘news’ or ‘non-news’
- Topic

2.4.1 File IDs

All plain text files were given a unique file ID including newspaper, year, news (N) or non-news (NN) (see **News vs. non-news**) and item number (in sequential order for the given year). For example:

S13N008 *Sydney Morning Herald*, 2013, News, 8th item

Abbreviations used for each newspaper are given in brackets in Table 2.

Some additional items were excluded after coding. As such, numbering is not always sequential. For example, there are files S15N004 and S15N006 but no article numbered ‘005’.

2.4.2 Metadata

Information that was removed during data sanitisation (year of publication, newspaper, author) was recorded in a spreadsheet.

2.4.3 News vs. non-news

Items were coded as 'news' or 'non-news' according to the following working definitions:

> News is defined as an item that describes an event, happening or issue concerning other participants and where the reported event is either new, recent, or a new/recent development. News includes hard news, soft news, research news, business news, tech news, health news, etc.
>
> Non-news includes opinion pieces, advice, arguments, analysis/expert views, editorials, personal recounts or 1st person narratives, interview-only items, advertisements, reviews, event announcements, eulogies/obituaries, letters/emails to the editor, reader comments, biographies, explainers, quizzes/tests, cartoons, wish lists, etc.

Out of 694 articles, 577 were coded as news and 117 were coded as non-news.

2.4.4 Topics

Each item was categorised by the research assistant for topic, which was recorded in a spreadsheet. Identified topics included causes, prevention, prevalence, treatment, symptoms, effects, secondary effects, cure, cost, fundraising, education, training, and Other. All topics occurred in news and non-news items, except for 'training' which was not present in non-news. The same item can be classified as more than one topic. Any problematic issues for topic classification were flagged by the research assistant and resolved by the lead researcher.

2.5 CQPweb version

The corpus described above is the version used by the research team offline. However, the corpus has also been made available to other researchers via the CQPweb interface. This online version is lemmatised and annotated with CLAWS (part-of-speech tags), USAS (semantic tags), and the Oxford simplified tagset (major word class tags). Relevant information about these annotation systems can be found under 'Corpus info' in CQPweb. The corpus can be accessed at: http://cqpw-prod.vip.sydney.edu.au/CQPweb/. Students, staff, and associates with a University of Sydney email address should already be able to access the corpus. Those that are not affiliated with the university can obtain a user name from Monika Bednarek for research/education purposes only. Her email address is listed here: https://sydney.edu.au/arts/linguistics/staff/profiles/monika.bednarek.php
Any errors in the corpus should be reported to her.

References

Bednarek M, Caple H. *The Discourse of News Values: How News Organisations Create Newsworthiness*. Oxford/New York: Oxford University Press. 2017.

Gabrielatos C, Baker P. Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press, 1996-2005. *Journal of English Linguistics* 2008; 36(1): 5-38.

Gabrielatos C, McEnery T, Diggle P, Baker P. The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics* 2012; 17(2): 151-175.

Gounder F, Ameer R. Defining diabetes and assigning responsibility: how print media frame diabetes in New Zealand. *Journal of Applied Communication Research* 2018; 46(1): 93-112, DOI: 10.1080/00909882.2017.1409907.

Notes

[1] The initial search also covered the *Brisbane Times*, but these results could not be included because of restricted account privileges.