FactCorp: A Corpus of Dutch Fact-checks and its Multiple Usages

Marten van der Meulen, W. Gudrun Reijnierse
Centre for Language Studies, Radboud University
PO Box 9103, 6500 HD Nijmegen, The Netherlands
m.vandermeulen@let.ru.nl, g.reijnierse@let.ru.nl

Abstract

Fact-checking information before publication has long been a core task for journalists, but recent times have seen the emergence of dedicated news items specifically aimed at fact-checking after publication. This relatively new form of fact-checking receives a fair amount of attention from academics, with current research focusing mostly on journalists' motivations for publishing post-hoc fact-checks, the effects of fact-checking on the perceived accuracy of false claims, and the creation of computational tools for automatic fact-checking. In this paper, we propose to study fact-checks from a corpus linguistic perspective. This will enable us to gain insight in the scope and contents of fact-checks, to investigate what fact-checks can teach us about the way in which science appears (incorrectly) in the news, and to see how fact-checks behave in the science communication landscape. We report on the creation of FactCorp, a 1,16 million-word corpus containing 1,974 fact-checks from three major Dutch newspapers. We also present results of several exploratory analyses, including a rhetorical moves analysis, a qualitative content elements analysis, and keyword analyses. Through these analyses, we aim to demonstrate the wealth of possible applications that FactCorp allows, thereby stressing the importance of creating such resources.

Keywords: fact-checks, corpus linguistics, science communication, genre analysis

1. Introduction

While the checking of facts has always been an integral part of the publication cycle of journalism (so-called 'internal' or 'ante hoc' fact-checking), fact-checks as separate news items (so-called 'external' or 'post hoc' fact-checking) seem to only to have been undertaken since the 2000s (Graves and Cherubini, 2016:6). Since then, however, post hoc fact-checks have quickly become a staple of news reporting. Nowadays, over 200 dedicated fact-checking websites are active worldwide (Stencel and Luther, 2019), and an even larger number of newspapers contain fact-checking articles on a regular or incidental basis (Graves, Nyhan and Reifler, 2015). Although fact-checking seems to gain traction around elections in particular, political claims are not the only source for fact-checking. There are also initiatives that focus on a broader range of (online) misinformation, or on more specific topics such as health- and science-related claims (Stencel, 2019).

The rise in fact-checks has been accompanied by a fair amount of academic interest in the phenomenon. To date, fact-checking research has primarily focused on three strands of research: the effects of fact-checks on readers, fact-checkers' motivations and practices, and the development of tools for automatic fact-checking. The first strand of research has investigated the effect of fact-checks on aspects such as the perceived accuracy of false claims (e.g., Garret, Nesbit and Lynch, 2013), and attitudes towards politicians (e.g., Nyhan et al., 2019; Wintersieck, 2017). Other empirical studies examined how social relationships between senders and receivers of fact-checks impact acceptance of fact-check-based corrections (e.g., Margolin, Hannak and Weber, 2017). In the second strand of research, studies have focused on fact-checkers' underlying motives for conducting fact-checks (e.g., Graves, 2017; Graves, Nyhan and Reifler, 2016). Others have focused on the accuracy of fact-checks by comparing fact-checks on the same topics from multiple fact-checkers (Lim, 2018). The third strand of research has yielded various tools to automatically check facts based on NLP and machine learning (e.g., Hassan et al., 2017).

Corpus-linguistic approaches to fact-checking, however, do not seem to have been undertaken. Although there is corpus-based research that investigates the language of various types of (fake) news messages (Rashkin et al., 2017) and the linguistic signals in user comments in fact-checked posts on social media (Jiang and Wilson, 2018), we are not aware of any studies that are specifically concerned with analyzing the contents of fact-checks. Yet, taking a corpus linguistic approach to fact-checks could shed light on the build-up, scientific focus and use of argumentation in these texts, to name but a few applications.

In the current paper, we report on Fact-Check Corpus (FactCorp), a corpus containing almost 2,000 fact-checks from three national Dutch newspapers. After introducing the corpus, we present several preliminary analyses we conducted on the data, highlighting possible avenues of research, and showing the value of such work for both the theory and practice of fact-checking in particular, as well as for journalism and science communication more generally.

