A GATEKEEPER AMONG GATEKEEPERS
News agency influence in print and online newspapers in the Netherlands

Kasper Welbers, Wouter van Atteveldt, Jan Kleinnijenhuis, and Nel Ruigrok

This paper investigates the influence of news agency Algemeen Nederlands Persbureau (ANP) on the coverage and diversity of political news in Dutch national newspapers. Using computational text analysis, we analyzed the influence on print newspapers across three years (1996, 2008, and 2013) and compared influence on print and online newspapers in 2013. Results indicate that the influence of ANP on print newspapers only increased slightly. Online newspapers, however, depend heavily on ANP and are highly similar as a result of such dependence. We draw conclusions relating to the gatekeeping role of news agencies in the digital age in general, and in the context of the Netherlands in particular. Additionally, we demonstrate that techniques from the field of information retrieval can be used to perform these analyses on a large scale. Our scripts and instructions are published open-source to stimulate the use of these techniques in communication studies.

KEYWORDS content analysis; gatekeeping; journalism; news agencies; newspapers; media diversity; online media

Introduction

There is a conundrum regarding the diversity of news in the digital age: in spite of the larger variety of news publishers, there does not appear to be a larger variety of news (Boczkowski and De Santos 2007). One of the explanations for this phenomenon is that a large part of the news across a wide range of news publishers can be traced back to the same news agencies, or wire services. These news agencies thereby act as powerful gatekeepers: their choices in gathering, filtering, and shaping news messages affect the information input of many news publishers, and thereby indirectly have a substantial impact on the information input of citizens (Shoemaker and Vos 2009).

As professional news brokers, news agencies offer news publishers a relatively cheap, reliable, and fast supply of information. Although this can be a boon for the availability of affordable news content, it also restricts the diversity of news in society to the gatekeeping choices of one or several news agencies. This diversity is crucial in a democratic society and a core element in communications policy, because it helps distinguish facts from falsehoods, and ensures that the diversity of viewpoints in society is represented (Napoli 1999; Van Cuijilenburg 2007). Accordingly, the prominent role of news agencies in the contemporary media landscape “raises issues for news diversity and free speech” (Johnston and Forde 2011, 195).

In this paper, we investigate whether the gatekeeping influence of news agencies has increased in the digital age. Specifically, we analyze to what extent print and online editions...
of five Dutch national newspapers rely on the Algemeen Nederlands Persbureau (ANP), which is the largest and currently only national news agency in the Netherlands. We focus on political news coverage, for which the diversity of information has the most direct implications for the democratic process. Our analysis consists of three parts.

First, we analyze whether ANP’s influence on print newspapers has increased over time. Studies suggest that newspapers are becoming more dependent on news agencies (Paterson 2005; Lewis, Williams, and Franklin 2008), which has mainly been attributed to economic cutbacks in newspapers (Frijters and Velamuri 2010). However, more research is needed on the influence of individual news agencies within specific national contexts (see, e.g., Johnston and Forde 2009, 2011; Boyd-Barrett 2010) and over time. We analyze the influence of ANP on the print newspapers in three disjoint years (1996, 2008, 2013) covering two periods of economic turmoil: the economic recession and financial crisis that emerged in 2000 and 2008, respectively.

Second, we compare the influence of ANP on the print and online editions of newspapers in the first half of 2013. Online news publishers seem to rely more on news agencies, mainly because of the difficulty of making profit from online news and the quick paced 24/7 news cycle (Klinenberg 2005; Johnston and Forde 2011). We could not find studies that compared print and online newspapers, which is an interesting comparison because it provides “a window into how new journalistic forms emerge in the context of existing ones” (Boczkowski and De Santos 2007, 167–168). Furthermore, we found no figures on just how quickly paced this news cycle is. We address this gap by measuring the average time for an online newspaper to adopt a news agency article. If this time is indeed short, it puts additional weight to the argument that time pressure is an important reason for online newspapers to rely on news agencies. In addition, this information is relevant for studies that aim to model intermedia dynamics.

Finally, we investigate how the shared dependence of Dutch national newspapers on ANP affects the diversity of news content. This consequence is often assumed, but not analyzed empirically. Whether this is the case depends on the range of information offered by ANP, the extent to which newspapers rely on ANP, and the news selection choices of the newspapers. We measure how often different newspapers are influenced by the same ANP articles.

For our analyses we used a computational text analysis, based on techniques from the fields of information retrieval (IR) and natural language processing (NLP). This allows us to measure the cross-time similarity of news content across news organizations at the level of events. The use of IR and NLP techniques to measure document similarity is well established (Bagga and Baldwin 1998; Salton and Harman 2003), and similar techniques have on several occasions been used in communication research (e.g., Landauer and Dumais 1997; Van Atteveldt 2008). Yet, the use of document similarity measures appears to have received little attention in communication studies as a tool to measure interactions between news organizations. We discuss the advantages of this approach compared to other approaches, and facilitate its application in communication studies by providing scripts and instructions to apply it using the open-source statistical package R.

The Gatekeeping Role of News Agencies

The influence of news agencies on society can be conceptualized in terms of gatekeeping. Shoemaker and Vos (2009, 1) define gatekeeping broadly as “the process of
culling and crafting countless bits of information into the limited number of messages that reach people each day. “The people that perform this culling and crafting are referred to as gatekeepers. By controlling society’s supply of news, gatekeepers have a strong influence on society’s perception of relevant developments and the interpretation of these developments.

The more news publishers rely on a news agency as a source of news, the more influence the news agency has as a gatekeeper (McNelly 1959). News publishers might filter, re-interpret, and add new elements to the messages they obtain from news agencies, but the news agencies largely determine the agenda. This raises several concerns. For one, it has been argued that this can harm the quality of news, because journalists often blindly rely on the facts presented in news agency messages, but news agencies do not always uphold journalistic standards for checking sources (Davies 2008; Forde and Johnston 2013). Another concern is that the shared reliance of newspapers on the same news agency can harm the diversity of news content. The importance of diversity in the news has been well established in the communication literature, and is a core element in communications policy (Napoli 1999; Van Cuilenburg 2007). In the Netherlands, where one news agency is dominant and used by almost all major newspapers, diversity could indeed be in peril.

Newspapers and News Agencies

In 2008, Davies (2008) released the book Flat Earth News, in which he painted an alarming picture of the increased reliance of United Kingdom (UK) newspapers on news agencies. Based on a study by Lewis, Williams, and Franklin (2008), he claimed that no less than 70 percent of the news stories in the five most prestigious Fleet Street titles (i.e., popular London newspapers) were direct copies (30 percent) or rewrites (19 percent) of news agency articles, or at least contained elements (21 percent) from them (Davies 2008, 74). The concern that newspapers have become too dependent on news agencies is also shared in other countries, for instance in Australia (Johnston and Forde 2009, 2011) and the Netherlands (Scholten and Ruigrok 2009).

The increased reliance on news agency copy has mainly been attributed to economic cutbacks (Frijters and Velamuri 2010). Journalists are pressured to spend less time gathering and investigating information, and instead just “take it off the wires and knock it into shape” (journalist quoted in Davies 2008, 75). A survey of UK journalists showed that they indeed “felt that the pressure to produce a high number of stories daily has intensified, and that this increased their reliance on recycling material rather than reporting independently” Lewis et al. (2008, 4).

The reliance on news agencies appears to be even stronger for online editions of newspapers (Johnston and Forde 2009), which is alarming given the increasing popularity of online news consumption. The literature offers two main explanations for this difference. Firstly, economic constraints are higher for online newspapers due to the difficulty of making money from online news. Many users are not willing to pay for online content (Chyi 2005). Online newspapers therefore often rely solely on advertising, believing that “the revenue they could gain from content charging would be less than what they would lose in advertising” (Herbert and Thurman 2007, 213). Some newspapers have experimented with paywalls, but this was often not a viable business model (Arrese 2015)—though recently there have been more successful cases such as the New York
Given these economic difficulties, the reliance on news agency copy is likely to be high to reduce expenses.

Secondly, the influence of news agencies on online newspapers is boosted by the speed of the online news cycle. Online news can be published 24/7, which has created “an informational environment in which there is always breaking news to produce, consume, and—for reporters and their subjects—react against” (Klinenberg 2005, 54). Johnston and Forde (2011, 195–196) argue that this acceleration leads to “an even greater reliance on news agency copy than perhaps at any other time in news media history.”

News Agencies in the Netherlands

In this paper we focus on a single news agency, ANP. ANP was founded in 1934 by the Association of the Dutch Daily Press (De Nederlandse Dagbladpers) as a joint effort of the national newspapers to create a quick and independent source of news facts. Since it became a private limited company in 2001, newspaper publishers gradually divested ownership. Since 2010 ANP has been owned by the investment company V-Ventures (Rutten and Slot 2011).

The disjunction of ANP and the newspaper publishers opened up the market for new competition. In 2001 the news agency Novum was founded. Together with GPD—which was founded in 1936, and mainly provided news for regional newspapers—there were now three national news agencies. This competition eventually proved fatal. GPD ended its long history of service in 2013 after it lost an important client. Novum was taken over by ANP in 2014.

This shows that, even if newspapers have become more dependent on external news gatherers, the digital age is certainly not a golden age for news agencies. One of the main problems is that digital technology has made it much easier for news publishers to monitor and use news agency content without paying for it, by monitoring other websites, possibly using Web crawlers and Rich Site Summary (RSS) feed (Rutten and Slot 2011). Copyright law provides limited protection against this indirect use of news agency content due to the press exception—an exception in Dutch copyright law that allows news organizations to use each other’s news, at least in terms of bare facts (Guibault 2012). This greatly harms the value of news agency subscriptions, which depend on the exclusivity of information.

Together with competition from Novum, this caused significant economic cutbacks for ANP, due to which a large part of the workforce was fired after 2009 (Rutten and Slot 2011; Ebisch 2012). Despite these developments, ANP remained to be the largest news agency in the Netherlands during our study, and is currently the only national news agency. Except for the newspaper NRC Handelsblad in 2013, all national newspapers subscribed to ANP during the years analyzed in this paper.

The Influence of ANP on Dutch Newspapers

Studies that looked for traces of news agency copy in Dutch national newspapers confirmed that news agencies are an important source of information (Heijmans et al. 2009; Scholten and Ruigrok 2009). Scholten and Ruigrok (2009) focused specifically on ANP, and found that for nine prominent Dutch newspapers in 2008, on average 27.6 percent of the articles were copies or rewrites of ANP content.
Scholten and Ruigrok (2009) also found some evidence that the influence of ANP increased between 2006 and 2008. Other than that, there have been no longitudinal studies that compare the influence of ANP over time. Based on the theory that the influence of news agencies has increased due to economic cutbacks, we expect that this influence has increased more over a longer period of time.

Newspaper companies in the Netherlands experienced two substantial economic cutbacks during the last two decades (Bakker and Scholten 2011). One is the economic recession that started in 2000, and the other is the financial crisis of 2008. Bakker (2016) reports that since 2000, newspapers’ circulation has decreased by 40 percent, harming both sales and advertising incomes. Bakker and Scholten (2011) report that in all major newspaper concerns hundreds of job positions were eliminated, especially between 2008 and 2011. To analyze the impact of these developments, we compare the influence of ANP on print newspapers across these periods, focusing on three years: 1996, 2008, and 2013.

H1: The influence of ANP on political news in the Dutch national print newspapers increased between 1996 and 2013.

Prior studies in the Netherlands did not measure the influence of ANP on the online editions of newspapers. In Australia, Johnston and Forde (2009) found that two online daily newspapers rely heavily on news agency copy. In the Breaking News section, all stories in newspaper The Age were news agency copy, and for the newspaper The Daily Telegraph this was the case for at least 80 percent of the news. In the United States, Paterson (2005) found that news from online content producers on average contained between 43 and 60 percent verbatim copy of news agency text. Overall, these indications of news agency influence are much higher compared to those for print newspapers.