2. Introducing FactCorp

2.1 Creating the Corpus

To create FactCorp, we collected Dutch news papers articles dedicated to fact-checking. While such articles also exist online in the Netherlands, both by dedicated websites (such as nieuwscheckers.nl) and general online news media (such as nu.nl), we chose to focus here on newspapers, both because of the availability of sources and the (presumed) journalistic standard of the writing and research.

To create the corpus, we ran a series of searches in the NexisUni database (formerly LexisNexis Academic; 2020) to collect all relevant items. This online database, which is available through university login, contains the full texts of news articles from different newspapers and other news sources, including many Dutch newspapers. We set the 'Group Duplication' option to 'on', to group similar items together. Each time, we indicated that we were only interested in news that had been published in Dutch newspapers. Our query was twofold. On the one hand, we searched for particular columns, as we knew from anecdotal evidence that three Dutch national newspapers

(*NRC*, *nrec.next* and *De Volkskrant*) publish(ed) fact-checks on a regular basis as particular columns. On the other hand, knowing that the data in NexisUni is indexed in varying ways, we also used a broad-strokes approach, employing the most general search terms related to our goal of identifying fact-checks (i.e. the word 'factcheck' and its spelling variations), to find any remaining fact-checks that we were not able to identify in the main searches. An overview of the search terms used to compile the corpus is presented in Table 1 below. Together, these searches yielded 3,078 results.

Column	Newspaper	Items
"Klopt dit wel" (does this make sense?)	*De Volkskrant*	376
"Met een korrel zout" (with a pinch of salt)	*De Volkskrant*	157
"Waar/niet waar" (true/false)	*De Volkskrant*	33
"nrc.checkt"	*NRC Handelsblad* *nrc.next*	1,168
"next.checkt"	*NRC Handelsblad* *nrc.next*	901
"factcheck" OR "fact check" OR "fact-check"	*De Volkskrant* *NRC Handelsblad* *nrc.next* *other newspapers*	443
	Total	3,078

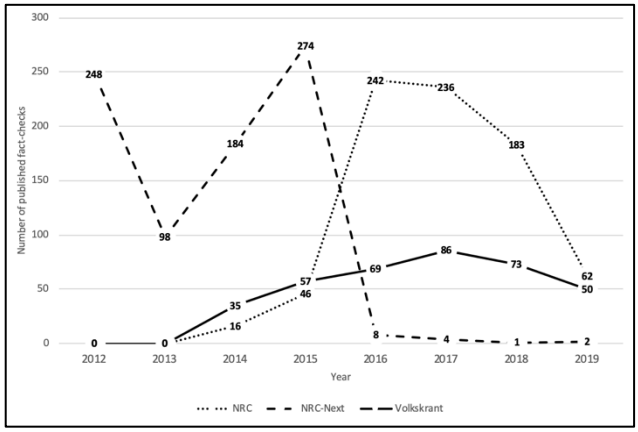
Table 1: Search terms for compiling FactCorp in NexisUni and results before clean-up

2.1.1 Corpus Clean-up

After the initial dataset of 3,078 items had been created, we manually cleaned this set up by removing double results, tables of content, opinion pieces and other articles that were not specifically aimed at fact-checking certain claims or statements. As in turned out, the number of fact checks outside of the three major newspapers was so small (<100) that we decided to remove these altogether at this stage of the research.

It is important to note at this point that between 2006 and 2015, *NRC Handelsblad* and *nrc.next* were two separate (but related) newspapers, each with their own editorial staff, different moments of publication (afternoon and morning, respectively), and different target audiences (general public and a younger audience, respectively). In 2015, the editorial staff of both newspapers merged, and since then, all news for *NRC* (i.e. *NRC Handelsblad* and *nrc.next*) is published online first. Each morning and each afternoon a print newspaper is created based on the available online articles, making the moment of publication the only remaining difference between the two newspapers. News items, including fact-checks, that are published in the morning paper, *nrc.next,* are also sometimes (verbatim) published in the afternoon paper, *NRC Handelsblad*. We made sure to include each fact-check from any of the two *NRC* newspapers only once in our final corpus. When a fact-check occurred in both newspapers, we kept the one from *NRC Handelsblad*, as this is the main newspaper, both in terms of number of readers and age of publication. As a result of this, from 2016 onwards the number of articles from *nrc.next* drops to almost zero, whereas the number of articles from NRC increases (see Figure 1).