In the Netherlands we expect to find similar results, mostly due to the difficult economic situation of Dutch online newspapers. In 2012, Christian van Thillo, CEO of media company De Persgroep, stated that the free model—generating income through advertising only—does not work (van Soest 2012). Despite announcements of experiments with paywalls (Van der Laan 2013), no lasting solutions appear to have been found, and during the period in which we analyzed the online newspapers (the first half of 2013) the free model was still used. This struggle to make online newspapers profitable, together with the theory that reliance on news agency copy is higher due to the 24/7 online news cycle (Johnston and Forde 2011), leads to the following hypothesis:

H2: The influence of ANP on political news in online editions of Dutch national newspapers is stronger than the influence on the print editions.

To look closer into the influence of the 24/7 news cycle, we also measure the time it takes for online newspapers to respond to ANP publications. We did not find prior research that measured this, but the theory suggests that online newspapers will copy news agency items as quickly as possible. We pose the following research question:

RQ1: What is the average time for online newspapers to adopt an ANP article?

If news publishers are influenced by the same news agency, this potentially affects the diversity of news content. Whether this is the case depends on the amount of news supplied by the news agency, the amount of news required by the newspaper, and the newspaper’s news selection criteria. Put simply, if the news agency publishes sufficient news
items, and newspapers select different items to cover, then their shared reliance on the news agency does not affect the diversity of news. We thus pose the following research question:

**RQ2:** To what extent are newspapers influenced by the same ANP articles?

**Data**

We collected the news articles of the print and online editions of five national newspapers from the Netherlands: *De Telegraaf*, *Algemeen Dagblad*, *De Volkskrant*, *Trouw*, and *NRC Handelsblad*. For the news agency (ANP) and the print newspapers we gathered all articles from the LexisNexis database for three disjoint years: 1996, 2008, and 2013. LexisNexis contains a complete archive of the national newspapers used in this study back to 1996, which we therefore used as the starting point for our longitudinal analysis. Only *De Telegraaf* was not available this far back, which is therefore only used for the comparison between print and online news in 2013. For the online newspapers, we gathered all articles for the first half of 2013 using a Web-scraping algorithm. If articles were updated, we used the initial publication time, since we are interested in the time at which an event is first covered. For the news agency, updates were also filtered out.

To focus the analysis on political news, we used a search query to select only news articles that mention Dutch political parties. All types of newspaper articles were sampled, including columns and briefs. We also included ANP articles that did not match the search query, but that addressed the same event as a newspaper article that did, because we found that newspapers often added quotes or statements from politicians to add a political context to news agency items.

In total 848,479 news articles were collected, of which 59,687 were selected as political news and used in the analysis. The number of articles per medium is presented in Table 1. There is a notable increase in ANP articles between 1996 and 2008. This is

|                     | All news | Political news |
|---------------------|----------|----------------|
| **News agency**     |          |                |
| ANP                 | 56,170   | 2174           |
| Print newspapers    |          |                |
| *Algemeen Dagblad*  | 46,859   | 2480           |
| *De Volkskrant*     | 33,789   | 2848           |
| *Trouw*             | 32,743   | 3118           |
| *NRC Handelsblad*   | 36,429   | 2783           |
| *De Telegraaf*      | 56,490   | 3577           |
| Online newspapers   |          |                |
| *Algemeen Dagblad*  | 39,997   | 1788           |
| *De Volkskrant*     | 32,477   | 2262           |
| *Trouw*             | 22,107   | 2014           |
| *NRC Handelsblad*   | 5364     | 744            |
| *De Telegraaf*      | 34,351   | 1761           |

Values for online newspapers are based on the first half of 2013 only.
consistent with Rutten and Slot (2011), who mentioned that in the five years leading up to 2011, ANP produced about 40 percent more output than before. The decrease between 2008 and 2013 is likely related to economic cutbacks (Rutten and Slot 2011; Ebisch 2012). Also notable is the low amount of news articles on the website of NRC Handelsblad. Unlike most online newspapers, NRC Handelsblad mainly provides longer background stories instead of short news updates.

**Methodology**

Ideally, one would be able to learn about the influence of a news agency on a newspaper article from explicit source references, such as author information or hyperlinks (see, e.g., Johnston and Forde 2009; Meraz 2009). The problem is that this information is often inaccurate. Lewis, Williams, and Franklin (2008) studied five United Kingdom (UK) newspapers, and found that while only 1 percent of news articles mention news agencies as a source, more than half of the news could actually be traced back to news agencies, with about 30 percent being near exact copies. Print newspapers in particular are often reluctant to mention news agencies, presumably because this “dilutes the authority of a newspaper” (Matheson 2004, 458). In the Netherlands, the number of articles in print newspapers that can be traced back to a news agency also appears to be higher than the number of articles that explicitly refers to ANP (Scholten and Ruigrok 2009).

Meraz (2011) used time-series analysis as an alternative to hyperlink analysis as a method to analyze the intermediary influence of blogs and newspapers (also see Hollanders and Vliegenthart 2008; Vliegenthart and Walgrave 2008). Influence is then measured in terms of Granger (1988) causality, or predictive causality; as the extent to which the attention for a news item in one medium can be predicted based on the recent attention for this item in another medium.

Similarly, to analyze to what extent newspapers adopt stories from news agencies, we want to measure whether newspaper coverage of specific events can be traced back to prior news agency coverage of these events. This introduces two complications. The first is content analysis. To code all news items at the level of events would require an enormous effort. All news items would have to be coded inductively, and coders would need to be able to distinguish a huge number of codes (one for each event). The second complication is that this data cannot be analyzed with common time-series models. Time-series analysis requires repeated measurements over time, but media attention for specific events generally only lasts one or a few days.

We therefore use an alternative approach. Using a document similarity measure, we measure for each newspaper article how similar it is to recent news agency articles (the similarity measure is discussed in the next section). This type of approach was also used by Scholten and Ruigrok (2009) and Paterson (2005), who measured the similarity of news articles as the percentage of overlapping word n-grams (i.e., sets of n consecutive words). Influence is then measured based on the extent to which newspapers contain verbatim copy from news agencies, which is akin to scanning for plagiarism.

An important difference of our approach is that we used a different similarity measure, one that allows us to also account for the fact that news agency influence does not necessarily result in verbatim copy. Journalists can also use the information from a news agency article to write an entirely new article. Furthermore, news agencies can set the agenda of journalists: influencing only what journalists consider to be “the
important issues for debate or consideration,” without determining specifically what to say (Boyle 2001). To measure influence on this more subtle level, we calculated the similarity of news articles based on the most distinguishing nouns and proper names in the headline and lead of the article. The resulting similarity score indicates whether articles address the same event. We use this to measure the influence of ANP on a newspaper as the proportion of newspaper articles for which the event was earlier covered by ANP.

A general limitation of content analysis based approaches for measuring influence in news diffusion is that content similarities can also result from journalists using the same sources. Thus, even if traces of news agency copy are found in an article, this does not prove that the article would not have been published if the news agency rejected it. This is particularly the case for press releases and public relations material, which journalists can often easily obtain without relying on a news agency.

Nevertheless, previous studies show that journalists also often rely on news agencies for press releases and public relations material (Lewis, Williams, and Franklin 2008). Also, even if a news agency is not the only possible source of certain information for journalists, it does make this information more accessible and lends legitimacy to it (Forde and Johnston 2013). Accordingly, notwithstanding the aforementioned limitation, traces of news agency content in newspaper articles provide useful insight into the gatekeeping influence of news agencies, as also demonstrated below in our validity tests.

Measuring Document Similarity

To measure the similarity of documents, we use the vector space model approach (Salton and Harman 2003). The first step of this approach is to decide which elements of documents are used to represent them as vectors. We used a bag-of-words approach, which means that we only look at word occurrence. More specifically, we only look at the nouns and proper names, which we extracted using an NLP technique called part-of-speech tagging.1 Proper names refer to unique, named entities, such as specific people, organizations, and locations, and thus contain much information to distinguish events in news content. We also used normal nouns, because news articles also often describe events using unnamed actors and things, such as in the sentence: “a young man stole a bicycle.” To ignore different word forms (e.g., singular versus plural) we used lemmatization to reduce words to their morphological root form.

To focus on the main event of a news article, we only used the headline and first five sentences, based on the domain knowledge that newspaper articles generally have an inverted pyramid structure—the who, what, and where are immediately introduced (Knobloch et al. 2004). We did not delete low-frequency words because these can be very informative about specific events. Of the high-frequency words, that occurred in more than 1 percent of all articles, we manually deleted words that were not informative about events, such as temporal location expressions (e.g., yesterday) and author information.

Next, we weight the vectors. Turney and Pantel (2010, 156) explains that “The idea of weighting is to give more weight to surprising events and less weight to expected events,” which is important because “surprising events, if shared by two vectors, are more discriminative of the similarity between the vectors than less surprising events.” Thus, we want to give more weight to rare words than common words. For this we use the term-frequency inverse document frequency (tf.idf), which is a classic weighting scheme and recommended standard in information retrieval (Monroe, Colaresi, and Quinn 2008).
The similarity of documents can now be measured based on how close they are together in the vector space. A common measure used in information retrieval is the cosine of the angle between vectors. Since there are no negative values in our document vectors, the cosine similarity measure ranges from 0 (zero similarity) to 1 (identical).

**Measuring Influence**

We compared each political newspaper article to all news agency articles that were published within two days before the newspaper article. If this similarity score is higher than a certain threshold (determined in the validity section) then the news agency article is considered to likely have influenced the newspaper article.

The two-day time-frame was used because we assume that if a newspaper is influenced by a news agency article this happens in the short term—which we also demonstrate for online news. For print newspapers, we took into account that the ANP article had to be published before the newspaper is pressed, which is midnight for most newspapers, and in the morning for the afternoon newspaper *NRC Handelsblad*. For online newspapers, we did not impose a similar publication delay, because we also found exact copies of ANP articles that were published simultaneously by ANP and the online newspaper. Furthermore, we subtracted one hour from the ANP publication time because we found that some articles that are certainly ANP copies—they were identical, and some also credited ANP—were published before the ANP publication time in our data.2

**Validity and Similarity Threshold**

To determine whether documents address the same event, we need to decide on a threshold for the similarity score. There is no logical default for this threshold, and the most useful threshold—that is best at measuring what we want to measure—will differ depending on the data and research question. To determine the most useful threshold for our study, we performed two tests to measure the validity of our similarity measure at various thresholds.

By validity we mean the extent to which the results of the computational approach correspond to a gold standard (i.e. results that are assumed to be correct). For the first gold standard, we drew six samples of 75 pairs of newspaper and ANP documents with different levels of similarity. These document pairs were manually coded by a coder who did not see the similarity scores. The coder had to select from one of three options: the documents address unrelated events, different but related events, or the same event. If documents address the same event, the coder also coded whether the documents are (partial) copies, which was assisted by highlighting identical seven-word phrases. Although for this study we are not interested in articles with different but related events, we added this category for additional insight into the performance of the similarity measure.

The results are presented in Figure 1. We see that document pairs with similarity scores above 0.4 very often address the same event, and above 0.6 are often (partial) copies. Similarity scores below 0.2 generally indicate that documents address different events, and between 0.2 and 0.4 the results are more ambiguous, with many documents addressing different but related events.

To determine the most suitable threshold, we calculated the precision and recall scores at different levels of similarity. Precision is the proportion of pairs with a similarity
score above the threshold that are actually similar (based on the gold standard). Recall is the proportion of actually similar pairs that have a similarity score above the threshold. For reference, we added Cohen’s Kappa, which is a common inter-coder reliability measure. The performance of the computational approach is only good if both scores are high, which can be measured as their harmonic mean, called the F1 score. These results are presented in Figure 2. For events, the F1 score is highest at a threshold of 0.4 (F1 = 0.89, Kappa = 0.78). Both values indicate that the computational measurement of events is good. The measurement of (partial) copy is also good at a threshold of 0.7 (F1 = 0.87, Kappa = 0.82). For reference, a threshold of 0.2 would provide a good measurement of whether documents at least address related events (F1 = 0.86, F1 = 0.85).