Next, we delved into the specific files, removing most of the metadata added by NexisUni. This included the section Classification, which contains information about the language of the item, its publication type, the Journal Code, an automatic analysis of the subject matter and other information such as the origins of pictures or illustrations used. We finally discarded a number of general statements,



such as 'Bekijk de oorspronkelijke pagina' ('view original page'), that were related to the online nature of the texts. Figure 1 displays the number of fact-checks per newspaper over time.

Figure 1: Number of fact-checks per newspaper over time

2.2 FactCorp: Final Corpus[1]

2.2.1 General Description

After the various clean-up operations, the final corpus, FactCorp, consisted of 1,974 fact-checks (Table 2), amounting to a total number of 1,160,636 words, with an average of 588 words per fact-check.

Newspaper	No. of fact-checks	No. of words
NRC Handelsblad	785	450,763
nrc.next	819	554,364
de Volkskrant	370	155,509
Total	1,974	1,160,636

Table 2: Number of fact-checks and words per newspaper

2.2.2 Basic Corpus Annotations

We annotated the texts in our corpus with a layer of basic annotations. This layer contained metadata such as the newspaper in which the fact-check had appeared, the title of the column, the section of the newspaper in which it was published, and the date of publication.

3. Possible Usages

We will now describe some exploratory analyses which we performed on certain subsets of the corpus. In this way, we showcase the rich potential of FactCorp as a research tool.

The first type of analysis that we present is concerned with the description of fact-checks as a genre, with a special focus on their structure. Then, we present a qualitative content analysis of the types of claims that are

[1] A preliminary version of the corpus is available from the website of the first author at https://martenvandermeulen.com/factcorp/. Note that work on this corpus is ongoing, and that updates will follow.

fact-checked. Thirdly, we explore a number of linguistic analyses. The final example of a possible usage of FactCorp is specifically concerned with fact-checks that refer to scientific findings. In this analysis we examine what fact-checks can teach us about science communication.

3.1 Fact-checks as a Genre

A first possible application is to study fact-checks as a genre. Post-hoc fact-checking is a relatively new phenomenon that has not yet been added to the 'genre circle' of established genres (Asbreuk, de Moor and Van der Veer, 2016:25). Because of the upsurge of fact-checks in the news, it is worthwhile to investigate whether they constitute a genre on their own, and should therefore have their own place in the genre circle.

At first glance the fact-check seems to share some of its characteristics with established journalistic genres such as the feature, the commentary, the news analysis, and the opinion piece. For example, the fact-check mimics the feature in that both often discuss (part of) an earlier news item in greater depth. Fact-checks also resemble the genres of news analysis and commentary because they all discuss a view on the news. Finally, fact-checks provide a 'final' verdict or sometimes even an opinion about the news, making them similar to opinion pieces (Jansen, Steehouder and Gijsen, 2006). At the same time, external fact-checks have their own column in certain newspapers – at least in the Dutch context – which suggests that they constitute a genre on their own (or are at least perceived as such).

A genre-analytical approach to FactCorp can help establishing to what extent fact-checks display a unique character in terms of communicative goals, composition or structure, and use of linguistic features that sets them apart from other genres. Here, we present a case study investigating the composition of fact-checks by means of a thematic analysis (Braun and Clarke, 2006), focusing on rhetorical moves (e.g., Swales, 1990).[2]

To investigate which rhetorical moves are typical of fact-checks, we carried out a 6-step qualitative thematic analysis of 24 randomly selected items from the corpus (8 per newspaper). To get acquainted with the data (step 1; Braun and Clarke, 2006), all items in the sample used for this analysis were read closely. Based on this close-reading of the fact-checks, a main division into three components could be distinguished:

(1) Occasion/Introduction
(2) Factuality analysis
(3) Final judgement/Conclusion

This pattern captured the build-up of 21 of the fact-checks. However, three fact-checks in the dataset displayed a somewhat deviating pattern of main components. Two of these turned out to be 'special editions' in which correspondents fact-checked common misunderstandings about the country in which they live and work. In the third deviating item, the fact-check was part of a larger news item. Because of these deviations, we decided to exclude these three items from the current analysis. Of course, they will be taken into consideration in future work.