The first gold standard shows how the computational similarity score relates to a human interpretation of similarity, but does not show how well this enables us to measure whether a newspaper article is actually based on an ANP article. To test this, we used the data for the online newspaper Trouw. Trouw appears to be reliably consistent in crediting sources, and 66 percent of their articles explicitly credit ANP. Thus, we use this explicit reference to ANP as our second gold standard.

The results are presented in the right-hand panel of Figure 2. Here we see that the results are good when using a threshold of 0.4 (F1 = 0.90, Kappa = 0.67) and even better...
when using a threshold of 0.7 (F1 = 0.91, Kappa = 0.75). This indicates that *Trouw* articles that credit ANP as a source are often (partial) copies. Interestingly, precision is clearly lower when using a threshold of 0.4, which indicates that there are quite a few *Trouw* articles that address an event that was earlier covered by ANP, but do not explicitly credit ANP as a source. As discussed in the section on measuring news agency influence, it is still possible that in these cases ANP was a source, or at least a factor in the news selection process. We found some indication of this: *Trouw* often covered these articles very shortly after they were published by ANP, as we show in the Results section. In addition, we also found some articles that were (partial) copies of ANP articles but did not credit ANP, indicating that source references in *Trouw* are not 100 percent reliable—which also means that the precision of our measurement is in truth higher.

In summary, the first validity test verifies that the similarity score is a valid measurement of whether a newspaper article contains an event or textual passages that previously occurred in a news agency article. The second test verifies that, at least for one online newspaper, this measurement corresponds to the actual use of ANP as a source. The 0.4 threshold appears to be the best measure for whether articles address the same event, and the 0.7 threshold indicates that one article is the (partial) copy of the other. We report our results using both thresholds, as two complementing measures of influence.

**Results**

The results are presented in three parts. First, the influence of ANP on print newspapers over time is analysed. Second, the influence of ANP on print and online news is compared. Third, homogeneity as a result of newspapers adopting the same news agency items is measured.

*Influence of ANP on Print Newspapers Over Time*

The left-hand panel in Figure 3 presents the influence of ANP on the print newspapers in 1996, 2008, and 2013, measured as the proportion of newspaper articles in which the event can be traced back to a news agency article. Overall, we see that this lies around 31 percent, ranging from 29 to 36 percent. In comparison to Scholten and Ruigrok (2009), who found an average of 27.6 percent ANP-based articles across

![FIGURE 3](image-url)

Proportion of articles that can be traced back to ANP per newspaper per year
newspapers in 2008, our results are slightly higher for that year. This is likely because we
look for similarity in terms of events, whereas Scholten and Ruigrok focused on verbatim
quotes.

If we look at the results using a cosine threshold of 0.7, we zoom in on newspaper
articles that are likely to be copies or rewrites of ANP articles. The average percentage
drops to about 9 percent, and differs more strongly across newspapers. Most noticeable
is a sharp decrease in influence on *NRC Handelsblad* in 2013. This makes sense, because
in 2010 this newspaper broke their contract with ANP, meaning that it can no longer
publish verbatim copies of ANP content. The articles that do match an ANP article at this
level of similarity are mainly short articles that contain the same quotes from politicians
as reported by ANP, meaning that its possible but not certain that ANP is an (indirect)
source.

Interestingly, even though *NRC Handelsblad* was no longer subscribed to ANP in
2013, many of the events it covered can be traced back to ANP, as seen in the results
using the 0.4 threshold. To some extent, this can simply be the result of coincidence:
ANP publishes faster than *NRC Handelsblad*, so if they independently cover the same
event then ANP is faster to report it. But, it should also be taken into account that *NRC Han-
delsblad* can indirectly rely on ANP by monitoring news publishers that do have a subscrip-
tion (which, as discussed, is also legal to do due to the press exception in copyright law).
Also, journalists tend to monitor the work of their colleagues to gather information and
to confirm their own sense of news (Gans 1979). Since *NRC Handelsblad* is an afternoon
newspaper, this can include news from the morning newspapers.

Looking at the changes over time, there is a clear increase between 1993 and 2008 in
all four newspapers in the proportion of articles traced back to ANP. Though essentially we
analysed the whole population, we also calculated whether the differences between pro-
portions are significant based on a binomial distribution. This is the case for both measure-
ments of influence: in terms of events \( p < 0.001 \) and in terms of partial copy \( p < 0.01 \).

Between 2008 and 2013 we did not see an increase, and in some cases even a signifi-
cant decrease. This could be related to economic cutbacks within ANP, as discussed in the
section on news agencies in the Netherlands. Note, for instance, that the number of political
ANP articles decreased during this period. In combination with many free online sources
that depend almost exclusively on ANP for information, this harms the exclusivity of
ANP. *NRC Handelsblad* stated this as a main reason for breaking off their contract with
ANP (Van Vulpen 2010). Other newspapers might have responded by looking for more
alternative affordable sources of information. Studies show, for instance, that journalists
increasingly use the internet for news gathering (Borden and Harvey 2013; Lecheler and
Kruikemeier 2015).

If we compare 1996 to 2013 for the measurement based on similar events, we still see
a significant increase in all newspapers \( p < 0.01 \). For the measurement based on partial
copy this is also the case for *Algemeen Dagblad* \( p < 0.001 \). Thus, we still find evidence
of an increase in ANP influence between 1993 and 2013, based on which we accept H1.

*Influence of ANP on Online Newspapers*

Figure 4 presents the influence of ANP on the print and online newspapers in the first
half of 2013. Looking at the results for similarity scores above 0.7, we see that the websites
often publish (near) exact copies of ANP articles. The only exception is *NRC Handelsblad*, but
this makes sense since it was not subscribed to ANP in 2013. These results provide strong support for H2: news agency reliance is stronger for online newspapers.

Next, we investigate the time it takes for online newspapers to respond to ANP publications. The results are presented in Figure 5. If a newspaper article matched with multiple ANP articles—for instance, if multiple ANP articles cover the same event—then only the strongest match was used to calculate the time difference. As explained, we subtracted one hour from the ANP publication time in the previous analyses. For the current analysis we used the original publication times, and if the newspaper article was published within one hour before ANP, then the response time was set to 0.

The results clearly show that online newspapers most often adopt an ANP article within one hour—at least 75 percent, except for NRC Handelsblad. For partial copies this was even above 85 percent. For all newspapers combined, the average response time (RQ1), measured as the median, is 14 minutes for same-event articles and 12 minutes for partial copies. Overall, this supports the role of ANP in the quick-paced online news cycle in the Netherlands.

**Homogeneity in Adopting ANP Articles**

To investigate the impact of ANP on content homogeneity, we analyzed what proportion of a newspaper’s articles can be traced back to the same ANP articles as another
newspaper's articles. For the sake of parsimony, we only report the results for the analysis at the level of events. The results are presented in Figure 6. Scores represent proportions for the newspapers in the rows. For example, the cell in the second row, first column indicates that 33 percent of the articles in the print edition of *De Volkskrant* can be traced back to ANP articles that also influenced the print edition of *Algemeen Dagblad*.

These results contain our answer to RQ2, and several findings are particularly interesting. Firstly, we see a clear cluster of strong proportions between the online newspapers, in particular between *Algemeen Dagblad*, *De Volkskrant*, and *Trouw*. It is notable that these three newspapers are all owned by De Nederlandse Persgroep. In 2011, a central editorial board was formed that would manage the general news for their online editions (Nu 2011). This largely explains why their similarity in the use of ANP articles is higher than 90 percent. Given that, except for *NRC Handelsblad*, all newspapers depend strongly on ANP, as seen in Figure 4, we conclude that their shared dependence on ANP indeed harms the diversity of their political news coverage.

Secondly, we see that among print newspapers these proportions are clearly lower. Most are below 30 percent, and the strongest proportions are found for *De Volkskrant* towards *Trouw* (41 percent) and *NRC Handelsblad* (42 percent). Overall, this signifies that print newspapers largely filter ANP news in different ways. Thus, the diversity of news in print newspapers does not appear to suffer much from their shared dependence on ANP.

**Conclusion**

In this paper we analyzed the influence of a single news agency, ANP, on political news coverage in print and online newspapers in the Netherlands. The first part of our analysis focused on changes in the influence of ANP on political news in print newspapers between 1996, 2008, and 2013. We observed an increase between 1996 and 2013, which
can be explained by economic cutbacks that force newspapers to cut back on news-gathering expenses. But we also found that its influence decreased between 2008 and 2013, despite severe economic cutbacks for newspapers within this period. A potential explanation for this decrease is that newspapers have become less satisfied with the exclusivity offered by their ANP subscription (Van Vulpen 2010; Rutten and Slot 2011). In response, print newspapers might have turned more to alternative affordable sources, in particular using the internet (Borden and Harvey 2013; Lecheler and Kruikemeier 2015). More research focusing on this period is required to find out whether this is the case, and could provide important insights in the seemingly fragile position of news agencies in the contemporary media landscape.

In the second part of our analysis we compared the influence of ANP on political news in print and online newspapers in the first half of 2013. Our results clearly verify that the online editions depend more on ANP than the print editions. For the four newspapers with ANP subscriptions, we found that between 50 and 75 percent of political news consisted of (partial) copies of ANP articles. Also, we found strong empirical support for a high-speed online news cycle: about 85 percent of (partial) copies were published within one hour after ANP. Note that in addition to theoretical implications, this finding has important methodological implications for time-series studies on the interactions of online news publishers. It underlines the need for models that are able to capture interactions at the level of minutes.

The third part of our analysis focused on how the shared dependence of newspapers on ANP affected the diversity of political news across newspapers. We found that print newspapers were often influenced by different ANP articles. Online newspapers, however, were often influenced by the same articles, which in combination with their strong dependence on ANP copy substantially harms the diversity of their political news coverage.

More generally, we believe that this signifies an important difference in the market logic for print and online news. Whereas diversity is an important area for competition between print newspapers, online diversity appears to be sacrificed for the sake of speed. Other studies already pointed out that in online newsrooms the pressure to be first suppresses the pressure to be right (Johnston and Forde 2009). Our study adds that it also suppresses the pressure to be diverse.

Based on these findings, we conclude that the news agency ANP has indeed become a more influential gatekeeper regarding political news in the Netherlands. In recent decades, its influence appears to have increased due to economic cutbacks in newspapers, and even more so as a result of the growing popularity of internet technology as a news medium. Notwithstanding the importance of ANP as a news gatherer, this raises concerns for the diversity of news.

It is important to note that we did not investigate the quality of journalistic work within ANP, nor did we investigate how well newspaper journalists check the reliability of news agency content. The harm to the the quality of news, as Davies (2008) claimed to observe in the United Kingdom, might not apply in the Netherlands. To conclude whether the strong influence of ANP also harms the quality of news content, additional studies are required that look into these journalistic practices.

In this paper, we used two complementary measures of influence, and each has an important limitation. Regarding the first measure: if an event is first covered by the news agency and later covered by a newspaper then there is not necessarily a causal relation. There can be alternative sources from which a news publisher could have learned about an
event, and it is generally not possible to take all possible sources into account. Furthermore, if multiple sources covered the event, it is unclear which—if any—causes the news publisher to cover it. To some extent, we can address this problem with the second measure. That is, by looking for explicit traces of influence found in how the article is written, either by using higher levels of document similarity (as in this paper) or by looking for verbatim quotes (see, e.g., Paterson 2005; Scholten and Ruigrok 2009). The limitation of this measure is that influence does not always leave these explicit traces. Also, even if explicit traces are found, it can still be the case that a journalist wrote the article independently, e.g., if the same quotes from politicians are used. To our best knowledge, these limitations cannot be overcome with only content-analysis data. Still, using both measures as complementary indications of influence appears to be a good way to address these limitations.