In step 2 of the thematic analysis, we analyzed the items in our dataset in detail by identifying information units and ascribing thematic labels that capture their contents. This yielded a list of labels that could be categorized into initial themes (step 3; resembling rhetorical moves; Swales, 1990). The initial themes were then reviewed (step 4) and finalized (step 5). The results of this thematic analysis are summarized in Table 3 (step 6). This analysis suggests the presence of a series of recurrent meaningful patterns that characterize the fact-check. Next to the main components (occasion, factuality analysis, final judgement), which are rather general, various rhetorical moves could be identified that occur more or less frequently across the dataset analyzed for this study.

Component	Thematic label	No. of occurrences
Occasion	Claim	35
	Source of claim	19
	Medium that published the claim	11
	Grounds for claim	9
	Source quoted	4
	General introduction	9
	Explanation of purpose	4
	Rhetorical question	7
Factuality analysis	Authority quoted	43
	Author statement	30
	Author critical remark	7
	Information exact	36
	General information	24
	Rhetorical question	11
	Source quoted	5
Final judgement	Verdict	13
	Summary factuality analysis	11
	Reader request	5
	Claim	3
	Authority quoted	4
	Concluding statement	4

Table 3: Rhetorical moves in fact-checks

Some of the moves occur on a frequent basis, and may therefore be considered obligatory moves in the genre of fact-checks. These are the 'claim' (that is being fact-checked) and reference to an 'authority' (to comment on the (in)correctness of the claim). Other frequently occurring moves are statements by the author of the fact-check, exact information (numbers, percentages, etc.) and more general information. It should be noted that some of these moves occur more than once in a single fact-check, suggesting that, for instance, multiple references to the claim may be made in this genre.

Some other moves are less frequent and may be considered optional (i.e., not prototypical for the genre). These include directly quoting the source of the claim that is being checked, rhetorical questions, and repetition of the claim in the final part of the fact-check. Somewhat unexpectedly, not all fact-checks in the dataset contain a

[2] This analysis is based on Madie van Ingen's BA thesis (Van Ingen, 2019), which was supervised by the second author of this article and prof. dr. W. Spooren. We thank Madie van Ingen for her contribution to this subproject of the article.

verdict. In these cases, the reader herself seems to be required to draw a conclusion about the status of the claim based on the factuality analysis provided in the article.

These findings can be considered a first attempt to characterize the genre of fact-checks by means of an analysis of meaningful patterns in the structure of fact-checking articles. An extension of the annotations to a larger part of the corpus will allow us to validate the results obtained in this pilot project in the future.

3.2 Content Elements Analysis

Next, we delved into the contents of the fact-checks.[3] This 'content elements' approach, as we call it, is part of a larger research project in which we aim to conduct qualitative and quantitative analyses to find out more about which types of claims are fact-checked, and how they are evaluated.

In the first part of this project, we examined the contents of the claims that were fact-checked, thus leaving the contents of the factuality analysis and final judgement (see section 3.1) for later stages. A different random sample of 35 fact-checks was analyzed: 10 from *de Volkskrant*, 10 from *nrc.next*, and 15 from *NRC Handelsblad*. This analysis reviewed four main aspects related to the contents of the claims:

1. Claim subjected to fact-check
2. Source of the claim (claimant)
3. Claim's location of publication
4. Spread of the claim before fact-checking'
5. Reason for fact-checking

Firstly, we examined the type of claims that are subjected to fact-checking. These cover a wide range of topics, ranging from politics to society and from science to 'fun facts'. In addition, the dataset suggests a predilection for checking number-based statements. Examples (1)-(4) illustrate this variety and the prevalence of numbers in the claims.