We found that the use of document similarity scores can be a powerful approach for tracing informational relations between news organizations on a large scale. In this line of research, we only encountered the use of this approach for the purpose of tracing verbatim quotes (Paterson 2005; Scholten and Ruigrok 2009). We expanded on this approach by using techniques from the fields of IR and NLP. For future studies we will further explore and improve this approach. Our computer scripts and instructions are available online as the RNewsflow package for the open-source statistical software R.5 We aim to keep developing this package as a free and accessible tool for the analysis of content homogeneity and news diffusion patterns.

**DISCLOSURE STATEMENT**

No potential conflict of interest was reported by the authors.

**NOTES**

1. For the NLP techniques used in this paper we used the Frog software (Van den Bosch et al. 2007), a free-to-use memory-based morphosyntactic tagger and parser for the Dutch language. Similar software is also freely available for other languages, such as CoreNLP for English (Manning et al. 2014).

2. Based on the validity tests using explicit source references in Trouw as a gold standard, we also verified that this increases the validity.

3. The Kappa and the F1 scores across thresholds were highly correlated (Pearson correlation = 0.95).

4. The median is more appropriate than the mean given the highly skewed distribution.

5. The R package is hosted on the Comprehensive R Archive Network (CRAN) under the name RNewsflow, and contains a vignette with detailed instructions: https://cran.r-project.org/web/packages/RNewsflow/vignettes/RNewsflow.html. The code is available open-source on GitHub: https://github.com/kasperwelbers/RNewsflow.

**REFERENCES**

Arrese, Ángel. 2015. “From Gratis to Paywalls: A Brief History of a Retro-Innovation in the Press’s Business.” Journalism Studies. doi:10.1080/1461670x.2015.1027788.
Bagga, Amit, and Breck Baldwin. 1998. “Entity-based Cross-Document Coreferencing Using the Vector Space Model.” Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics-volume 1, 79–85.

Bakker, Piet. 2016. “Dagbladen verliezen ruim 40% van printoplage sinds 2000.” Journalism lab, April 18. http://www.journalismlab.nl/dagbladen-verliezen-ruim-40-van-printoplage-sinds-2000.

Bakker, Piet, and Otto Scholten. 2011. Communicatiekaart van Nederland. Overzicht van media en communicatie (achtste geheel herziene druk). Amsterdam: Kluwer.

Boczkowski, Pablo J., and Martin De Santos. 2007. “When More Media Equals Less News: Patterns of Content Homogenization in Argentina’s Leading Print and Online Newspapers.” Political Communication 24 (2): 167–180.

Borden, Diane L., and Kerric Harvey, eds. 2013. The Electronic Grapevine: Rumor, Reputation, and Reporting in the New On-line Environment. New York: Routledge.

Boyd-Barrett, Oliver, ed. 2010. News Agencies in the Turbulent Era of the Internet. Barcelona/Generalitat de Catalunya: Lexikon.

Boyle, Thomas P. 2001. “Intermedia Agenda Setting in the 1996 Presidential Election.” Journalism & Mass Communication Quarterly 78 (1): 26–44.

Chyi, Hsiang I. 2005. “Willingness to Pay for Online News: An Empirical Study on the Viability of the Subscription Model.” Journal of Media Economics 18 (2): 131–142.

Cook, Jonathan E., and Shahzeen Z. Attari. 2012. “Paying for What was Free: Lessons from the New York Times Paywall.” Cyberpsychology, Behavior, and Social Networking 15 (12): 682–687.

Davies, Nick. 2008. Flat Earth News: An Award-Winning Reporter Exposes Falsehood, Distortion and Propaganda in the Global Media. London: Random House.

Ebisch, Bart. 2012. “Anp, Novum en de digitale paradox.” Persinnovatie, Februari 22. http://www.persinnovatie.nl/5092/ml/anp-novum-en-de-digitale-paradox.

Forde, Susan, and Jane Johnston. 2013. “The News Triumvirate: Public Relations, Wire Agencies and Online Copy.” Journalism Studies 14 (1): 113–129.

Frijters, Paul, and Malathi Velamuri. 2010. “Is the Internet Bad News? The Online News Era and the Market for High-Quality News.” Review of Network Economics 9 (2). doi:10.2202/1446-9022.1187.

Gans, Herbert. 1979. Deciding What’s News. New York: Vintage Books.

Granger, Clive. 1988. “Some Recent Development in a Concept of Causality.” Journal of Econometrics 39 (1): 199–211.

Guibault, Lucie. 2012. “The Press Exception in the Dutch Copyright Act.” In A Century of Dutch Copyright Law: auteurswet 1912–2012, edited by Bernt Hugenholtz, Antoon Quaedvlieg and Dirk Visser, 443–475. Amsterdam: deLex.

Heijmans, Ellen, Kees Buiks, Pytrik Schaafraad, Anne Marie Frye, Serena Daalmans, and Daphne Ten Haaf. 2009. “Nieuwsbronnen en de kwaliteit van de journalistiek.” In Journalistiek in diskrediet, edited by Bert Ummelen, 23–40. Diemen: AMB.

Herbert, Jack, and Neil Thurman. 2007. “Paid Content Strategies for News Websites: An Empirical Study of British Newspapers’ Online Business Models.” Journalism Practice 1 (2): 208–226.

Hollanders, David, and Rens Vliegenthart. 2008. “Telling what Yesterday’s News Might be Tomorrow: Modeling Media Dynamics.” Communications 33 (1): 47–68.
Johnston, Jane, and Susan Forde. 2009. “’Not Wrong for Long’: The Role and Penetration of News Wire Agencies in the 24/7 Landscape.” Global Media Journal-Australian Edition 3 (2). http://www.commarts.uws.edu.au/gmjau/v3_2009_2/johnson_forde_RA.html.

Johnston, Jane, and Susan Forde. 2011. “The Silent Partner: News Agencies and 21st Century News.” International Journal of Communication 5: 195–214.

Klinenberg, Eric. 2005. “Convergence: News Production in a Digital Age.” The Annals of the American Academy of Political and Social Science 597 (1): 48–64.

Knobloch, Silvia, Grit Patzig, Anna-Maria Mende, and Matthias Hastall. 2004. “Affective News Effects of Discourse Structure in Narratives on Suspense, Curiosity, and Enjoyment While Reading News and Novels.” Communication Research 31 (3): 259–287.