(1) 70 miljard lagere investeringen uit VS veroorzaakt door maatregelen kabinet ('the 70 billion decrease in investments from the US are caused by measures of the cabinet') [NRC_576] (politics)
(2) Trein Amsterdam naar Enschede ging in 1949 sneller dan in 2014 ('the train between Amsterdam and Enschede ran faster in 1949 than in 2014') [NRC_775] (society)
(3) Bier bestrijdt pijn beter dan paracetamol ('beer combats pain better than paracetamol') [Volkskrant_120] (science)
(4) Johan Cruijff, "Nederlands beste voetballer ooit" ('Johan Cruijff, the Netherlands best football player ever') [NRCNext_349] (fun fact)

Then, we identified whose claim was checked. This part of the analysis showed that the claimant was mentioned by name in about two-thirds of the fact-checks. They might be politicians, such as US president Donald Trump or Dutch Prime Minister Mark Rutte, senior executives of businesses such as IKEA or the Dutch railway company NS, union leaders, or journalists. In about one-thirds of the cases, the claimant was not mentioned by name, and a more general description of the origin of the claim was given. These descriptions include reference to other media (e.g., a news item in a newspaper or a voice-over in a television programme), non-profit organizations (e.g., the Dutch society for the protection of birds), and universities (e.g., the University of Chicago). Sometimes, newspapers also fact-check news items they themselves published before.

The location where the claim was made displays an equally varied selection, including television programmes, (online) opinion pieces, (personal) websites, press releases, scientific journals, and the newspaper that conducts the fact-check.

The fourth aspect that we analyzed related to the claims, was the spread of the claim before fact-checking. We consider this a relevant aspect of the fact-check, as it can be indicative of what might lead to a claim being picked up for fact-checking: are facts only checked when they reach a certain distribution? It also yields more information about the focus of Dutch journalists (adding to our claim-analysis) on domestic or foreign sources. Results from our study show that in about half of the cases reference is made to other media that published the claim. These references range from very general descriptions ('various news websites', 'various media') to specific references to national and international newspapers ('*The Guardian*, *De Telegraaf*, and also *De Volkskrant* published an item about it'), online news sources, and (online) magazines. Future cross-referencing this spread with the subject matter discussed in the fact-check can uncover the relative salience of certain subject matters specifically, and areas of scientific enquiry in general.

Finally, we examined whether reference was made to the reasons for checking a particular claim. Results suggest that the occasion leading to checking a claim is only mentioned in a small number of cases. When it is mentioned, the reason for conducting a fact-check was always the fact that a reader of the newspaper asked whether the claim they read/heard somewhere is correct. This suggests that what is fact-checked may in part be determined by the interest of the readers of a particular media outlet.

3.3 Linguistic Analysis

The third way in which FactCorp can be used is to explore a number of linguistic characteristics of the corpus. Here, we present a keyword analysis and a token-based analysis.

To explore the relative keyness of certain words in our corpus, we compared FactCorp with a general news corpus of Dutch news articles, the VU-DNC (Vis, 2011). The VU-DNC is a diachronic corpus consisting of 3,006 news articles from different sections of five major Dutch newspapers. Because news language changes over time (e.g., Esser and Umbricht, 2014; Vis 2011), we only used the 2002 section of the corpus for our comparison. This subsection amounts to 1,036,423 words. The keyword analysis was carried out using AntConc (Anthony, 2019). The results are presented in Table 4.

[3] This subproject was part of a research internship at Radboud University Nijmegen, supervised by both authors of this paper.

We thank Vincent Silvold for his help with annotating parts of FactCorp.

Rank	Keyword	English	Keyness
1.	nrc	nrc	3.041.221
2.	next	next	2.690.383
3.	klopt	is right	2.388.126
4.	section	section	2.366.151
5.	cijfers	numbers	1.727.616
6.	checkt	checks	1.661.954
7.	onderzoek	research	1.659.613
8.	procent	percentage	1.575.006
9.	bewering	claim	1.571.680
10.	blz	page	1.436.030
11.	één	one	1.392.772
12.	beoordelen	judge	1.358.778
13.	stelling	statement	1.312.186
14.	gebaseerd	based on	1.235.534
15.	per	per	1.177.305
16.	aanleiding	occasion	1.090.228
17.	conclusie	conclusion	1.069.989
18.	dus	so	1.048.359
19.	we	we	975.044
20.	dat	that	959.538

Table 4: Keywords of FactCorp compared to VU-DNC

Aside from some interference from metadata we retained for the purpose of analysis ('NRC', 'next', 'section', and 'page'), three meaningful clusters of words emerged from the keyword analysis. The first cluster contains words that relate to the form of the fact-check, notably 'to judge', 'occasion' and 'is true'. Secondly, there is a closely related group of words related to statements and research, such as 'research', 'claim', 'statement', 'based on', 'so' and 'conclusion'. Finally, a few words relate to numbers, namely 'numbers', 'one' and 'percentage'. The keyness of this last word is particularly interesting, as Van der Meulen and Van der Sijs (submitted) showed that this word is also key in newspaper language of 2002 as opposed to 1951. Reasons for this trend remain unexplored.

As part of our linguistic analysis of the corpus, we also performed simple token-based analysis by comparing the mentions of Dutch political parties in fact-checks with their relative presence in the Dutch parliament. As two general elections took place during the timespan of our corpus (in 2012 and 2017), we took the total number of seats in the Tweede Kamer (literally: Second Chamber; House of Representatives) in these two periods for each political party to calculate their relative presence. Results of this analysis are presented in Table 5.

It is noteworthy that the ranking of the presence of political parties matched fairly closely between reality and FactCorp. The PVV (Partij voor de Vrijheid, 'Freedom Party') drops two positions to the benefit of CDA (Christen-Democratisch Appèl, 'Christian Democrats') and D'66 (Democraten 66, 'Democrats 66') in FactCorp compared to its relative presence in the House. It is tempting to see this as a reflection of the historical status of these two latter parties, as they have been part of the political landscape of the Netherlands for decades, whereas the PVV first joined the elections in 2006. Conflicting evidence comes, however, from the fact that FvD (Forum voor Democratie, 'Forum for Democracy') is mentioned more often in the corpus than one would expected based on its relative presence in parliament, even though this party only joined the elections for the first time in 2017. Finally, the SGP (Staatkundig Gereformeerde Partij, 'Reformed Political Party') punches well above its weight: it is possible that this is an effect of their relatively outspoken religious standpoints being at odds with science.

Political party	Relative presence House in %	Relative presence FactCorp in %
VVD	24,7	20,3
PvdA	15,7	13,3
PVV	11,7	10,3
CDA	10,7	11,8
D66	10,3	11,9
SP	9,7	10,0
GroenLinks	6,0	7,8
ChristenUnie	3,3	2,6
PvdD	2,3	2,8
SGP	2,0	4,6
50Plus	2,0	2,6
Denk	1,0	0,2
FvD	0,7	1,9

Table 5: Relative presence of political parties in the House of Representatives versus relative mention in FactCorp

3.4 Tracking Science Communication

A final application of FactCorp that we present in this paper is concerned with the way in which fact-checks can provide insight into how science is communicated to the general public, as well as to how fact-checks can serve as a means of developing scientific literacy. By evaluating the accuracy of science-based claims, fact-checkers can correct inaccurate representations of science using substantiated considerations. This, in turn, may lead to increased awareness of, and knowledge about, science and scientific processes in the general audience, thus creating or increasing scientific literacy (e.g., Priest, 2013).

Our content elements analysis (section 3.2) showed that science-related claims are frequently fact-checked in Dutch media. The corpus allows the investigation of the types of pitfalls in science-public communication that are remarked on most by journalists. One particular way in which potential problems can be studied is by examining the presence of statistical terms in FactCorp. We discuss a pilot study in this direction in section 3.4.1.

From the factuality analysis (section 3.1), we furthermore know that authorities – including scientists – are often used as sources for a claim that is fact-checked, or to provide arguments for or against the factuality of a claim. By studying which sources are used, and in which way, FactCorp can shed light on the type of (scientific) information on which journalists/fact-checkers base themselves to carry out the fact-check. A pilot study examining this issue is discussed in section 3.4.2.

3.4.1 Statistical Terms in FactCorp

When checking facts, reference to numbers and statistical information is frequent (see also section 3.2). Such presence may be explained by the fact that many of the fact-checks investigate whether or not number-based claims are correct. In addition, numbers and reference to statistics may be used to specify information that was described in a

vague or unclear way in the original claim (i.e., the claim that is being fact-checked).

To examine the distribution of statistical terms in FactCorp, we created a non-exhaustive list of terms and examined their token count in the corpus (Table 6). Results of this analysis give a first idea about the statistical terms that are used in Dutch fact-checks.