Landauer, Thomas K., and Susan T. Dumais. 1997. “A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge.” Psychological Review 104 (2): 211–240.

Lecheler, Sophie, and Sanne Kruikemeier. 2015. “Re-Evaluating Journalistic Routines in a Digital Age: A Review of Research on the Use of Online Sources.” New Media & Society 18 (1): 156–171.

Lewis, Justin, Andrew Williams, and Bob Franklin. 2008. “A Compromised Fourth Estate? UK News Journalism, Public Relations and News Sources.” Journalism Studies 9 (1): 1–20.

Lewis, Justin, Andrew Williams, Bob Franklin, James Thomas, and Nicholas Mosdell. 2008. The Quality and Independence of British Journalism: Tracking the Changes Over 20 Years. MediaWise, http://www.cardiff.ac.uk/jomec/resources/QualityIndependenceofBritishJournalism.pdf.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. “The Stanford CoreNLP Natural Language Processing Toolkit.” Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations, 55–60.

Matheson, Donald. 2004. “Weblogs and the Epistemology of the News: Some Trends in Online Journalism.” New Media & Society 6 (4): 443–468.

McNelly, John T. 1959. “Intermediary Communicators in the International Flow of News.” Journalism & Mass Communication Quarterly 36 (1): 23–26.

Meraz, Sharon. 2009. “Is there an Elite Hold? Traditional Media to Social Media Agenda Setting Influence in Blog Networks.” Journal of Computer-Mediated Communication 14 (3): 682–707.

Meraz, Sharon. 2011. “Using Time Series Analysis to Measure Intermedia Agenda-Setting Influence in Traditional Media and Political Blog Networks.” Journalism & Mass Communication Quarterly 88 (1): 176–194.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict.” Political Analysis 16 (4): 372–403.

Napoli, Philip M. 1999. “Deconstructing the Diversity Principle.” Journal of Communication 49 (4): 7–34.

Nu. 2011. “Een webredactie voor AD, Parool, Trouw en Volkskrant.” NU.nl, Januari 13. http://www.nu.nl/media/2422096/webredactie-ad-parool-trouw-en-volkskrant.html.

Paterson, Chris. 2005. “News Agency Dominance in International News on the Internet. Converging Media, Diverging Politics.” In Converging Media, Diverging Politics, edited by David Skinner, James Compton, and Mike Gasher, 145–163. Lanham, MD: Lexington Books.

Rutten, Paul, and Mijke Slot. 2011. “Zijn de persbureaus te verslaan? de positie van de Nederlandse persbureaus in de nieuwsketen.” TNO. https://www.rijksoverheid.nl/binaries/
rijksoverheid/documenten/rapporten/2011/12/06/rapport-de-positie-van-nederlandse-persbureaus-in-de-nieuwsketen/rapport-de-positie-van-nederlandse-persbureaus-in-de-nieuwsketen.pdf.

Salton, Gerard, and Donna Harman. 2003. *Information Retrieval*. Chichester, UK: John Wiley and Sons.

Scholten, Otto, and Nel Ruigrok. 2009. “Bronnen in het nieuws. nieuwsvoorziening landelijke dagbladen steeds afhankelijker van ANP.” Mediamonitor. http://www.mediamonitor.nl/gastauuteurs/otto-scholten-en-nel-ruigrok-2009/.

Shoemaker, Pamela, and Tim Vos. 2009. *Gatekeeping Theory*. New York: Routledge.

van Soest, Thijs. 2012. “Van Thillo: Het gratis content model online heeft zijn langste tijd gehad.” Trouw, June 27. http://www.trouw.nl/tr/nl/5133/Media-technologie/article/detail/3277900/2012/06/27/Van-Thillo-Het-gratis-content-model-online-heeft-zijn-langste-tijd-gehad.dhtml.

Turney, Peter, and Patrick Pantel. 2010. “From Frequency to Meaning: Vector Space Models of Semantics.” *Journal of Artificial Intelligence Research* 37 (1): 141–188.

Van Atteveldt, Wouter. 2008. *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content (dissertation)*. Charleston, SC: BookSurge.

Van Cuijlenburg, Jan. 2007. “Media Diversity, Competition and Concentration: Concepts and Theories.” In *Media between Culture and Commerce*, edited by Els de Bens, 25–54. Bristol: Intellect Books.

Van den Bosch, Antal, Bertjan Busser, Walter Daelemans, and Sander Canisius. 2007. “An Efficient Memory-Based Morphosyntactic Tagger and Parser for Dutch.” Selected papers of the 17th computational linguistics in the Netherlands Meeting.

Van der Laan, Servaas. 2013. “Nederlandse krantensites verdwijnen achter betaalmuur.” Elsevier, Januari 10. http://www.elsevier.nl/Cultuur–Televisie/nieuws/2013/1/Nederlandse-krantensites-verdwijnen-achter-betaalmuur-1143515W/.

Van Vulpen, Maarten. 2010. “NRC zegt contract met ANP op.” Publishr, December 23. http://www.publishr.nl/2010/12/nrc-zegt-contract-met-anp-op/.

Vliegenthart, Rens, and Stefaan Walgrave. 2008. “The Contingency of Intermedia Agenda-Setting. A Longitudinal Study in Belgium.” *Journalism and Mass Communication Quarterly* 85 (4): 860–877.

---

**Kasper Welbers** (author to whom correspondence should be addressed), Department of Communication Science, Vrije Universiteit van Amsterdam, The Netherlands. E-mail: k.welbers@vu.nl

**Wouter van Atteveldt**, Department of Communication Science, Vrije Universiteit van Amsterdam, The Netherlands. E-mail: wouter@vanatteveldt.com

**Jan Kleinnijenhuis**, Department of Communication Science, Vrije Universiteit van Amsterdam, The Netherlands. E-mail: j.kleinnijenhuis@vu.nl

**Nel Ruigrok**, Department of Media and Communication, Erasmus University of Rotterdam, The Netherlands. E-mail: nelruigrok@nieuwsmonitor.org