Word	English	Tokencount
procent	percent	5079
gemiddeld	on average	830
gemiddelde	average	651
percentage	percentage	545
effect	effect	404
statistisch	statistical	77
significant	significant	63
experiment	experiment	66
procentpunt	percentage point	39
correlatie	correlation	21
causaliteit	causality	12
absolute cijfers	absolute numbers	9
niet-significant	non-significant	8
robuust	robust	7
relatieve cijfers	relative numbers	1

Table 6: Occurrences of statistical terms in FactCorp

Results of this token-count analysis show that the more lower-level terms from our list, such as 'percent/percentage' and 'average/on average', are more frequent in the corpus than more sophisticated statistical terms such as '(non)-significant', 'correlation', 'causality', and 'robust'. Concepts such as 'standard deviation' or 'validity' are not mentioned at all.

Research suggests that issues related to correlation and causation, as well as the robustness of scientific findings are some of the main pitfalls in communicating science to a general audience (e.g., Koetsenruijter and Berkenbosch, 2006). Our preliminary analysis of the prevalence of these (and related) terms suggests that these issues are indeed discussed in fact-checks, but possibly not as frequently as might be expected based on the literature. One possible reason for this asymmetry may be that members of the public who read fact-checks are not always familiar with the intricacies of these statistical notions. Closer inspection of the use of statistical terms is needed to study their occurrence and use in more detail.

3.4.2 Reference to (Scientific) Authority

The final pilot study that we conducted examined the reference to academic authorities/sources in the corpus. Of particular interest to us is the manner in which sources are presented as authorities. This dimension is highly relevant, as we found elsewhere that the presence of an authority may play a role in the credibility of misrepresented science news (Reijnierse & Van der Meulen, 2019).

Reference to academic sources/authorities occurs at different points in the fact-check, and takes a variety of forms. First, the person or organization making the original (i.e., to be fact-checked) claim may be a researcher or a university (see also section 3.2). Second, fact-checkers also refer to academic sources when tracing the origins of the claim that is being fact-checked. Finally, reference to academic research(ers) is found in the factuality analysis.

In all these situations, the specific background of the researcher, or the location where the research was conducted or published, are described in a variety of ways, as is illustrated in examples (5)-(11) below.

(5) ze verwijzen daarbij naar een Zuid-Afrikaans onderzoek met marathonlopers ('they refer to South African research with marathon runners') [VK_257]
(6) Volgens Amerikaans onderzoek uit de jaren negentig ('according to American research from the nineties') [NRCNext_54]
(7) De stukken verwijzen naar een onderzoek van augustus dit jaar door de State University of New York in Buffalo ('these documents refer to research from August of this year by the State University of New York in Buffalo') [NRC_251]
(8) Het International Journal of Obesity concludeerde dat ook ('The International Journal of Obesity also came to this conclusion') [NRC_384]
(9) de bevindingen van sportwetenschapper Richard Pulsford in de International Journal of Epidemiology ('the findings of sports scientist Richard Pulsford in the International Journal of Epidemiology') [VK_281]
(10) De informatie komt van hoogleraar milieuchemie en ecotoxicologie Guang-Guo Ying ('this information stems from professor of environmental chemistry and ecotoxicology Guang-Guo Ying') [NRC_117]
(11) Henk van der Kolk, politicoloog aan de Universiteit van Twente ('Henk van der Kolk, politicologist at the University of Twente') [VK_108]

These preliminary results show the diverse range of ways in which journalists can refer to scientific authorities. We aim to further study the different formats in more detail to establish, for instance, whether the level of specificity varies as a function of the type of source that is being referred to or their position in the fact-check (claimant, source for checking, etc.).

4. Conclusion and Future Work

In this paper, we have introduced FactCorp, a corpus of Dutch newspaper fact-checks. Aside from detailing the construction of this novel language resource for Dutch, we also illustrated its potential uses as a research tool, through a series of (preliminary) analyses that we conducted on the data. For example, we provided corroborating evidence that (Dutch) fact-checks share certain structural as well as linguistic characteristics which sets them apart from other genres. Moreover, we showed how analysing fact checks can shed light on the focus of science reporting and on (perceptions about) scientific literacy.

As we indicated repeatedly, we are currently conducting and planning several qualitative and quantitative follow-up studies to elaborate and fine-tune the analyses reported here. As Dutch is a high-resource language, FactCorp may also be used for other purposes. These include comparisons with fact-checks from other media (such as magazines and/ or websites), domain-specific fact-checks (such as medical news) and with fact-checks from other countries. Additionally, the position of

fact-checks as a genre between news and science could be investigated (following Biber & Conrad, 2009). Finally, knowing the properties of fact checks and having a reference corpus of them could help NLP-researchers into automatic fact-checking in checking content.

5. Bibliographical References

Asbreuk, H., De Moor, A., and Van der Meer, E. (2016). *Basisboek journalistiek schrijven*. Groningen: Noordhoff Uitgevers BV.

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2) :77–101. doi:10.1191/1478088706qp063oa

Esser, F., and Umbricht, A. (2014). The evolution of objective and interpretative journalism in the Western press: Comparing six news systems since the 1960s. *Journalism & Mass Communication Quarterly,* 91:229-249. doi:10.1177/1077699014527459

Garrett, R. K., Nisbet, E. C., and Lynch, E. K. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory: Undermining corrective effects. *Journal of Communication*, 63(4):617–637. doi:10.1111/jcom.12038

Graves, L. (2017). Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture & Critique*, 10(3):518–537. doi:10.1111/cccr.12163

Graves, L. and Cherubini, F. (2016). *The rise of fact-checking sites in Europe*. Oxford: Reuters Institute for the Study of Journalism.

Graves, L., Nyhan, B., and Reifler, J. (2016). Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking. *Journal of Communication*, 66(1):102–138. doi:10.1111/jcom.12198

Hassan, N., Nayak, A. K., Sable, V., Li, C., Tremayne, M., Zhang, G., …, and Kulkarni, A. (2017). ClaimBuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948. doi:10.14778/3137765.3137815

Jansen, C., Steehouder, M., and Gijsen, M. (2006). *Professioneel communiceren: taal- en communicatiegids*. Houten: Martinus Nijhoff.

Jiang, S., and Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), 82:1–82:23. doi:10.1145/3274351

Koetsenruijter, W. and Berkenbosch, R. (2006). *Cijfers in het nieuws*. Amsterdam: Boom.

Lim, C. (2018). Checking how fact-checkers check. *Research & Politics*. doi:10.1177/2053168018786848

Margolin, D. B., Hannak, A., and Weber, I. (2018). Political fact-checking on Twitter: When do corrections have an effect? *Political Communication*, 35(2):196–219. doi:10.1080/10584609.2017.1334018

Nyhan, B., Porter, E., Reifler, J., and Wood, T. J. (2019). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*. doi:10.1007/s11109-01909528-x

Priest, S. (2013). Critical science literacy: What citizens and journalists need to know to make sense of science. *Bulletin of Science, Technology & Society*, 33(5–6):138–145. doi:10.1177/0270467614529707

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937. doi:10.18653/v1/D17-1317

Reijnierse, W.G., and Van der Meulen, M. (2019). Trust the professor? The influence of authority on the plausibility of science news. Presentation given at the *Future of Journalism 2019 Conference*, Cardiff, 12-13 September, 2019.

Stencel, M. (2019, June 11). Number of fact-checking outlets surges to 188 in more than 60 countries. Retrieved from https://reporterslab.org/

Stencel, M. & Luther, J. (2019, October 21). Reporters' Lab fact-checking tally tops 200. Retrieved from https://reporterslab.org/

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.

Van der Meulen, M. and and Van der Sijs, N. (submitted). Tokenaanwezigheid van leenwoorden in Nederlandse kranten 1951-2002. *Nederlandse Taalkunde.*

Van Ingen, M. (2019). *De opkomst van de factcheck. Een thematische analyse*. Unpublished BA thesis. Retrieved from: https://theses.ubn.ru.nl/handle/123456789/7250

Vis, K. (2011). *Subjectivity in news discourse: A corpus linguistic analysis of informalization*. PhD dissertation. Oisterwijk: Uitgeverij BOXPress.

Wintersieck, A. L. (2017). Debating the truth: The impact of fact-checking during electoral debates. *American Politics Research*, 45(2):304–331. doi:10.1177/1532673X16686555

6. Language Resource References

Anthony, L. (2019). *AntConc* (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Nexis Uni. (2020) Online at https://advance.lexis.com/